Bayesian one-inflated models for population size estimation

Tiziana Tuoto* Davide Di Cecco † Andrea Tancredi‡

Abstract

We present a Bayesian approach to a class of counting models for capture–recapture in presence of "one–inflation". One–inflation has received has received an increasing attention in capture–recapture literature in recent years, particularly in estimating the size of illegal populations. The phenomenon consists in the observation of an excess of individuals captured exactly once. If we do not explicitly model this aspect in the counting distribution, we can overestimate the population size. Bayesian model selection and the role of prior distributions are discussed. Applications to real data for the estimate of the size of some illegal populations are used to illustrate the proposed methodology.

Key Words: Bayesian Model Selection; Capture-recapture; Illegal populations; One–inflated Count Data Models

1. Introduction

A popular methodology to estimate the size of an elusive population is the capture-recapture method. An important distinction in the methodologies concerns the nature of the data: When capture occasions for each units are inherently different, for example, when different sources/mechanisms report the presence of each unit, data consists of individual capture histories, and we refer to a "Multiple Systems Estimate". When capture occasions are considered equally informative (typically, when the captures/observations are continuously collected in a fixed interval of time, and time is considered uninfluential), we only register the total number of captures for each unit. In this paper we focus on this second case, usually referred to as "repeated counting data". To estimate the population size, one needs to model the counting process of observation/capturing.

In recent years, a series of paper (see, e.g., [11], [9], [10], [3], [2]), has been devoted to the phenomenon of "one–inflation" in repeated counting data. We observe an excess of "ones" in the counting distribution, i.e., more units than expected are captured exactly once. The excess of "ones" is usually evaluated with respect to a chosen family of counting distributions. In [11] the authors considered one–inflation with respect to a "base" Poisson model. Godwin then extended the work to more complex counting distribution: Negative Binomial in [9] and finite mixture of Poissons in [10].

One–inflation can occur for different reasons; for instance we observe it when some units of the population cannot be captured anymore after the first capture. This may be the case of some wild animal populations, when the animals that experienced the capture once, find it so unpleasant that some of them develop the desire and ability to avoid subsequent captures. A similar reasoning can be applied also to human populations, particularly when the first capture consists of law enforcement,

*Istat, via Cesare Balbo 16, 00184 Roma

†Università di Roma La Sapienza, Department of Economics, viale del Castro Laurenziano, 9, 00100 Roma

‡Università di Roma La Sapienza, Department of Economics, viale del Castro Laurenziano, 9, 00100 Roma

involves imprisonment or reveals an undesirable characteristic/behaviour. See [11] for a rich discussion on the justifications and conditions for one–inflation in capture–recapture, also including an interpretation of one–inflation as limiting case of the so–called "trap shy" behavioural model (see, e.g., pg. 37 of [13] or pg. 119 of [5]).

One–inflation deserves specific attention due to its effect on usual estimators for the population size. In fact, when not accounted for, one–inflation causes overestimation of the total population size. This is true even for the well–known lower–bound Chao estimator, as discussed in [7] and [3].

In this paper we propose a Bayesian approach to count data models with one–inflation. The properties of our models are analyzed by both simulation studies and real data applications. In particular, we apply our models to real data for estimating the size of some illegal population active in Italy in 2014 and to some real data available from the literature on capture-recapture, where the issue of one–inflation has been recognised.

The paper is organized as follows: in Section 2 we introduce the notation for repeated count data and the Bayesian inference for population size, and describes the passages of a Gibbs sampler. Section 3 specifies the results under the Poisson assumption, and introduces a Bayesian test of the one–inflation assumption. In Section 4 we consider the Negative Binomial distribution and its one–inflated counterpart, and analyze the associated boundary problem. In Section 5 we show the results of our approach on data on prostitution exploitation in Italy and on some popular datasets in capture–recapture literature. Section 6 concludes the paper with some remarks.

2. Bayesian inference for population size

According to the standard formulation, consider a closed population (no birth, death or migration) of size N. For each unit in the population, let Y be a random variable taking value $j = 0, 1, 2, \dots$ if the individual is observed/captured j times. We only observe the n individuals, $n \leq N$, which are captured at least once. Let $\mathbf{y} = (y_1, \dots, y_n)$ be the vector of the individual number of captures. Note that $\mathbf{y}$ will denote the result of the capture-recapture experiment that comprises both the number n of captured individuals and the number of captures for each observed individual.

Let n_j denote the number of individuals observed j times, that is, n_j is the frequency of count j in sample $\mathbf{y}$. Our interest is to estimate the number of uncaptured units n_0, and, consequently, the total population size $N = n + n_0$, on the basis of some model for the observed n_j.

Bayesian inference for the population size N can be obtained by standard Markov Chain Monte Carlo (MCMC) algorithms. In fact, let $f(y|\theta) = P(Y = y|\theta)$ for $y = 0, 1 \dots$, be the probability distribution function for Y. The generic expression for the likelihood $f(\mathbf{y}|\theta, N)$ is

$$f(\mathbf{y}|\theta, N) = \binom{N}{n} f(0|\theta)^{N-n} \prod_{i=1}^{n} f(y_i|\theta). \tag{1}$$

Assuming independent priors for θ and N, i.e., $p(\theta, N) = p(\theta)p(N)$, the posterior distribution $p(\theta, N|\mathbf{y})$ can be easily drawn, for example, by updating the conditional distributions

$$p(\theta|N, \mathbf{y}) \propto f(0|\theta)^{N-n} \prod_{i=1}^{n} f(y_i|\theta)\, p(\theta)$$

and

$$p(N|\theta, \mathbf{y}) \propto \binom{N}{n} f(0|\theta)^{N-n}\, p(N).$$

We can generate from those posteriors via Gibbs or Metropolis-Hastings steps, according to the parametric family for Y and the prior for N. In particular, we note that

i) by assuming $p(N) \propto 1/N$, the full conditional distribution of $n_0 = N - n$ is Negative Binomial with size parameter n and probability $f(0|\theta)$ whatever the model for Y can be;

ii) the full conditional of θ corresponds to its posterior distribution when also the zero counts are known.

For example, when Y is Poisson(λ) and a priori we take the conjugate prior for λ which is Gamma($\alpha_\lambda, \beta_\lambda$) the latter step consists only in the generation of the Gamma posterior with parameters given by $\alpha_\lambda + s$ and $\beta_\lambda + n + n_0$, where s is the sum of the observed captures.

2.1 One–inflated models

We assume that in our population a specific behavioural mechanism is acting. That is, an individual that without that mechanism would face multiple captures, now has a positive probability ω of being captured just once.

Let Y denote the observed number of captures for a unit, and Y^* the latent value we would observe without the behavioural mechanism. The two variables are linked by means of the following infinite transition matrix:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & \omega & 1-\omega & 0 & 0 & \cdots \\ 0 & \omega & 0 & 1-\omega & 0 & \cdots \\ 0 & \omega & 0 & 0 & \ddots & \\ \vdots & \vdots & \vdots & \vdots & & \end{pmatrix},$$

where the (k, j)–th element represents the conditional probability $P(Y = j - 1 \mid \omega, Y^* = k - 1)$. When $k > 1$ these conditional probabilities can be written as

$$P(Y = j \mid \omega, Y^* = k) = \omega^{(1-\delta_k(j))}(1-\omega)^{\delta_k(j)} \quad j = 1, k.$$

where $\delta_k(j)$ is Kronecker delta.

Let $f(k|\theta) = P(Y^* = k \mid \theta)$ be the probability distribution, depending on some parameter θ, of the number of captures without the behavioural effect, and let $F(\theta)$ denote the associated c.d.f. Then, the resulting distribution for Y is the one–inflated model defined as follows:

$$P(Y = j \mid \theta, \omega) = \begin{cases} f(0|\theta) & \text{if } j = 0; \\ (1-\omega)f(1|\theta) + \omega(1 - f(0|\theta)) & \text{if } j = 1; \\ (1-\omega)f(j|\theta) & \text{if } j > 1. \end{cases}$$

The conditional distribution of Y^* when $Y = j$ is concentrated on j when $j \neq 1$, while, when $j = 1$, we have:

$$P(Y^* = k \mid Y = 1, \theta, \omega) = \begin{cases} 0 & \text{if } k = 0; \\ \dfrac{f(1|\theta)}{f(1|\theta) + \omega(1 - F(1|\theta))} & \text{if } k = 1; \\ \dfrac{\omega f(k|\theta)}{f(1|\theta) + \omega(1 - F(1|\theta))} & \text{if } k > 1. \end{cases} \tag{2}$$

2.2 Gibbs sampler for one–inflated models

Bayesian inference for the one–inflated models can be obtained by simulating the posterior distribution of $\theta, \omega, N, y_1^*, \ldots, y_n^*$ given the observed data $\mathbf{y}$, where $y_1^*, \ldots, y_n^*$ indicate the unknown captures that the n observed units would have faced without the behavioural mechanism. Let us assume that the parameters θ, ω and N are a priori independent and let $p(\theta, \omega, N) = p(\omega)p(\theta)p(N)$ denote the prior distribution. The general expression for the posterior distribution of one–inflated models augmented with the vector $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ is

$$\begin{aligned} p(\theta, \omega, N, \mathbf{y}^* | \mathbf{y}) &\propto p(\mathbf{y}|\theta, \omega, N, \mathbf{y}^*) p(\mathbf{y}^*, \theta, \omega, N) \\ &\propto \prod_{i=1}^{n} P(Y_i = y_i | y_i^*, \omega) p(\mathbf{y}^* | N, \theta) p(\theta) p(\omega) p(N) \\ &\propto \binom{N}{n} f(0|\theta)^{N-n} \prod_{i=1}^{n} P(Y_i = y_i | y_i^*, \omega) f(y_i^* | \theta) p(\theta) p(\omega) p(N). \end{aligned}$$

To describe our approach to simulate the posterior distribution of one–inflated models, we introduce an additional latent binary variable Z_i indicating the presence/absence of the behavioural mechanism which causes the one–inflation in unit i, i.e., Z_i is the indicator function of the event $\{Y_i \neq Y_i^*\}$. Then, we have that:

$$P(Z_i = 1 \mid Y_i \neq 1) = 0,$$

and, from (2), we have

$$P(Z_i = 1 \mid Y_i = 1) = \frac{\omega(1 - F(1|\theta))}{f(1|\theta) + \omega(1 - F(1|\theta))}.$$

Then , since $Z_i = 1$ implies $Y_i^* > 1$, we have

$$P(Y_i^* = k \mid Z_i = 1) = \begin{cases} \dfrac{f(k \mid \theta)}{1 - F(1 \mid \theta)} & \text{if } k > 1; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Now we can outline a Gibbs sampler looping over the full conditionals of Y^* and ω, N and θ. The updating of θ will depend on the model assumption for Y^* and may require a Metropolis–within–Gibbs step, whereas the updating of Y^*, ω and N can always be performed by the following exact Gibbs steps:

i) The simulation of the full conditional of $Y_1^*, \ldots, Y_n^*$ can be obtained in two steps, by first updating $Z_1, \ldots, Z_n$. In fact, let $n_z = \sum_{i=1}^{n} Z_i$ be the number of units affected by one–inflation, then, conditional on the current value of ω and θ, we can generate a value for n_z from

$$Binom\left(n_1^j\,,\ \frac{\omega(1 - F(1|\theta))}{f(1|\theta) + \omega(1 - F(1|\theta))}\right).$$

Then, for each of the n_z units, we can generate a value of Y^* by simply simulating a number of captures from the truncated count distribution (3).

ii) Consider the prior

$$\omega \sim Beta(\alpha_\omega, \beta_\omega),$$

and let $n_{z,k}$ be the number of units among the n_z for which $Y^* = k$, such that $\sum_k n_{z,k} = n_z$. Then, we can write the full conditional of ω, $p(\omega \mid -)$ as:

$$p(\omega \mid -) \propto \omega^{\alpha_\omega - 1}(1-\omega)^{\beta_\omega - 1} \prod_{k>1} \big[\omega f(k \mid \theta)\big]^{n_{z,k}} \cdot \big[(1-\omega) f(k \mid \theta)\big]^{n_k}.$$

That is, we can directly draw ω from

$$Beta\left(\alpha_\omega + n_z\,,\, \beta_\omega + \sum_{k>1} n_k\right).$$

iii) The full conditional distribution of N is given by

$$p(N \mid -) \propto \binom{N}{n} f(0|\theta)^{N-n} p(N).$$

and, by assuming the improper prior $p(N) \propto 1/N$ we can directly draw n_0 from the following Negative Binomial

$$\binom{N-1}{n-1} f(0|\theta)^{N-n}(1 - f(0|\theta))^n.$$

Finally, as we have said, the updating of θ will depend on the model assumption for Y^*. The general expression for the full conditional of θ is:

$$p(\theta \mid -) \propto f(0|\theta)^{N-n} \prod_{i=1}^{n} f(Y_i^*|\theta) p(\theta).$$

3. One–inflated Poisson

If we assume that our count data Y^* follows a Poisson distribution, i.e., $f(\theta)$ represents a Poisson density with parameter λ, the model proposed for the observed Y in previous section 2.1 corresponds to the one presented in [11].

The estimating procedure is based on the Gibbs sampler described in Section 2.1, where, in order to complete the analysis framework, we assume a Gamma(α_λ,β_λ) prior for λ, α_λ and β_λ being shape and rate parameters. Let n_k^* be the total number of units captured k times after updating n_0, n_z and Y^*, that is,

$$n_k^* = \begin{cases} n_0 & \text{for } k = 0; \\ n_1 - n_z & \text{for } k = 1; \\ n_k + n_{z,k} & \text{if } k > 1. \end{cases}$$

and let $\{n^*\}$ denote the set of all values n_k^* for $k = 0, 1, ...$ Then, we can generate the updated value for λ from its full conditional

$$Gamma\left(\alpha_\lambda + \sum_{k>0} k\, n_k^*\,,\, \beta_\lambda + N\right).$$

3.1 Testing the one–inflation assumption

To test the one–inflation assumption, we may adopt a fully Bayesian comparison of the Poisson and the One–Inflated Poisson (OIP, hereafter) models. That is, we give the two competing models (M_1 and M_2 respectively) equal prior probabilities, and evaluate the Bayes factor (BF) in favour of the OIP

$$BF = \frac{P(M_2 \,|\, \mathbf{y})}{P(M_1 \,|\, \mathbf{y})} = \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)},$$

where $P(M_i \,|\, \mathbf{y})$ indicates the posterior probability of model M_i given the data, and $p(\mathbf{y}|M_i)$ is the marginal likelihood that can be generally written as

$$p(\mathbf{y}|M_i) = \int \sum_{N=n}^{\infty} f(\mathbf{y} \mid \theta_i, N, M_i) p(\theta_i, N \mid M_i)\, d\theta_i,$$

where $\theta_1 = \lambda$ and $\theta_2 = (\lambda, \omega)$ denote respectively the parameters of the Poisson and the OIP models. Note that assuming the non informative prior $p(N) = c/N$ would produce marginal likelihoods depending on the constant c. However, the parameter N has got the same meaning across the two models, hence the use of the same improper prior $p(N) = c/N$ is justified and the constant c cancels out in the Bayes factor, (see [12]).

An analytical evaluation of the marginal likelihoods $p(\mathbf{y}|M_i)$ is not possible, then we use Chib's approximation introduced in [6], which can be easily obtained as a by-product of the Gibbs algorithm both for the Poisson and the OIP model.

To validate the use of the BF in this context, we design a simulation study for the model selection criterion. We generate datasets from each of the two models and compared the performance of each using the BF. Specifically, we set up three scenarios: in the first scenario we generate from a Poisson, in the second scenario we generate from a one–inflated Poisson with a moderate inflation rate, in the third scenario the data generating process is a one–inflated Poisson with a substantial inflation rate. The scenarios and the values of the different parameters we tested are summarised in Table 1. We set the parameters using values similar to those from the real cases analysed in Section 5.

Table 1: Simulation scenarios and data generation models, parameters' values, and the expected sample size $E[n]$ (note that the expected values of n are common to all three scenarios)

Scenario I	Scenario II	Scenario III	N	λ	$E[n]$
Poisson	OIP ($\omega = 0.2$)	OIP ($\omega = 0.5$)	1000	1	632
				2	865
			5000	1	3161
				2	4323

In each scenario, we generated 100 datasets of N units from the relative generating model, and remove the 0-counts. The sample size n varies at each iteration, and in Table 1 we report the average value for each scenario. In each simulation we computed the posterior mean of n_0 under both models, and the BF in favour of OIP. Table 2 reports the evidence of the BF in favor of the OIP model under different scenarios. We adopt the categories proposed in [12] to describe the evidence in favor of the statistical model.

Table 2: Evidence of the Bayes Factor in favor of the One–inflated Poisson model in the three scenarios

		Evidence in favour of OIP				
		Against	Anecdotal	Moderate	Strong	Extreme
Scenario I		≤ 1	1-3.2	3.2-10	10-100	≥ 100
N=1000	λ=1	97	2	1	0	0
	λ=2	100	0	0	0	0
N=5000	λ=1	98	0	1	1	0
	λ=2	99	1	0	0	0
Scenario II		Against	Anecdotal	Moderate	Strong	Extreme
N=1000	λ=1	29	23	17	23	8
	λ=2	0	0	0	4	96
N=5000	λ=1	0	0	2	8	90
	λ=2	0	0	0	0	100
Scenario III		Against	Anecdotal	Moderate	Strong	Extreme
N=1000	λ=1	0	1	2	7	90
	λ=2	0	0	0	0	100
N=5000	λ=1	0	0	0	0	100
	λ=2	0	0	0	0	100

Clearly, the BF favors the true data-generating model in all scenarios and parameter combinations, with the only exception of Scenario II, with $N = 1000$, and $\lambda = 1$. Note that in this case, the sample size we observe is small ($E[n] = 632$), as is the one–inflation ω. The behaviour of the BF in this particular setting can be better interpreted by analysing the simulation results in terms of parameter estimates.

Figure 1 shows the results of the simulation study in terms of % relative bias in the estimation of the zero counts n_0. The % relative bias is calculated as the relative difference between the true value and the posterior mean of the parameter.

The posterior mean we obtain with a one–inflated model is always lower than that obtained with the Poisson model. Consequently, when the OIP is the true generating process, the posterior mean deriving from the Poisson model severely and systematically overestimates N. On the converse, if we generate data from the Poisson, the bias deriving from considering the posterior mean of the OIP model is less severe and, on average, we moderately underestimate N.

In conclusion, as expected, the OIP model encompasses the Poisson model and, when one–inflation is not present, the slight underestimation of N decreases as n increases.

4. One–inflated Negative Binomial

The Negative Binomial distribution (NB) is often adopted as a two-parameters generalization of the Poisson that can account for over-dispersed count data. Its use is known in capture–recapture, and has been also investigated in the presence of one–inflation in [9].

Here we assume that the unobserved count Y^* follows a NB model with the

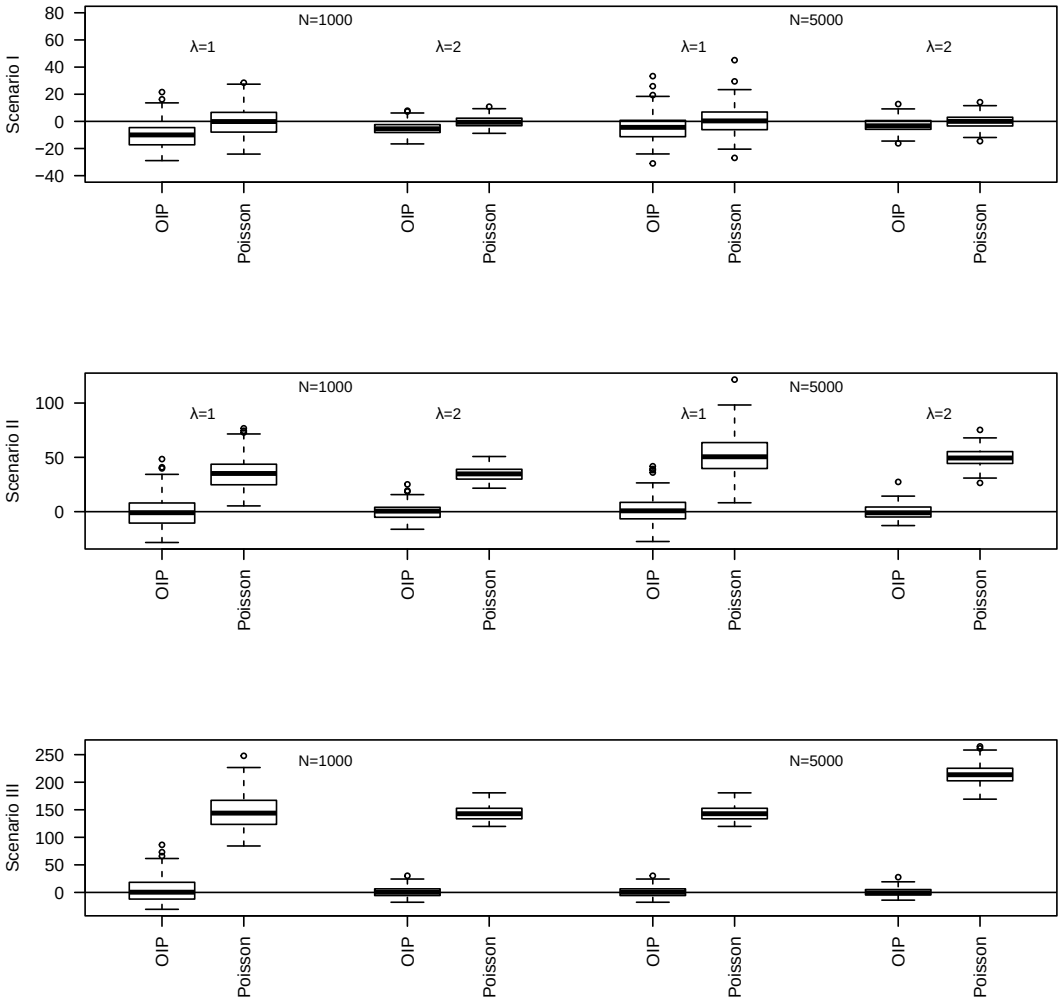

Figure 1: Relative Bias (%) of n_0 estimates in different simulation settings, when using Poisson and one–inflated Poisson (OIP) models.

following parameterization in terms of r and p:

$$P(Y^* = k \,|\, r, p) = \frac{\Gamma(k+r)}{\Gamma(r)k!} p^r (1-p)^k, \tag{4}$$

and we will call the resulting model for Y, One–inflated Negative Binomial (OINB). In our Bayesian approach, we set two independent priors on the parameters p and r. For p we take a $Beta(\alpha_p, \beta_p)$ prior, while for r we compare Gamma and Inverse Gamma priors in order to evaluate the different tail behaviour of these distributions on the posterior summaries.

The Gibbs sampler we developed follows the same passages presented in section 2.1, where $f(\theta)$ takes the form (4). Recall that n_k^* represents the number of units captured k times after updating n_0, Z and Y^*. Then, generating from the full conditional of p presents no difficulties, as it results to be:

$$[p \,|\, -] \sim Beta\left(\alpha_p + Nr \,,\, \beta_p + \sum_{k>0} k\, n_k^*\right).$$

To update r, we compare two different approaches: a Gaussian random-walk Metropolis-Hastings step, and the two-stages Gibbs sampler proposed by [19].

4.1 Metropolis Hastings

The full conditional of r results in:

$$P(r \mid -) \propto p^{Nr} \prod_{k=0,1,\ldots} \left(\frac{\Gamma(k+r)}{\Gamma(r)k!} \right)^{n_k^*} \frac{r^{\alpha_r - 1}}{e^{r\beta_r}}.$$

Then, if we consider a Gaussian random walk Metropolis-Hastings, we accept a proposed value r' with probability equal to the minimum between 1 and

$$\exp\Big\{ \sum_k n_k^* \left[\log\Gamma(r'+k) - \log\Gamma(r') - \log\Gamma(r+k) + \log\Gamma(r) \right] + N(r'-r)\log(p) + \Psi \Big\},$$

where

$$\Psi = \begin{cases} (\alpha_r - 1)\log(r'/r) + \beta_r(r - r') & \text{if } r \sim Gamma(\alpha_r, \beta_r); \\ (\alpha_r - 1)\log(r/r') + \beta_r(1/r - 1/r') & \text{if } r \sim InvGamma(\alpha_r, \beta_r). \end{cases}$$

4.2 Two-stages Gibbs sampler

[19] approach exploits a representation of the Negative Binomial as a compound Poisson distribution (a result that dates back to [14]):

$$Y_i^* \sim \text{NB}(r, p) \quad \Longleftrightarrow \quad Y_i^* = \sum_{j=1}^{l_i} u_{i,j}$$

where

$$l_i \sim Poisson(-r\log(p)) \qquad \text{and} \qquad u_{i,j} \overset{iid}{\sim} Logarithmic(1-p).$$

They found the explicit distribution of the full conditional of l_i to be the Chinese Restaurant Table (CRT) distribution with concentration parameter r. Then, the two Gibbs steps are the following:

i) We sample the latent counts, l_i, associated to each observed count y_i^*, which can be generated as:

$$l_i = \sum_{j=1}^{y_i^*} v_j, \qquad v_j \sim Bernoulli\left(\frac{r}{r+j-1} \right).$$

ii) We sample r from its full conditional that, given the conjugacy between the Gamma prior for r and the Poisson distribution, results in

$$[r \mid -] \sim Gamma\left(\alpha_r + \sum_{i=1}^{n} l_i \ , \ \beta_r - N\log(p) \right). \tag{5}$$

Note that, since the total number of captures is often in the order of thousands, and in (5) we are just interested in generating the sum of the l_i, we can simply adopt a Gaussian approximation in the first step. That is,

$$\sum_i l_i \sim N\left(\sum_i E[l_i], \sum_i Var[l_i] \right).$$

4.3 Boundary problem

The use of the NB in capture–recapture is limited by the so called "boundary problem" (see, e.g., [1]). That is, when the estimate of r approaches zero, the Horvitz–Thompson estimation of the population size diverges. More generally, when in the observed (truncated) data the mean number of captures is close to one (which is typically the case in the presence of one–inflation), the NB model severely overestimate N, sometimes by several order of magnitudes, even in simulated data generated by the NB itself. As pointed out in [9], accounting for one–inflation alleviate this phenomenon, but does not completely avoid it.

We can confirm that, even in our Bayesian approach to the OINB model, we encounter the boundary problem. In general, we noted a great sensitivity of the estimates of N to small differences in the value of parameter r, particularly when $r < 1$, and, accordingly, a great sensitivity of the estimates to specification of the prior distribution over r.

We see this phenomenon as an opportunity to investigate the usefulness of the Bayesian approach to further alleviate the boundary problem under the OINB. To this purpose, we conducted a simulation study to assess the effect of different prior specifications on the parameter r. We generate 100 replications of random values drawn from an OINB with parameters $p = 0.35$, $r = 0.5$, and $\omega = 0.5$. N is set to 5000. The observed sample size n varies at each replication; its expected value over the 100 replications is 2040. The values of these parameters are comparable to the values studied in [9], in the frequentist setting. In addition, they are akin to some values actually observed in the real cases analysed in section 5.

We test some prior specifications on the r parameter, considering both the Gamma and the Inverse Gamma distributions. For the estimation of r, we apply both the Metropolis-Hasting step and the two-stages Gibbs sampler proposed by [19], observing negligible differences in the results. The outcomes presented in this Section are obtained using the Metropolis-Hasting approach. Finally, we compare the results with the maximum likelihood estimates for one-inflated Negative Binomial.

Table 3 shows the % relative bias and the % mean square error (MSE) of the population size estimates, considering the difference between the true value and the mean of the posterior distribution obtained by the MCMC simulations. Table 3 also reports the number of cases, in percentage, in which we encountered the boundary problem. In fact, we can define the boundary problem on both $\hat{r}$ and $\hat{N}$. We adopt the following convention for the occurrence of the boundary problem: on $\hat{r}$, we consider to have run into the boundary problem if $\hat{r} < 0.25$; on $\hat{N}$, we consider to have run into the boundary problem if $\hat{N} > 5N$. Finally, in the last row, Table 3 reports the results of the maximum likelihood approach (MLE), obtained using the model proposed by [9] and the R code provided by him as Supporting Information.

The Bayesian procedure implements the algorithm described in section 4.1, setting the number of replications of the MCMC algorithm to $2 \cdot 10^6$. We set, a priori, $p(N) \propto 1/N$, and $Beta(1, 1)$ for both ω and p.

From Table 3, it can be seen that a weakly informative prior specification for r, like $Gamma(1, 1)$ is already useful in reducing the boundary problem, when compared to the MLE approach. A stronger limitation of the boundary problem is achieved by using the Inverse Gamma as prior distribution for r. In the simulation, the Inverse Gamma prior has the double advantage of reducing both the boundary problem and the MSE of the estimates, at the cost of introducing a negative bias

Table 3: Boundary cases for $\hat{r}$ and $\hat{N}$, %bias and %MSE of $\hat{N}$ for some prior specifications of r. Results from MLE in the last row, for comparison

Prior distribution over r	% Boundary cases for r	for N	% bias of $\hat{N}$	% MSE of $\hat{N}$
Gamma(0.1,0.1)	33	30	218.59	1618.82
Gamma(1,1)	11	11	97.64	859.51
InvGamma(0.1,0.1)	0	0	-10.52	6.71
InvGamma(0.5,0.5)	0	0	-15.58	5.13
InvGamma(1,1)	0	0	-19.06	5.27
InvGamma(1,2)	0	0	-26.70	7.91
MLE	16	3	91.75	2217.32

(underestimation) of the population size N.

Note that we used the convention of defining the occurrence of the boundary problem when $\hat{r} < 0.25$, while in [9] the boundary problem is fixed at $\hat{r} < 0.05$. We believe that $\hat{r} < 0.25$ is already enough to indicate the presence of the boundary problem, since as clear from table 3 it corresponds approximately to an estimate of N 5 times larger than its true value.

5. Results on estimating illegal populations

Illegal activities by their nature are difficult to measure because the people involved have obvious reasons to hide these activities. In this Section, we apply our models to estimate the number of people implicated in the exploitation of prostitution, in Italy in 2014. In addition, in Section 5.1 we illustrate the results obtained on some well-known data-sets in capture-recapture literature.

In Italy, prostitution is neither persecuted nor regulated, but trafficking, exploitation, and aiding and abetting of prostitution is a crime, disciplined by law and prosecuted. This activity is mostly managed from foreign organizations, e.g. Chinese, African and East-European. In this study we exploit administrative records from the Ministry of Justice, which report criminal complaints for which the judicial authority has initiated criminal proceedings.

Records in the registers of the Public Prosecutor's offices, contain soft identifiers of the denounced subjects, namely date and place of birth and gender, as well as some characteristics of the denounced subjects and the crimes, like age at the moment of the crime, nationality, the association with other subjects and previous crimes. On the basis of soft identifiers (date, country of birth and gender), perpetrators can be identified and followed over a given time span, that is one year in this application. In this way, the administrative source can be viewed as a list of potential prostitution exploiters and we can observe the number of times an individual is charged. Obviously, we cannot observe the units not captured by the Justice system. We aim to estimate the hidden part of the population, i.e., the size of those unreported to the Public Prosecutor's offices. Capture-recapture models have already been used to investigate prostitution and sex workers, see for instance [16] which estimate the number of street prostitutes in 1986/1987, in Vancouver and [15] which estimate their clients. In this paper, we aim to estimate the size of prostitution procurers, rather than the number of prostitutes or their clients.

Figure 2 shows our data. The total number of observed prostitution exploiters is $n = 2740$, and the number of individuals captured once is $n_1 = 2269$. Counts larger than 5 are quite low, 12 is the maximum number of observed captures.

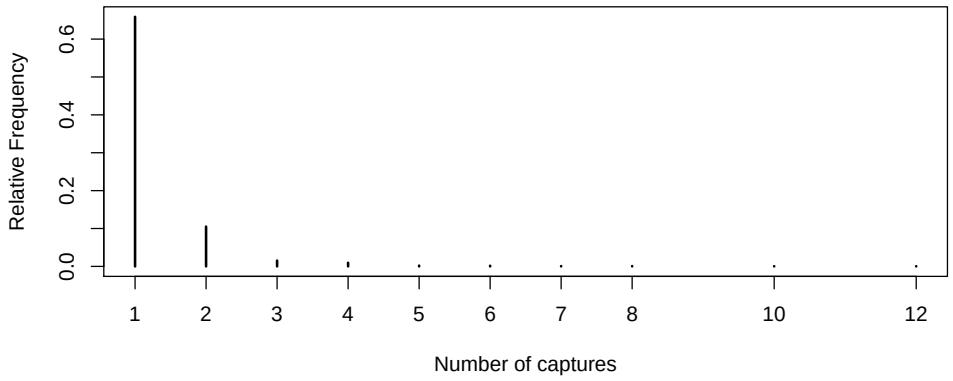

Figure 2: Relative frequencies of observed counts for prostitution exploitation data in Italy in 2014

The presence of one–inflation seems quite evident from figure 2. To test the one–inflation assumption, we calculated the Bayes Factor obtaining a value for the logarithm equal to 101, i.e. a decisive evidence in favor of one–inflation.

Hence, we apply the OIP model proposed in Section 3 to estimate the unknown population size N. We set, a priori, $\lambda \sim Gamma(0.01, 0.01)$, $\omega \sim Beta(1, 1)$ and $p(N) \propto 1/N$. Different values of the parameters for the Gamma prior were also tested, e.g. $\alpha_\lambda = \beta_\lambda = 1$, obtaining very similar results. The number of replications of the MCMC algorithm is 10^6 with a thinning of 20 observations.

Figure 3 shows the estimated posterior distributions of the unobserved population size n_0, λ and ω. The regular shape of the posterior distributions is evident from Figure 3, so the differences in adopting the posterior mode, median or mean are quite negligible. The regularity of the posterior distributions has been consistently observed in all applications and simulations presented in this paper.

In Table 4, we compare our results with other popular approaches. On the upper part of the Table we report the estimates that ignore the one–inflation, i.e., Chao's lower bound estimator, the Zelterman estimator, the Poisson maximum likelihood estimator, (ML.Poisson), and the Poisson Bayesian estimator, (B.Poisson). In the lower part of the Table we report results from models that account for one–inflation, i.e., the maximum likelihood OIP estimator proposed by [11], (ML.OIP), and our Bayesian proposal, (B.OIP).

As expected, if we ignore the one–inflation, we risk severely overestimating the population size, even when using the non-parametric Chao's lower-bound estimator, which is known to be robust to other types of heterogeneity in the data. It should be noted that the use of non-informative priors in the Bayesian context produces estimates that are very close to the ML ones, in both cases, with and without accounting for one–inflation. The data on prostitution exploitation do not show over-dispersion, so the Negative Binomial distribution is not appropriate in this case. In the following Section, we consider some applications where the OINB

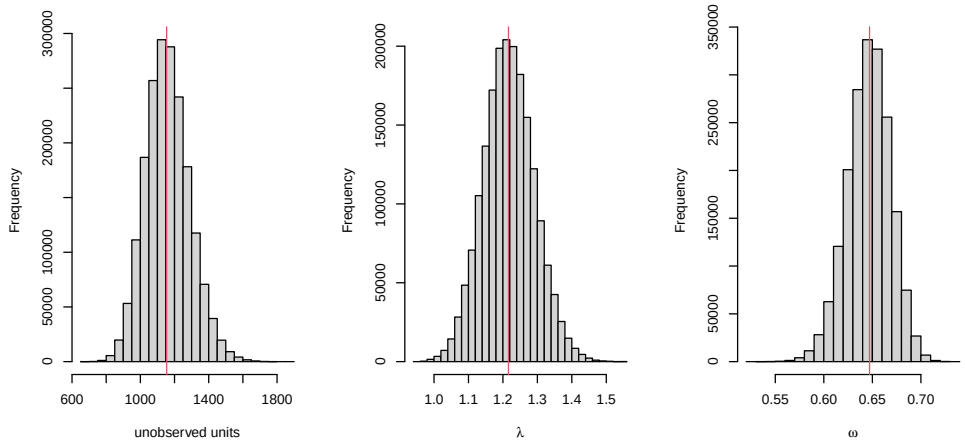

Figure 3: Posterior estimates of the unobserved units n_0, and the parameters λ and ω under one–inflated Poisson for prostitution exploitation data

Table 4: Comparison of population size estimates N, confidence/credible intervals, and the parameters λ and ω for prostitution exploitation data

Estimator/Model	$\hat{N}$	CI.$\hat{N}$	$\hat{\lambda}$	
Ignoring one–inflation				
Chao	9851	8961 - 10868		
Zelterman	10030	9033 - 11027	0.319	
ML.Poisson	7234	6858 - 7680	0.476	
B.Poisson	7214	6783 - 7693	0.476	
Modeling one–inflation				$\hat{\omega}$
ML.OIP	3890	3678 - 4156	1.219	0.648
B.OIP	3889	3652 - 4155	1.212	0.647

Table 5: The posterior mode and credible intervals for the population size N, posterior mean for ω and model parameters, for some popular real cases analysed

1. Prostitutes in Vancouver		$\hat{N}$	HPD($\hat{N}$)	$\hat{\omega}$	$\hat{\lambda}$	$\hat{r}$	$\hat{p}$
Model	Poisson	1237	1178 – 1301		1.253		
	OIP	1016	980 – 1057	0.439	2.037		
	OINB	1157	1017 – 1753	0.327		4.175	0.6862
2. Opiate users in Rotterdam		$\hat{N}$	HPD($\hat{N}$)	$\hat{\omega}$	$\hat{\lambda}$	$\hat{r}$	$\hat{p}$
Model	Poisson	2929	2832 – 3038		1.174		
	OIP	2500	2418 – 2587	0.336	1.663		
	OINB	3753	2965 – 5488	0.100		1.666	0.629
3. Heroin users in Bangkok		$\hat{N}$	HPD($\hat{N}$)	$\hat{\omega}$	$\hat{\lambda}$	$\hat{r}$	$\hat{p}$
Model	Poisson	9453	9427 – 9477		4.134		
	OIP	9364	9349 – 9380	0.207	5.004		
	OINB	10850	10619 – 11109	0.055		1.616	0.301

model is more appropriate.

5.1 Results on some popular case-studies

In this section, we apply the Bayesian model on some well-known datasets in the capture–recapture literature. We consider the following real cases:

1 **street prostitutes in Vancouver:** counts of prostitution arrests made by the Vancouver Police Department Vice Squad for engaging in prostitution in 1986/1987, initially presented and analysed in [16];

2 **opiate users in Rotterdam:** numbers of applications for a methadone treatment program made by opiate users in Rotterdam in 1994, first reported and analysed in [8];

3 **heroin users in Bangkok:** counts of treatment episodes by heroin users in Bangkok in 2002, available in [18] and [4].

In the Vancouver prostitutes dataset, we observe $n = 886$ individuals and the number of units captured once is $n_1 = 541$. The Rotterdam opiate users dataset contains $n = 2029$ units and $n_1 = 1206$. The Bangkok heroins users dataset reports $n = 9302$ observations with $n_1 = 2176$. These data sets have been widely examined in capture–recapture literature, also under the one–inflation hypothesis, (see [11] and [9]). The Bayesian test for model selection introduced in 3.1 is strongly in favor of one–inflation, with a BF greater than 100 in all cases.

We apply our models to the three case–studies, with the following priors setting: For the Poisson and OIP models we set, a priori, $\omega \sim Beta(1, 1)$ and $\lambda \sim Gamma(0.1, 0.1)$. In the OINB model we set $r \sim Gamma(0.1, 0.1)$ and $p \sim Beta(1, 1)$. In all our applications, the number of replications of the MCMC algorithm is 10^6 with a thinning of 20 observations. Standard diagnostics tools confirmed the convergence of the algorithm. The results for all three datasets are summarized in Table 5, which reports the posterior modes and credible intervals of N, and the posterior medians of the model parameters.

The presence of one–inflation in these datasets is less severe than in the prostitution exploitation data analysed in the previous Section. However, as expected,

the Poisson estimates are always larger than the OIP estimates, confirming that we might be overestimating the population size if we ignore one–inflation. In the cases of Vancouver prostitutes and Rotterdam opiate users, for example, the estimate of n_0 under the Poisson model is more than twice that under the OIP model. As for the OINB model, the boundary problem is not an issue with these datasets, since the estimates of r are fairly larger than 1. In the cases considered, the one–inflation rate estimates under the OINB model are always lower than the estimates obtained from the OIP model. It appears that by using the OINB, part of the one-inflation component identified by the OIP is instead explained through the two parameters of the Negative Binomial. Also, OINB's credible intervals are always larger than OIP's, and barely overlap.

Results of Table 5 can be compared with non Bayesian results reported in [11] and [9]. We note that, using weakly informative priors leads to results that are close to the frequentist approach.

6. Concluding remarks and future works

In this paper we presented a Bayesian approach to the analysis of the one–inflated Poisson and one–inflated Negative Binomial in capture–recapture. A fully Bayesian test for the one–inflation assumption has been developed for the Poisson distribution. We discussed the boundary problem of the Negative Binomial distribution, and showed how weakly informative priors can help in stabilizing the estimation procedure.

Currently, we are investigating more general classes of counting distribution and their one–inflated counterparts, to model unobserved heterogeneity.

Moreover, we are investigating different source of one–inflation, deriving from record linkage errors. In fact, when dealing with sensible data which do not share a unique identifier, like the prostitution exploitation data, we may encounter record linkage problems. In this case it would be important to take account for the record linkage process uncertainty in population size estimation, (see [17]).

References

[1] D. Böhning. Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. *Metron*, 73(2):201–216, 2015.

[2] D. Böhning and H. Friedl. Population size estimation based upon zero-truncated, one-inflated and sparse count data. *Statistical Methods & Applications*, pages 1–21, 2021.

[3] D. Böhning, P. Kaskasamkul, and P.G.M. van der Heijden. A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika*, 82(3):361–384, 2018.

[4] D. Böhning, B. Suppawattanabodee, W. Kusolvisitkul, and C. Viwatwongkasem. Estimating the number of drug users in Bangkok 2001: A capture–recapture approach using repeated entries in one list. *European journal of epidemiology*, 19(12):1075, 2004.

[5] D.L. Borchers, S.T. Buckland, W.E. Stephens, and W. Zucchini. *Estimating animal abundance: closed populations*, volume 13. Springer Science & Business Media, 2002.

[6] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[7] Chun-Huo Chiu and A. Chao. Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ*, 4:e1634, 2016.

[8] M. J. Cruyff and P. G. van der Heijden. Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal*, 50:1035–1050, 2008.

[9] R.T. Godwin. One-inflation and unobserved heterogeneity in population size estimation. *Biometrical Journal*, 59(1):79–93, 2017.

[10] R.T. Godwin. The one-inflated positive Poisson mixture model for use in population size estimation. *Biometrical Journal*, 61(6):1541–1556, 2019.

[11] R.T. Godwin and D. Böhning. Estimation of the population size by using the one-inflated positive Poisson model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(2):425–448, 2017.

[12] RE Kass and AE Raftery. Bayes factor and model uncertainty. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[13] R.S. McCrea and B.J.T. Morgan. *Analysis of capture-recapture data*. CRC Press, 2014.

[14] M.H. Quenouille. A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics*, 5(2):162–164, 1949.

[15] J.M. Roberts Jr and D.D. Brewer. Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture–recapture method. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):745–756, 2006.

[16] D.K. Rossmo and R. Routledge. Estimating the size of criminal populations. *Journal of quantitative criminology*, 6(3):293–314, 1990.

[17] A. Tancredi and B. Liseo. A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, 5(2B):1553–1585, 2011.

[18] C. Viwatwongkasem, R. Kuhnert, and P. Satitvipawee. A comparison of population size estimators under the truncated count model with and without allowance for contaminations. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(6):1006–1021, 2008.

[19] M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.