

Bayesian analysis to correct false-negative errors in capture–recapture photo-ID abundance estimates

Cibele Q. da-Silva

Campus Universitário Darcy Ribeiro

Abstract. Capture–recapture methods are largely used for estimating the size of some cetacean populations. The application of those methods for photo-identification data of recognizable individuals is very common. Poor quality photographs may lead the analyst to identify two sightings of the same individual as being different (false-negative errors). This kind of matching error inflates population size estimates. We develop a Bayesian approach to obtain bias corrected estimates of the population size N . The method can be used for M_t type capture–recapture models (Otis et al. *Wildlife Monographs* **62** (1978) 1–135) involving two or more sampling occasions. We used the methodology for simulated data.

1 Introduction

Population size estimation using capture–recapture models for photo-identification data may present some problems related to the identifiability of the individuals. As photo-ID data used for animal recognition is based on natural markings, such as scars, colour pattern, or any trait related to species characteristics that can be used to distinguish the individuals, it could be the case that identification of a poorly marked individual may be overlooked. Consider, for example, a particular poorly marked individual that has already been captured in a previous sampling occasion. If that animal is photographed on a future occasion, but the analyst fails to recognize the correct match, two possibilities of mismatch may occur: (i) the photo may be associated to another individual already included in the data set, leading to a *false-positive error*, or (ii) the analyst can consider the photo as being from a new individual, resulting in a *false-negative error*, and, in this case, a new line in the data set will be created. Abundance estimates based on data which incorporate false-negative errors tend to be positively biased.

Some work to evaluate relative rates and sources of errors in capture–recapture studies using photographic and genetic data were pioneered by Stevick, Palsboll and Allen (1998). The authors worked with data of doubly tagged humpback whales, *Megaptera vaeangliae*. The identification tags used in the study were the natural markings of the whales and a genetic marker. The results of their experiment revealed the occurrence of no false-positive errors either in the genetic or the

Key words and phrases. Capture–recapture, false-negative errors, Bayesian models, multiple imputation.

Received December 2006; accepted February 2008.

photographic data. However, false-negative errors were identified in both kinds of data. The matching protocols make false-positive errors highly unlikely. The study also revealed that photographic error rates were strongly correlated with image quality.

Working with a double-marking experiment using photo-ID data and microsatellite genetic markers to identify humpback whales, Stevick et al. (2001) proposed a correction for abundance estimates based on a Petersen–Lincoln type estimator. Since both photo-ID and genetic data were available for a group of individuals, evaluating false-negative errors was possible. This kind of error was taken into account in order to bias correct abundance estimates.

In this work we suggest a Bayesian approach to correct the positive bias in abundance estimates caused by false-negative errors. The main advantage of the proposed methodology compared to the one suggested by Stevick et al. (2001) is the possibility of allowing for data from more than two sampling occasions.

In Section 2, we introduce some notation. In Section 3, we present two Bayesian procedures meant to bias correct population size estimates of a closed population. The structural model that served as a basis for this work was due to Darroch (1958), representing a M_t type model (see Otis et al. (1978)). In Section 4, we describe a simulation experiment aimed to produce capture–recapture data in several different scenarios. In Section 5, we analyze the data using the proposed methodology.

2 Notation

We now introduce some notation:

- N : population size.
- p_j : the capture probability at time j .
- γ : the probability of correctly identifying a captured individual.
- t : number of sampling occasions.
- X_j : number of good quality photos of naturally marked individuals at time j .
- n_j : observed sample size at time j .
- b_j : actual sample size at time j .
- a_j : true number of resightings in sample j .
- r : observed number of different individuals captured over the experiment.
- h : actual number of different individuals captured over the experiment.
- ω : a subset of $\{1, \dots, s\}$, representing a given capture history.
- u_ω : is the number of individuals with capture history ω .

In this work, the probability $\gamma = 1 - P(\text{false-negative error})$ of correctly identifying a captured individual is considered known and fixed in the sampling occasions.

3 Correcting for identification bias

Consider that for a capture–recapture data analysed using a M_t type model, it is assumed that all individuals are uniquely identified. Thus, according to Darroch (1958), the vector $\{u_\omega\}$ can be modeled using the following expression which depends only on the sufficient statistics h and $\{b_j\}$,

$$P[\{u_\omega\}|\{p_j\}, N] = \frac{N!}{(N-h)! \prod_\omega u_\omega!} \prod_{j=1}^t p_j^{b_j} (1-p_j)^{N-b_j}. \quad (1)$$

In the case of photo-ID data, instead of observing $\{b_j\}$ and h , no rare we observe, respectively, $\{n_j\}$ and r , with $n_j \geq b_j$ and $r \geq h$. This may happen because a given individual that is photographed more than once may not be recognized as so. As a result, a new line in the data set for a supposedly new capture is created, causing positively biased estimates for the population size. The statistics h and $\{b_j\}$ are, actually, *latent data* (unknowns). The same is true for the vector $\{a_j\}$, relating the true number of resightings in the sampling occasions.

The analysis we propose for correcting inflated estimates of N is to consider the latent data as missing observations, using statistical models to fill in likely values for these unobserved data. These generated data will then be taken into account in a capture–recapture model aimed to produce bias corrected estimates of N . Our approach is an adaptation of multiple imputation methods (Rubin and Schenker (1986); Tanner and Wong (1987); Heitjan and Landis (1994)).

The methodological innovation in the present study is the idea of treating the sets of unobserved true values or *missing values* ($\{a_j\}, h$) and ($\{b_j\}, h$) as incomplete data. For each of these sets, different model strategies are suggested for filling in likely values for these unobserved data and attacking the problem of bias correcting N . This methodology is described in Sections 3.1.1 and 3.1.2.

The approach we are considering for bias correcting N presents some similarities with the methodology proposed by Givens, Smith and Tweedie (1997). In this work the authors used a data augmentation methodology within a Bayesian hierarchical model to estimate and adjust for publication bias (expressed by the log of a relative risk) caused either by a researcher that would not submit a nonstatistically significant result for publication or a given study which goes unpublished because it does not contain statistically significant results. In their study, the authors augmented the observed data by simulating the outcomes for the missing data (latent data). The joint posterior distribution involving the parameters in the model and the latent data made it possible to obtain posterior distributions for quantities of interest which were then marginalized across the latent variables. The idea was to obtain improved estimates of the parameters of interest by combining information from the observed data along with the effect of the estimated publication bias.

3.1 Multiple imputation strategy

According to Rubin and Schenker (1986), let Y_{obs} be the set of observed values and Y_{mis} be the set of missing values. The posterior density of a population quantity θ can be written as

$$P(\theta|Y_{\text{obs}}) = \int g(\theta|Y_{\text{obs}}, Y_{\text{mis}}) f(Y_{\text{mis}}|Y_{\text{obs}}) dY_{\text{mis}}, \quad (2)$$

where $f(\cdot)$ is the posterior density of the missing values and $g(\cdot)$ is the complete-data posterior density of θ . Multiple imputations are simulated from the posterior distribution of missing data $f(\cdot)$.

Bearing in mind the scheme above, bias corrected estimates for N will be obtained by using a two-step data augmentation approach as suggested by model (2). In the *first step* we obtain a Bayesian estimate of $Y_{\text{mis}} = (\{a_j\}, h)$ (or $Y_{\text{mis}} = (\{b_j\}, h)$) and, in the *second step*, we obtain a Bayesian estimate of $\theta = (N, \{p_j\})$ using model (1). In this missing data problem a Markov chain is created by iterating the two steps a large number of times. Using a notation introduced in Schafer (1997), at the t th iteration the steps can be defined as:

- Step I (*imputation step*)—Draw $Y_{\text{mis}}^{(t+1)}$ from $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$;
- Step II (*posterior step*)—Draw $\theta^{(t+1)}$ from $P(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$.

Considering some initial value $\theta^{(0)}$, we generate a sequence $(\{Y_{\text{mis}}^{(t)}, \theta^{(t)}\}, t = 1, 2, \dots)$. According to Tanner and Wong (1987), data patterns generated across iterations will converge to the $P(\theta|Y_{\text{obs}})$.

For $Y_{\text{mis}} = (\{a_j\}, h)$ and $Y_{\text{mis}} = (\{b_j\}, h)$, a Poisson model and a binomial model will be used, respectively, for describing $p(Y_{\text{mis}}|Y_{\text{obs}})$. The advantages of each model strategy will be discussed in Section 5.

It is advisable to draw a sample of size k from $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$ each time step I is performed. In this case $Y_{\text{mis}}^{(t+1)}$ would be the average of these k values. Using $k > 1$ (a multiple imputation approach) may improve the inferences. In this work we used $k = 5$.

3.1.1 Modeling $Y_{\text{mis}}|Y_{\text{obs}}$ using a binomial model. Let $Y_{\text{mis}} = (\{b_j\}, h)$ be the vector of latent data, and let the conditional joint distribution of the latent data given the observed data, $Y_{\text{obs}} = (\{X_j\}, \{n_j\}, r, \gamma)$, be such that

$$\begin{aligned} p(Y_{\text{mis}}|Y_{\text{obs}}) &= P(\{b_j\}, h|\{X_j\}, \{n_j\}, r, \gamma) \\ &= P(h|\{n_j\}, \{b_j\}, r, \gamma) \times P(\{b_j\}|\{n_j\}, \gamma) \\ &= \frac{\binom{r}{h} \gamma^h (1-\gamma)^{r-h} \times I[h \in \{\max\{b_j\}, \dots, r\}]}{\sum_{k=\max\{b_j\}}^r \binom{r}{k} \gamma^k (1-\gamma)^{r-k}} \\ &\times \prod_{j=1}^t \frac{\binom{n_j}{b_j} \gamma^{b_j} (1-\gamma)^{n_j-b_j} \times I[b_j \in \{1, \dots, n_j\}]}{\sum_{i=1}^{n_j} \binom{n_j}{i} \gamma^i (1-\gamma)^{n_j-i}}. \end{aligned} \quad (3)$$

In model (3) the truncated binomial distributions reflect the restrictions among $\{b_j\}$, $\{n_j\}$, h , and r . Each correctly identified new individual is considered a *success*, which may happen considering a specific sampling occasion, or considering the whole capture–recapture study.

Step 1: In this step our target distribution is the joint posterior distribution of the latent data, implied by the assumed priors and the conditional likelihood given by the model in expression (3), $\pi(\{b_j\}, h|\{X_j\}, \{n_j\}, r, \gamma)$.

Since $1 \leq b_j \leq n_j$, $j = 1, \dots, t$ and $\max\{b_j\} \leq h \leq r$, we consider uniform prior distributions for b_j , $j = 1, \dots, t$ and for h defined on these respective intervals. Therefore, considering that the b_j 's are independent a priori, since sample size values observed in earlier times should not inform anything about these quantities in future times, the full conditional distributions for b_j , $j = 1, \dots, t$ and for h are, respectively,

$$\begin{aligned} \pi(b_j|\{n_j\}, \gamma) &= p(Y_{\text{mis}}|Y_{\text{obs}})\pi(b_j) \\ &\propto \prod_{j=1}^t \binom{n_j}{b_j} \gamma^{b_j} (1-\gamma)^{n_j-b_j} \times I[b_j \in \{1, \dots, n_j\}], \end{aligned} \quad (4)$$

$$\begin{aligned} \pi(h|\{b_j\}, r, \gamma) &= p(Y_{\text{mis}}|Y_{\text{obs}})\pi(h) \\ &\propto \binom{r}{h} \gamma^h (1-\gamma)^{r-h} \times I[h \in \{\max\{b_j\}, \dots, r\}]. \end{aligned} \quad (5)$$

Step 2: Now our target distribution is the joint posterior distribution associated to $\{p_j\}$ and N , $\pi(\{p_j\}, N|\{u_\omega\}, \{b_j\}, h)$, implied by the convenient priors and model (1). Consider the prior distributions $p_j \sim \text{Beta}(c_1, c_2)$, $j = 1, \dots, t$, and $\pi(N) \propto \frac{1}{N}$, where c_1 and c_2 are taken to express vague prior information. Thus, the full conditional distributions of $N - h$ and p_j are, respectively,

$$N - h|\{p_j\}, h \sim \text{Neg-Bin}\left(h, 1 - \prod_{j=1}^t (1 - p_j)\right), \quad (6)$$

$$p_j|N, b_j, c_1, c_2 \sim \text{Beta}(b_j + c_1, N - b_j + c_2), \quad j = 1, \dots, t. \quad (7)$$

We use the Gibbs sampler to draw observations from all the distributions involved in the Bayesian formulation.

3.1.2 Modeling $Y_{\text{mis}}|Y_{\text{obs}}$ using a Poisson model. Now, let $Y_{\text{mis}} = (\{a_j\}, h)$ be the vector of latent data, and let the conditional joint distribution of the latent data given the observed data, $Y_{\text{obs}} = (\{X_j\}, \{n_j\}, r, \gamma)$, be such that

$$\begin{aligned} p(Y_{\text{mis}}|Y_{\text{obs}}) &= P(\{a_j\}, h|\{X_j\}, \{n_j\}, r, \gamma) \\ &= P(h|\{X_j\}, \{a_j\}, \{n_j\}, r, \gamma) \times P(\{a_j\}|\{X_j\}, \{n_j\}, r, \gamma) \\ &= \frac{(e^{-\phi} \phi^h / h!) I(h \geq \max\{X_j - a_j\})}{\sum_{I[h \in \{\max\{X_j - a_j\}, \dots, r\}]} e^{-\phi} \phi^h / h!} \times \left[\prod_{j=1}^t \frac{e^{-\lambda_j} \lambda_j^{a_j}}{a_j!} \right], \end{aligned} \quad (8)$$

where $\lambda_j = \frac{X_j - n_j}{\gamma}$, $j = 1, \dots, t$, and $\phi = \gamma r$.

In model (8) the truncated Poisson distribution reflects the restrictions among $\{a_j\}$, $\{n_j\}$, h and r . Besides that, following Stevick et al. (2001), who estimated a_j , by $\hat{a}_j = \frac{X_j - n_j}{\gamma}$, $j = 1, 2$, we formulated the conditional distribution of $Y_{\text{mis}}|Y_{\text{obs}}$ so that λ_j , the *average rate of the true number of resightings in sample j* , would be represented by a plausible guess since we are describing λ_j as \hat{a}_j . Additionally, since there is no other information that suggests a better guess for the *average rate of the true number of distinct individuals captured over the experiment*, ϕ , it was assumed to equal γr , where $\gamma = 1 - P(\text{false-negative error})$.

The full conditional posterior distributions for the current approach are obtained in a two-step procedure:

Step 1: In this step, our target distribution is the joint posterior distribution of the latent data, implied by the assumed priors and the conditional likelihood given by the model in expression (8), $\pi(\{a_j\}, h|\{X_j\}, \{n_j\}, r, \gamma)$. Since $X_j - n_j \leq a_j \leq X_j$, $j = 1, \dots, t$ and $\max\{X_j - a_j\} \leq h \leq r$, we consider uniform prior distributions for a_j , $j = 1, \dots, t$ and for h defined on these respective intervals. Therefore, considering that the a_j 's are independent a priori, since the true number of resightings observed in earlier times should not inform anything about these quantities in future times, the full conditional distributions for a_j , $j = 1, \dots, t$ and for h are, respectively,

$$\begin{aligned} \pi(a_j|\{X_j\}, \{n_j\}, r, \gamma) &= p(Y_{\text{mis}}|Y_{\text{obs}})\pi(a_j), \\ &\propto \frac{e^{-\lambda_j} \lambda_j^{a_j}}{a_j!} \times I(a_j \in \{X_j - n_j, \dots, X_j\}), \end{aligned} \quad (9)$$

$$\begin{aligned} \pi(h|\{X_j\}, \{a_j\}, r, \gamma) &= p(Y_{\text{mis}}|Y_{\text{obs}})\pi(h) \\ &\propto \frac{e^{-\phi} \phi^h}{h!} \times I(h \in \{\max\{X_j - a_j\}, \dots, r\}). \end{aligned} \quad (10)$$

Step 2: Now our target distribution is the joint posterior distribution associated to $\{p_j\}$ and N , $\pi(\{p_j\}, N|\{u_\omega\}, \{b_j\}, h)$, implied by the convenient priors and model (1). Since $b_j = X_j - a_j$, $j = 1, \dots, t$, then consider again the prior distributions $p_j \sim \text{Beta}(c_1, c_2)$, $j = 1, \dots, t$ and $\pi(N) \propto \frac{1}{N}$, where c_1 and c_2 are taken to express vague prior information. The full conditional distributions of $N - h$ and p_j are respectively given by expressions (6) and (7).

We used vague beta prior $\text{Beta}(0.5, 0.5)$ for p_j , $j = 1, \dots, t$, since it is known to have a small impact in the Bayesian inferences of capture–recapture data (see da-Silva et al. (2003) and Smith (1991)).

4 Simulation experiment

Capture–recapture simulated data for this study were produced by considering first a *true simulated data set*, TSDDS, where both false-positive or false-negative errors

in the identification process of the individuals were not allowed. In a second instance, the first data set is then debased when one allows the occurrence of false-negative errors only. The resulting data set will be called *error simulated data set*, ESDS. A routine written in S-PLUS was developed to simulate the data.

The TSDS were produced under the following conditions:

- (i) Each individual in a sample had an average of two photos.
- (ii) The probability of taking a good photo of any individual was fixed at 0.75.
- (iii) Four sampling occasions were considered and the capture probabilities were fixed at 0.25, 0.15, 0.15, and 0.10.
- (iv) Population size was fixed at three levels: $N = 1000$, $N = 5000$, and $N = 10,000$.
- (v) For each set of conditions above, 50 TSDS sets were generated.

The ESDS were obtained by corrupting the data sets produced above by considering:

- (a) Three levels of false-negative errors: 0.02, 0.05, and 0.10. Therefore, a total of nine ESDS cases were described by the combination of the three levels of population size and those three levels of false-negative errors.
- (b) Each photo erroneously assigned as being of a *new* individual was never matched to the photos of the same animal or to the photos of any other individual at any time. Such restriction was to prevent the occurrence of positive errors.
- (c) For each TSDS generated under a given set of conditions, 50 ESDS sets were then generated. Therefore, for each of the nine cases above, 2500 ESDS data sets were generated. This procedure enabled us to evaluate the variance of the estimated latent data.

5 Results

Using the methodology presented in Section 3, for each of the 2500 ESDS data sets of each of the nine cases, we obtained the corresponding *bias corrected*, BC, estimates of N , the *true*, TR, estimates of N , which were calculated using the original simulated datasets, and the *error*, ER, estimates of N , calculated using the corrupted datasets. The codes used in the analysis were developed in FORTRAN77.

5.1 Results obtained with the binomial model

Table 1 summarises some results for the binomial model. The notation BC–TR denotes the bias between the estimated N calculated using bias corrected and TSDS generated data. More specifically, for each case and each of the 50 TSDS we calculated the discrepancies between the estimated value of N , calculated from a *true simulated data set*, and each of the 50 corresponding bias corrected estimated values of N . Considering the derived 50 discrepancies, we calculated some summary

Table 1 Summary statistics related to the bias correcting methodology using the binomial model

$P(\text{error})$	Summary statistics	1000		5000		10,000	
		BC-TR	ER-TR	BC-TR	ER-TR	BC-TR	ER-TR
0.02	\bar{x}	19.1	39.7	98.5	202.7	194.0	402.2
	<i>s.d.</i>	2.5	3.7	5.6	7.7	9.8	14.6
	<i>c.v.</i>	0.131	0.09	0.06	0.04	0.05	0.04
0.05	\bar{x}	46.7	101.9	235.3	509.8	471.4	1023.0
	<i>s.d.</i>	5.8	8.4	15.5	22.7	19.1	31.8
	<i>c.v.</i>	0.124	0.08	0.06	0.04	0.04	0.03
0.10	\bar{x}	88.7	210.0	445.0	1050.0	879.0	2086.0
	<i>s.d.</i>	11.3	19.3	28.2	52.3	41.5	65.8
	<i>c.v.</i>	0.127	0.09	0.06	0.05	0.05	0.03

Notes: ER-TR is the bias between estimated N calculated using the ESDS and TSDS generated data. BC-TR is the bias between estimated N calculated using bias corrected and TSDS generated data.

statistics like the average, \bar{x} , standard deviation, SD , and coefficient of variation, CV . In Table 1 notation ER-TR denotes the bias between estimated N calculated using the ESDS and TSDS generated data. That is, for each case and each of the 50 TSDS we calculated the discrepancies between the estimated value of N , calculated from a *true simulated data set*, and each of the 50 corresponding estimated values of N calculated from the debased datasets.

As the values of the summary statistics described in Table 1 relating the BC-TR biases indicate, for all of the nine cases, the methodology suggested in Section 3.1.1 produced improved estimates of N . However, despite the improvement, some positive bias in the bias corrected estimates of N are still observed, though they are considerably smaller than the ones produced by the uncorrected estimates of N . A comparison between the individual biases BC-TR and ER-TR, for each of the nine cases, is shown in Figure 1. As we can observe from such figure, for all population sizes and false-negative error levels, the ER-TR biases are larger than the BC-TR ones. As can be observed from Table 1, as it could be expected, the larger the values of the probabilities of false-negative errors, the larger the observed biases in the estimated values of N . Relatively larger variability among the calculated discrepancies was observed when calculated under the smaller values (0.02 and 0.05) of false-negative error probabilities.

5.2 Results obtained with the Poisson model

Similarly to the analyses presented in the last section, Table 2 summarises some results for the Poisson model. The whole notation remains the same.

As the values of the summary statistics in Table 2 relating the BC-TR and ER-TR biases indicate, for a probability of false-negative error of 0.02, regardless of

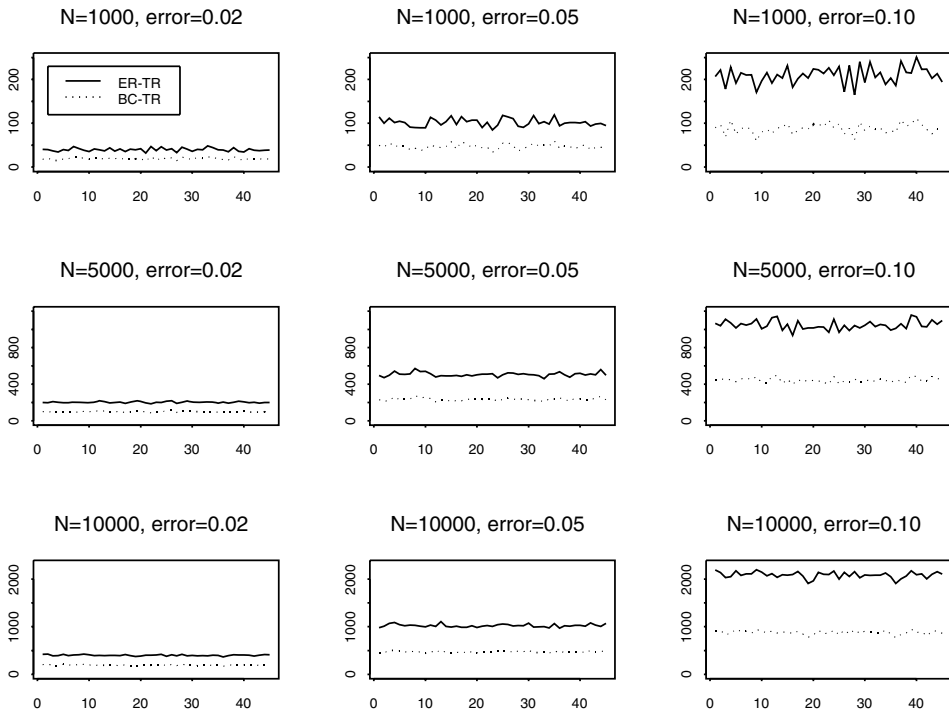


Figure 1 Comparison among the estimated values of N -binomial model.

Table 2 Summary statistics related to the bias correcting methodology—Poisson model

$P(\text{error})$	Summary statistics	1000		5000		10,000	
		BC-TR	ER-TR	BC-TR	ER-TR	BC-TR	ER-TR
0.02	\bar{x}	166.8	39.7	339.9	202.7	490.1	402.2
	<i>s.d.</i>	48.0	3.7	33.1	7.7	33.8	14.6
	<i>c.v.</i>	0.29	0.09	0.10	0.04	0.07	0.04
0.05	\bar{x}	185.7	101.9	305.8	509.8	370.5	1023.0
	<i>s.d.</i>	49.7	8.4	42.5	22.7	31.50	31.8
	<i>c.v.</i>	0.28	0.08	0.14	0.04	0.14	0.03
0.10	\bar{x}	131.6	210.0	136.6	1050.0	198.7	2086.0
	<i>s.d.</i>	39.8	19.3	68.7	52.3	102.9	65.8
	<i>c.v.</i>	0.30	0.09	0.06	0.05	0.51	0.03

Notes: ER-TR is the bias between estimated N calculated using the ESDS and TSDS generated data. BC-TR is the bias between estimated N calculated using bias corrected and TSDS data.

the true population size, the bias correcting methodology using the Poisson model is not profitable. It even accentuates the positive bias in the estimated values of N .

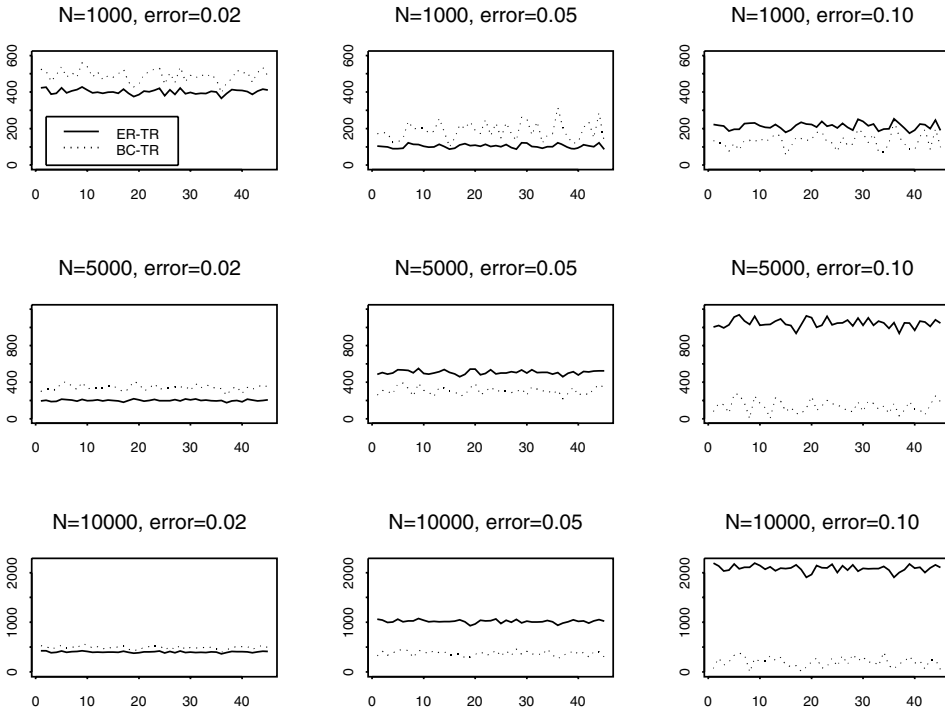


Figure 2 Comparison among the estimated values of N -Poisson model.

Figure 2 illustrates such aspects, since the curves described by the BC–TR biases are always above the ones described by the ER–TR ones. For a probability of false-negative error of 0.05, the proposed methodology is useful only for the population sizes $N = 5000$ and $N = 10,000$, and, in the case of $N = 10,000$, the methodology is more effective than the one obtained when we use the binomial model. Finally, for a probability of false-negative error of 0.10, the bias correcting methodology using the Poisson model is effective for all the population sizes in study. Moreover, for $N = 5000$ and $N = 10,000$, the benefit of using the Poisson model as opposed to the binomial model is evident. For these cases, the comparison of the profiles described in Figures 1 and 2 elucidates these observations, since the Poisson BC–TR biases vary around zero while the binomial BC–TR biases describe profiles varying around y -values vary far from zero. For a given case, an approximate $100(1 - 2\alpha)\%$ credible set may be found using the percentile method which consists of ordering the $B = 2500$ estimates of N , from smallest to largest, and taking as the lower and upper limits $\hat{N}_{(k)}$ and $\hat{N}_{(k')}$, respectively, where $k = (B + 1)\alpha$ and $k' = (B + 1)(1 - \alpha)$, both rounded to the nearest integer value. We obtained 95% credible sets for N . Besides that, for all the cases we estimated the coverage of the credible sets constructed using the described methodology.

Table 3 95% credible sets for N when Poisson and binomial bias correction are employed

Error	Data type	1000	5000	10,000
0.02	<i>error</i>	(881.04; 1203.04)	(4905.07; 5564.92)	(9809.06; 10,891.88)
	%	96	84	74
0.02	<i>B.C.-Poi.</i>	(949.81; 1405.72)	(4901.81; 5608.74)	(9857.77; 10,994.10)
	%	63	66	91
0.02	<i>B.C.-Bin.</i>	(863.34; 1177.04)	(4806.97; 5456.61)	(9612.95; 10,670.81)
	%	98	95	94
0.02	<i>true</i>	(849.98; 1147.99)	(4718.53; 5325.44)	(9429.00; 10,430.00)
	%	95	98	93
0.05	<i>error</i>	(943.46; 1283.88)	(5113.24; 5938.07)	(10,478.64; 11,712.38)
	%	82	28	4
0.05	<i>B.C.-Poi.</i>	(978.93; 1432.12)	(4862.99; 5588.05)	(9746.77; 10,877.13)
	%	58	72	77
0.05	<i>B.C.-Bin.</i>	(895.91; 1219.32)	(4854.44; 5639.68)	(9953.87; 11,124.19)
	%	95	88	62
0.05	<i>true</i>	(860.08; 1143.14)	(4683.16; 5382.58)	(9534.90; 10,572.36)
	%	99	97	95
0.10	<i>error</i>	(1011.11; 1445.50)	(5552.25; 6622.68)	(11,364.57; 12,691.28)
	%	39	0	0
0.10	<i>B.C.-Poi.</i>	(954.32; 1353.67)	(4682.73; 5426.35)	(9556.66; 10,648.63)
	%	71	93	100
0.10	<i>B.C.-Bin.</i>	(912.09; 1301.57)	(4994.51; 5953.28)	(10,231.86; 11,413.78)
	%	90	54	17
0.10	<i>true</i>	(863.32; 1211.47)	(4619.25; 5447.96)	(9429.00; 10,430.00)
	%	94	87	93

Notes: Each interval is based on a sample of 2500 elements, for the cases (1) *error*—generated data with error. (2) *B.C.*—bias corrected. (3) *true*—generated data without error. The percentages refer to the observed coverage based on the 2500 samples.

As can be observed from Table 3, for a probability of false-negative error of 0.02, regardless of the population size, of whether or not bias correction was employed or the model (binomial or Poisson), all the credible sets included the true value of N . The smallest estimated coverage value was 63%, and it was due to a case where the samples were bias corrected using the Poisson model. Nonetheless, the result indicates that the effort to obtain bias corrected estimates of N , when the negative error is that little, may not be worthwhile, even though some improvement could be observed when the binomial model was used.

For a false-negative error of 0.05, Table 3 shows that for $N = 5000$ and $N = 10,000$, the credible sets based on the uncorrected samples did not contain the true N . The respective estimated coverage were very far from the nominal level of 95% (24% and 4%). The use of the bias correcting methodology, either using the Poisson or the Binomial model, considerably improved the estimated coverage, that changed from 28% to 72% (88%) and from 4% to 77% (62%), respectively.

For a probability of false-negative error of 0.10, Table 3 shows that for $N = 5000$ and $N = 10,000$, the respective estimated coverage of the credible sets based on the uncorrected samples was null in both cases. Again, the use of the bias corrected methodology, either using the Poisson or the binomial model, considerably improved the estimated coverage. However, in the present situation, the Poisson approach is undoubtedly superior to the binomial one since for the former model the produced estimated coverages are closer to the nominal level of 95%.

The convergence of the MCMC procedure was verified by Gelman and Rubin (1992)'s convergence diagnostics available in the software CODA. In order to perform the diagnostic, two sequences with 21,000 elements were generated using the procedures described above.

6 Discussion

In this paper, we suggest a Bayesian approach to correct the positive bias in abundance estimates caused by false-negative errors present in capture–recapture data obtained from a demographically closed population. The suggested methodology was developed for the case where the capture probabilities vary only due to temporal effects and the false-negative error probability, γ , is constant and known. Based on the binomial and the Poisson models, two approaches are proposed for bias correcting the estimated values of N .

The main advantage of the proposed methodology compared to the one developed by Stevick et al. (2001) is the possibility of dealing with capture–recapture data from more than two sampling occasions. Besides that, a Bayesian approach allows, through prior specification, that interdependent variables that are constrained to a particular range of values be easily considered in the analysis. Such feature contributed, for example, to a straightforward formulation of the dependencies between the “true” sample sizes, $\{b_j\}$ and the “true” number of distinct individuals observed in the sampling occasions, h .

The proposed methodology is useful when the size of the false-negative error probability is not too small. For false-negative error probabilities of approximately 0.05 or larger, the methodology seems very promising, yielding credible sets close to the nominal level. For population sizes of 5000 and 10,000 and a false-negative error probability of 0.10, the Poisson approach is superior to the binomial one.

Future work includes (1) a more complete study about the guesses for λ_j and ϕ and their influences on the results. (2) The use of hierarchical Poisson models for the formulations described by equations (9) and (10) considering non-informative prior distributions. (3) A study of the impact of the probability of taking a good quality photo (specially for small such probabilities) on the bias corrected estimates of N . (4) The development of methodology for dealing with the estimation of the false-negative error, γ , under distinct sampling occasions and as a function of photo quality. (5) Another important issue is the development of adequate

methodology for bias correcting N when other sources besides the temporal ones affect the capture probabilities.

Acknowledgments

We would like to thank Professor Antonio Eduardo Gomes and the anonymous referees for their careful reading of the manuscript and constructive suggestions which led to several improvements in this work.

References

- Darroch, J. H. (1958). The multiple-recapture census. I: Estimation of a closed population. *Biometrika* **45** 343–359. [MR0119360](#)
- da-Silva, C. Q., Rodrigues, J., Leite, J. G. and Milan, L. A. (2003). Bayesian estimation of the size of a closed population using photo-id data with part of the population uncachable. *Communications in Statistics. Simulation and Computation* **32** 677–696. [MR1998236](#)
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7** 457–511. [MR1196276](#)
- Givens, G. H., Smith, D. D. and Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* **12** 221–250.
- Heitjan, D. F. and Landis, J. R. (1994). Assessing secular trends in blood pressure: A multiple-imputation approach. *Journal of the American Statistical Association* **89** 427, 750–759.
- Otis, D. L., Burnham, K. P., White, G. C. and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* **62** 1–135.
- Rubin, D. B. and Schenker, N. (1986). Multiple imputation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81** 366–374. [MR0845877](#)
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, New York. [MR1692799](#)
- Smith, P. J. (1991). Bayesian analysis of multiple capture–recapture model. *Biometrika* **78** 399–401. [MR1131174](#)
- Stevick, P. T., Palsboll, P. J. and Allen, J. M. (1998). Comparison of capture–recapture results using photographic and genetic identification: Relative rates and sources of errors. WMMSC, Monaco. Abstract volume.
- Stevick, P. T., Palsboll, P. J., Smith, T. D., Bravington, M. V. and Hammond, P. S. (2001). Errors in identification using natural markings: Rates, sources, and effects on capture–recapture estimates of abundances. *Canadian Journal of Fisheries and Aquatic Sciences* **58** 1861–1870.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82** 528–550. [MR0898357](#)

Departamento de Estatística-UnB
 Campus Universitário Darcy Ribeiro
 70910-900—Brasília—DF
 Brazil
 E-mail: cibeleqs@unb.br