Estimating population size with repeated observations of non-invasive sampling in presence of misidentification

Rémi Fraysse[1], Rémi Choquet[1], and Roger Pradel[1]

https://doi.org/10.5802/fake.doi

Abstract

Although non-invasive sampling is increasingly used in capture-recapture (CR) experiments, it carries a risk of misidentification that, if ignored, causes an overestimation of population size. Models that deal with misidentification have been proposed. However, these models assume that only one sample can be collected per individual at one occasion. This is not true for several experiments based on DNA, for example for those that extract the DNA from faecal samples. The models do not take repeated observations into account, leading to biased estimates.

In this paper, we develop an approach that extends the latent multinomial model (LMM) of Link et al., 2010 using a Poisson distribution to account for simultaneous multiple sampling of the same individual. We then conduct simulations to test how our new model performs. As an illustration, we applied the new Poisson model to a collection of Eurasian otter faeces Lampa et al., 2015.

Our findings indicate that repeated observations can be modelled without bias. The application on otters shows that our model is necessary to estimate properly the population size in presence of misidentification and repeated observations.

Keywords: CMR, Natural tags, Faecal DNA, Misidentifications, Multiple captures, Latent multinomial

[1]CEFE, Univ Montpellier, CNRS, EPHE, IRD – Montpellier, France

Correspondence
machin@example.edu

Introduction

One question that must often be answered in conservation management is "How many individuals are there in a population ?". To answer that question, capture-recapture models are powerful methods that require that previously captured individuals be recognised as such. However, unlike in traditional protocols where marks are posed on individuals upon initial capture, with non-invasive protocols such as DNA-sampling of faeces, hair or feathers, this necessarily means individual identification. Studies using DNA-sampling are common and have been carried out on bears (Dreher et al., 2007), bobcats (Morin et al., 2018; Ruell et al., 2009), pronghorns (Woodruff et al., 2016), elephants (Laguardia et al., 2021) or otter (Lampa et al., 2015), for example. While non-invasive sampling allows the study of free-ranging, elusive species without the need to capture, handle or, in some cases, even observe them, compared to traditional tagging methods, it carries a much higher risk of incorrect individual identification (Taberlet et al., 1999). If misidentifications are ignored, then standard models would overestimate population size, potentially by up to five times (Creel et al., 2003).

Proposals to reduce misidentifications from DNA samples range from adapted field methods and improved laboratory techniques (Paetkau, 2003; Waits and Paetkau, 2005) to pre-analysis software that helps filter out data likely to contain errors (McKelvey and Schwartz, 2005). In addition, various approaches have been put forward to model misidentifications when estimating population size (Link et al., 2010; Lukacs and Burnham, 2005; Wright et al., 2009; Yoshizaki et al., 2011). Today, the most common practice is still to filter out under-amplified DNA samples.

However, discarding a non-negligible percentage of poor quality data may result in too little data being retained for reliable estimation of the parameters of interest. In such cases, it may be beneficial to allow a small degree of uncertainty in identification, as suggested by Lukacs and Burnham, 2005 (around 1-5%), as it is now possible to model this error rate (Link et al., 2010; Lukacs and Burnham, 2005; Wright et al., 2009; Yoshizaki et al., 2011). If the cost of adding a parameter (the error rate) is offset by the additional number of samples that can be retained, then the trade-off is of interest.

Of the approaches that incorporate the misidentification process into the analysis, Yoshizaki's model (Yoshizaki et al., 2011) is the simplest, at the cost of discarding some data. It consists of removing all single-capture histories, allowing more straightforward treatment of the uncertainty in the remaining dataset. However, instead of deleting data, we want to model misidentifications to keep as much data as possible. So this model appears of little use when the amount of data that would be discarded is large.

Wright's model (Wright et al., 2009) uses the full dataset, but requires genotype replicates to estimate an error rate. This approach increases the costs, especially for large populations for which many samples may be obtained.

The Latent Multinomial Model (LMM, Link et al., 2010) is a malleable framework that has received attention and has been extended by other publications (Bonner et al., 2015; McClintock et al., 2014; Schofield and Bonner, 2015). It estimates, in a Bayesian framework, the misidentification rate without the need of additional information beyond the capture histories. However, one of the main hypothesis of the LMM is that individuals can only be seen once per occasion. In studies where DNA is extracted from faeces to identify the individuals, the individuals can be 'captured' multiple times at the same occasion. Using the LMM to estimate the population size would lead to an overestimation.

In this paper, we develop an extension of the LMM to account for repeated observations of an individual on the same occasion. To illustrate the problem and our solution, we take the example of a study of the Eurasian otter (*Lutra lutra*) in Upper Lusatia, Saxony, Germany (Lampa et al., 2015). We first present the otter dataset data as well as simulated data that cover more general situations. We then present the original model from Link et al., 2010 and our extension.

Material and methods

Otter data

Otters are nocturnal and elusive and pose challenges for (live-)trapping (Kruuk, 2006). Otter faeces (spraints) on the other hand are particularly conspicuous as they are used for intraspecific communication. Otters produce up to 30 spraints a day, tending to defecate on frequently visited visible terrestrial sites at specific locations throughout their home range. Thus, these spraints can be easily sampled for DNA-identification of the individuals. Lampa et al., 2015 have collected spraints over five consecutive days ($T = 5$) from 2006 to 2012 (excluding 2009). Sampling was conducted in March (2006, 2010, 2011, 2012), April (2007) and May (2008). The datasets from each year were considered independently so the population can be considered closed (no birth, death, immigration, emigration) over 5 five days. The authors considered it unlikely that repeated PCR could completely eliminate all genotyping errors due to relatively high genotyping error rates and low genotyping success rates. Thus, they used the error-incorporating misidentification model proposed by Lukacs and Burnham, 2005 (hereafter, model L&B), implemented in the MARK program (White and Burnham, 1999). However, this model does not adequately address ghost capture histories resulting from misidentification (Link et al., 2010; Yoshizaki et al., 2011).

Simulations

To validate our approach for a broader range of situations than the otter case, we conducted a simulation study. We simulated datasets for a population of $N = 500$ individuals. The captures were sampled from Poisson distributions with time-dependent parameters λ_t. We aimed for four different capture rates (with the meaning 'seen at least once' or 'not seen at all' on one occasion) of 0.1, 0.2, 0.3 and 0.4, so we set $\lambda_t = 0.11, 0.23, 0.36, 0.51$ respectively (so on average, with $\lambda_t = 0.11$, 10% of the individuals will be seen at least once and 90% will remain unseen). Three different simulated identification rates were used with values $\alpha = 0.8, 0.9, 0.95$. We simulated 10 datasets for these 12 scenarios.

Closed population capture recapture: model $\mathbf{M_t}$

When estimating population size N in a closed capture-recapture experiment (i.e. the population is assumed not to change), with the model known as M_t (Darroch, 1958; Otis et al., 1978), individuals are assumed to be observed ('captured') with probability p_t at occasion t for $t = 1, 2, \ldots, T$ and identified individually through tags/tracking devices or natural markings. Capture events are assumed independent between individuals and over time.

For each occasion, an individual is assigned 0 if it was not captured, or 1 if it was. This leads to 2^T possible distinct histories, including the unobservable all-zero history. They are represented by the sequence $\boldsymbol{\omega}_i = (\omega_{i,1}, \ldots, \omega_{i,T})$. Here, y_i is the number of individuals with history ω_i and

$\mathbf{y} = (y_1, y_2, ..., y_{2^T-1})$. $\mathbf{y}$ follows a multinomial distribution with index N and cell probability

$$\pi_i = \prod_{t=1}^{T} \left[p_t^{I(\omega_{i,t}=1)} (1-p_t)^{I(\omega_{i,t}=0)} \right] \tag{1}$$

where $I(test)$ is 1 if *test* is true, and is 0 otherwise.

Modelling misidentifications: model $\mathbf{M_{t,\alpha}}$

To account for individual misidentifications, Yoshizaki et al., 2011 proposed the model $M_{t,\alpha}$, in which captured individuals are correctly identified with probability α. Misidentifications are assumed to create a new individual (a 'ghost'). An individual cannot be mistaken as another, and two misidentifications cannot create the same ghost. Link et al., 2010 developed a latent structure for the model $M_{t,\alpha}$, which allows for Bayesian parameter estimation. In this structure, misidentifications are represented by 2 in the latent error histories $\boldsymbol{\nu}_j = (\nu_{j,1}, ..., \nu_{j,T})$. The frequency of the latent error history ν_j is noted x_j, and the vector of all latent error frequencies is $\mathbf{x} = (x_1, ..., x_{3^T})$. To make future developments of the model easier, the observation process of the likelihood is broken down into two parts, the capture process and the identification process. We followed Bonner et al., 2015 by introducing latent capture histories $\boldsymbol{\xi}_k = (\xi_{k,1}, ..., \xi_{k,T})$. These are the true capture histories, i.e. in the absence of individual misidentifications, composed of 0 and 1. The frequency of the latent capture history ξ_k is noted z_k, and the vector of all latent capture frequencies is $\mathbf{z} = (z_1, ..., z_{2^T})$.

In this model framework, the observed frequencies vector $\mathbf{y}$ is a known linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$ of the latent error frequencies $\mathbf{x}$ for a given matrix $\mathbf{A}$. The constraint matrix $\mathbf{A}$ is $(2^T - 1) \times 3^T$ with a 1 in row i and column j if the latent error history j gives rise to the observed i. All the other entries are zeros. The latent capture frequencies vector $\mathbf{z}$ is another linear transformation $\mathbf{z} = \mathbf{B}\mathbf{x}$ for a given matrix $\mathbf{B}$. $\mathbf{B}$ is $2^T \times 3^T$ with 1 at row k and column j if the latent capture history ξ_k and the latent error history ν_j have the same capture pattern.

The joint likelihood of $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{z}$ is

$$[\mathbf{y}|\mathbf{x}, \mathbf{z}, N, p, \alpha] = I(\mathbf{y} = \mathbf{A}\mathbf{x})\, [\mathbf{x}|\mathbf{z}, \alpha]\, [\mathbf{z}|N, \mathbf{p}] \tag{2}$$

The probability of the capture process $[\mathbf{z}|N, \mathbf{p}]$ is the same as the closed CR model M_t, using histories ξ and frequencies $\mathbf{z}$. The capture likelihood is the following multinomial product where π_k is computed as in Equation 1, using history ξ instead of ω:

$$[\mathbf{z}|N, \mathbf{p}] = \frac{N!}{\prod_k z_k!} \prod_k \pi_k^{z_k} \tag{3}$$

Bonner et al., 2015 gives the likelihood of the identification process, knowing the real captures:

$$[\mathbf{x}|\mathbf{z}, \alpha] = I(\mathbf{z} = \mathbf{B}\mathbf{x}) \frac{\prod_k z_k!}{\prod_j x_j!} \prod_j \left[\prod_{t=1}^{T} \alpha^{I(\nu_{j,t}=1)} (1-\alpha)^{I(\nu_{j,t}=2)} \right]^{x_j} \tag{4}$$

The ratio of factorials accounts for the relabelling of the marked individuals that would produce the same counts in $\mathbf{x}$ and $\mathbf{z}$.

The full likelihood is obtained by summing the conditional $[\mathbf{y}|\mathbf{x}, \mathbf{z}, N, p, \alpha]$ over all values of $\mathbf{x}$ belonging to the set $\mathcal{F}_{\mathbf{y}} = \{\mathbf{x}|\mathbf{y} = \mathbf{A}\mathbf{x}\}$:

$$[\mathbf{y}|N, p, \alpha] = \sum_{x \in \mathcal{F}_{\mathbf{y}}} [\mathbf{y}|\mathbf{x}, \mathbf{z}, N, p, \alpha] \tag{5}$$

Note that there is no summation over $\mathbf{z}$ in this equation because $\mathbf{z}$ is the result of a deterministic function of $\mathbf{x}$ ($\mathbf{z} = \mathbf{B}\mathbf{x}$).

The feasible set $\mathcal{F}_{\mathbf{y}}$ is complicated to enumerate, which makes the likelihood (Equation 5) almost intractable in terms of computation. Maximum likelihood estimation is therefore not practical. Conveniently, Link et al., 2010 and Bonner et al., 2015 show how a Markov chain Monte Carlo (MCMC) can be used in this context. The Markov chain allows for the estimation of the posterior density:

$$[N, \mathbf{p}, \alpha|\mathbf{y}] \propto [\mathbf{y}|N, \mathbf{p}, \alpha]\,[N]\,[\mathbf{p}]\,[\alpha], \tag{6}$$

where $[N]$, $[\mathbf{p}]$ and $[\alpha]$ denote the priors on population size, capture probability and identification probability.

Link et al., 2010 and Bonner et al., 2015 show how a Markov chain Monte Carlo (MCMC) can be used in this context. The Markov chain allows for the estimation of the posterior density. The complete algorithm is detailed in Section A of the Supplementary Material.

Extension to repeated observation: model $M_{\lambda,\alpha}$

In this section, we consider the case in which several captures of the same individual can occur at the same occasion. In this case, the observable history ω of the individual is composed of counts representing the number of times an individual was observed and identified at each occasion. For example, we might observe the following individual history: $(0, 2, 0, 3, 1)$, where 2 is the number of times the individual is detected at occasion 2. In an experiment in which individuals can be 'captured' several times on one occasion, one of the captures might result in a misidentification and the creation of a ghost history, while the other captures are correct identifications. In contrast to the model $M_{t,\alpha}$, an individual producing a ghost at a given occasion can be seen at the same occasion if another capture of it results in a correct identification. Due to the new structure of the data, we must modify the notations used for misidentifications: we cannot use '2' for indicating a misidentification. In addition, a history could contain several misidentifications for the same occasion. To represent latent error histories ν_j, we note the total number of observations of an individual on each occasion with the number of these observations that resulted in misidentification as a superscript. For example, the previous observed history might have been generated by the latent error history $(1^{(1)}, 2, 0, 3, 3^{(2)})$. In this example, the observation at the first occasion and two observations at the fifth occasion were misidentified, resulting in zero observations at occasion 1 and one observation at occasion 5. The latent capture history ξ_k is the same as the latent error history without the superscripts. In our example, the latent capture history is $(1, 2, 0, 3, 3)$.

In this section, we make the same assumptions about the misidentifications as in the model $M_{t,\alpha}$. Identifications are independent, misidentifications always result in ghosts (i.e. false individuals), and ghosts cannot be seen again (i.e. ghosts have exactly one observation in their history).

As in the model $M_{t,\alpha}$, the observed histories of an experiment can be summarized in the frequency vector $\mathbf{y}$. The different latent error histories possibly responsible for the observed

	ν_j	$(1,2)$	$(0,3)$	$(1,0)$	$(2^{(1)},2)$	$(1^{(1)},3)$	$(1^{(1)},0)$
ω_i	$\mathbf{y}/\mathbf{x}$	2	2	0	1	0	0
$(1,2)$	3	1	0	0	1	0	0
$(0,3)$	2	0	1	0	0	1	0
$(1,0)$	1	0	0	1	1	1	1

Table 1 – The first and second columns are, respectively, the observed histories and the observed history frequencies. The first and second rows are, respectively, the possible latent error histories and the possible latent error history frequencies. The rest of the table is the matrix **A**.

	ν_j	$(1,2)$	$(0,3)$	$(1,0)$	$(2^{(1)},2)$	$(1^{(1)},3)$	$(1^{(1)},0)$
ξ_k	$\mathbf{z}/\mathbf{x}$	2	2	0	1	0	0
$(1,2)$	2	1	0	0	1	0	0
$(0,3)$	2	0	1	0	0	1	0
$(1,0)$	0	0	0	1	1	1	1
$(2,2)$	1	1	0	0	1	0	0
$(1,3)$	0	0	1	0	0	1	0

Table 2 – The first and second columns are, respectively, the latent capture histories and the latent capture history frequencies. The first and second rows are, respectively, the latent error histories and the latent error history frequencies. The rest of the table is the matrix **B**.

histories can be summarized in the frequency vector $\mathbf{x}$. As in the model $M_{t,\alpha}$, the vector $\mathbf{y}$ is the linear transformation of the vector $\mathbf{x}$: $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a known matrix. Table 1 shows an example of $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{A}$ in the case where the three histories $(1,2)$, $(0,3)$ and $(1,0)$ were observed, and there is only one history $(1,0)$. In this case, there is only one possible ghost, but we do not know which history might have generated it.

Similarly, the latent capture histories can be summarized in the frequency vector $\mathbf{z}$. Also, $\mathbf{z}$ is a linear transformation of $\mathbf{x}$: $\mathbf{z} = \mathbf{B}\mathbf{x}$. Table 2 illustrates the example from above, showing possible $\mathbf{z}$ and $\mathbf{B}$.

Here, $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_T)$ is the set of parameters involved in the capture process, modelled by a Poisson process (each λ_t being a parameter of the Poisson process at occasion t). The likelihood of the model, jointly to $\mathbf{x}$ and $\mathbf{z}$, is given by the Equation 7.

$$[\mathbf{y}, \mathbf{x}, \mathbf{z}|N, \boldsymbol{\lambda}, \alpha] = I(\mathbf{y} = \mathbf{A}\mathbf{x})\,[\mathbf{x}|\mathbf{z}, \alpha]\,[\mathbf{z}|N, \boldsymbol{\lambda}] \tag{7}$$

We now describe the two elements that change compared to model $M_{t,\alpha}$, i.e. $[\mathbf{z}|N, \boldsymbol{\lambda}]$ and $[\mathbf{x}|\mathbf{z}, \alpha]$.

- $[\mathbf{z}|N, \boldsymbol{\lambda}]$

We model the true number of observations of an individual at an occasion with a Poisson distribution. The probability π_k that an individual has a given latent capture history ξ_k is given by:

$$\pi_k = \prod_{t=1}^{T} \frac{\boldsymbol{\lambda}_t^{\xi_{k,t}}}{\xi_{k,t}!} e^{-\lambda_t} \tag{8}$$

The capture likelihood has a multinomial form:

$$[\mathbf{z}|N, \boldsymbol{\lambda}] = \frac{N!}{\prod_k z_k!} \prod_k \pi_k^{z_k} \tag{9}$$

- $[\mathbf{x}|\mathbf{z}, \alpha]$

All actual captures are potentially subject to misidentifications. If $o_{j,t}$ is the number of good identifications for individuals with history ν_j at occasion t, then, knowing the true number of captures, the probability that it was correctly identified $o_{j,t}$ times is binomial. Thus:

$$[\mathbf{x}|\mathbf{z}, \alpha] = \prod_j \prod_{t=1}^{T} \binom{\nu_{j,t}}{o_{j,t}} o_{j,t}^{\alpha} (\nu_{j,t} - o_{j,t})^{1-\alpha} \tag{10}$$

In addition to this model, we also adapted the algorithm to estimate the parameters. The full algorithm is presented in Section B of the Supplementary Material.

Implementation and MCMC configuration

We used NIMBLE (Valpine et al., 2017) to implement the models. The advantage of NIMBLE is that it allows writing all the samplers of the MCMC (mandatory here) and new distributions. Thus, we wrote the likelihood of the model and the sampler of **x**. We were also able to rewrite all the Gibbs samplers of the MCMC for maximum computational efficiency. To improve efficiency, all observable histories with a count of zero were ignored, i.e. their corresponding rows and columns in the matrices **A** and **B** were deleted, as suggested in Schofield and Bonner, 2015.

For the simulations MCMC, the $M_{\lambda,\alpha}$ model was run over 1E6 iterations after a burn-in period of 100,000 iterations, and the chains were thinned by a factor of 1/100. Yoshizaki's model was run over 10,000 iterations after a burn-in period of 1000 iterations. The M_t model was run over 100,000 iterations after a burn-in period of 10,000 iterations, and the chains were thinned by a factor of 1/10.

For the otter MCMC, the $M_{\lambda,\alpha}$ model was run over 10,000 iterations after a burn-in period of 1,000 iterations, and the chains were thinned by a factor of 1/10 in order to limit memory usage. Yoshizaki's model was run over 10,000 iterations after a burn-in period of 1000 iterations. The M_t model was run over 10,000 iterations after a burn-in period of 1,000 iterations, and the chains were thinned by a factor of 1/10.

We always ran two chains with two different starting points. The first starting point had **x** initialized as the set of observed histories, as if there were no errors. In the second starting point, we arbitrarily added 40 random errors for the simulations and 5 for the otter. Previous tests indicated that two chains were sufficient to see convergence.

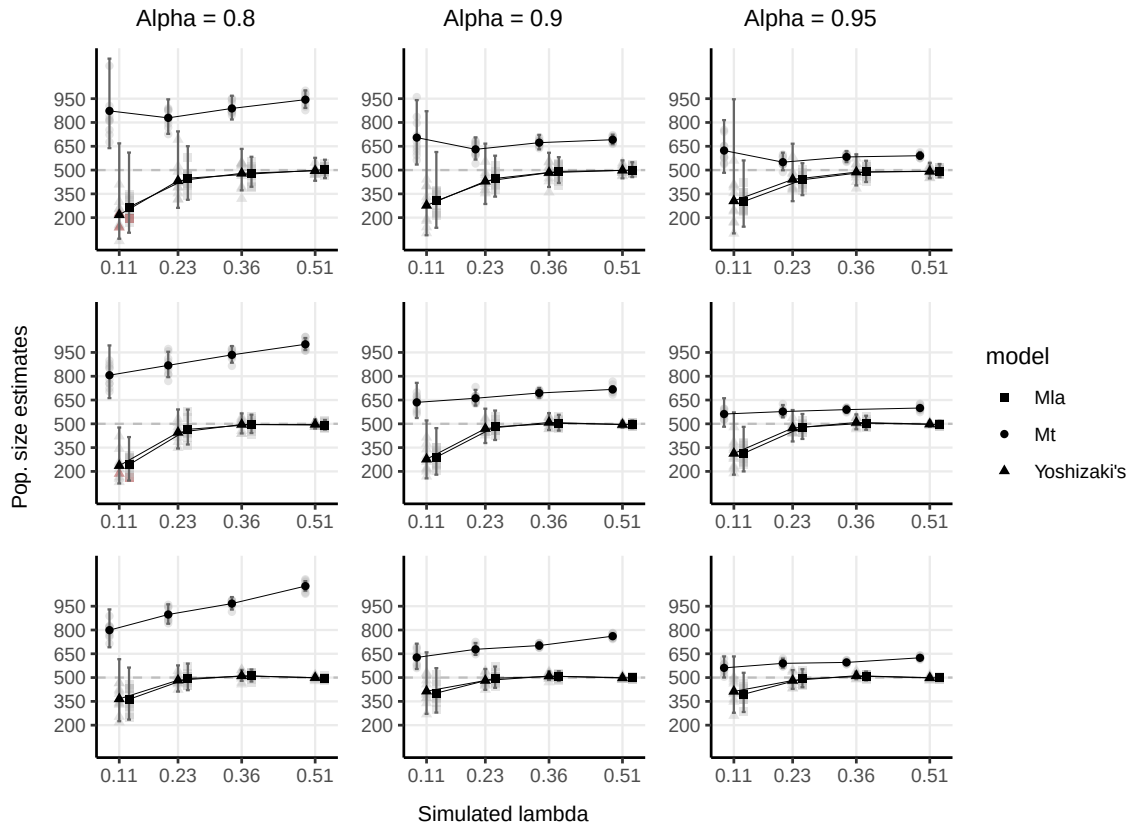
Results

Simulations results

We evaluated our extension of the LMM against the model proposed by Yoshizaki et al., 2011, in which all histories with a single capture are excluded. We also tested the standard capture-recapture model M_t to demonstrate the importance of disregarding misidentifications.

Running two chains of 1,100,000 iterations for the model $M_{\lambda,\alpha}$ on a 3.0GHz Intel processor took less than 10 minutes. We checked the convergence graphically by looking at the N chains, as N is the slowest moving parameter. We also looked at the $\hat{R}$ and the resulting effective sample size. The model M_t always converged perfectly and there was no evidence of any problems. Yoshizaki's model converged for all simulations, but some chains for $p = 0.1$ seemed to be constrained by a lower bound. For the $M_{\lambda,\alpha}$, the MCMC converged in all scenarios. The convergence was always faster for higher identification rates.

Figure 1 – Multi-capture CMR population size estimations (y axis) depending on capture probability (x axis), identification probability (columns), number of capture occasions (lines) and model used (dot shape). The model $M_{\lambda,\alpha}$ is noted Mla. Horizontal dashed lines indicate true population size. Grey and red symbols show simulation-specific estimates of the population-size posterior mean. Red symbols indicate that the N chains had $\hat{R} > 1.1$ or an effective sample size under 50. Black symbols connected by lines show the mean estimates averaged across simulations. Error bars show the estimate of the 97.5% and 2.5% quantiles averaged across simulations.



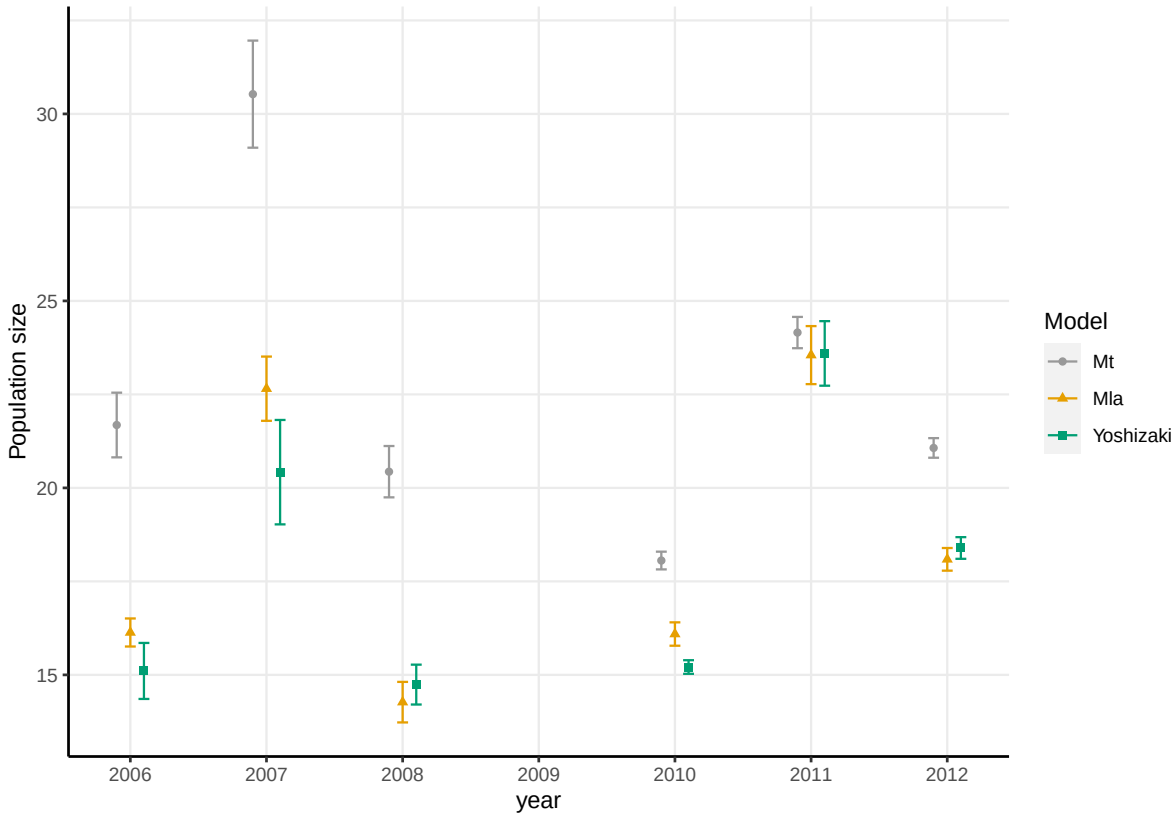
The population size estimates for all models are shown in Figure 1. The model M_t always overestimates the population size, up to a factor of two. Lower identification rates and higher capture rates both lead to greater overestimation. Yoshizaki's model and the $M_{\lambda,\alpha}$ model have the same trend of estimates. For very low capture rates, the chains seem to minimize the population size, resulting in an unlikely very low identification probability estimate. For a capture rate of 0.2, there is only a small bias with seven or fewer occasions, and no bias with nine occasions. There was no bias for higher capture rates. The credible interval estimated by the model $M_{\lambda,\alpha}$ is smaller than that estimated by Yoshizaki's model for most scenarios with seven or fewer occasions. They are slightly larger with nine occasions, but the true population size is more often in the credible interval estimated by the model $M_{\lambda,\alpha}$.

Otter results

We used three models to analyse the data: our new $M_{\lambda,\alpha}$ model, Yoshizaki's model (discarding single capture histories) and the standard model M_t that ignores misidentifications.

The estimates are shown in Table 3 and on Figure 2. For 2011, all models estimate a similar population size. For all other years, the M_t model overestimates the population size compared to the other models. For most years, the $M_{\lambda,\alpha}$ model and Yohizaki's model have similar estimates.

Figure 2 – Otter population size estimations.



	Model	2006	2007	2008	2010	2011	2012
N	Yoshizaki's	15.1 ± 0.8	20.4 ± 1.4	14.7 ± 0.5	15.2 ± 0.2	23.6 ± 0.9	18.4 ± 0.3
	$M_{\lambda,\alpha}$	16.1 ± 0.4	22.7 ± 0.9	14.3 ± 0.5	16.1 ± 0.3	23.6 ± 0.8	18.1 ± 0.3
	M_t	21.7 ± 0.9	30.5 ± 1.4	20.4 ± 0.7	18.1 ± 0.2	24.2 ± 0.4	21.1 ± 0.3
	L&B	19.0 ± 2.6	24.0 ± 3.5	19.4 ± 2.2	15.4 ± 2.1	25.1 ± 1.8	21.5 ± 2.2
$N\alpha$	$M_{\lambda,\alpha}$	5.8	7.5	7.8	4.9	2.5	5.9
	L&B	8.3	12.9	4	17.6	3.9	13

Table 3 – Closed population model estimates of the otter population in Upper Lusatia (Saxony, Germany). N is the population size, and here $N\alpha$ indicates the number of misidentifications. The values are the mean estimate $\pm$ the standard error.

The minimum population size (numbers of observed individuals that cannot be ghosts considering the hypothesis of the LMM) are 16, 22, 14, 16, 23, 18 for each year respectively. For 2006, 2007 and 2010, the population size estimated by Yoshizaki's model is under this minimum. Considering the hypothesis of the LMM, the maximum number of misidentifications were 6, 8, 8, 5, 3 and 6 for each year respectively. For all years, the model $M_{\lambda,\alpha}$ estimated that all the histories with a single capture were probably ghosts. In contrast, for most years, the mean estimate identification rates from model L&B indicated more misidentifications than our assumptions allowed. Other histories than those with only one capture would need to contain misidentifications.

Discussion

In this paper, we treat the ecological question of population estimation with capture-recapture when the individuals are identified with DNA extracted from faeces samples. We develop an extension of the LMM Link et al., 2010 to account for repeated observations on an occasion. We apply our model to an otter dataset and conduct a simulation study to demonstrate its validity and usefulness.

For all years except 2011 the otter dataset certainly contained misidentifications since Yoshizaki's model and our $M_{\lambda,\alpha}$ model agreed. Ignoring the misidentifications leads to an estimate up to 30% higher than when accounting for misidentifications. This example, and the overestimation of the population size in the simulations, should be sufficient to entice researchers into accounting for eventual misidentifications.

Yoshizaki's model relies on deleting unsure data but it also lead to delete sure and available data. For example, histories of individuals seen at only one occasion but several times (like "03000") cannot be ghosts. However these histories are still removed from the data. All additional captures, are also removed. For example, the history "20130" is simplified into "10110". This loss of information can lead to an underestimation of the population size, as it happened with the 2006, 2007 and 2010 datasets. It might also be responsible for higher uncertainty observed on the simulation study. In contrast, our model can make use of all the available information to get better estimates of the population size.

Compared to the simulations we conducted, the otter dataset was on the favourable side for good estimates. The capture rates were higher than what were tested in the simulations, and the identification rate was around the highest tested (95%). However, the population size was very small compared to the simulations.

The problem addressed in this paper is that of population size estimation. However, many studies focus on estimating survival rather than population size and face the same problem of potential misidentifications. Therefore, it seems interesting and important to develop an extension of the LMM that would estimate survival with repeated observation.

The use of DNA from samples such as faeces is a great tool used in many capture recapture studies. The risk of misidentification, most times avoided through discarding samples can now be efficiently modelled. Using such a model can greatly benefit, from simply confirming the absence of misidentification to modelling them and keeping data that would have been removed otherwise.

Acknowledgements

We would like to thank Daniel B. Turek at Williams College Department of Mathematics and Statistics, without whom the NIMBLE implementation would have taken much more time. Preprint version xxx[change to the correct number] of this article has been peer-reviewed and recommended by Peer Community In XYZ[change to the name of the PCI] (https://doi.org/10.24072/pci.xxx[replace by the doi of the recommendation]; **fake**[replace by the citation of the recommendation]).

Fundings

This study was supported as a part of the French National Agency for Research project MoVe–ADAPT.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data, script, code, and supplementary information availability

Otter data from Lampa et al., 2015 are available online (https://doi.org/10.1371/journal.pone.0125684.s002)
Script and codes are available online (https://github.com/RemiFray/reprodPaperLMMRepObs [create DOI and replace if accepted)
Supplementary information is available online (https://doi.org/10.24072/fake3)

References

Bonner S, Schofield M, Noren P, Price S (2015). *Extending the latent multinomial model with complex error processes and dynamic markov bases*. *The Annals of Applied Statistics* **10**. https://doi.org/10.1214/15-AOAS889.

Creel S, Spong G, Sands JL, Rotella J, Zeigle J, Joe L, Murphy KM, Smith D (2003). *Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes*. *Molecular Ecology* **12**, 2003–2009. https://doi.org/10.1046/j.1365-294X.2003.01868.x.

Darroch JN (1958). *The Multiple-recapture census: I. estimation of a closed population*. *Biometrika* **45**, 343–359. https://doi.org/10.2307/2333183.

Dreher BP, Winterstein SR, Scribner KT, Lukacs PM, Etter DR, Rosa GJM, Lopez VA, Libants S, Filcek KB (2007). *Noninvasive estimation of black bear abundance incorporating genotyping errors and harvested bear*. *The Journal of Wildlife Management* **71**, 2684–2693. https://doi.org/10.2193/2006-398.

Kruuk H (2006). *Otters: Ecology, behaviour and conservation*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198565871.001.0001.

Laguardia A, Bourgeois S, Strindberg S, Gobush KS, Abitsi G, Bikang Bi Ateme HG, Ebouta F, Fay JM, Gopalaswamy AM, Maisels F, Simira Banga Daouda ELF, White LJT, Stokes EJ (2021). *Nationwide abundance and distribution of African forest elephants across Gabon using non-invasive SNP genotyping*. *Global Ecology and Conservation* **32**, e01894. https://doi.org/10.1016/j.gecco.2021.e01894.

Lampa S, Mihoub JB, Gruber B, Klenke R, Henle K (2015). *Non-Invasive Genetic Mark-Recapture as a Means to Study Population Sizes and Marking Behaviour of the Elusive Eurasian Otter (Lutra lutra)*. *PLOS ONE* **10**. Publisher: Public Library of Science, e0125684. https://doi.org/10.1371/journal.pone.0125684.

Link WA, Yoshizaki J, Bailey LL, Pollock KH (2010). *Uncovering a latent multinomial: analysis of mark-recapture data with misidentification*. *Biometrics* **66**, 178–185. https://doi.org/https://doi.org/10.1111/j.1541-0420.2009.01244.x.

Lukacs PM, Burnham KP (2005). *Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error*. *The Journal of Wildlife Management* **69**, 396–403. https://doi.org/https://doi.org/10.2193/0022-541X(2005)069<0396:EPSFDC>2.0.CO;2.

McClintock BT, Bailey LL, Dreher BP, Link WA (2014). *Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentification*. *arXiv e-prints* **1401**, arXiv:1401.3290.

McKelvey KS, Schwartz MK (2005). *dropout: a program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework*. *Molecular Ecology Notes* **5**, 716–718. https://doi.org/10.1111/j.1471-8286.2005.01038.x.

Morin DJ, Waits LP, McNitt DC, Kelly MJ (2018). *Efficient single-survey estimation of carnivore density using fecal DNA and spatial capture-recapture: a bobcat case study*. *Population Ecology* **60**, 197–209. https://doi.org/10.1007/s10144-018-0606-9.

Otis DL, Burnham KP, White GC, Anderson DR (1978). *Statistical inference from capture data on closed animal populations*. *Wildlife Monographs*. Publisher: [Wiley, Wildlife Society], 3–135.

Paetkau D (2003). *An empirical exploration of data quality in DNA-based population inventories*. *Molecular Ecology* **12**, 1375–1387. https://doi.org/10.1046/j.1365-294X.2003.01820.x.

Ruell EW, Riley SPD, Douglas MR, Pollinger JP, Crooks KR (2009). *Estimating bobcat population sizes and densities in a fragmented urban landscape using noninvasive capture–recapture sampling*. *Journal of Mammalogy* **90**, 129–135. https://doi.org/10.1644/07-MAMM-A-249.1.

Schofield MR, Bonner S (2015). *Connecting the latent multinomial*. *Biometrics* **71**, 1070–1080. https://doi.org/https://doi.org/10.1111/biom.12333.

Taberlet P, Waits LP, Luikart G (1999). *Noninvasive genetic sampling: look before you leap*. *Trends in Ecology & Evolution* **14**, 323–327. https://doi.org/10.1016/S0169-5347(99)01637-7.

Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R (2017). *Programming with models: writing statistical algorithms for general model structures with NIMBLE*. *Journal of Computational and Graphical Statistics* **26**, 403–413. https://doi.org/10.1080/10618600.2016.1172487.

Waits LP, Paetkau D (2005). *Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection*. *The Journal of Wildlife Management* **69**, 1419–1433. https://doi.org/10.2193/0022-541X(2005)69[1419:NGSTFW]2.0.CO;2.

White GC, Burnham KP (1999). *Program MARK: survival estimation from populations of marked animals*. *Bird Study* **46**. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00063659909477239, S120–S139. https://doi.org/10.1080/00063659909477239.

Woodruff SP, Lukacs PM, Christianson D, Waits LP (2016). *Estimating Sonoran pronghorn abundance and survival with fecal DNA and capture—recapture methods*. *Conservation Biology* **30**, 1102–1111.

Wright JA, Barker RJ, Schofield MR, Frantz AC, Byrom AE, Gleeson DM (2009). *Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples*. *Biometrics* **65**, 833–840. https://doi.org/10.1111/j.1541-0420.2008.01165.x.

Yoshizaki J, Brownie C, Pollock K, Link W (2011). *Modeling misidentification errors that result from use of genetic tags in capture–recapture studies*. *Environmental and Ecological Statistics* **18**, 27–55. https://doi.org/10.1007/s10651-009-0116-1.