Using informative priors to account for identifiability issues in occupancy models with identification errors

Célian Monchy[1], Marie-Pierre Etienne[#,2], and Olivier Gimenez[#,1]

Abstract

Camera traps, autonomous recording units and environmental DNA metabarcoding have become important non-invasive monitoring techniques to collect environmental data with the objective to better understand species distribution. These new data have fueled the development of statistical models to suit with specific sampling designs and get reliable ecological inferences.
Among others, site occupancy models enable the estimation of occurrence patterns for one or more target species, considering that their presence may be observed or not. In that context, it is useful to distinguish a) the detection process, which is unobserved and corresponds to the situation where the species of interest leaves signs of its presence (by triggering a camera-trap for example), and b) the identification process, which allows to recognize the species of interest from its signs of presence. Detection and identification are both imperfect processes in which false-negative and false-positive errors may occur, especially when collected data are massive and require automated treatment. Misclassification at both steps can lead to significant biases in ecological estimates, and several extensions have been proposed to correct for these potential errors.
The naive occupancy model does not account for false detection since it results in identifiability issues, hence the existence of alternative models that combine several data sources to allow the identification of error parameters. These models however are more data-hungry and require to include a source of data without false-positives. As an alternative, we propose to reuse available data from the identification process in a Bayesian framework through an informative prior, in order to overcome the identifiability issue and reduce estimation bias. We compare through a simulation study these different approaches considering various monitoring designs (e.g., sensors or eDNA data, for species monitoring or disease ecology).
Overall, what is at stake is enhancing statistical methods together with sampling non-invasive technologies, in a way to provide ecological outcomes suitable for conservation decision-making.

Keywords: species occupancy, false-positive, mis-identification, identifiability, Bayesian modelling, informative priors, environmental DNA, camera traps, non-invasive sampling

[1]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France, [2]IRMAR - Institut de Recherche Mathématique de Rennes, [#]Equal contribution

Introduction

One of the key aims of ecologists and conservation scientists is to understand how populations and communities are distributed in space and time. Many animal species, plants and even pathogens are monitored by collecting information on their presence and ideally absence, in order to evaluate the proportion of an area occupied by the species.

Different methods may be used to assess the presence or absence of an organism in an environment during a monitoring survey. Recently, non-invasive monitoring methods have emerged like camera traps (e.g. Hofmeester et al., 2019; Parsons et al., 2017), autonomous acoustic recording units (e.g. Shonfield and Bayne, 2017; Wrege et al., 2017) or environmental DNA sampling (e.g. Da Silva Neto et al., 2020; Griffin et al., 2020). The use of passive sensors to monitor wildlife is growing in conservation projects, and thanks to technical improvements, they allow to study non-invasively and simultaneously multiple species, involving cryptic species, in large areas and challenging environments (Burton et al., 2015). When some discrete taxa remain silent, do not trigger camera traps or leave few tracks of their passage (either at macroscopic or microscopic level) (Belmont et al., 2022; Goldman et al., 2023), it becomes necessary to account for the probability of detecting them, irrespective of the observation method. Importantly, failure to consider detection as an imperfect process may result in a biased estimation of the probability of presence (MacKenzie et al., 2002; Tyre et al., 2003).

MacKenzie et al. (2002, see also Tyre et al., 2003) introduced a site occupancy model to estimate the proportion of sites occupied by a species while correcting for the probability of detection (MacKenzie et al., 2002). However, only one visit on one site does not allow to disentangle the case where the species is present but not detected, from true absence on this site. MacKenzie et al. (2002) showed that repeated visits on a same site allow to estimate the false-negative error rate, defined by a taxon present at a site but remaining undetected during a visit.

Within the context of sensor-based assessment method, data are massive and needs to be processed before being analyzed. This consists in identifying the taxon of interest in a large amount of collected data, referred to as identification or classification, which can be carried out either manually by one or several operators (Swanson et al., 2015; Welbourne et al., 2015), either automated using deep learning algorithms (Duggan et al., 2021; Tabak et al., 2019), or a combination of both (Augustine et al., 2023; Campos-Cerqueira and Aide, 2016). For images and acoustic data, combining manually and automatically-generated data processing helps to control classification errors; when one species is mistaken for another (Barré et al., 2019). Importantly, in a context where such classification errors, or mis-identifications, are likely to occur, it may result in biased estimates of the actual proportion of occupied sites by the species of interest (Spiers et al., 2022). The identification process follows the detection process, and allows for two types of errors: the species is detected (for example the camera is triggered), but not correctly identified, it is a false-negative mis-identification; or the species is not detected but the manual or automatic analysis is erroneous and the species is accidentally identified, this is usually referred to as false-positive mis-identification.

Environmental DNA studies also generate large amounts of data from which to extract presence data (Hunter et al., 2015; Schmidt et al., 2013; Thomsen et al., 2012). Samples are molecular data, analyzed with PCR-based approaches and species identification efficiency depends on laboratory procedures (e.g. temperature, number of replicates, primers specificity) (Doi et al., 2019; Willoughby et al., 2016). Using site occupancy model in eco-epidemiology studies relies

on estimating the occurrence of pathogens responsible of wildlife disease in a sample unit, in order to infer some spatial patterns and disease dynamics (McClintock et al., 2010b). Finally, the challenge for wildlife disease surveys is comparable with the one of camera-trapping for wildlife conservation, since it involves estimating occupancy parameters of a pathogen detected in a population, based on an imperfect diagnostic test (Lachish et al., 2012; McClintock et al., 2010b; Thompson, 2007).

The standard site occupancy model accounts for false-negative errors by estimating the imperfect detection probability, but this model does not account for cases where a species is incorrectly detected at a site it does not occupy. However, false-positive errors can have negative impacts on occupancy estimates if not accounted for, such as overestimation of occupancy probability (McClintock et al., 2010a; Miller et al., 2011; Royle and Link, 2006). As a consequence, these authors have proposed to extend MacKenzie's site occupancy model by accounting for false detection. Most of these extensions raise identifiability issues (Chambert et al., 2015) generally adressed by integrating different sources of data including one without errors. For example, Miller et al., 2011 proposed a multiple detection state model in which certain and ambiguous data are used at each site. Then Chambert et al., 2015 built on to this proposal by considering "reference sites" exempt from detection error, and McKibben et al., 2023 returned to the notion of detection ambiguity introduced by Miller et al., 2011 by scoring the level of confidence of observers. These studies provide solutions to account for detection errors and in particular false positive ones but require to combine different sources of data, which represents a strong constraint that cannot always be met. Indeed, great logistics and human efforts are often needed to design sampling protocols, collect and/or verify data, and to finally get several sources of data with some of them guaranteed to be error free. Apart from data collected with a special design sampling protocol to answer a particular question, some knowledge about the system may still exist (e.g., expert beliefs, calibration experiments or performance metrics), then eliciting informative prior distribution may be an alternative to the combination of several sources of data (Cruickshank et al., 2019; Guillera-Arroita et al., 2017). The use of Bayesian statistics in ecological researches, allows to integrate additional information through informative prior, which has been shown to increase confidence in the results (Choy et al., 2009; McCarthy and Masters, 2005). In occupancy studies with sparse data, a precise choice of priors influences trend occupancy estimates (Outhwaite et al., 2018). However, those informative priors must be chosen carefully, in accordance with the data, otherwise the parameter estimates could be biased (Morris et al., 2015).

We propose a hierarchical model, which builds up on the classical occupancy model, to take into account identification errors and applicable to different types of data. We first provide a probabilistic description of the model. We then discuss the limitations of a frequentist approach to fit this model to data. Finally, we propose to overcome these limitations using a Bayesian framework that allows incorporating available information through informative priors.

Site Occupancy model with identification errors

We start by describing the occupancy model as developed by MacKenzie et al., 2002. Then we explain how it can be extended to handle with data containing identification errors.

Detection and non-detection data on a species arise from visiting S sites J times. Repeated visits enable to discriminate sites with only true absences from sites with some absences resulting from non-detection on occupied sites. In the hierarchical formulation of the occupancy model (Royle and Kéry, 2007)the latent occupancy state of a site i is a Bernoulli distributed random variable of parameter ψ, hence the species is present on a site i ($Z_i = 1$) with a probability ψ:

$$Z_i \overset{i.i.d}{\sim} Bernoulli(\psi) \tag{1}$$

$$Y_{ij}|Z_i \overset{i.i.d}{\sim} Bernoulli(Z_i \times p) \tag{2}$$

It is assumed that the presence of the species at a site is independent of its presence on the other sites, such as Z_i (with i from 1 to S) are independent. Conditionally on its presence on site i, Y_{ij} outlines the detection state of the species in a visit j. It follows a Bernoulli distribution with parameter p, such as the species may be detected with a probability p during the jth visit on the occupied site i, and missed with probability $1 - p$.

In this model, each visit is considered as an observation, the species being detected or not. Conditionally on the presence ($Z_i = 1$), the history of detection is a set of independent observations for a site, represented by a vector of detections (1) and non-detections (0). In contrast here, on top of the detection process represented by the random variable Y_{ij} as a latent variable, we denote the identification process in the hierarchical occupancy model through an additional level. In this way, an observation may correspond to a record (acoustic or image) on which the species is identified (correctly or incorrectly), while detection is an unknown variable and refers to the sensor triggering and the recording of the presence of the species. The species may be detected but the quality of the recorded file may be insufficient to classify it as the correct species (Findlay et al., 2020), then it is impossible to deduct the detection state from the record. In the case of environmental DNA for example, the species is considered detected if it is present in the DNA sample, and it is considered identified if its DNA is observed in a PCR analysis replicate (Schmidt et al., 2013). In disease ecology, an infected sample unit may be detected (or not) through an infected (or not) sampled individual, and then the laboratory assay may identify or not the pathogen.

We denote W_{ij} as the species identification at site i on visit j. W_{ij} equals 1 if the species is identified and 0 otherwise. The identification process is imperfect and suffers from two types of error related to the detection or non-detection of the species, each with its own probability (Fig. 1). The probability to identify the species in the jth visit from site i if it has been detected is w_A, and it is equivalent to the probability of correctly classifying a record in the class associated with the species it presents. This is related to the true positive probability, also known as sensitivity or *recall*. Otherwise, the probability to falsely identify the species while it has not been detected is $1 - w_B$, usually referred to as the false positive rate (also known as *fall-out*), and corresponding to the probability of associating the record to the wrong species. In the site occupancy model from MacKenzie et al., 2002 the identification process is not developed which amounts to write $w_A = 1$ and $w_B = 0$. In other words, the probability to not identify a species which has been detected is null, as well as the probability to confuse a not detected species with a detected one.

The conditional probability of identifying a species $W_{ij} = 1$ given that it is detected or not is written as:

$$W_{ij}|Y_{ij} = \begin{cases} W_{ij}|(Y_{ij}=0) \sim Bernoulli(1-w_{B,ij}) \\ W_{ij}|(Y_{ij}=1) \sim Bernoulli(w_{A,ij}) \end{cases} \tag{3}$$

In this hierarchical model, Z_i and Y_{ij} are latent variables respectively related to occupancy state and detection state of the target species at site i during visit j, and where W_{ij} is the observation data related to identification (Fig. 1).

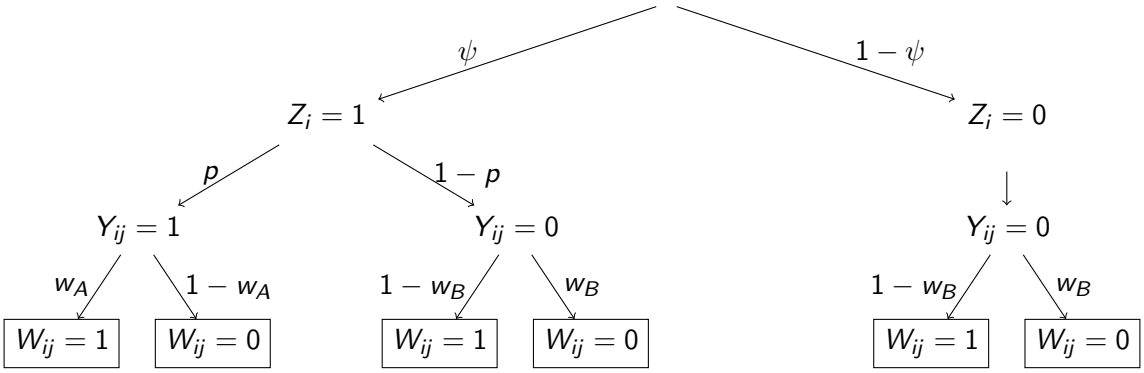

Figure 1 – Tree diagram representing data generation, underlying latent states, and associated parameters; The latter are the probabilities of the events indicated by the corresponding branches, e.g. w_A is the probability that the species is identified during a time period when the species was indeed present and detected.

For each site, the identification record of the target species is compiled on the basis of visits. We can derive the probability to observe w ($w = 0$ or 1) at visit j on site i by considering the different possible states for Y_{ij}:

$$\begin{aligned} \pi_1(w) &:= Pr\left(W_{ij} = w \,|Z_i = 1\right) \\ &= Pr\left(W_{ik} = w,\ Y_{ij} = 1 \,|Z_i = 1\right) + Pr\left(W_{ij} = w, Y_{ij} = 0 \,|Z_i = 1\right) \\ &= Pr\left(W_{ij} = w \,|Y_{ij} = 1\right) Pr\left(Y_{ij} = 1 \,|Z_i = 1\right) + Pr\left(W_{ij} = w \,|Y_{ik} = 0\right) Pr\left(Y_{ij} = 0 \,|Z_i = 1\right) \\ &= w_A^w (1-w_A)^{1-w} p + (1-w_B)^w (w_B)^{1-w} (1-p) \end{aligned} \tag{4}$$

$$\begin{aligned} \pi_0(w) &:= Pr\left(W_{ij} = w \,|Z_i = 0\right) \\ &= Pr\left(W_{ij} = w,\ Y_{ij} = 0 \,|Z_i = 0\right) \\ &= Pr\left(W_{ij} = w \,|Y_{ij} = 0\right) Pr\left(Y_{ij} = 0 \,|Z_i = 0\right) \\ &= (1-w_B)^w (w_B)^{1-w} \end{aligned} \tag{5}$$

For example, for a site visited 3 times where the species is identified only during the second visit, we get the following identification history 010. Out of 3 visits, the occupancy state of the site is unknown but the species was identified once so we combine equations 4, 5, which account for the site's occupancy state. This may be a true identification; in which case the species is present on the site but not easily identifiable. Otherwise, because this model includes false-positives, the species may have been wrongly identified and the site would not be occupied (Fig. 1). Without including false-positives in the identification process, the site would have been necessarily considered occupied.

Conditionally on the site occupancy status and given that the visits are assumed to be independent, the probability to observe the identification history $W_i = (0, 1, 0)$ is given by:

$$\begin{aligned} Pr\left(W_i = (0,1,0)\right) &= Pr\left(W_i = (0,1,0),\ Z_i = 1\right) + Pr\left(W_i = (0,1,0),\ Z_i = 0\right) \\ &= Pr\left(W_i = (0,1,0)\,|Z_i = 1\right) Pr\left(Z_i = 1\right) + Pr\left(W_i = (0,1,0)\,|Z_i = 0\right) Pr\left(Z_i = 0\right) \\ &= \psi\pi_1(w_{i1})\pi_1(w_{i2})\pi_1(w_{i3}) + (1-\psi)\pi_0(w_{11})\pi_0(w_{12})\pi_0(w_{13}) \end{aligned} \tag{6}$$

For S independent sites with J independent visits of which j_i^* with positive identification, and given that parameters are constant between visits and sites, we obtain to the following model likelihood:

$$\begin{aligned} L(w_A, w_B, p, \psi\,|data) &= \prod_{i=1}^{N} Pr(W_i) = \prod_{i=1}^{N} \left(Pr(W_i, Z_i = 1) + Pr(W_i, Z_i = 0)\right) \\ &= \prod_{i=1}^{N} \Big[\psi\left[(1-w_B)(1-p) + w_A p\right]^{j_i^*} \left[w_B(1-p) + (1-w_A)p\right]^{J-j_i^*} + \\ &\qquad (1-\psi) w_B^{J-j_i^*} (1-w_B)^{j_i^*}\Big] \end{aligned} \tag{7}$$

Classical estimation from the identification layer

In this section, we assess the quality of estimates obtained by maximum likelihood using a simulation study. First, we investigate whether adding the identification process and the 2 types of error associated with it, leads to more reliable estimates. Second, we study how the number of visits made on a site affects precision of the estimates.

We carried out simulations by generating 1000 data sets with N=30 sites and J=12 or 36 visits. The parameter values used to create the matrices of observations were $\psi = 0.8$, $p = 0.5$, $w_A = 0.9$ and $w_B = 0.7$ (chosen according to resulting values from a site occupancy study of Eurasian Lynx *(Lynx lynx)* population in France; Gimenez et al., 2022). We obtained maximum likelihood estimates for all parameters from the minimization of the negative log-likelihood (Equ. 7). We compared the precision of estimates between data sets with 12 or 36 occasions.

Occupancy $\hat{\psi}$ is overestimated by the model without identification layer since all sites have at least one positive identification, whereas we get the correct value in model by adding an identification layer.

It has been showed that parameter estimates were biased if false-positive detections are not accounted for. In particular, the detection probability is underestimated, and occupancy is overestimated (McClintock et al., 2010a; Miller et al., 2011; Royle and Link, 2006).

We used the standard deviation as a measure of accuracy, which decreases as the number of occasions increases (from 0.22 for 12 visits to 0.08 for 36 visits for occupancy probability estimates $\hat{\psi}$)(Figure 2). Despite the amount of data available with 36 visits, estimates $\hat{p}$ and $\hat{w_A}$ are biased like $Bias(\hat{p}) = 0.17$ and $Bias(\hat{w_A}) = -0.15$.

If one of the two parameters, w_A or p, is fixed, then the other is estimated without bias (Supplementary Figure 5). Such parameter redundancy in the likelihood function is at the root of model identifiability issues (Gimenez et al., 2004).

In a context of lack of identifiability in a model with mis-detection Royle and Link, 2006 suggested to impose constraints in the model by setting the probability to correctly detect a

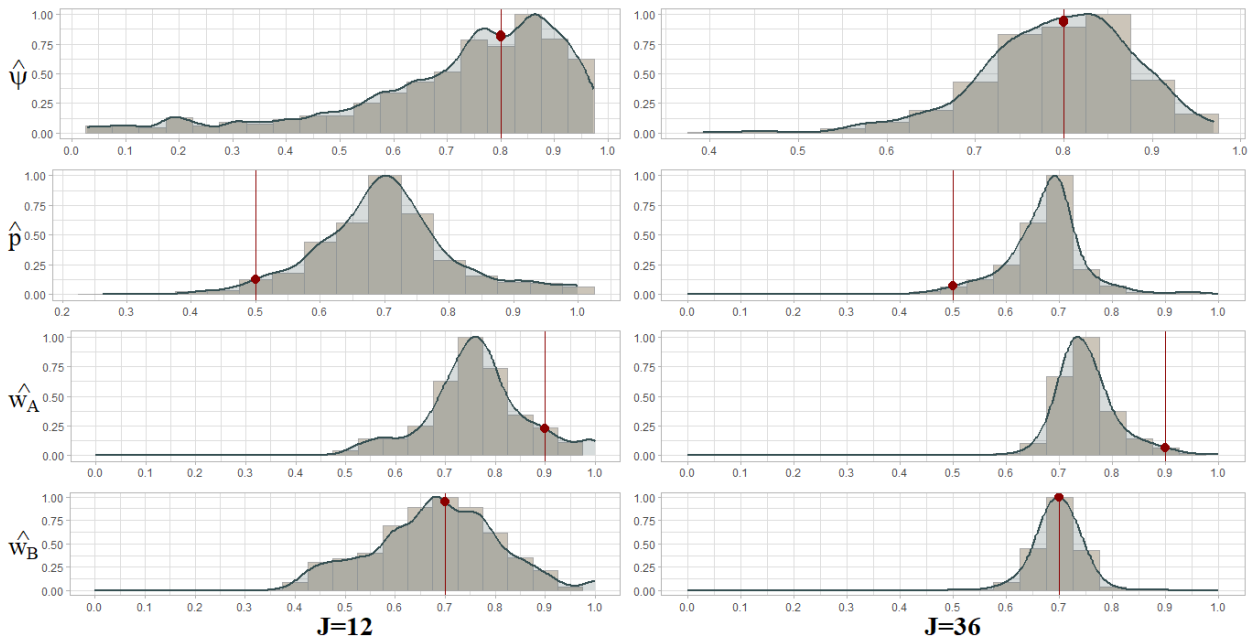

Figure 2 – Site Occupancy Model accounting for false-positive and false-negative errors in the identification layer. Histogram and kernel estimates of the distribution of maximum-likelihood estimates for 1000 simulations for J=12 (left column) or J=36 (right column) visits on N=30 sites, and the initial parameter value use to create datasets (in red). Estimates are the occupancy probability $\hat{\psi}$, the detection probability $\hat{p}$, the positive identification probability $\hat{w_A}$ and the negative identification probability $\hat{w_B}$.

present species higher than the probability to incorrectly detect it when it is absent. We first explore this suggestion, while remaining within a frequentist approach, then in the next section we move to a Bayesian approach using informative priors in order to get rid of these identifiability issues. In our model, a constraint is applied to the identification probabilities such as $w_A > 1 - w_B$.

We simulate 1000 observation datasets with values for the positive identification probability w_A and the negative identification probability w_B ranging between 0.5 and 0.95. Then, we estimate the parameters from our site occupancy model accounting for both types of error in the identification layer, by minimizing the log-likelihood function with and without the constraint.

The constraint provides less biased estimates of $\hat{p}$ for values of w_A and w_B around 0.5 (Figure 7). Whatever the initial value of w_A, estimates of $\hat{w_A}$ are concentrated around 0.7 (Figure 2), hence the decreasing slope of the associated bias. As the value of w_A or w_B increases, the estimates produced with and without constraints become more similar. Indeed, usually sensitivity (the true-positive rate), modeled by w_A, is higher than the false-positive rate determined by $1 - w_B$ (Guillera-Arroita et al., 2017), so this constraint might not be strong enough to solve the identifiability problem (Figure 3).

Using an informative prior to address identifiability issues

In this section we lift the identifiability issue of the model by using knowledge about the risk of mis-identifications without necessarily having access to other data sources. We adopt a Bayesian approach and use this knowledge by providing informative priors.

In many situations, it is possible to have a good knowledge of the false-negative rate in the identification process. For example, one may simply use a dataset that has been annotated during a previous manipulation. Otherwise, when using sensor data processed automatically by

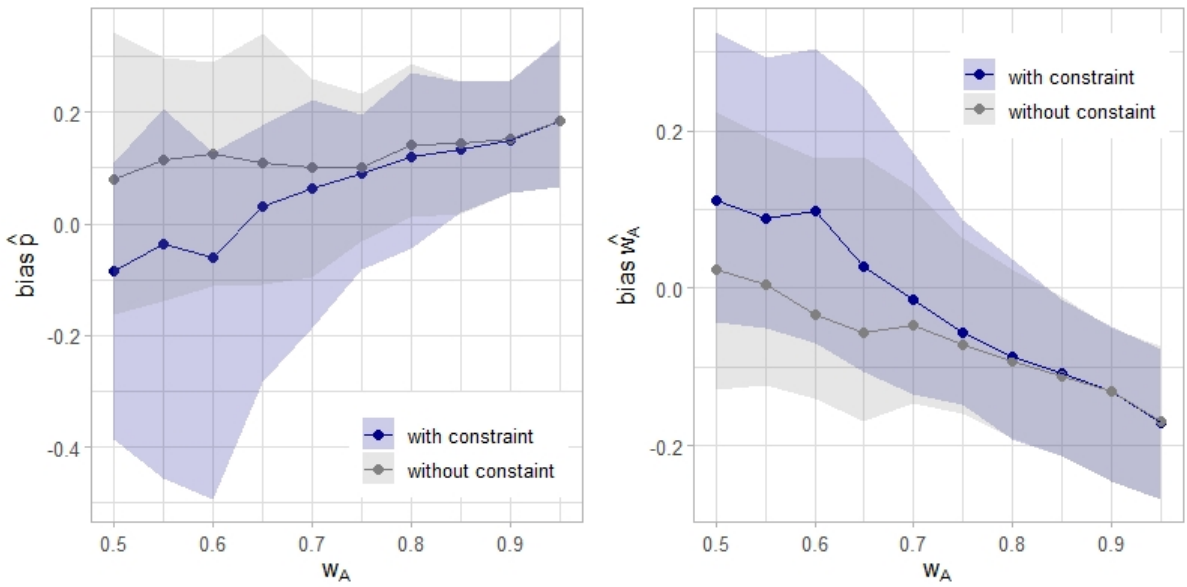

Figure 3 – Bias trend as a function of the value of w_A used to simulate the data. In blue, the estimator bias after optimization under the constraint, and in gray the bias without the constraint. The bias is associated with the median of the maximum-likelihood estimates, and values between the 0.1 and 0.9 quantile.

neural networks, metrics associated with the performance of the classification algorithm are obtained after validation, including its sensitivity (also named recall). Or else in the context of eDNA metabarcoding, the sensitivity of the identification process corresponds to the proportion of PCR replicates in which the species is present and whose DNA is identified.

Here we examine how the contribution of this external information, integrated into the elicitation of a prior, can be used to address identifiability issues and reduce bias in parameters estimates.

We attempt to construct the most suitable prior distribution given the available knowledge about the identification process, and more particularly on the sensitivity of this process modeled by the parameter w_A, i.e., the probability that the species will be identified when it is detected.

We study 3 types of priors for parameter w_A: A vague prior and two informative priors from uniform and beta distributions around one value.

We introduce 2 different prior distributions for the probability to correctly identify the species while it is detected and therefore present:

- In a context of classifying data generated by sensors like camera traps, the use of deep learning algorithms informs about the global rate of objects with the target species correctly identified and it constitutes our prior knowledge on the identification process. Therefore, prior elicitation brought us to consider a uniform distribution whose range depends on the accuracy of the neural network for classifying species (i.e. the percentage of correct classification). Tests of wildlife image classifiers generally indicate high accuracy in identifying animal species (Tabak et al., 2019; Vélez et al., 2023).
- In the case of molecular data, sensitivity may be expressed as a probability to correctly identify a species, or a pathogen, present in a replicate. Carrying out independent calibration experiments with water samples to realize PCR blanks may inform about false-positive rate (Guillera-Arroita et al., 2017). Realizing additional unambiguous surveys may

inform about the probability to identify a species actually present in the eDNA sample, or to positively diagnose an infected individual. To elicit an informative prior distribution reporting on the number of successful identifications out of the total number of replicates, we used a beta distribution which is defined between $[0; 1]$. We used a matching method to choose the beta distribution parameters with the appropriate values arisen from the survey results (Denham and Mengersen, 2007; Falconer et al., 2022).

To illustrate sensitivity (rate of positive identifications among true ones) we first chose a uniform distribution centered tightened around 0.9, such as w_A prior is defined with $\mathcal{U}(0.8, 1)$. In the case of molecular data, sensitivity may be expressed as a probability to correctly identify a species present in the DNA sample, then we match a beta distribution with a mean around 0.9 and a standard deviation of 0.1, which gives a prior distribution $\mathcal{B}(45, 5)$.

We simulated 100 observation datasets and we estimated model parameters in a Bayesian framework (using JAGS) for each prior.

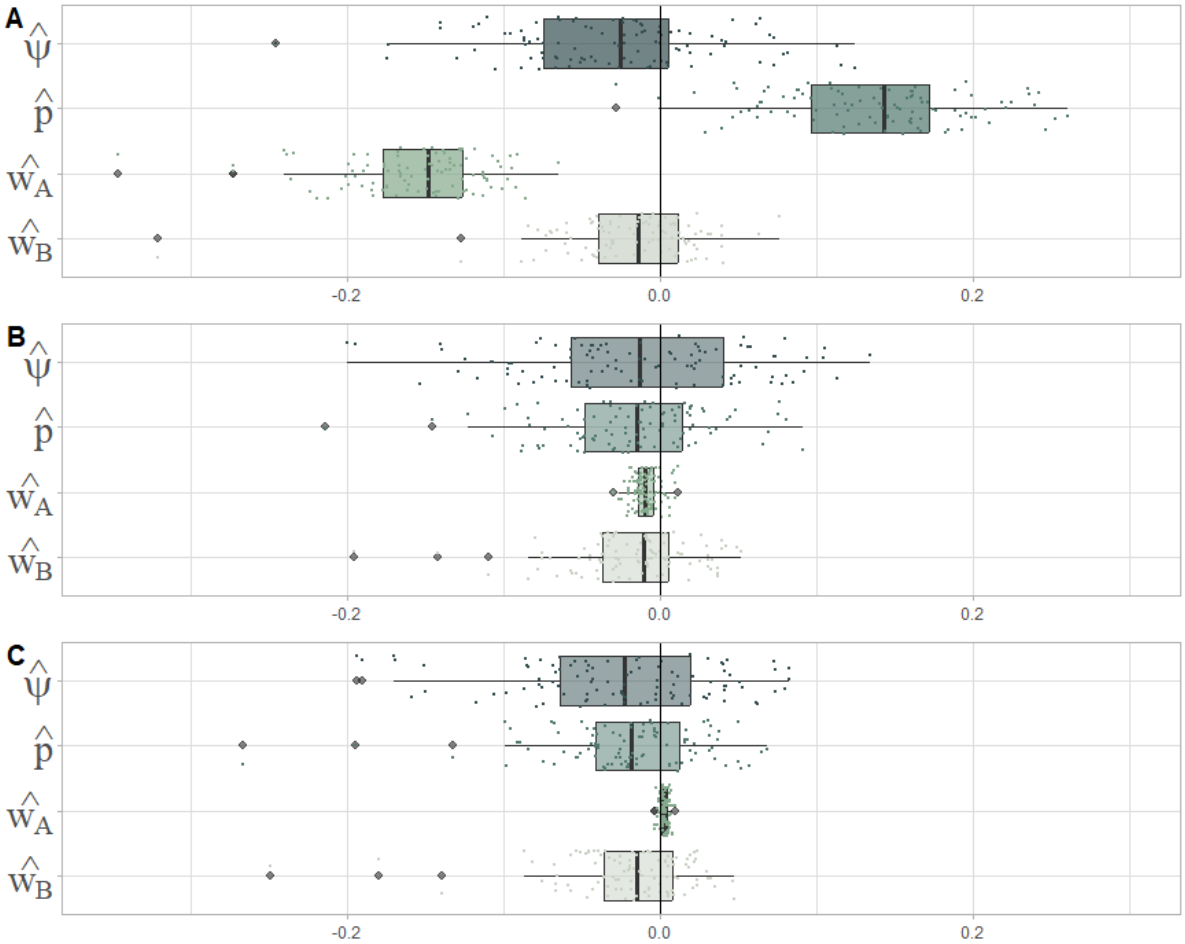

Figure 4 – Boxplot of the relative difference between posterior median inferences in a bayesian framework and the value set for data simulations ($\psi = 0.8$, $p = 0.5$, $w_A = 0.9$, $w_B = 0.7$). The informative prior for the sensitivity w_A follows a uniform distribution like $\mathcal{U}(0, 1)$ **(A)**, a uniform distribution like $\mathcal{U}(0.8, 1)$ **(B)**, and a beta distribution like $\mathcal{B}(45, 5)$ **(C)**.

Using non-informative priors for identification parameters drives to biased posterior distributions. The inference on the detection probability $\hat{p}$ is improved with setting a prior for sensitivity (Fig. 4) . Informative priors chosen are two different approaches to integrate information about

the identification process, and they behave in a comparable way. Actually, the median posterior values obtained for 100 simulations are only weakly influenced by the type of prior.

Discussion

We proposed a single-species occupancy model usable with different types of data (images, acoustic records and molecular data) and acknowledging the two-step structure in detection and identification of the observation process. We presented a hierarchical occupancy model, considering both detection and identification processes, which can be independent source of errors. We considered, on the one hand, false-negatives in detection thanks to the detection parameter p, and on the other hand, identification errors, whether in favor of the target species or not, with parameters w_A and w_B.

We first implemented our model in the maximum-likelihood framework and found that some of the estimates were biased due to model mis-specifications and identifiability issues. We showed that by shifting to a Bayesian approach and using informative priors based on performance metrics of the identification process, such as sensitivity, we could address these identifiability issues.

Sensors deployment and eDNA sampling generate more data than conventional sampling, and above all they are not specific, so the data must be sorted to identify the target species. This second stage can be carried out by one or more observers in the case of sensor data notably through crowd-sourcing analysis project (e.g. *www.chimpandsee.org*,*www.chicagowildlifewatch.org*, *www.snapshotserengeti.org*) for images or with the assistance of experts for acoustic (e.g. Shonfield and Bayne, 2017; Zwart et al., 2014). Species recognition can then be automated, but without human verification which can be particularly time-consuming (Barré et al., 2019; Spiers et al., 2022), identification errors can arise and distort inferences (Ferguson et al., 2015; Lonsinger et al., 2023; McClintock et al., 2010a). Accounting for those identification errors in addition to detection errors requires to develop different versions of the site occupancy model. Firstly, the model developed by Nichols et al., 2008 considered multiple detection methods at the sampling occasion scale, and so introduced the idea we are following, that a visit on a site may be a set of observations. In a certain way, dividing a visit in two different detection events is equivalent to assuming a two-stage survey protocol as suggested by Guillera-Arroita et al., 2017, on which we rely on.

Finally, by reducing data processing time, potential identification errors are introduced, which, especially false positives, may have a severe impact on inferences. The formalization of models thus evolves to best adapt to the different sampling levels, but this comes at a cost, as some parameters are no longer estimable. Some authors have therefore suggested combining several sources of information (Chambert et al., 2015; Guillera-Arroita et al., 2017; Miller et al., 2011) to solve the problem of identifiability. Because multiplying the sources of collected data is costly, we propose to use the performance metrics directly assessed from the dataset, through prior elicitation. The main limitation of our approach lies in the fact that we need feedback on the performance of the identification process in order to provide a relevant informative prior. Also, simulations show that even with a weakly informative prior, our model provides reliable posterior estimates of the presence parameter ψ. Cruickshank et al., 2019 avoid identifiability issues when accounting for false-positive errors by integrating informative prior based on realistic assumptions about volunteer data for monitoring. Similarly, our approach involving inputs on the

detection process in the form of a prior provides relevant posterior estimates, while offering an alternative to approaches using additional datasets.

Passive sensors like camera traps and autonomous recording units for acoustics provide opportunities to address wide array of questions in ecological applications and conservation management. In combination with other approaches such as environmental DNA, these technological developments free ecologists from constraints in collecting data at large spatial scales or fine temporal resolution, and allow to study species that are difficult to observe (Ross et al., 2023; Sahu et al., 2023). For such complex and massive data, making reliable taxonomic identification is challenging, but taking into account the noise resulting from this processing remains essential. In this context, our proposal can be included in the ecologist's toolbox for analyzing sensor and molecular biological data to address questions in conservation biology, wildlife management and disease ecology.

Appendix A. Supplementary Results

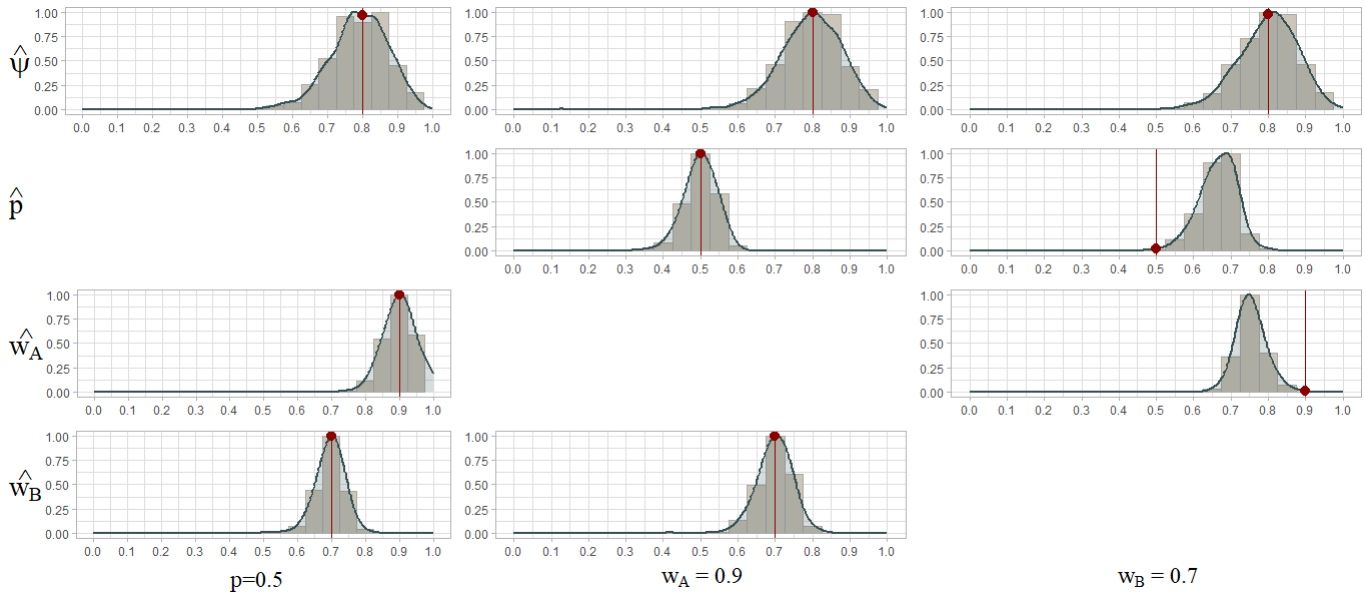

Figure 5 – Distribution of maximum-likelihood estimates for 1000 simulations when a parameter is set to a constant value (in columns).

$\hat{w}_A$ or $\hat{p}$ are estimated without bias when the other parameter is set to a fixed value in the expression of likelihood. This result reflects parameter redundancy in the likelihood function.

We consider the profile deviance on p to investigate model identifiability.

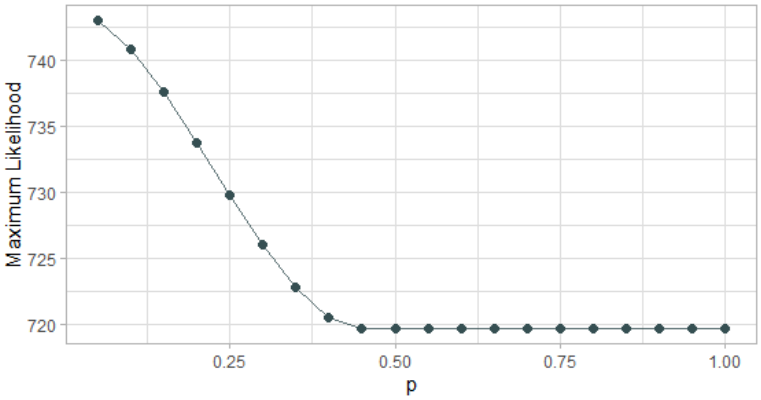

Figure 6 – Profile deviance on p

Deviance ($-2Log - Likelihood$) is constant for p greater than 0.45, beyond this value the model is not identifiable, which means that $\hat{p}$ and $\hat{w}_A$ cannot be distinguished.

The model is not globally identifiable (Cole et al., 2010) since there are different sets of parameters that give rise to the same likelihood function value (Table 1).

As pointed out by Royle and Link, 2006, including false positives leads to model identifiability concerns. The authors proposed to impose a constraint in the likelihood optimization in order to solve this parameter redundancy issue. Specifically, the idea is to constrain the probability to truly detect a species as higher than the probability to falsely detect it. Translating this constraint to our model with an identification layer means that truly identifying the target species is more likely to occur than falsely identifying the target species when it has not been detected.

Table 1 – Profile deviance on p

$\hat{\psi}$	p	$\hat{w_A}$	$\hat{w_B}$	**Likelihood**
0.7419	0.9	0.6592	0.6777	**719.6974**
0.7421	0.8	0.7013	0.6777	**719.6974**
0.7419	0.7	0.7554	0.6778	**719.6974**
0.7422	0.6	0.8276	0.6776	**719.6974**
0.7419	0.5	0.9286	0.6777	**719.6974**

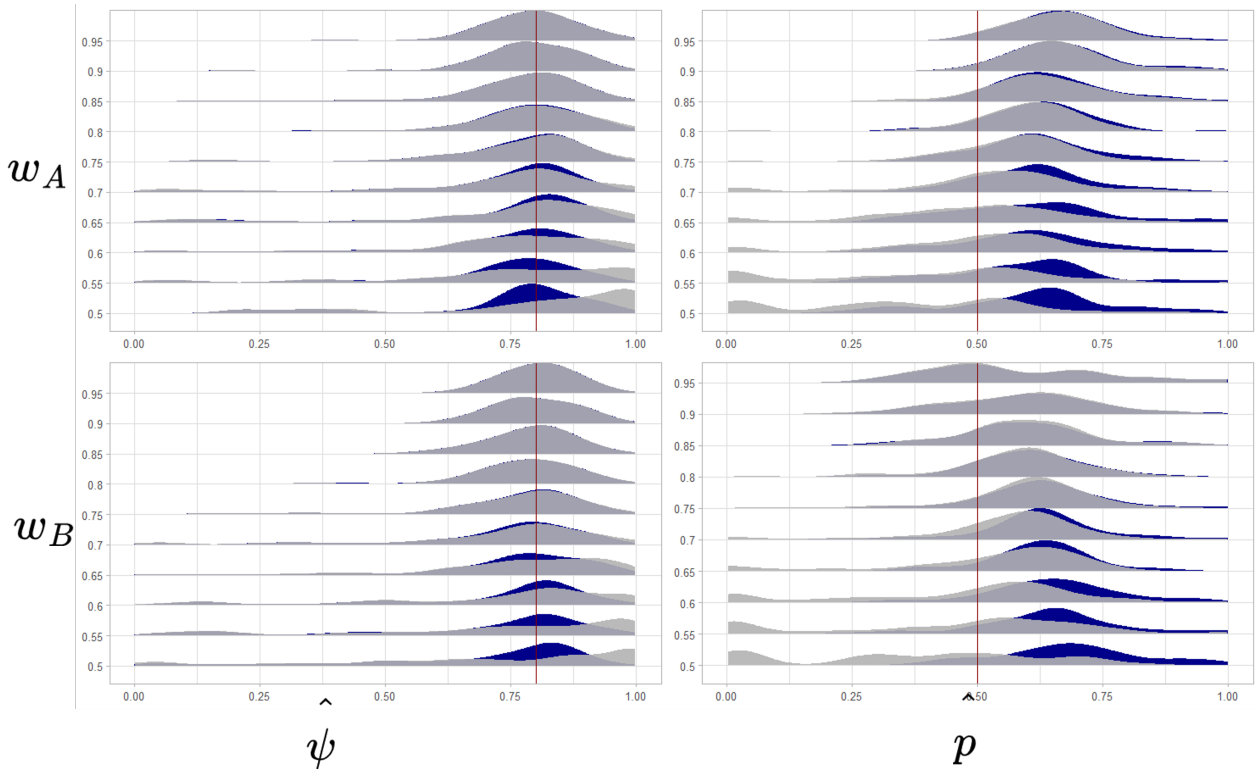

Figure 7 – Distribution of $\hat{\psi}$ and $\hat{p}$ for 1000 simulated data sets with w_A set between 0.5 and 0.95 (top) and w_B set between 0.5 and 0.95 (bottom). Optimization under the constraint $w_A > 1 - w_B$ (in gray) and without it (in blue).

The constraint proposed does not help to fix the estimation issue in the detection probability, however for small values of w_A or p, close to 0.5, occupancy estimates are reliable.

Acknowledgements

Fundings

The authors declare that they have received no specific funding for this study.

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

Script and codes are available online (https://zenodo.org/doi/10.5281/zenodo.11121903; Monchy et al., 2024)

References

Augustine BC, Koneff MD, Pickens BA, Royle JA (2023). *Towards estimating marine wildlife abundance using aerial surveys and deep learning with hierarchical classifications subject to error*. preprint. Ecology. https://doi.org/10.1101/2023.02.20.529272.

Barré K, Le Viol I, Julliard R, Pauwels J, Newson SE, Julien JF, Claireau F, Kerbiriou C, Bas Y (2019). *Accounting for automated identification errors in acoustic surveys*. *Methods in Ecology and Evolution* **10**, 1171–1188. https://doi.org/10.1111/2041-210X.13198.

Belmont J, Miller C, Scott M, Wilkie C (2022). *A new statistical approach for identifying rare species under imperfect detection*. *Diversity and Distributions* **28**, 882–893. https://doi.org/10.1111/ddi.13495.

Burton AC, Neilson E, Moreira D, Ladle A, Steenweg R, Fisher JT, Bayne E, Boutin S (2015). *REVIEW: Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes*. *Journal of Applied Ecology* **52**. Ed. by Phil Stephens, 675–685. https://doi.org/10.1111/1365-2664.12432.

Campos-Cerqueira M, Aide TM (2016). *Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling*. *Methods in Ecology and Evolution* **7**, 1340–1348. https://doi.org/10.1111/2041-210X.12599.

Chambert T, Miller DAW, Nichols JD (2015). *Modeling false positive detections in species occurrence data under different study designs*. *Ecology* **96**, 332–339. https://doi.org/10.1890/14-1507.1.

Choy SL, O'Leary R, Mengersen K (2009). *Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models*. *Ecology* **90**, 265–277. https://doi.org/10.1890/07-1886.1.

Cole DJ, Morgan BJT, Titterington DM (2010). *Determining the parametric structure of models*. *Mathematical Biosciences* **228**, 16–30. https://doi.org/10.1016/j.mbs.2010.08.004.

Cruickshank SS, Bühler C, Schmidt BR (2019). *Quantifying data quality in a citizen science monitoring program: False negatives, false positives and occupancy trends*. *Conservation Science and Practice* **1**, e54. https://doi.org/10.1111/csp2.54.

Da Silva Neto JG, Sutton WB, Spear SF, Freake MJ, Kéry M, Schmidt BR (2020). *Integrating species distribution and occupancy modeling to study hellbender (Cryptobranchus alleganiensis) occurrence based on eDNA surveys*. *Biological Conservation* **251**, 108787. https://doi.org/10.1016/j.biocon.2020.108787.

Denham R, Mengersen K (2007). *Geographically assisted elicitation of expert opinion for regression models*. *Bayesian Analysis* **2**. https://doi.org/10.1214/07-BA205.

Doi H, Fukaya K, Oka Si, Sato K, Kondoh M, Miya M (2019). *Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model*. *Scientific Reports* **9**, 3581. https://doi.org/10.1038/s41598-019-40233-1.

Duggan MT, Groleau MF, Shealy EP, Self LS, Utter TE, Waller MM, Hall BC, Stone CG, Anderson LL, Mousseau TA (2021). *An approach to rapid processing of camera trap images with minimal human input*. *Ecology and Evolution* **11**, 12051–12063. https://doi.org/10.1002/ece3.7970.

Falconer JR, Frank E, Polaschek DLL, Joshi C (2022). *Methods for Eliciting Informative Prior Distributions: A Critical Review*. arXiv:2112.07090 [stat].

Ferguson PF, Conroy MJ, Hepinstall-Cymerman J (2015). *Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys*. *Methods in Ecology and Evolution* **6**, 1395–1406. https://doi.org/10.1111/2041-210X.12442.

Findlay MA, Briers RA, White PJC (2020). *Component processes of detection probability in camera-trap studies: understanding the occurrence of false-negatives*. *Mammal Research* **65**, 167–180. https://doi.org/10.1007/s13364-020-00478-y.

Gimenez O, Viallefont A, Catchpole EA, Choquet R, Morgan BJT (2004). *Methods for investigating parameter redundancy*. *Animal Biodiversity and Conservation*.

Gimenez O, Kervellec M, Fanjul JB, Chaine A, Marescot L, Bollet Y, Duchamp C (2022). *Trade-off between deep learning for species identification and inference about predator-prey co-occurrence*. *Computo*. https://doi.org/10.57750/yfm2-5f45.

Goldman MR, Shinderman M, Jeffress MR, Rodhouse TJ, Shoemaker KT (2023). *Integrating multiple sign types to improve occupancy estimation for inconspicuous species*. *Ecology and Evolution* **13**, e10019. https://doi.org/10.1002/ece3.10019.

Griffin JE, Matechou E, Buxton AS, Bormpoudakis D, Griffiths RA (2020). *Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors*. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **69**, 377–392. https://doi.org/10.1111/rssc.12390.

Guillera-Arroita G, Lahoz-Monfort JJ, Rooyen AR, Weeks AR, Tingley R (2017). *Dealing with false-positive and false-negative errors about species occurrence at multiple levels*. *Methods in Ecology and Evolution* **8**, 1081–1091. https://doi.org/10.1111/2041-210X.12743.

Hofmeester TR, Cromsigt JPGM, Odden J, Andrén H, Kindberg J, Linnell JDC (2019). *Framing pictures: A conceptual framework to identify and correct for biases in detection probability of camera traps enabling multi-species comparison*. *Ecology and Evolution* **9**, 2320–2336. https://doi.org/10.1002/ece3.4878.

Hunter ME, Oyler-McCance SJ, Dorazio RM, Fike JA, Smith BJ, Hunter CT, Reed RN, Hart KM (2015). *Environmental DNA (eDNA) Sampling Improves Occurrence and Detection Estimates of Invasive Burmese Pythons*. *PLOS ONE* **10**, e0121655. https://doi.org/10.1371/journal.pone.0121655.

Lachish S, Gopalaswamy AM, Knowles SCL, Sheldon BC (2012). *Site-occupancy modelling as a novel framework for assessing test sensitivity and estimating wildlife disease prevalence from imperfect diagnostic tests*. *Methods in Ecology and Evolution* **3**, 339–348. https://doi.org/10.1111/j.2041-210X.2011.00156.x.

Lonsinger RC, Dart MM, Larsen RT, Knight RN (2023). *Efficacy of machine learning image classification for automated occupancy-based monitoring*. *Remote Sensing in Ecology and Conservation*, rse2.356. https://doi.org/10.1002/rse2.356.

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Andrew Royle J, Langtimm CA (2002). *ESTIMATING SITE OCCUPANCY RATES WHEN DETECTION PROBABILITIES ARE LESS THAN ONE*.

Ecology **83**, 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.

McCarthy MA, Masters P (2005). *Profiting from prior information in Bayesian analyses of ecological data*. *Journal of Applied Ecology* **42**, 1012–1019. https://doi.org/10.1111/j.1365-2664.2005.01101.x.

McClintock BT, Bailey LL, Pollock KH, Simons TR (2010a). *Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections*. *Ecology* **91**, 2446–2454. https://doi.org/10.1890/09-1287.1.

McClintock BT, Nichols JD, Bailey LL, MacKenzie DI, Kendall WL, Franklin AB (2010b). *Seeking a second opinion: uncertainty in disease ecology: Uncertainty in disease ecology*. *Ecology Letters* **13**, 659–674. https://doi.org/10.1111/j.1461-0248.2010.01472.x.

McKibben FE, Abadi F, Frey JK (2023). *To model or not to model: false positive detection error in camera surveys*. *The Journal of Wildlife Management* **87**, e22365. https://doi.org/10.1002/jwmg.22365.

Miller DA, Nichols JD, McClintock BT, Grant EHC, Bailey LL, Weir LA (2011). *Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification*. *Ecology* **92**, 1422–1428. https://doi.org/10.1890/10-1396.1.

Morris WK, Vesk PA, McCarthy MA, Bunyavejchewin S, Baker PJ (2015). *The neglected tool in the Bayesian ecologist's shed: a case study testing informative priors' effect on model accuracy*. *Ecology and Evolution* **5**, 102–108. https://doi.org/10.1002/ece3.1346.

Nichols JD, Bailey LL, O'Connell Jr. AF, Talancy NW, Campbell Grant EH, Gilbert AT, Annand EM, Husband TP, Hines JE (2008). *Multi-scale occupancy estimation and modelling using multiple detection methods*. *Journal of Applied Ecology* **45**, 1321–1329. https://doi.org/10.1111/j.1365-2664.2008.01509.x.

Outhwaite CL, Chandler RE, Powney GD, Collen B, Gregory RD, Isaac NJB (2018). *Prior specification in Bayesian occupancy modelling improves analysis of species occurrence data*. *Ecological Indicators* **93**, 333–343. https://doi.org/10.1016/j.ecolind.2018.05.010.

Parsons AW, Forrester T, McShea WJ, Baker-Whatton MC, Millspaugh JJ, Kays R (2017). *Do occupancy or detection rates from camera traps reflect deer density?* *Journal of Mammalogy* **98**, 1547–1557. https://doi.org/10.1093/jmammal/gyx128.

Ross SRPJ, O'Connell DP, Deichmann JL, Desjonquères C, Gasc A, Phillips JN, Sethi SS, Wood CM, Burivalova Z (2023). *Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions*. *Functional Ecology* **37**, 959–975. https://doi.org/10.1111/1365-2435.14275.

Royle JA, Kéry M (2007). *A Bayesian State-Space Formulation of Dynamic Occupancy Models*. *Ecology* **88**, 1813–1823. https://doi.org/10.1890/06-0669.1.

Royle JA, Link WA (2006). *Generalized Site Occupancy Models Allowing for False Positive and False Negative Errors*. *Ecology* **87**, 835–841. https://doi.org/10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2.

Sahu A, Kumar N, Pal Singh C, Singh M (2023). *Environmental DNA (eDNA): Powerful technique for biodiversity conservation*. *Journal for Nature Conservation* **71**, 126325. https://doi.org/10.1016/j.jnc.2022.126325.

Schmidt BR, Kéry M, Ursenbacher S, Hyman OJ, Collins JP (2013). *Site occupancy models in the analysis of environmental DNA presence/absence surveys: a case study of an emerging amphibian*

pathogen. *Methods in Ecology and Evolution* **4**, 646–653. https://doi.org/10.1111/2041-210X.12052.

Shonfield J, Bayne EM (2017). *Autonomous recording units in avian ecological research: current use and future applications*. *Avian Conservation and Ecology* **12**, art14. https://doi.org/10.5751/ACE-00974-120114.

Spiers AI, Royle JA, Torrens CL, Joseph MB (2022). *Estimating species misclassification with occupancy dynamics and encounter rates: A semi-supervised, individual-level approach*. *Methods in Ecology and Evolution* **13**, 1528–1539. https://doi.org/10.1111/2041-210X.13858.

Swanson A, Kosmala M, Lintott C, Simpson R, Smith A, Packer C (2015). *Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna*. *Scientific Data* **2**, 150026. https://doi.org/10.1038/sdata.2015.26.

Tabak MA, Norouzzadeh MS, Wolfson DW, Sweeney SJ, Vercauteren KC, Snow NP, Halseth JM, Di Salvo PA, Lewis JS, White MD, Teton B, Beasley JC, Schlichting PE, Boughton RK, Wight B, Newkirk ES, Ivan JS, Odell EA, Brook RK, Lukacs PM, et al. (2019). *Machine learning to classify animal species in camera trap images: Applications in ecology*. *Methods in Ecology and Evolution* **10**, 585–590. https://doi.org/10.1111/2041-210X.13120.

Thompson KG (2007). *Use of Site Occupancy Models to Estimate Prevalence of Myxobolus cerebralis Infection in Trout*. *Journal of Aquatic Animal Health* **19**, 8–13. https://doi.org/10.1577/H06-016.1.

Thomsen PF, Kielgast J, Iversen LL, Wiuf C, Rasmussen M, Gilbert MTP, Orlando L, Willerslev E (2012). *Monitoring endangered freshwater biodiversity using environmental DNA*. *Molecular Ecology* **21**, 2565–2573. https://doi.org/10.1111/j.1365-294X.2011.05418.x.

Tyre AJ, Tenhumberg B, Field SA, Niejalke D, Parris K, Possingham HP (2003). *Improving Precision and Reducing Bias in Biological Surveys: Estimating False-Negative Error Rates*. *Ecological Applications* **13**, 1790–1801. https://doi.org/10.1890/02-5078.

Vélez J, McShea W, Shamon H, Castiblanco-Camacho PJ, Tabak MA, Chalmers C, Fergus P, Fieberg J (2023). *An evaluation of platforms for processing camera-trap data using artificial intelligence*. *Methods in Ecology and Evolution* **14**, 459–477. https://doi.org/10.1111/2041-210X.14044.

Welbourne D, Macgregor C, Paull D, Lindenmayer D (2015). *The effectiveness and cost of camera traps for surveying small reptiles and critical weight range mammals: A comparison with labour-intensive complementary methods*. *Wildlife Research* **42**, 414–425. https://doi.org/10.1071/WR15054.

Willoughby JR, Wijayawardena BK, Sundaram M, Swihart RK, DeWoody JA (2016). *The importance of including imperfect detection models in eDNA experimental design*. *Molecular Ecology Resources* **16**, 837–844. https://doi.org/10.1111/1755-0998.12531.

Wrege PH, Rowland ED, Keen S, Shiu Y (2017). *Acoustic monitoring for conservation in tropical forests: examples from forest elephants*. *Methods in Ecology and Evolution* **8**, 1292–1301. https://doi.org/10.1111/2041-210X.12730.

Zwart MC, Baker A, McGowan PJK, Whittingham MJ (2014). *The Use of Automated Bioacoustic Recorders to Replace Human Wildlife Surveys: An Example Using Nightjars*. *PLoS ONE* **9**. Ed. by Stephen G. Willis, e102770. https://doi.org/10.1371/journal.pone.0102770.