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

Non-invasive monitoring techniques like camera traps, autonomous recording units and environmental DNA are increasingly used to collect data for understanding species distribution. These methods have prompted the development of statistical models to suit with specific sampling designs and get reliable ecological inferences.
Site occupancy models estimate species occurrence patterns, accounting for the possibility that the target species may be present but unobserved. Here, two key processes are crucial: detection, when a species leaves signs of its presence, and identification where these signs are accurately recognized. While both processes are prone to error in general, wrong identifications are often considered as negligible with in situ observations. When applied to passive bio-monitoring data, characterized by datasets requiring automated processing, this second source of error can no longer be ignored as misclassifications at both steps can lead to significant biases in ecological estimates. Several model extensions aim to address these potential errors.
We propose an extended occupancy model that accounts for the identification process in addition to detection. Similar to other recent attempts to account for false positives, our model may suffer from identifiability issues, which usually require another source of data with perfect identification to resolve them. As an alternative when such data are unavailable, we propose leveraging existing knowledge of the identification process within a Bayesian framework by incorporating this knowledge through an informative prior. Through simulations, we compare different prior choices that encode varying levels of information, ranging from cases where no prior knowledge is available, to instances with accurate metrics on the performance of the identification, and scenarios based on generally accepted assumptions. We demonstrate that, compared to using a default prior, integrating information about the identification process as a prior reduces bias in parameter estimates. Overall, our approach mitigates identifiability issues, reduces estimation bias, and minimizes data requirements.
In conclusion, we provide a statistical method applicable to various monitoring designs, such as camera trap, bioacoustics, or eDNA surveys, alongside non-invasive sampling technologies, to produce ecological outcomes that inform conservation decisions.

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

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

Introduction

A primary objective for ecologists and conservation scientists is to understand how populations and communities are distributed across space and time. Monitoring animal species, plants, and even pathogens typically involves collecting data on their presence, and ideally, their absence, in order to evaluate their distribution area. Occupancy models have been developed by MacKenzie et al. (2002, see also Tyre et al., 2003) to account for potential undetected presence. These models estimate the proportion of sites occupied by a species while accounting for the imperfect detection of the species during field surveys (MacKenzie et al., 2002). Since a single visit is not sufficient to distinguish between a present but undetected species and its true absence from a site, MacKenzie et al. (2002) showed that repeated visits to the same site enable the estimation of the false-negative error rate, defined as the probability that a species present at a site remains undetected during a visit. Over the last decade, the development of new, non-invasive monitoring techniques such as 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) and environmental DNA sampling (e.g. Da Silva Neto et al., 2020; Griffin et al., 2020) has deeply changed data collection for biodiversity monitoring. The integration of passive sensor technologies into conservation projects is expanding, driven by technical improvements that facilitate the efficient monitoring of multiple species, including cryptic taxa, across large areas and challenging environments (Burton et al., 2015). However, these emerging methods are not exempt from imperfect detection. Indeed, certain discrete taxa may remain silent, do not trigger camera traps, or leave minimal detectable traces (Belmont et al., 2022; Goldman et al., 2023), so it remains essential to consider the probability of detecting them, regardless of the observation method used.

Within the context of sensor-based assessment method, data are massive and needs to be processed before being analyzed. In particular, this involves identifying the taxon of interest in a large amount of collected data, either manually by operators (Swanson et al., 2015; Welbourne et al., 2015), through automated 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). This step raises many statistical challenges (Hartig et al., 2024). For images and acoustic data, combining manual and automated processing helps to control classification errors; such as misidentifying one species as another (Barré et al., 2019). Similarly, environmental DNA studies also generate large datasets from which presence data must be extracted (Hunter et al., 2015; Schmidt et al., 2013; Thomsen et al., 2012). Detecting an organism's presence from its DNA in the environment is subject to various sources of variability, including the molecular techniques employed, laboratory procedures, and the amount of DNA collected (Doi et al., 2019; Willoughby et al., 2016). Despite the sensitivity of molecular techniques, once data is processed, distinguishing between real absences and those resulting from poor sampling or identification errors remains challenging (Goldberg et al., 2016). Thus, it is essential to consider both mis-identification and mis-detection in eDNA surveys. In eco-epidemiology studies, site occupancy models are used to estimate the occurrence of pathogens responsible for wildlife diseases within a sample unit, providing insights into spatial patterns and disease dynamics (McClintock et al., 2010b). The challenge for wildlife disease surveys is similar to that in camera-trapping for conservation, as both involve estimating occupancy parameters based on imperfect diagnostic tests (Lachish et al., 2012; McClintock et al., 2010b; Thompson, 2007).

The challenges of studies based on new biomonitoring technologies stem from the sequential nature of the detection and identification processes, each of which introduces two types of errors. A false-negative mis-identification occurs when a species is detected (e.g., the camera is triggered) but not correctly identified. Conversely, a false-positive mis-identification occurs when a species is not detected, but an error in data processing leads to its accidental identification. This two-step process increases the likelihood of errors in eDNA or sensor-based studies, compared to conventional surveys (Hartig et al., 2024). Failure to account for these identification errors can result in biased estimates of the actual proportion of occupied sites (MacKenzie et al., 2002; Spiers et al., 2022; Tyre et al., 2003). The standard site occupancy model accounts for false-negative errors by estimating the probability of imperfect detection, but it does not account for the possibility of false-positive detections, where a species is incorrectly identified at a site it does not occupy. False-positive errors, if unaddressed, can lead to overestimating occupancy probability (McClintock et al., 2010a; Miller et al., 2011; Royle and Link, 2006). Consequently, several authors have proposed extending MacKenzie's site occupancy model by accounting for false detection, although these extensions face identifiability issues (Chambert et al., 2015) often resolved by incorporating additional data sources, including one without errors. For example, Miller et al. (2011) proposed a multiple detection state model in which both certain and ambiguous data are used at each site. Building on this, Chambert et al. (2015) introduced the concept of "reference sites" exempt from detection error, and McKibben et al. (2023) revisited the notion of detection ambiguity introduced by Miller et al. (2011) by scoring observer confidence levels.

While these studies offer solutions for addressing detection errors, especially false positives, they rely on the integration of different data sources, 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. Although error-free data are rarely available, some knowledge about the reliability of the identification process may still be accessible (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 allows the integration of 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 available knowledge, otherwise the parameter estimates could be biased (Morris et al., 2015). Here, we propose a hierarchical model that builds on the classical occupancy model to account for identification errors across different types of data. We first provide a probabilistic description of the model, discuss the limitations of a frequentist approach for fitting this model, and then propose to overcome these limitations using a Bayesian framework that allows incorporating available information through informative priors. Through simulations, we compare the effectiveness of the different approaches.

Model Description

Standard Occupancy model

Detection and non-detection data on a species are collected from S sites, visited J times. These repeated visits help differentiate between sites where the species is truly absent and those

where the species is present but not detected. In the hierarchical formulation of the occupancy model (MacKenzie et al., 2002) 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:

$$Z_i \overset{i.i.d}{\sim} Bernoulli(\psi) \tag{1}$$
$$Y_{ij}|Z_i \overset{i.i.d}{\sim} Bernoulli(Z_i \times p)$$

Furthermore, it is assumed that species presence at one site is independent of its presence at other sites, meaning that Z_i (with i from 1 to S) are independent. Given the species is present at site i, Y_{ij} represents the detection state during 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).

While this model is appropriate for traditional field observations, it can be adapted according to the monitoring method. For some species, passive biomonitoring techniques offer a cost-effective alternative to field observations, but introduce new challenges. Unlike direct field observations, sensor data must be processed to determine species presence, and this introduces potential errors in detection history, including false positives, which are not accounted for in the standard occupancy model.

Extended model to identification level

To address these challenges, we extend the original model by introducing an additional identification process that accounts for potential errors in species identification. This step is particularly important when working with data where species identification can be ambiguous.

In this extended model, the potential detection becomes a latent variable Y_{ij} and we add a second layer to account for potential error in the identification process: an observation may correspond to a record (acoustic or image) where the species is identified (either correctly or incorrectly). Detection, however remains an unknown variable, referring to the sensor triggering and capturing the species' presence. In some cases, the species may be detected, but the quality of the recorded file may be too poor to reliably identify the species (Findlay et al., 2020). Thus it is impossible to deduce the detection state from the record alone.

To formalize this, 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 identify the detected species. 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 an observation to the wrong species.

In contrast to the standard model from MacKenzie et al. (2002), where the identification errors are not considered, assuming that $w_A = 1$ and $w_B = 1$, this extended model explicitly

accounts for the possibility of false identifications. In other words, the probability of failing to identify a species that has been detected is zero, as is the probability of confusing an undetected species with a detected one.

Given this extended framework, the conditional probability of identifying a species $W_{ij} = 1$ given that it is detected or not is written as:

$$W_{ij}|Y_{ij} = \begin{cases} W_{ij}|(Y_{ij} = 0) \sim Bernoulli(1 - w_{B,ij}) \\ W_{ij}|(Y_{ij} = 1) \sim Bernoulli(w_{A,ij}) \end{cases} \tag{2}$$

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).

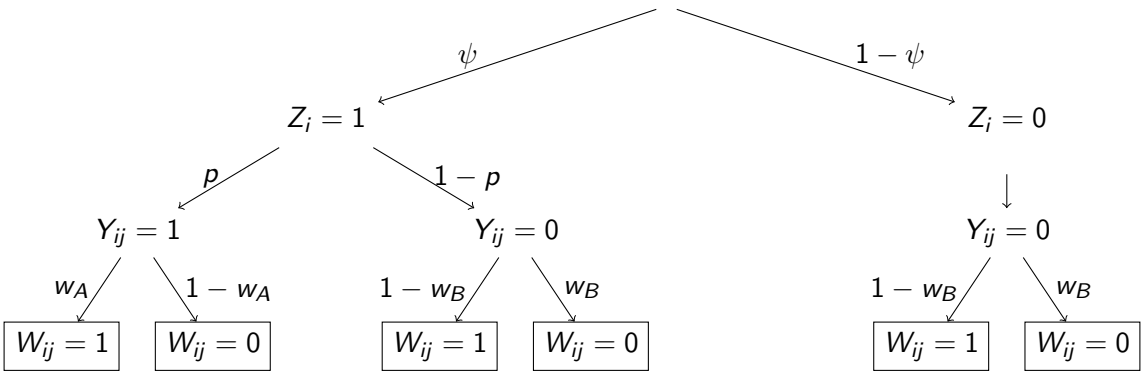


Figure 1 – Tree diagram illustrating the structure of the extended hierarchical model accounting for identification in occupancy. The nodes represent the possible events for the latent occupancy and detection variables, Z and Y, respectively associated with the occurrence probabilities ψ and p, defined along the branches. The leaves indicate the observed data, W_{ij}, recorded during visit j at site i, which depend on the detection state Y_{ij} and the associated identification probability : w_A if the species is detected ($Y_{ij} = 1$), and w_B otherwise. The detection of the target species ($Y_{ij} = 1$) occurs with probability ψ at an occupied site i (i.e $Z_i = 1$).

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}:

$$\begin{aligned} \pi_1(w) &:= Pr\left(W_{ij} = w \,|Z_i = 1\right) \\ &= Pr\left(W_{ik} = w,\ Y_{ij} = 1 \,|Z_i = 1\right) + Pr\left(W_{ij} = w, Y_{ij} = 0 \,|Z_i = 1\right) \\ &= Pr\left(W_{ij} = w \,|Y_{ij} = 1\right) Pr\left(Y_{ij} = 1 \,|Z_i = 1\right) + Pr\left(W_{ij} = w \,|Y_{ik} = 0\right) Pr\left(Y_{ij} = 0 \,|Z_i = 1\right) \\ &= w_A^w (1 - w_A)^{1-w} p + (1 - w_B)^w (w_B)^{1-w} (1 - p) \end{aligned} \tag{3}$$

$$\begin{aligned} \pi_0(w) &:= Pr\left(W_{ij} = w \,|Z_i = 0\right) \\ &= Pr\left(W_{ij} = w,\ Y_{ij} = 0 \,|Z_i = 0\right) \\ &= Pr\left(W_{ij} = w \,|Y_{ij} = 0\right) Pr\left(Y_{ij} = 0 \,|Z_i = 0\right) \\ &= (1 - w_B)^w (w_B)^{1-w} \end{aligned} \tag{4}$$

For example, at a site visited three times, where the species is identified only during the second visit, the identification history would be 010. Out of these three visits, the occupancy

state of the site is unknown but the species was identified once so we combine equations 3, 4, 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:

$$\begin{aligned} Pr\left(W_i = (0,1,0)\right) &= Pr\left(W_i = (0,1,0),\ Z_i = 1\right) + Pr\left(W_i = (0,1,0),\ Z_i = 0\right) \\ &= Pr\left(W_i = (0,1,0)\,|Z_i = 1\right) Pr\left(Z_i = 1\right) + Pr\left(W_i = (0,1,0)\,|Z_i = 0\right) Pr\left(Z_i = 0\right) \\ &= \psi\pi_1(w_{i1})\pi_1(w_{i2})\pi_1(w_{i3}) + (1-\psi)\pi_0(w_{11})\pi_0(w_{12})\pi_0(w_{13}) \end{aligned} \tag{5}$$

Finally, for S independent sites, each with J independent visits - where j_i^* denotes positive identification - and assuming constant parameters across visits and sites, the model likelihood can be expressed as :

$$\begin{aligned} L(w_A, w_B, p, \psi\,|data) &= \prod_{i=1}^{N} Pr(W_i) = \prod_{i=1}^{N} \left(Pr(W_i, Z_i = 1) + Pr(W_i, Z_i = 0)\right) \\ &= \prod_{i=1}^{N} \Big[\psi\left[(1-w_B)(1-p) + w_A p\right]^{j_i^*} \left[w_B(1-p) + (1-w_A)p\right]^{J-j_i^*} + \\ &\qquad (1-\psi) w_B^{J-j_i^*} (1-w_B)^{j_i^*}\Big] \end{aligned} \tag{6}$$

Simulation study

Classical estimation with a frequentist approach

In this section, we assess the quality of estimates obtained through maximum likelihood using a simulation study. Specifically, we aim to assess two key aspects: first, whether incorporating the identification process and accounting for its two types of error leads to more reliable estimates; second, how the number of site visits affects the precision of these estimates.

In order to investigate these points, we carried out simulations by generating 1000 data sets with N=30 sites and J=12 or 36 visits according to our proposed model defined in Equations (1), (2). The parameter values used to create the matrices of observations were $\psi = 0.8$, $p = 0.5$, $w_A = 0.9$ and $w_B = 0.7$. These values were chosen based on a site occupancy study of the Eurasian lynx (Lynx lynx) population in France (Gimenez et al., 2022). After generating the datasets, we applied maximum likelihood estimation by minimizing the negative log-likelihood function to obtain parameter estimates (Equ. 6). To examine the influence of the number of visits, we compared the precision of estimates between datasets with 12 visits and those with 36 visits.

The results reveal that the occupancy parameter, ψ, tends to be overestimated when using the original model without the identification. This overestimation occurs because, in the absence of the identification process, all sites with at least one positive identification are assumed to be occupied (mean estimates for 1000 simulations with the original model for 36 visits : $\bar{\hat{\psi}} = 1$).

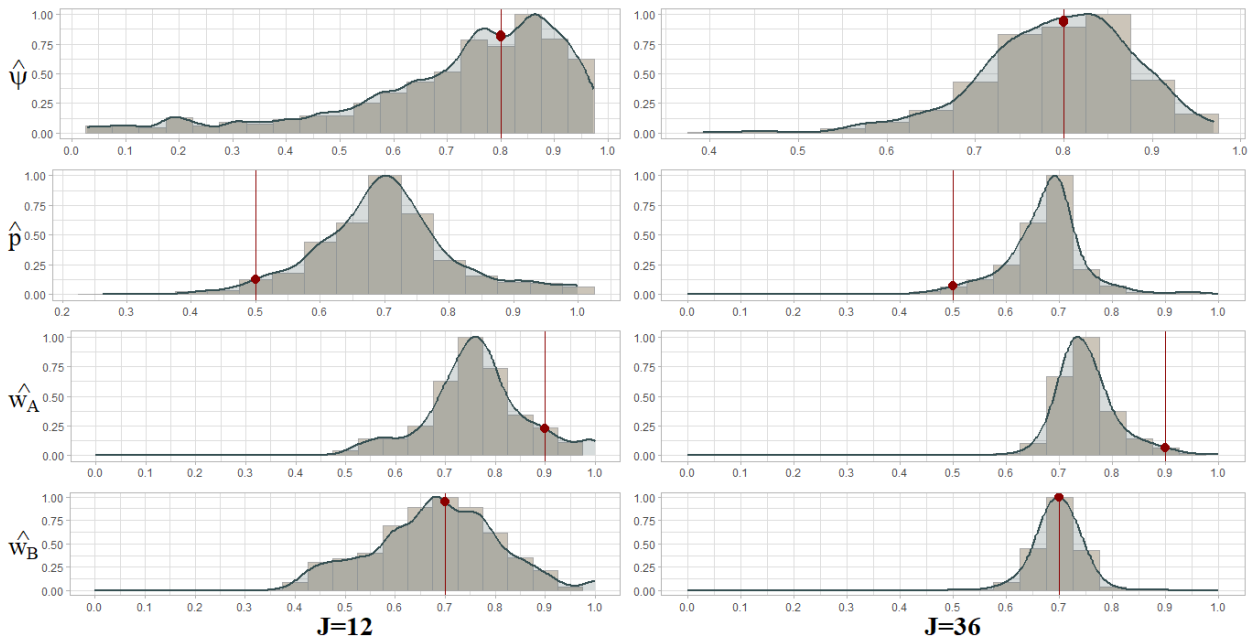


Figure 2 – Site Occupancy Model accounting for false-positive and false-negative errors in the identification layer. Histogram and kernel estimates of the distribution of maximum-likelihood estimates for 1000 simulations for J=12 (left column) or J=36 (right column) visits on N=30 sites, and the initial 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 $\hat{w_B}$.

Identifiability issues

Previous studies have demonstrated that parameter estimates become biased if false-positive detections are not properly 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).

In our analysis, we used the standard deviation of estimates 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 $\hat{\psi}$)(Fig. 2). However, despite the increase in available data from 36 visits, the estimates for the detection probability, $\hat{p}$, and the positive identification probability, $\hat{w_A}$, remain biased ($Bias(\hat{p}) = 0.17$ and $Bias(\hat{w_A}) = -0.15$).

One way to address these biases is to fix one of the two parameters, w_A or p, then the other can be estimated without bias (Supplementary A.1). Such parameter redundancy in the likelihood function is at the core of model identifiability issues (Supplementary A.2, A.1)(Gimenez et al., 2004).

Addressing identifiability issues with a constraint

To further address the lack of identifiability in models that incorporate misdetection, Royle and Link (2006) suggested to impose constraints on the model. They proposed to set the probability to correctly detect a present species higher than the probability to incorrectly detect it when it is absent. We first explore this recommendation using a frequentist approach, before turning on a Bayesian approach using informative priors in order to solve these identifiability issues.

To adapt the recommended constraint to our model, we apply it on the identification probabilities, such that $w_A > 1 - w_B$. This ensures that the probability of correctly identifying the species is higher than the probability of making a false positive identification.

To evaluate the impact of this constraint, we simulated 1000 datasets with values for the true-positive identification probability w_A and the true-negative identification probability w_B ranging between 0.5 and 0.95. We then estimated the parameters of our site occupancy model accounting for both types of error in the identification layer, using maximum likelihood estimation with and without the constraint.

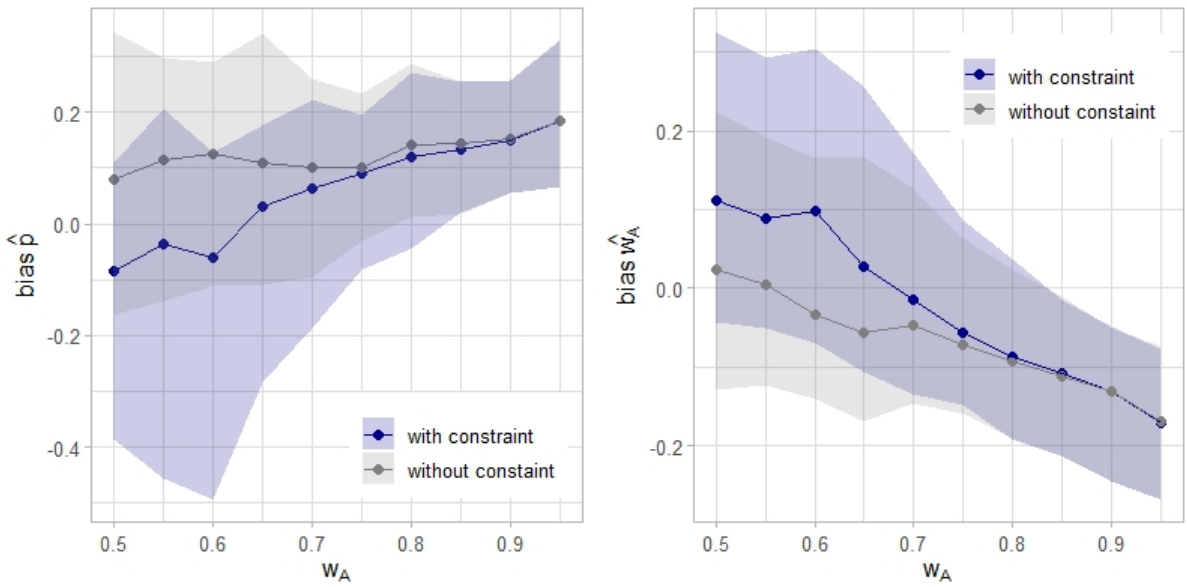


Figure 3 – Bias trend as a function of the probability of correctly identifying the species. The focus is on parameters likely to be biased by identifiability issues : the detection estimates $\hat{p}$ (on the left), and the correct identification estimates $\hat{w_A}$ (on the right). The bias is contrasted between two optimization cases: under the constraint *(in blue)* stating that the probability of correctly identifying the species is higher than the probability of incorrectly identifying the species, and without the constraint *(in gray)*. The bias is assessed according to the true value of w_A used in the data simulation, and is calculated based on the median and the range between the 0.1 and 0.9 quantiles of the maximum-likelihood estimates.

The results show that applying the constraint reduces the bias in the detection probability estimates ($\hat{p}$ for values of w_A and w_B around 0.5 ; Supplementary A.3). Moreover, regardless of the initial value of w_A, the estimates of $\hat{w_A}$ are concentrated around 0.7, which leads to a reduction in bias as the value of $\hat{w_A}$ (Fig. 2). As w_A and w_B approach higher values, the estimates produced with and without the constraint become more similar. Nevertheless, while the constraint helps reduce bias, it may not be strong enough to completely eliminate the identifiability issue (Fig. 3). This is because, in practice, the true-positive rate, w_A, is generally higher than the false-positive rate $1 - w_B$ (Guillera-Arroita et al., 2017).

Using an informative prior to address identifiability issues

In this section we address the issue of the model identifiability by leveraging knowledge about the risk of misidentifications, even in the absence of additional data sources. We adopt a Bayesian approach, incorporating this knowledge through the use of an informative prior.

In many situations, it is possible to have a good knowledge of the false-negative rate in the identification process. In particular, we are interested in utilising prior knowledge regarding the sensitivity of the identification process as a means of addressing the redundancy between detection and positive identification parameters, previously described. As the process of species identification is inherently imperfect, its performance is evaluated through the implementation of tests which compare the predicted identifications to the actual outcomes of a verified dataset. Insofar as the underlying truth of the data is not accessible, these performance tests must be carried out beforehand, thus facilitating the acquisition of knowledge regarding the risk of misidentifications. Therefore, the inclusion of additional data sources free of one kind of misidentification is not necessary.

In the context of sensor data classified by a deep learning algorithm, labelled data are used to evaluate the performance of the classifier before employing it for the classification of unlabeled data (Pichler and Hartig, 2023). Performance tests are designed to compute metrics that quantify both types of misclassifications. These include the recall defined as the true positive rate (or sensitivity) for each class, and which is of particular interest in the context of identifying one target species (Pichler and Hartig, 2023). This information is often accessible in the confusion matrix of a classifier, and the transfer learning ensures the consistency of the classifier's performance on other datasets (Norouzzadeh et al., 2021; Tabak et al., 2019; Vélez et al., 2023). Those performance metrics, including sensitivity, may constitute prior knowledge that is more or less informative. 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 parameter 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.

A highly informative knowledge is characterised by a precise definition of the sensitivity with a median value enhanced by a confidence interval. Consequently, the sensitivity can be expressed as a density distribution with a mean and a standard deviation (e.g. Griffin et al., 2020 with 0.81 [0.71,0.90] and Tabak et al., 2020 provide the recall values and 95% confidence intervals for each studied species with *MLWIC2*). In this context, a beta distribution is the most appropriate distribution to elicit a prior on the probability of correctly identifying a species present (Banner et al., 2020). In the case of lesser but still informative knowledge, sensitivity can be defined as a unique value without any confidence interval (e.g Schneider et al., 2024 give the confusion matrices from their open species recognition models, and the *Wildlife Insights* (2024) platform gives its classifier's performance metrics for many species). We then specified a spread beta distribution as a diluted informative prior. In the absence of information concerning the sensitivity of the identification process, it may be reasonably argued that the probability of correctly identifying the target species in an occupancy study is greater than the probability of incorrectly identifying it. This vague knowledge justifies the consideration of a flat uniform distribution ranging from 0.5 to 1 for the positive identification parameter.

Based on Banner et al. (2020) proposition, we study 4 different types of prior for parameter w_A, whose default prior, according to the available knowledge on the sensitivity of the identification process (Supplementary A.4) :

- a uniform distribution from 0 to 1, as a default non-informative prior for a probability,

- a flat uniform distribution ranging from 0.5 to 1, as a vague non-informative prior adapted to the context of identification for occupancy,
- a spread-out beta distribution, as a weakly informative prior,
- a tight beta distribution, as a highly informative prior.

The beta prior distribution was elicited using a matching method to accurately define its parameters (Denham and Mengersen, 2007; Falconer et al., 2022). Following the approach proposed by Wu et al. (2008) we constructed a unimodal beta distribution through a two-step process. First, we aligned the sensitivity value with the mode of the beta distribution, which represents the most frequent value. Here the sensitivity value is 0.9 according to the values used for the simulations and as a reference to Gimenez et al. (2022). Subsequently, we integrated the probability density function by utilizing the confidence interval of the sensitivity as the distribution's range. We simulated 100 observation datasets and we estimated model parameters in a Bayesian framework (using NIMBLE v1.2.0; de Valpine et al., 2024) for each prior distributions of w_A (the distribution priors of all the others parameters are default prior i.e $\mathcal{U}(0,1)$). We used a block sampler accounting for the correlation between the detection p, and the positive identification w_A, parameters. The model convergence was analysed for different values of positive identification probability as a simulation parameter (Supplementary A.5, A.6).

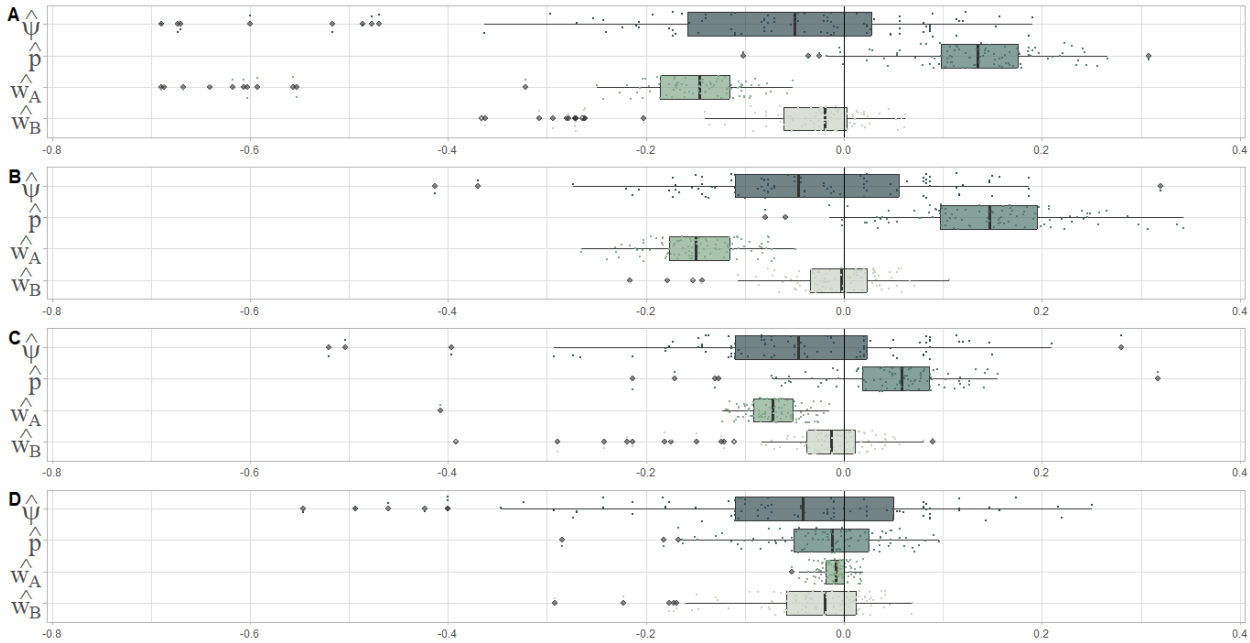


Figure 4 – Boxplot of the difference between the median values of the posterior distributions and the parameter values calculated from simulated datasets. Occupancy parameters are set to fixed values to simulate 100 datasets : $\psi = 0.8$, $p = 0.5$, $w_A = 0.9$, $w_B = 0.7$. The sensitivity parameter (w_A) is introduced as **(A)** a default non-informative prior with a uniform distribution $\mathcal{U}(0,1)$, **(B)** a vague prior with a uniform distribution like $\mathcal{U}(0.5,1)$, **(C)**a weakly informative prior with a beta distribution like $\mathcal{B}(8.8,1.9)$, and **(D)** a highly informative prior with a beta distribution like $\mathcal{B}(45,5)$.

Using non-informative priors for identification parameters leads to biased posterior distributions, especially for the detection and positive identification parameters. The mean bias associated with the median of the posterior for $\hat{p}$ and $\hat{w_A}$ are 0.13 and -0.19, respectively, when using a default non-informative prior for sensitivity. Notably the negative bias on the positive identification parameter, w_A, is not fully compensated by the bias on the detection parameter. The

inference for the detection probability $\hat{p}$ improves when an informative prior for sensitivity is applied. In this case, the mean bias associated with the median of the posterior for $\hat{p}$ decreases to -0.02 with a highly informative prior (Fig. 4). A vague non-informative prior slightly reduces the mean mean bias in the median of the posteriors of $\hat{\psi}$. The informative priors used represent two different approaches to integrate information about the identification process, and both perform comparably. Actually, the median values of $\hat{\psi}$ posteriors, obtained for 100 simulations are only weakly affected by the type of prior.

Discussion

We proposed a single-species occupancy model that can be applied to various data types, including images, acoustic recordings, and molecular data. This model acknowledges the two-step structure of the observation process, consisting of detection and identification. Our hierarchical occupancy model considers both detection and identification processes, which are independent source of errors. On the one hand, we account for false negatives in detection using the detection parameter p, and on the other hand, we address identification errors, whether in favor of the target species or not, with parameters w_A and w_B. Initially, we implemented our model within a maximum-likelihood framework, but we encountered biases in some estimates due to model mis-specifications and identifiability issues. By shifting to a Bayesian approach and using informative priors based on identification performance metrics, such as sensitivity, we successfully mitigated these identifiability issues.

The deployment of sensors and molecular techniques generates more data than conventional sampling methods, and because these data are not inherently specific to any species, they require further sorting to identify the target species. Particularly with sensor data, this secondary stage may involve multiple observers, through crowd-sourced projects (e.g. *Zooniverse* 2024) for images classification, or expert analysis for acoustic data (e.g. Shonfield and Bayne, 2017; Zwart et al., 2014). Automated species recognition can reduce processing time, but without human verification which is time-consuming (Barré et al., 2019; Spiers et al., 2022), identification errors can distort inferences (Ferguson et al., 2015; Lonsinger et al., 2023; McClintock et al., 2010a). Accounting for these identification errors in addition to detection errors requires developing 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 essence, dividing a visit into two different detection events is equivalent to the two-stage survey protocol proposed by Guillera-Arroita et al. (2017), which we rely on. Finally, by reducing data processing time through automation and the absence of human validation, potential identification errors are introduced, which, especially false positives, may have a severe impact on inferences. As the number of model parameters increases to better accommodate different sampling levels, the price to pay is that some parameters become difficult to estimate. Several authors have therefore suggested combining multiple sources of information (Chambert et al., 2015; Guillera-Arroita et al., 2017; Miller et al., 2011) to overcome the problem of identifiability. However, since increasing data sources is costly, we propose using performance metrics from the identification process to inform priors.

In the context of molecular data, a species is detected if its DNA is present in the sample, and it is identified if its DNA is observed in a PCR analysis replicate (Schmidt et al., 2013). Sensitivity

is thus defined as the probability of correctly identifying the species, or pathogen, in the replicate. Unlike acoustic or camera trap methods, where detection and identification can be separated, this distinction is more challenging in eDNA surveys, where the sample composition remains unknown until molecular and bioinformatics analysis are performed (Goldberg et al., 2016). Some studies use additional surveys to verify species presence and calibrate eDNA sensitivity, while others rely on experimental or statistical methods (e.g. Griffin et al., 2020; Mathieu et al., 2020). The use of positive control involving foreign DNA, can help to identify PCR inhibition and provide information on the false-positive rate (e.g. Furlan et al., 2016; Goldberg et al., 2016)(Hyatt et al., 2007). Nevertheless, quantifying sensitivity remains challenging across studies using similar methodologies due to high variability in taxa, environmental, and experimental conditions (Gold et al., 2023; Keller et al., 2022; Thomsen et al., 2012). Despite this, eDNA is generally more sensitive than other sampling methods (Darling and Mahon, 2011), though this heightened sensitivity may increase the likelihood of false positives (Cristescu and Hebert, 2018). Taking into account the identification process is therefore crucial, although the positive identification rate (w_A) must be close enough to 1 to guarantee the convergence of the model.

The main limitation of our approach lies in the fact that we need to gather knowledge on the performance of the identification process to construct a relevant informative prior. While this knowledge is necessary, it is still less costly than incorporating additional data sources, especially if sensitivity information is provided by another study, or as a parameter of the identification tool (e.g. Tabak et al., 2020, Rigoudy et al., 2023). Indeed, we suggest that when using deep learning algorithms for species classification, or following a molecular and bioinformatics pipeline for eDNA, the performance metrics of the methods should be made accessible. Simulations indicate that even with non-informative priors, our model produces reliable posterior estimates of the presence parameter (ψ). When only presence is of interest, we recommend using this model with non-informative priors to handle misidentifications and detection errors while disregarding identifiability issues in the detection parameter. However, when the detection parameter is of concern, using an informative is necessary to address parameter redundancy. Cruickshank et al. (2019) successfully avoided identifiability issues related to false-positive errors by integrating informative prior based reasonable assumptions from volunteer-collected monitoring data. Similarly, our approach, which incorporates prior information about the identification process, produces robust posterior estimates and provides an alternative to approaches requiring additional datasets. Also, as in many studies using a Bayesian approach, the choice of a wrong prior for a parameter may causes bias in the definition of the posterior distribution for this parameter (Northrup and Gerber, 2018).

Passive sensors like camera traps and autonomous recording units offer valuable opportunities for addressing a wide range of ecological and conservation questions. Combined with approaches like eDNA sampling, these technologies enable ecologists to collect data at large spatial scales or fine temporal resolutions and study cryptic species (Ross et al., 2023; Sahu et al., 2023). For such large and complex datasets, accurate taxonomic identification is challenging, but accounting for the noise generated during processing is essential.. In this context, our proposed model 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

A.1. Identifiability issues

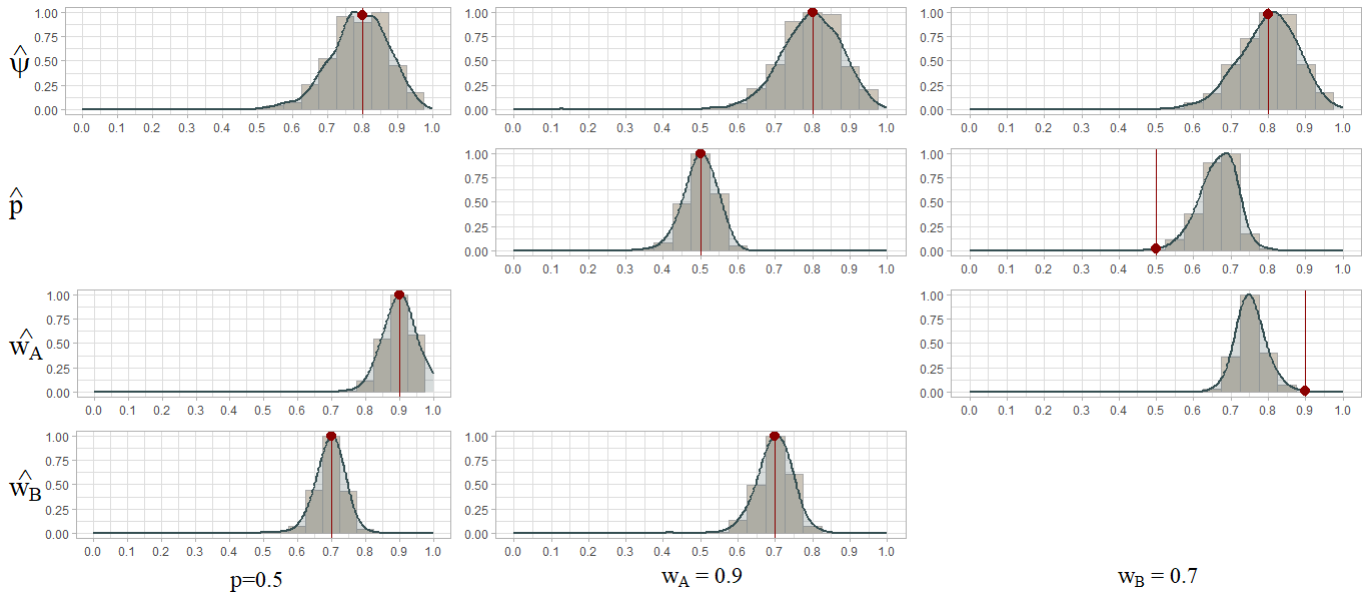


Figure A.1 – Distribution of maximum-likelihood estimates for 1000 simulations when a parameter is set to a constant value (in columns).Detection (p) and identification parameters (w_A and w_B) are successively excluded from the estimation, since their value are fixed in the expression of the likelihood function.

$\hat{w_A}$ 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.

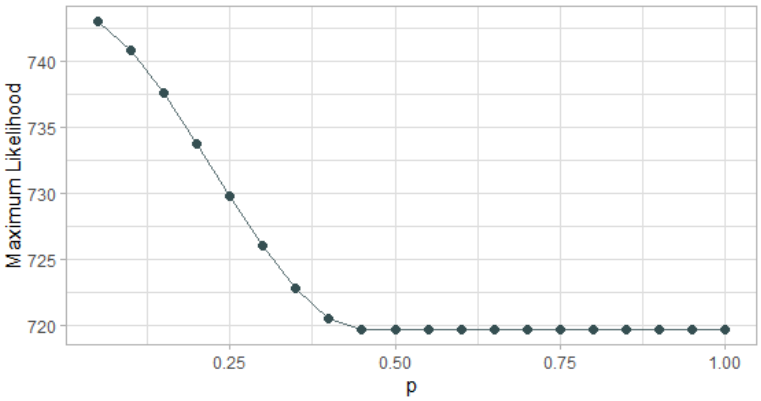


Figure A.2 – Profile deviance on p

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.

As pointed out by Royle and Link (2006), including false positives raises concerns about model identifiability . To address this issue of parameter redundancy, the authors proposed to set

Table A.1 – Profile deviance on detection parameter 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**

a constraint during likelihood optimization. Specifically, they suggest ensuring that the probability of correctly detecting a species is higher than the probability of falsely detecting it. Applying this constraint to our model with an identification layer means that correctly identifying the target species is more likely than falsely identifying it when it has not been detected.

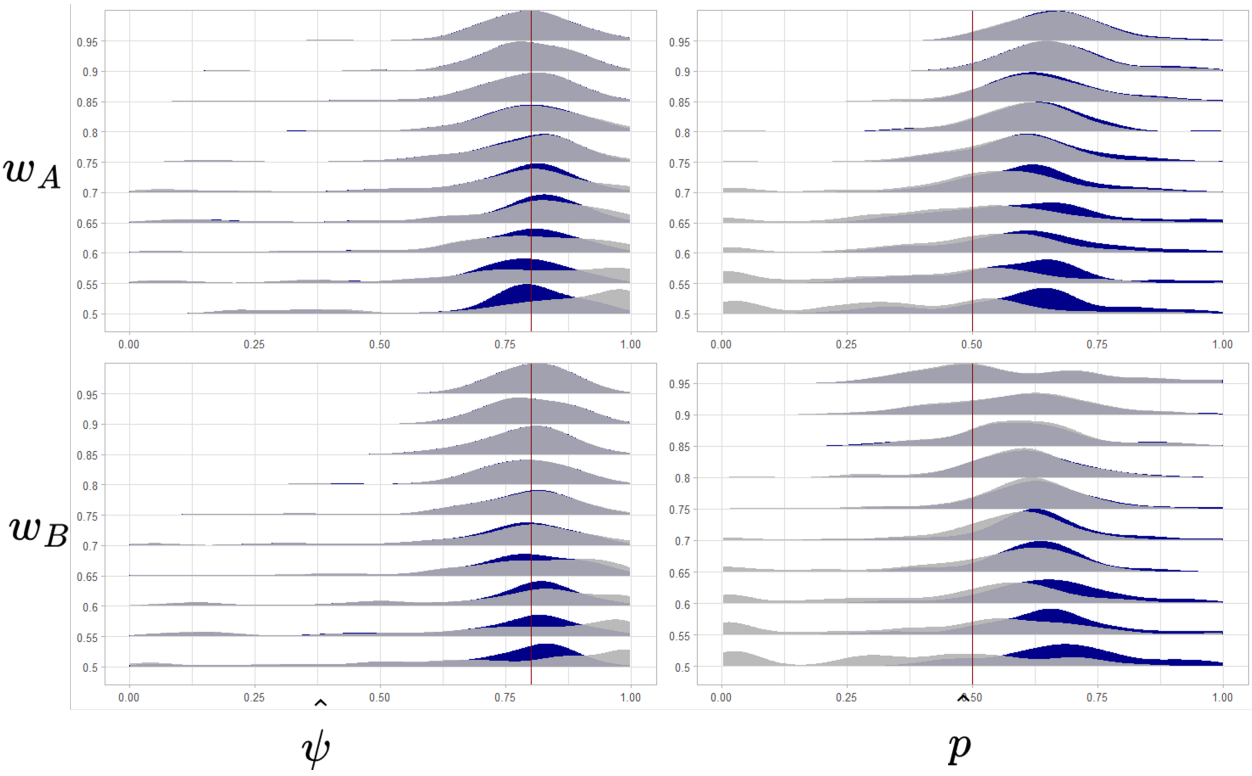


Figure A.3 – Distribution of $\hat{\psi}$ and $\hat{p}$ for 1000 simulated data sets for different values of identification parameters in the simulated data. With w_A set between 0.5 and 0.95 (top) and w_B set between 0.5 and 0.95 (bottom). Distributions of occupancy ($\hat{\psi}$) and detection $\hat{p}$ parameters are the results of optimization under the constraint $\hat{w_A} > 1 - \hat{w_B}$ (in gray) and without it (in blue). The true value of parameters are indicated by the red vertical bar.

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.

A.2. Using an informative prior to address identifiability issues

We evaluate the posterior distributions of the occupancy estimates according to four priors with different level of informativeness for the positive identification parameter, w_A, called sensitivity.

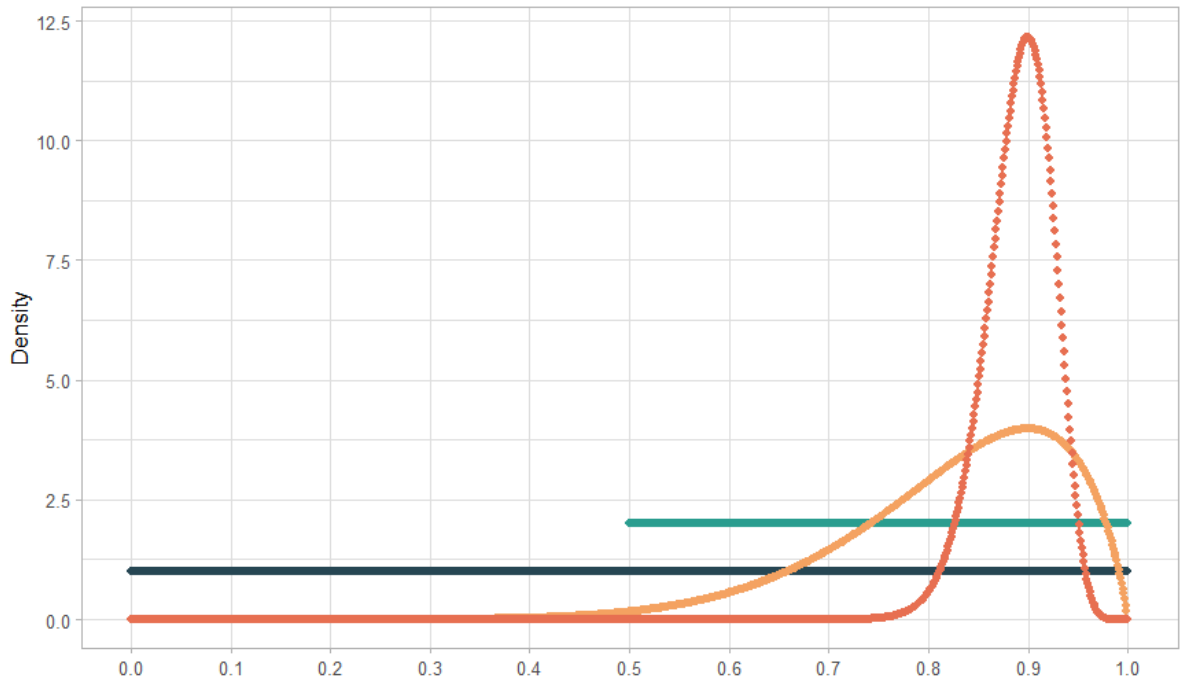


Figure A.4 – Prior distributions for the positive identification parameter or sensitivity w_A**.** Non informative prior (in blue) are uniform distributions : from 0 to 1 (in dark blue) and from 0.5 to 1 (in light blue). Informative priors (in orange) are beta distributions such as $\mathcal{B}(8.8, 1.9)$ is weakly informative (in light orange) and $\mathcal{B}(76, 9.3)$ is highly informative (in dark orange).

We elicited the beta priors by solving a 2 equations system explicating the mode and the density probability function with the beta distribution parameters, α and β, unknown (in the manner of the location and intervals method of Wu et al. (2008)) :

$$mode = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{7}$$

$$\int_0^R \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathcal{B}(\alpha, \beta)} dx - 0.01 = 0 \text{ for } P(x < R) = 0.01 \tag{8}$$

For both priors the mode is set to 0.9 which is the value chosen to simulate data. R is defined as the threshold below which the probability to find the value of sensitivity is nearly null : it is 0.5 in the case of a weakly informative prior and 0.8 in the case of the highly informative one. We ran with NIMBLE (v1.2.0; de Valpine et al., 2024) 2 chains on 4000 iterations following a 1000 iterations burn-in period. We assessed the model convergence through the R-hat and the trace and density plots (MCMCvis R package v0.16.3; Youngflesh, 2018), for each alternative priors .

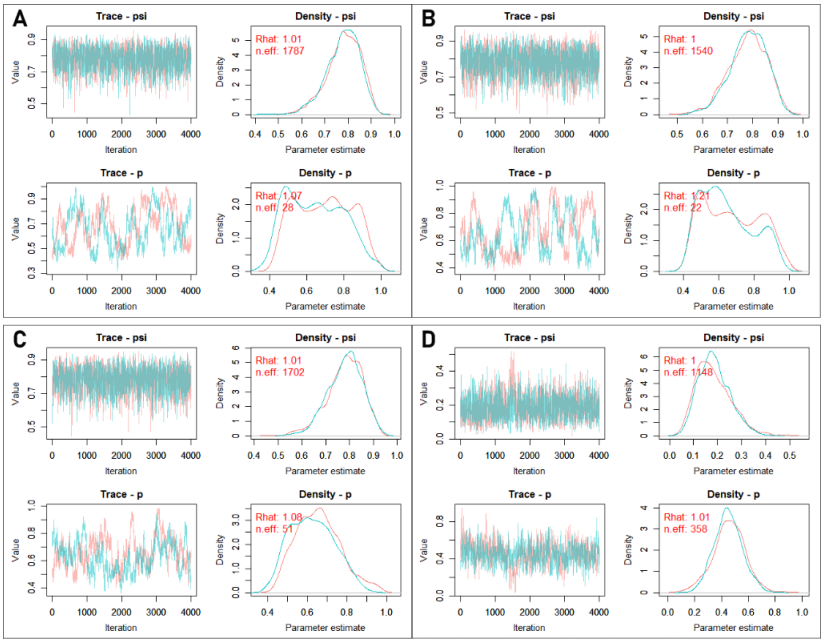


Figure A.5 – Chain trace and density plots of occupancy, $\hat{\psi}$, and detection, $\hat{p}$, posterior distribution, according 4 different priors on sensitivity parameter, w_A. On each of the 4 panels, the trace plots (on the left) represent the evolution of both chains on 4000 iterations, and the density plots (on the right) represent the posterior distribution for each chain. The distribution priors on w_A are (**A**) $\mathcal{U}(0, 1)$, (**B**) $\mathcal{U}(0.5, 1)$, (**C**) $\mathcal{B}(8.8, 1.9)$ and (**D**) $\mathcal{B}(76, 9.3)$.

Chains convergence is reached for ψ whatever the prior on w_A, however only the most informative prior enable a satisfying mix of chains for the detection parameter p (R-hat=1.01).

Finally, we drove a sensitivity analysis for 3 values of w_A (0.2, 0.5 and 0.8) used to simulate data. We used a highly informative prior in order to evaluate the impact of the value of w_A on the convergence. The chains for the occupancy estimates do not converge when the positive identi-

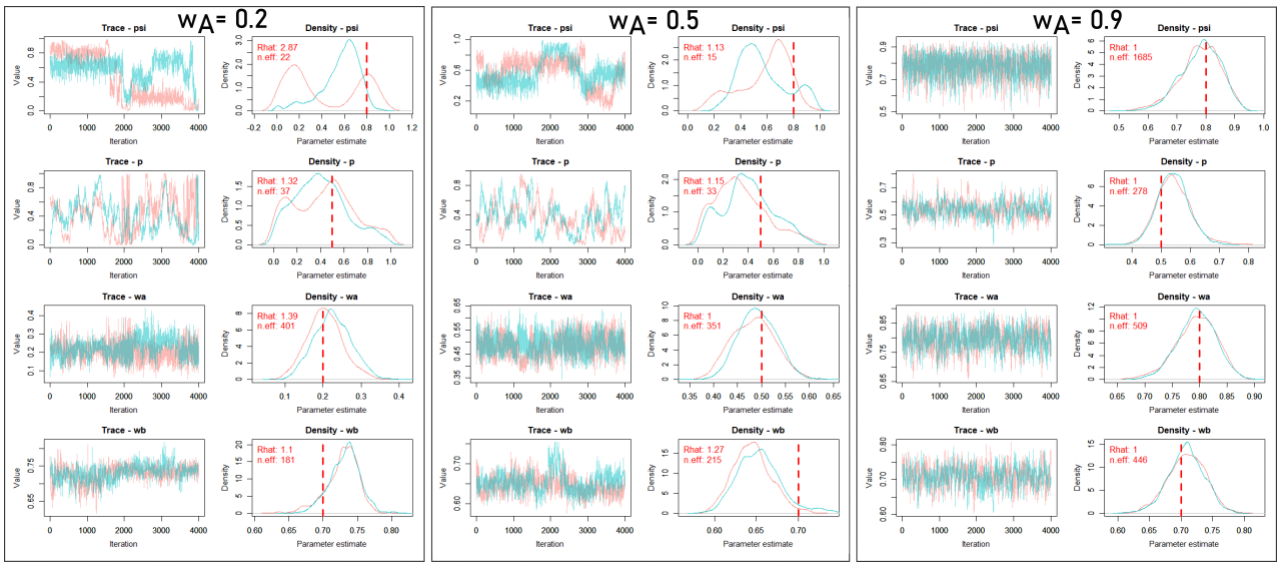


Figure A.6 – Sensitivity analysis of the extended occupancy model using an highly informative on the positive identification parameter, w_A. Data are simulated for 30 sites visited 36 times with fixed generative values (red dashed line) except for w_A.

fication rate is below 0.5, though this scenario seems unrealistic. This model should only be used when the sensitivity of the identification process is high (greater than 0.75). Indeed, if sensitivity falls below this threshold, the identification process should be considered too underperforming for use in occupancy studies.

Acknowledgements

We would like to acknowledge the assistance of ChatGPT, a language model developed by OpenAI, in improving the clarity and quality of the writing in this manuscript.

Fundings

This project has received financial support from the CNRS through the MITI interdisciplinary programs.

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.

Banner KM, Irvine KM, Rodhouse TJ (2020). *The use of Bayesian priors in Ecology: The good, the bad and the not great*. *Methods in Ecology and Evolution* **11**, 882–889. https://doi.org/10.1111/2041-210X.13407.

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**, 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.

Cristescu ME, Hebert PDN (2018). *Uses and Misuses of Environmental DNA in Biodiversity Science and Conservation*. *Annual Review of Ecology, Evolution, and Systematics* **49**, 209–230. https://doi.org/10.1146/annurev-ecolsys-110617-062306.

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.

Darling JA, Mahon AR (2011). *From molecules to management: Adopting DNA-based methods for monitoring biological invasions in aquatic environments*. *Environmental Research* **111**, 978–988. https://doi.org/10.1016/j.envres.2011.02.001.

de Valpine P, Paciorek C, Turek D, Michaud N, Anderson-Bergman C, Obermeyer F, Wehrhahn Cortes C, Rodrìguez A, Temple Lang D, Paganin S (2024). *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling*. Version 1.2.0. R package version 1.2.0. https://doi.org/10.5281/zenodo.1211190.

Denham R, Mengersen K (2007). *Geographically Assisted Elicitation of Expert Opinion for Regression Models*. *Bayesian Analysis* **2**, 99–136.

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*. *Decision Analysis* **19**, 189–204. https://doi.org/10.1287/deca.2022.0451.

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.

Furlan EM, Gleeson D, Hardy CM, Duncan RP (2016). *A framework for estimating the sensitivity of eDNA surveys*. *Molecular Ecology Resources* **16**, 641–654. https://doi.org/10.1111/1755-0998.12483.

Gimenez O, Viallefont A, Catchpole EA, Choquet R, Morgan BJT (2004). *Methods for investigating parameter redundancy*. *Animal Biodiversity and Conservation* **27.1**, 561–572.

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.

Gold Z, Koch MQ, Schooler NK, Emery KA, Dugan JE, Miller RJ, Page HM, Schroeder DM, Hubbard DM, Madden JR, Whitaker SG, Barber PH (2023). *A comparison of biomonitoring methodologies for surf zone fish communities*. *PLOS ONE* **18**, e0260903. https://doi.org/10.1371/journal.pone.0260903.

Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, McKee A, Oyler-McCance SJ, Cornman RS, Laramie MB, Mahon AR, Lance RF, Pilliod DS, Strickler KM, Waits LP, Fremier AK, Takahara T, Herder JE, Taberlet P (2016). *Critical considerations for the application of environmental DNA methods to detect aquatic species*. *Methods in Ecology and Evolution* **7**, 1299–1307. https://doi.org/10.1111/2041-210X.12595.

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.

Hartig F, Abrego N, Bush A, Chase JM, Guillera-Arroita G, Leibold MA, Ovaskainen O, Pellissier L, Pichler M, Poggiato G, Pollock L, Si-Moussi S, Thuiller W, Viana DS, Warton DI, Zurell D, Yu DW (2024). *Novel community data in ecology-properties and prospects*. *Trends in Ecology & Evolution* **39**, 280–293. https://doi.org/10.1016/j.tree.2023.09.017.

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.

Hyatt A, Boyle D, Olsen V, Boyle D, Berger L, Obendorf D, Dalton A, Kriger K, Hero M, Hines H, Phillott R, Campbell R, Marantelli G, Gleason F, Colling A (2007). *Diagnostic assays and sampling protocols for the detection of Batrachochytrium dendrobatidis*. *Diseases of Aquatic Organisms* **73**, 175–192. https://doi.org/10.3354/dao073175.

Keller AG, Grason EW, McDonald PS, Ramón-Laca A, Kelly RP (2022). *Tracking an invasion front with environmental DNA*. *Ecological Applications* **32**, e2561. https://doi.org/10.1002/eap.2561.

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*, 56–71. 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.

Mathieu C, Hermans SM, Lear G, Buckley TR, Lee KC, Buckley HL (2020). *A Systematic Review of Sources of Variability and Uncertainty in eDNA Data for Environmental Monitoring*. *Frontiers in Ecology and Evolution* **8**. https://doi.org/10.3389/fevo.2020.00135.

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.

Norouzzadeh MS, Morris D, Beery S, Joshi N, Jojic N, Clune J (2021). *A deep active learning system for species identification and counting in camera trap images*. *Methods in Ecology and Evolution* **12**, 150–161. https://doi.org/10.1111/2041-210X.13504.

Northrup JM, Gerber BD (2018). *A comment on priors for Bayesian occupancy models*. *PLOS ONE* **13**, 1–13. https://doi.org/10.1371/journal.pone.0192819.

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.

Pichler M, Hartig F (2023). *Machine learning and deep learning—A review for ecologists*. *Methods in Ecology and Evolution* **14**, 994–1016. https://doi.org/10.1111/2041-210X.14061.

Rigoudy N, Dussert G, Benyoub A, Besnard A, Birck C, Boyer J, Bollet Y, Bunz Y, Caussimont G, Chetouane E, Carriburu JC, Cornette P, Delestrade A, De Backer N, Dispan L, Le Barh M, Duhayer J, Elder JF, Fanjul JB, Fonderflick J, et al. (2023). *The DeepFaune initiative: a collaborative effort towards the automatic identification of European fauna in camera trap images*. *European Journal of Wildlife Research* **69**, 113. https://doi.org/10.1007/s10344-023-01742-7.

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, 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.

Schneider D, Lindner K, Vogelbacher M, Bellafkir H, Farwig N, Freisleben B (2024). *Recognition of European mammals and birds in camera trap images using deep neural networks*. *IET Computer Vision*. https://doi.org/10.1049/cvi2.12294.

Shonfield J, Bayne EM (2017). *Autonomous recording units in avian ecological research: current use and future applications*. *Avian Conservation and Ecology* **12**, 14. 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, Newton EJ, Boughton RK, Ivan JS, Odell EA, Newkirk ES, Conrey RY, Stenglein J, Iannarilli F, Erb J, Brook RK, Davis AJ, Lewis J, Walsh DP, Beasley JC, VerCauteren KC, Clune J, Miller RS (2020). *Improving the accessibility and transferability of*

machine learning algorithms for identification of animals in camera trap images: MLWIC2. *Ecology and Evolution* **10**, 10374–10383. https://doi.org/10.1002/ece3.6692.

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.

Wildlife Insights (2024). URL: https://www.wildlifeinsights.org/about-wildlife-insights-ai (visited on 09/23/2024).

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.

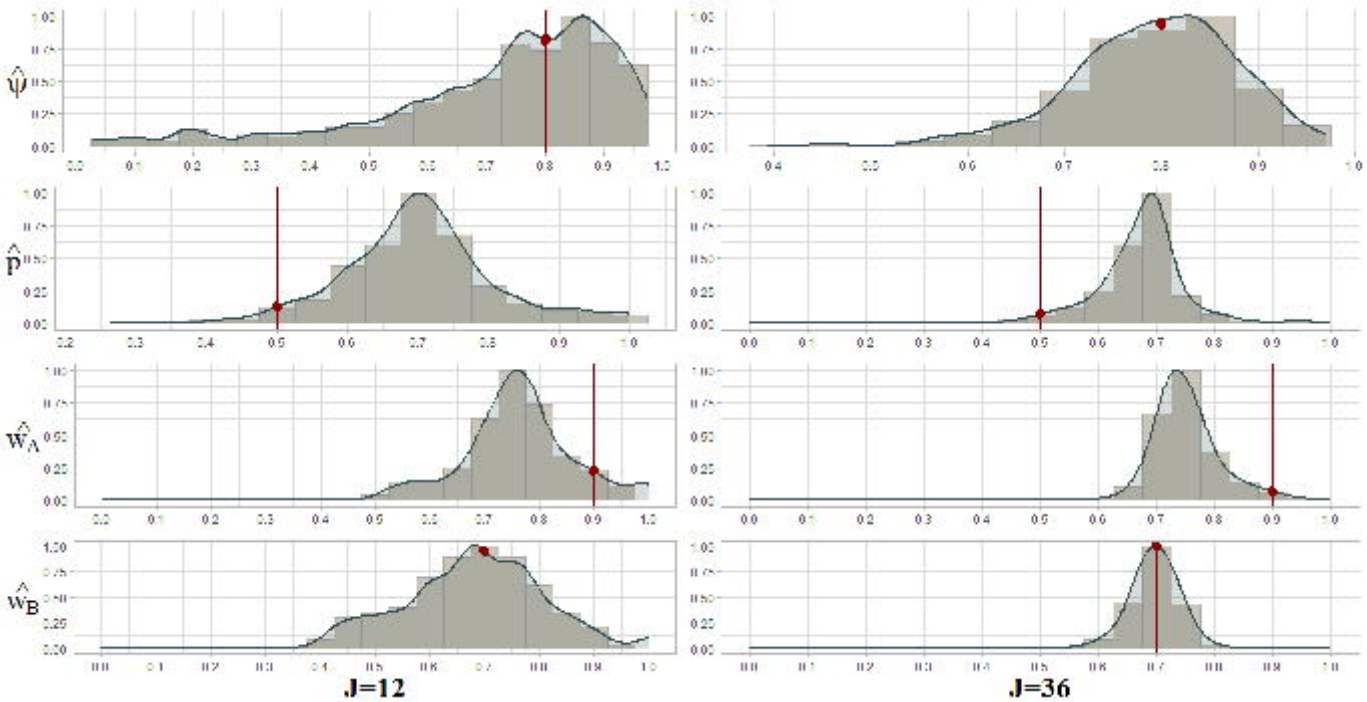
Wrege PH, Rowland ED, Keen S, Shiu Y (2017). *Acoustic monitoring for conservation in tropical forests: examples from forest elephants*. *Methods in Ecology and Evolution* **8**, 1292–1301. https://doi.org/10.1111/2041-210X.12730.

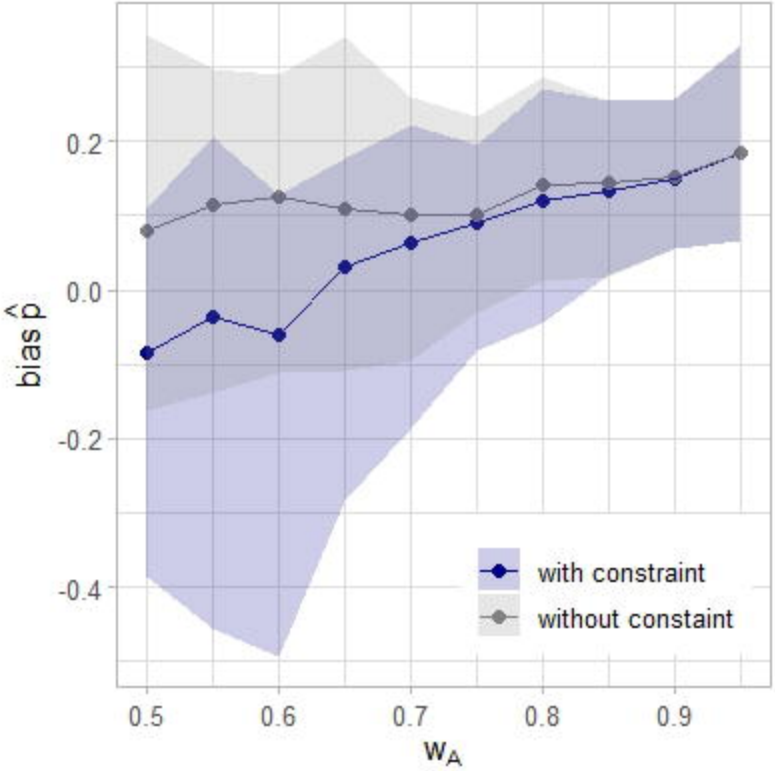
Wu Y, Shih WJ, Moore DF (2008). *Elicitation of a Beta Prior for Bayesian Inference in Clinical Trials*. *Biometrical Journal* **50**, 212–223. https://doi.org/10.1002/bimj.200710390.

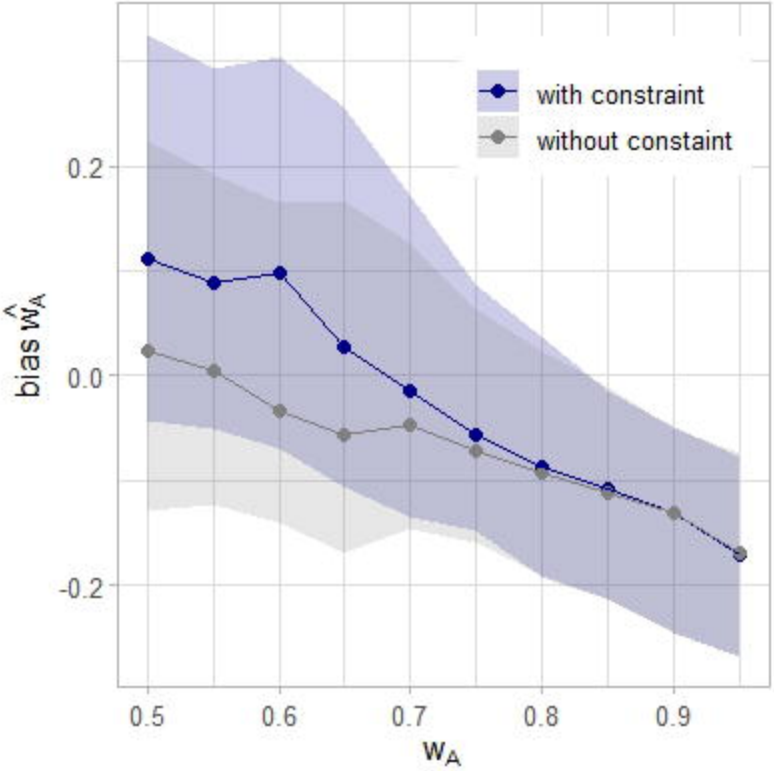
Youngflesh C (2018). *MCMCvis: Tools to visualize, manipulate, and summarize MCMC output*. *Journal of Open Source Software* **3**, 640. https://doi.org/10.21105/joss.00640.

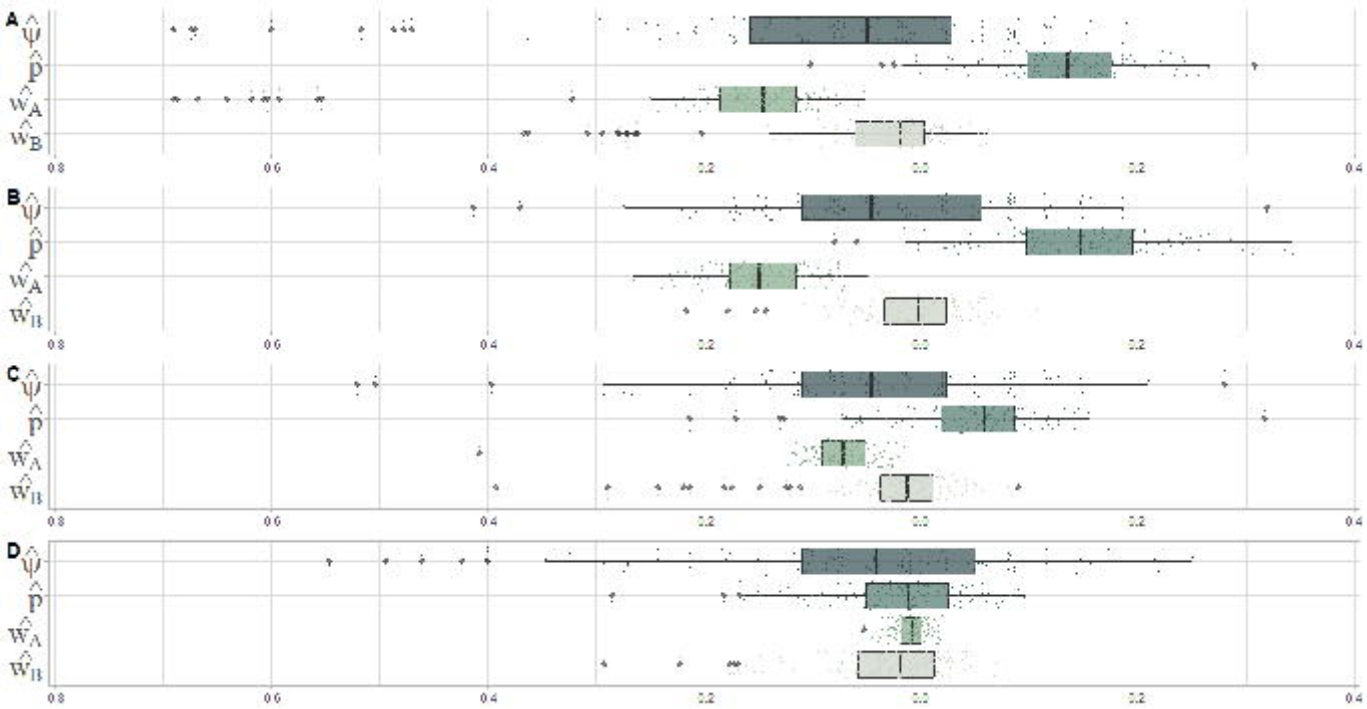
Zooniverse (2024). URL: https://www.zooniverse.org/ (visited on 09/20/2024).

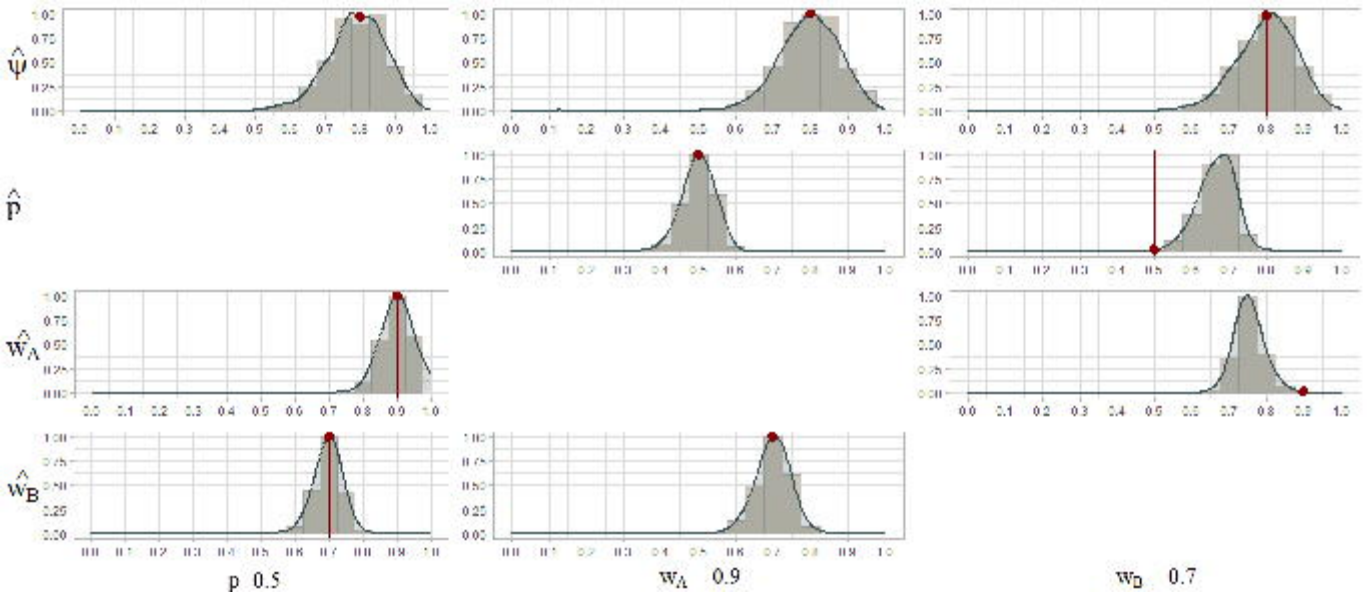
Zwart MC, Baker A, McGowan PJK, Whittingham MJ (2014). *The Use of Automated Bioacoustic Recorders to Replace Human Wildlife Surveys: An Example Using Nightjars*. *PLoS ONE* **9**, e102770. https://doi.org/10.1371/journal.pone.0102770.

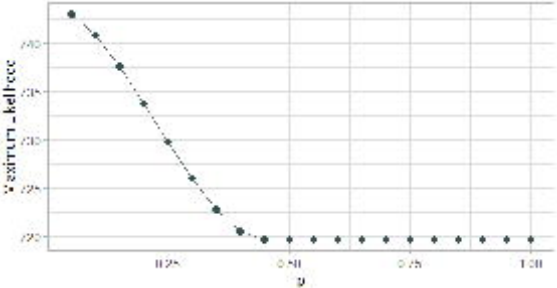

$\hat{\psi}$
$\hat{p}$
$\hat{w}_A$
$\hat{w}_B$
J=12
J=36

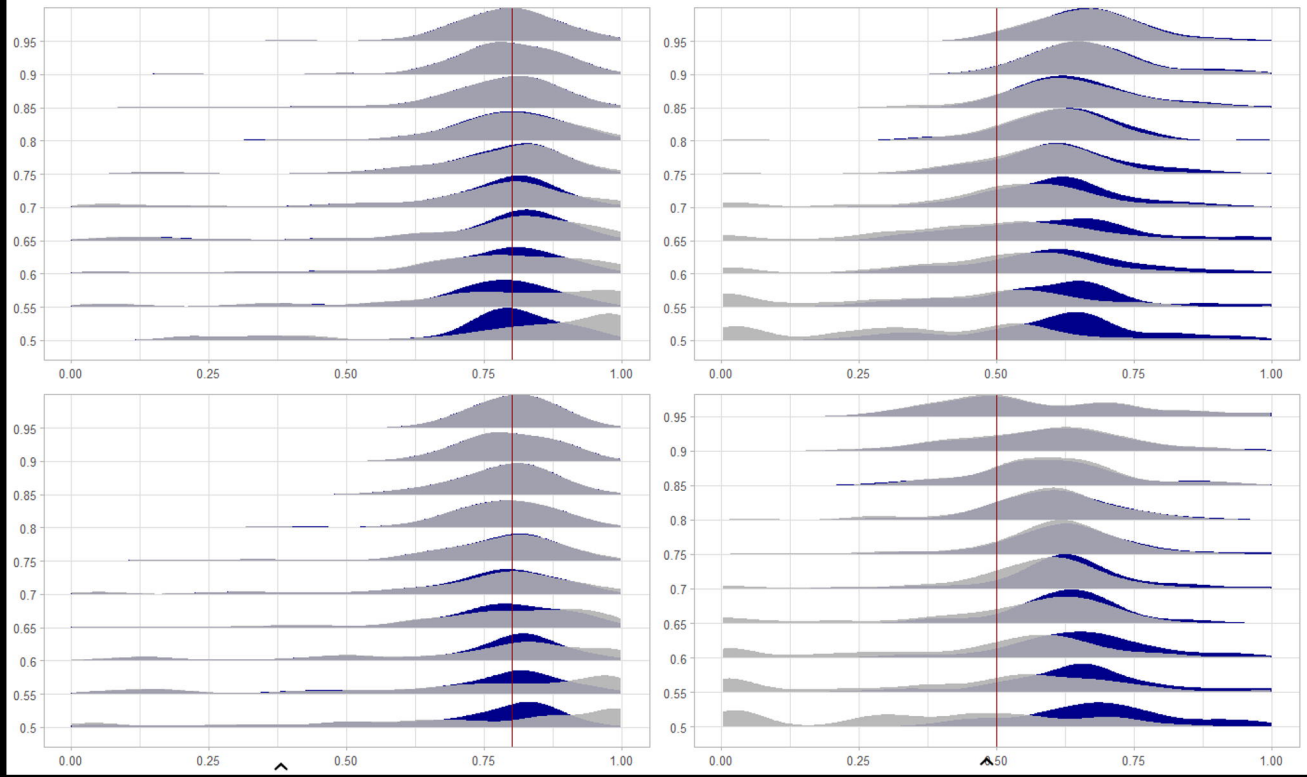

0.2
0.0
-0.2
-0.4
bias $\hat{p}$
with constraint
without constaint
0.5
0.6
0.7
0.8
0.9
w_A

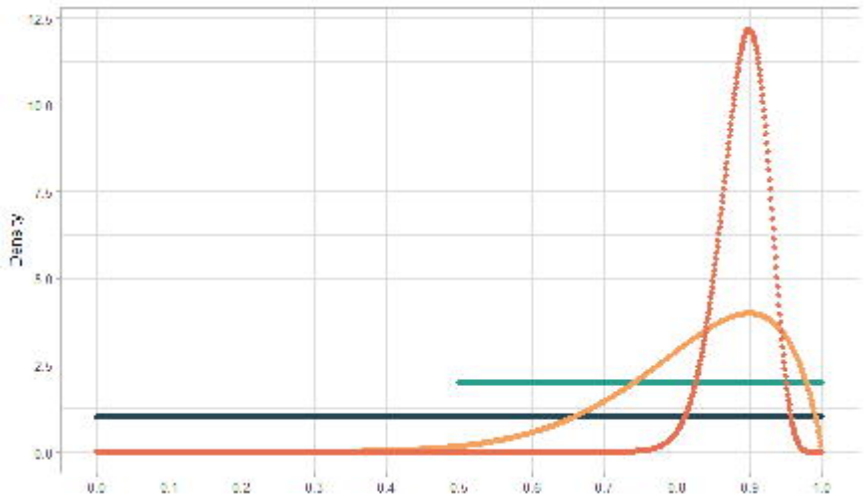

0.2
0.0
-0.2
bias $\hat{w}_A$
with constraint
without constaint
0.5
0.6
0.7
0.8
0.9
w_A

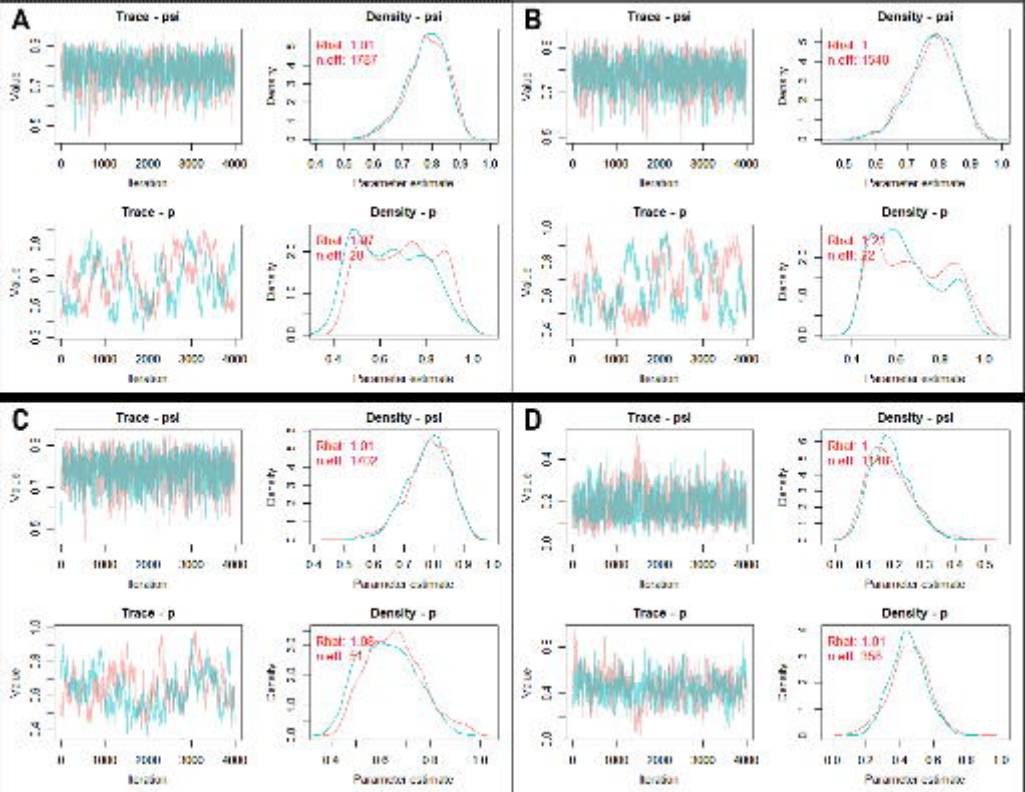

A
B
C
D
$\hat{\psi}$
$\hat{p}$
$\hat{w}_A$
$\hat{w}_B$
0.8
0.6
0.4
0.2
0.0
0.2
0.4

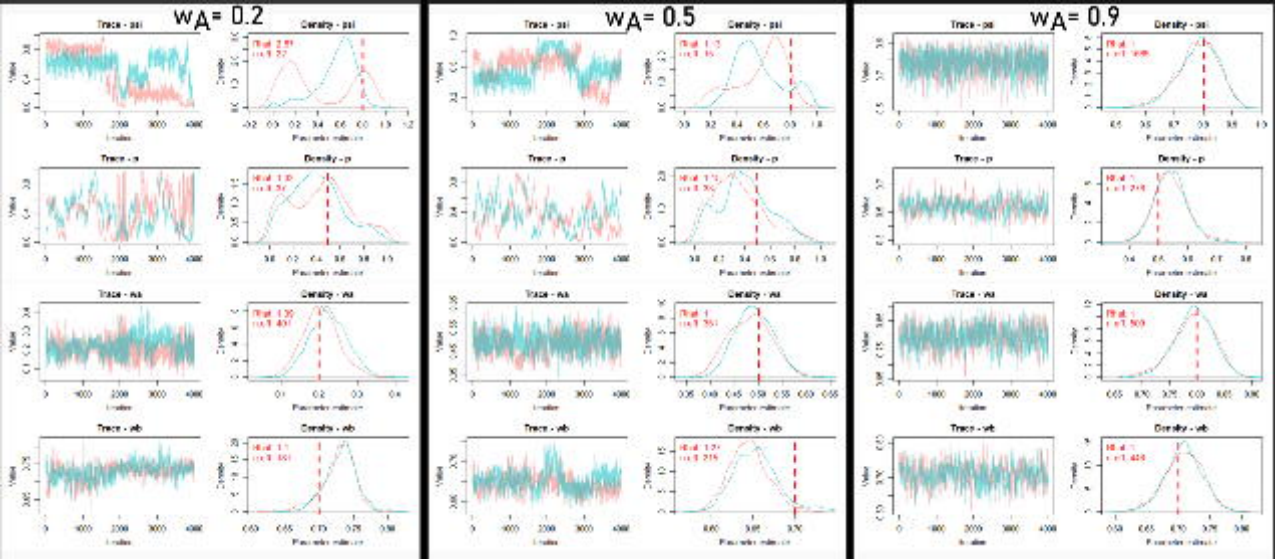

$\hat{\psi}$
$\hat{p}$
$\hat{w}_A$
$\hat{w}_B$
p 0.5
w_A 0.9
w_B 0.7




0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.00
0.25
0.50
0.75
1.00


Density
12.5
10.0
7.5
5.0
2.5
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


A
Trace - psi
Density - psi
Rhat: 1.01
n.eff: 1737
Trace - p
Density - p
Rhat: 1.02
n.eff: 20
B
Trace - psi
Density - psi
Rhat: 1
n.eff: 1540
Trace - p
Density - p
Rhat: 1.21
n.eff: 22
C
Trace - psi
Density - psi
Rhat: 1.01
n.eff: 1402
Trace - p
Density - p
Rhat: 1.06
n.eff: 51
D
Trace - psi
Density - psi
Rhat: 1
n.eff: 1148
Trace - p
Density - p
Rhat: 1.01
n.eff: 358


wA= 0.2
wA= 0.5
wA= 0.9
Trace - psi
Density - psi
Trace - p
Density - p
Trace - wa
Density - wa
Trace - wb
Density - wb