Multi-species occupancy models: an effective and flexible framework for studies of insect communities.

05 Jan 2021

Abstract

1. Entomological studies often aim to estimate species distribution, community composition, or species-richness patterns. False absences can, however, bias these estimates and should consequently not be overlooked in insect studies. Multi-species occupancy models (MSOMs) afford a flexible solution to cover the main topics in ecological entomology while dealing with detectability issues.
2. We sampled Orthoptera communities at 81 mountain grasslands sites in France, using three sampling techniques: sighting, listening, and sweep netting. Five plots were sampled per site. This sampling design allowed MSOMs to be used to estimate richness, occupancy, and detection probabilities while accounting for the effect of covariates. We also used MSOMs to evaluate the efficiency of the survey design and to assess the effects of sampling optimisation.
3. The estimates obtained for altitudinal distribution were reliable, with known species distributions confirming the relevance of MSOMs to model the effects of covariates on Orthoptera communities. The species-specific detection probability was often less than one and varied with the detection technique used and the grass height, confirming the need to deal with detection issues in orthopteran studies.
4. We estimated an inventory completeness superior to 0.80 for 93

Key words. Hierarchical model, imperfect detection, orthoptera communities, sampling optimisation, sampling efficiency, species distribution modelling.

1. Introduction

The study of species distribution and its determinants is of central interest in theoretical and applied ecology (Rushton *et al.*, 2004; Guisan & Thuiller, 2005; Vaughan & Ormerod, 2005). By acquiring information on species occupancy patterns in a set of sampling locations and/or sampling periods (Guillera-Arroita, 2017), ecologists are, for instance, able to make inferences about the environmental drivers behind species distribution (Guisan & Thuiller, 2005; Zipkin *et al.*, 2009). This information can in turn be used to plan appropriate management actions given conservation aims or to understand species range dynamics (Moritz *et al.*, 2008; Pecchi *et al.*, 2019).

Over the last decade, species distribution models (SDMs), relying on the modeling of occurrence data together with environmental covariates, have become the central tool used to study species distribution range or dynamics (Guisan & Thuiller, 2005). Yet most classically used SDMs are based on presence-only data (Guisan & Zimmermann, 2000; Guillera-Arroita *et al.*, 2015). They should be used with caution as their results could be highly biased with sampling effort (Phillips *et al.*, 2009) or with imperfect detection (Guillera-Arroita *et al.*, 2015). Besides these models allow inference about indices of species occurrence and not about probability of species presence, which is the common purpose of ecologists (Guillera-Arroita *et al.*, 2015). SDMs based on presence-absence data are considered as more flexible, but inferences about probability of species presence depend on the assumptions that the species detectability is almost perfect and remains constant between sites (Guillera-Arroita *et al.*, 2015). Violating those assumptions may induce high bias, especially for species characterised by a limited detectability (Lahoz-Monfort *et al.*, 2014). However, decades of studies on detection issues in ecological monitoring have shown that species probabilities of detection are often less than one and may vary with environmental features (Tyre *et al.*, 2003; Lahoz-Monfort *et al.*, 2014).

Detection issues generate false absences that in turn may result in strong bias when modeling species distribution (Tyre *et al.*, 2003; Lahoz-Monfort *et al.*, 2014). For instance, detection issues can lead to the systematic underestimation of the distribution range (MacKenzie *et al.*, 2002) or can overestimate occupancy turnover (MacKenzie *et al.*, 2003). In the early 2000s, so-called 'site-occupancy models' were specifically developed to solve this issue by simultaneously estimating detection and species occupancy probability (MacKenzie *et al.*, 2002; Tyre *et al.*, 2003). Since then, they have been rapidly extended to allow modeling the effects of covariates on occupancy and detection probabilities (MacKenzie *et al.*, 2006), to model range dynamics (Moritz *et al.*, 2008), or to take advantage of information on species states at study locations (Nichols *et al.*, 2007). Initially developed to model occupancy of one focal species, site-occupancy models have been extended to study the occupancy patterns of several species simultaneously (Dorazio & Royle, 2005). These models, called multi-species occupancy models (MSOMs), increase precision in occupancy estimations compared to single-species models, especially for rare species, by borrowing information from data-rich species (Zipkin *et al.*, 2009; Ovaskainen & Soininen, 2011). The hierarchical structure of these models also allows making inferences about the true species richness at study locations, a result that cannot be achieved by species-by-species analysis (Dorazio & Royle, 2005; Guillera-Arroita *et al.*, 2019). MSOMs thus appear to have potential as robust tools for biodiversity analysis or biological assessment (Mata *et al.*, 2014; Devarajan *et al.*, 2020; Tingley *et al.*, 2020). However, despite these models were developed 15 years ago, only 106 published studies have relied on their use, according to a recent review (Devarajan *et al.*, 2020). Among them, MSOMs have mainly been used for vertebrates, and only marginally in other taxa such as plants (see Roth *et al.* (2018)) or insects (Devarajan *et al.* (2020); but see Mata *et al.* (2014); Brodie *et al.* (2019); Dorazio *et al.* (2006)), even though these taxa may pose specific detection issues because of their ecology (phenology, for instance). Accordingly, the relevance of MSOMs appears to be still overlooked for a large range of taxa.

As has been previously highlighted by other authors (Mata *et al.*, 2014; Brodie *et al.*, 2019), since many entomological studies look for changes in insect communities, it is crucial that they take into account imperfect detection. In this context, MSOMs present some clear advantages. First, this approach makes it possible to estimate the occupancy probability even for data-poor taxa,

considering for instance the numerous rare, cryptic or elusive species in insect communities, often difficult to detect in the field (Coddington *et al.*, 2009; Silva *et al.*, 2019). These taxa are usually excluded in common analyses, such as multivariate methods, in which species found at less than 5% of sites are usually removed (Ter Braak & Smilauer, 2002; Pierik *et al.*, 2017). Secondly, both the species- and community- levels can be simultaneously studied with MSOMs (Mata *et al.*, 2014), thanks to the hierarchical structure of their models. The biodiversity estimators provided by MSOMs, such as species richness, have the advantage of accounting explicitly for the effects of survey-, site- and species-level covariates that may affect detectability, in contrast with most usual estimators (Tingley *et al.*, 2020). This is of particular interest in entomological studies, since the activity rate and the density of insects—and thus detection probability—are strongly affected by survey and site conditions (Wolda, 1988; Bale *et al.*, 2002).

Orthoptera is an insect group intensively studied in ecology and regularly used as ecological indicator (Marini *et al.*, 2009; Bazelet & Samways, 2011). Detectability issues are expected to occur in orthopteran field surveys because of their small size and the strong variations in abundance related to their phenology (Badenhausser *et al.*, 2009). Detectability is also expected to vary greatly among species (Badenhausser *et al.*, 2009) due to the high diversity in their ecological traits (e.g. mobility, singing activity, mimicry, etc.) and in their habitat preferences (e.g. closed forests vs open grasslands). Orthoptera detection also strongly depends on the sampling techniques used, the effectiveness of which is often influenced by species-specific traits. For instance, highly mobile Orthoptera may flush out when the observer approaches, becoming easily detectable by sight, but hardly detectable using the sweep net or the box quadrat methods. Conversely, sweep netting and box quadrats may increase the detectability of cryptic Orthoptera living close to the ground, often hardly detectable simply by sight. Despite these detection issues, only two studies conducted on single orthopteran species have explicitly dealt with imperfect detection by using site-occupancy models (MacKenzie *et al.*, 2003; Veran *et al.*, 2015). A third research applied site-occupancy models on orthopteran communities, but using single-species site-occupancy models for each species, without using a standardized sampling design (Malinowska *et al.*, 2014).

Site-occupancy models require a specific survey design, usually based on temporal replication conducted on a set of sampling sites (MacKenzie *et al.*, 2006). Replication at the site scale can be obtained through repeated visits, but also by using spatial replicates, multiple observers or multiple sampling techniques depending on the study (Guillera-Arroita, 2017). A constraint of this method is that this replication increases the sampling effort and the associated costs, precluding the use of this approach by practitioners, notably when the monitoring budget is limited (Field *et al.*, 2005). On the other hand, this replication is needed to explicitly deal with detection issues and thus to develop robust monitoring (Yoccoz *et al.*, 2001). Hence, there is a crucial need to develop monitoring that optimizes the trade-off between sampling effort, techniques and effectiveness when designing site-occupancy surveys (Field *et al.*, 2005).

Most existing grasshopper-sampling protocols are based on estimates of the abundance index or raw presence–absence (Gardiner *et al.*, 2005). To our knowledge, none were designed to use multispecies site-occupancy models. In this study, we therefore investigated the effectiveness of MSOMs in estimating the occupancy probability of Orthoptera at community level, while accounting for imperfect detection. We also evaluated the ability of MSOMs to assess the efficiency of the survey design under different sampling optimization scenario.

2. Materials and methods

2.1 Sampling method and design

Field surveys were conducted between the 9th August and the 5th October 2018 in the Mercantour National Park, located in the southern French Alps. We selected 81 sampling sites among 179 locations already studied between 1983 and 1988 (Gueguen, 1990), in order to encompass all the altitudinal and exposure gradients, ranging from 928 m to 2614 m (mean = 1869 m) above the sea level. Each sampling site consisted in a circle (70 m radius) placed in relatively homogeneous grasslands (Figure 1), and distanced at least 30 m from woodland areas in order to avoid edge effects (Bieringer & Zulka, 2003).

Within each sampling site, we defined five spatial replicates (hereafter 'plots'), placed on two lines perpendicular to the mountain slope and spaced at least 30 m from each other, to avoid potential double counts among plots (Figure 1). In few cases (N = 5 sites), this spatial design had to be slightly adapted depending on the sampling site configuration, notably when the grassland area was too small. Thus, some replicates were placed less than 30 m apart or from the forest edge. However, as a minimum distance of 20 m was maintained between plots and from the forest edge, we considered close plots as independent and no effect of the forest proximity on species composition.

[Figure 1 about here]

Each plot consisted of a 30 m2-area, measured by means of a cord, circular or rectangular in shape, depending on ground cover and slope steepness within the sampling site. In particular, circular plots were preferred in sites characterised by gentle slopes and/or shrubby vegetation, while rectangular plots were surveyed on steep slopes and/or in short-sward sites.

Samplings were carried out by a single trained observer (Y.B.), when the weather conditions were optimal for diurnal Orthoptera activity, *i.e.* no rain, low to moderate wind speed, and sunshine (or temperature exceeding 18°C if cloudy). Species identification was conducted in the field for almost all species, except for *Anonconotus occidentalis* Carron & Wermeille, 2002 and *Anonconotus ligustinus* Galvagni, 2002 which can not be distinguished without the examination of genitalia morphology. Hence, individuals that could be both *A. occidentalis* or *A. ligustinus* were captured and identified in the laboratory, only *A. occidentalis* was detected in our inventories. In each plot, orthopteran assemblages were surveyed following three successive steps: (1) one minute of listening to species stridulating in the plot by standing close to its edge, (2) six minutes of sighting species by walking across the entire plot, and (3) two 45-second sweep netting sessions across the entire plot. We chose this execution order for the different sampling techniques because we expected that it would optimize the number of species encountered. Beginning by the listening step lead to record singing species before disturbing them. Then, we expected that walking accross the plot will made the mobile species flush away making them easily detectable by sight. Finally, sweep netting sessions aimed to capture the remanant less mobile species that did not flush away in step 2. By means of this sampling design, detection/non-detection data were available for each sampling technique in each plot at each site.

2.2 Multi-species occupancy modeling

We then modeled the detection/non-detection data of the 15 replicates per sampling site (*i.e.* 5 plots $\times$ 3 sampling techniques) using site-occupancy models (MacKenzie *et al.*, 2002; Tyre *et al.*, 2003). These models estimate the occupancy probability while modeling imperfect detection through the use of sampling replication conducted at each site. The sampling replication at one site can be achieved with multiple visits, multiple detection methods or multiple observers, as well as with spatial replicates if samplings occur at different locations within a site in a single visit (MacKenzie *et al.*, 2006). In addition, different types of replication are available, depending on the characteristics of the target species, the study area and the objectives (Guillera-Arroita, 2017). Among these options, the spatial replicates method was selected in this study, due to field constraints related to the mountain environment (which required long travel to access the sampling sites). When based on spatial replicates, the model assumes that the occupancy status of sampling sites does not change between site replicates (in our case, plots), *i.e.* closure assumption (Kendall & White, 2009). Because the plots were closed in space and set up in sites of homogeneous vegetation cover, we considered that this assumption was met. We also assumed no false positives, *i.e.* only zero can be reported for a species at a site where it is absent, considering the observer's identification skills.

Site-occupancy models disentangle the ecological process, *i.e.* the true occupancy state for a species in a site, from the observational process, *i.e.* the detection/non-detection of a species at a site given it is present. In order to distinguish a true absence from a non-detection, we modeled the raw data for species i at replicate k of site j, denoted $X_{i,j,k}$, as the outcome of a Bernoulli random variable, defined by:

$$X_{i,j,k} \sim Bernoulli(p_{i,j,k} \times Z_{i,j})$$

where $p_{i,j,k}$ is the detection probability of species i at replicate k of site j, and $Z_{i,j}$ is a binary variable corresponding to the occupancy state of site j by species i (latent state). The model for occurrence is specified as:

$$Z_{i,j} \sim Bernoulli(\psi_{i,j})$$

where $\psi_{i,j}$ is the probability that species i occurs at site j.

We estimated the occupancy ($\psi_{i,j}$) and the detection ($p_{i,j,k}$) probabilities of all the species using a multi-species site-occupancy model (MSOM) with single-species occupancy models as building blocks (Dorazio & Royle, 2005; Zipkin *et al.*, 2009). The species-specific intercepts and slopes from occupancy and detection models were drawn from a shared, community-level distribution via random effects. Through such hierarchical structure, species with less data "borrow" information from other species that are data-rich, which improves precision in estimates (Zipkin *et al.*, 2009; Ovaskainen & Soininen, 2011). The linking of species via random effects also allows inferences to be made about the number of species N_j present at each site j, including species never detected (Dorazio *et al.*, 2006; Guillera-Arroita *et al.*, 2019). In this modelisation process, we followed the 'data augmentation' method described by Royle *et al.* (2007), also including N_0 hypothetical species to the dataset, all with zero detection.

The species-specific effect on the occupancy and the detection probabilities was integrated into the model using the logit link function. Furthermore, covariates supposed to influence the occupancy and the detection probabilities were also considered (altitude and grass height), and the occurrence

probability for species i at site j was modelled by incorporating site-specific characteristics.[su] We developped a relatively simple model including just few covariates in order to show MSOMs potential to study Orthoptera distribution but not explaining it. [/su] In particular, linear and quadratic effects of altitude, varying across species, were included to study the altitudinal distribution of Orthoptera.[su] We could assumed that grass height affect Orthoptera occurrence, but we did not add this occupancy covariate as it did not change significantly our results and interpretations, and complexified the model (see ESM 4 for results comparison).[/su] The occupancy model was defined as:

$$logit(\psi_{i,j}) = \alpha_{0_i} + \alpha_{1_i} \times \text{altitude}_j + \alpha_{2_i} \times \text{altitude}_j{}^2$$

where α_{0_i} is the species-level intercept and ($\alpha_{1_i},\ \alpha_{2_i}$) are the species-specific covariate effects.

The detection probability for species i was assumed to vary depending on the detection technique used (sighting, listening or sweep netting). We coded the sampling technique covariates as dummy variables, with the intercept corresponding to the sighting technique. We also added the linear effect of grass height on the probability of detecting species using the sighting technique. The grass height was standardized to have mean equal to zero:

$$logit(p_{i,j,k}) = \beta_{0_i} + \beta_{1_i} \times \text{listening}_{j,k} + \beta_{2_i} \times \text{netting}_{j,k} + \beta_{3_i} \times \text{height}_{j,k} \times \text{sighting}_{j,k}$$

with β_{0_i}, β_{1_i} and β_{2_i} the species-specific effects of the sampling techniques, and β_{3_i} the species-specific covariate effects.

Each random parameter (α's and β's) was modeled as drawn from a normal distribution described by the community mean (μ) and the variance between species (σ^2):

$$\alpha_{0_i} \sim N(\mu_{\alpha_0}, \sigma^2_{\alpha_0}),\ \alpha_{1_i} \sim N(\mu_{\alpha_1}, \sigma^2_{\alpha_1}),\ ...$$

A latent variable W_i was incorporated in the model to estimate overall species richness. It represents whether species i belongs or not to the community of N species. This binary variable was modeled as the outcome of a Bernoulli random variable, defined by:

$$W_i \sim Bernoulli(\Omega)$$

where Ω describes the probability of belonging to the community. Species detected at least once during the study belong to the community ($W_i = 1$, with i from 1 to N_{obs}), but species never encountered (i from N_{obs} to $N_{obs} + N_0$) could occupy the sampling area and remain undetected ($W_i = 1$ and $\sum X_{i,j,k} = 0$) or not belong to the community ($W_i = 0$). Therefore, the variable W_i is incorporated in the occupancy model (**??**):

$$Z_{i,j} \sim Bernoulli(\psi_{i,j} \times W_i)$$

We implemented the model in a Bayesian framework using the BUGS language and running it in JAGS (Plummer *et al.*, 2003), through the *jagsUI* package (Kellner, 2018) in the R software (R Core Team, 2018). The code is available in Electronic Supplementary Material 1 (ESM 1). Given the lack of prior knowledge of a parameter's true value, parameters and hyper-parameters were implemented

with non-informative priors, following common practice. We used uniform distributions from 0 to 1 for the community level parameter Ω, and for the species-level intercepts of occurrence and detection probabilities (α_0 and β_0). We used wide normal priors (with mean 0 and variance 1000) for the means of hyper-distributions of the site-specific and survey-specific effects (the μ_α's and μ_β's). We used inverse-gamma priors (*Inv-gamma(0.1, 0.1)*) for the community variances (the σ^2's) of all these parameters. We ran the analysis for three chains of 15,000 iterations with a burn-in of 15,000 iterations and a thinning rate of 15. Convergence was assessed by examining the Gelman-Rubin statistic ($\hat{R}$) for each parameter estimate, with $\hat{R} > 1.1$ suggesting a lack of convergence (Gelman & Hill, 2006). Model fit was checked graphically and using the Bayesian p-value (Kéry & Schaub, 2011).

2.3 Sampling effectiveness and optimization

May we reduce the number of spatial replicates?

We investigated the effectiveness of the sampling design in reaching inventory completeness using three, four or five plots. To this end, the inventory completeness ($C_{j,K}$) at site j after K plots ($K = \{3, 4, 5\}$) was calculated as the ratio between the observed number of species ($N_{obs_{j,K}}$, raw data) and the true number of species estimated by the MSOM ($\hat{N}_j$). We used the median values of the posterior distributions of the estimated number of species at each sampling site as point estimates for the true species richness. We reduced the number of plots per site by randomly selecting one to four plots among the five samples from the raw data to assess how reducing the number of plots affected the completeness.

We assessed the effectiveness of the sampling design in detecting species through the site-level probability of detecting a species given it is present. We computed for each iteration of the Markov chain Monte Carlo (MCMC) sample the overall detection probability ($P_{i,K}$) for species i to be detected at least once in K plots using the three detection techniques:

$$P_{i,K} = Pr(\sum_{k=1}^{K} X_{i,j,k} > 0 | Z_{i,j} = 1)$$

$$P_{i,K} = 1 - (1 - p_i(\text{sighting}))^K \times (1 - p_i(\text{listening}))^K \times (1 - p_i(\text{nettting}))^K$$

where $p_i(technique)$ is the plot-level detection probability for species i using the technique cited. We also calculated the site-level detection probability for an "average" species, using estimates of the community-level parameters.

Are the three detection techniques necessary?

The completeness and the overall detection probability (i.e. combining the five plots and the three techniques) were also used to investigate if the sampling techniques were complementary or redundant. Therefore, in order to assess if the sampling protocol could be optimized, the species-specific detection probabilities at site level for each technique were calculated, also verifying the effect of omitting the sweep netting technique. The detection probability without sweep netting

was obtained from the expression of $P_{i,K}$ above, in which we removed the term involving sweep netting. We did not try to investigate the effects of removing the listening technique because this is necessary to identify certain species that are tricky to distinguish visually, even for Orthoptera experts (Walker, 1964), e.g. *Chorthippus sp* Fieber, 1852.

3. Results

As a result of field surveys, we collected 2222 presence data at the plot level (from the three sampling techniques combined) in 405 plots within the 81 sampling sites, belonging to 56 Orthoptera species (ESM 2, Table S1). *Eight (8)* species represented more than 50% of the presence data, while almost half (N = 26) of the observed Orthoptera were detected on less than 5% of the plots. Besides, the true number of species estimated by the model was 75 ($CrI_{95\%}$ = [61; 98]).

Species-by-species results concerning estimations of occupancy and detection probabilities can be consulted on the shinyapps web application (see details in ESM 3).

3.1. Effects of environmental parameters on the occupancy and the detection probabilities

The MSOM method allowed us to investigate the effects of environmental covariates on the occupancy and detection probabilities at both community and species level (Figure 2; see details in ESM 3; ESM 2 Table S2 and Table S3). The overall trends at the community-level are represented by the average species response curves (Figure 2, top panels). They informed us that species tend to have their distribution optimum in the middle altitudinal range (Figure 2, a) and their detectability influenced positively, but not significantly, by the grass height (Figure 2, d). MSOM enabled us to characterize species response along the altitudinal gradient (Figure 2, left panels; see details in ESM 3), giving us predictions of the altitudinal optimum of each species. Although the precision decreased for species with low presence data (wider credible intervals: Figure 2, c), the prediction was still informative in terms of optimum position. In addition, the effect of grass height on the probability of detecting a species by sight was also estimated (Figure 2, right panels; ESM3 for details), highlighting variations among species, with some Orthoptera more detectable in tall grass (Figure 2, e), and others less detectable with increasing grass height (Figure 2, f).

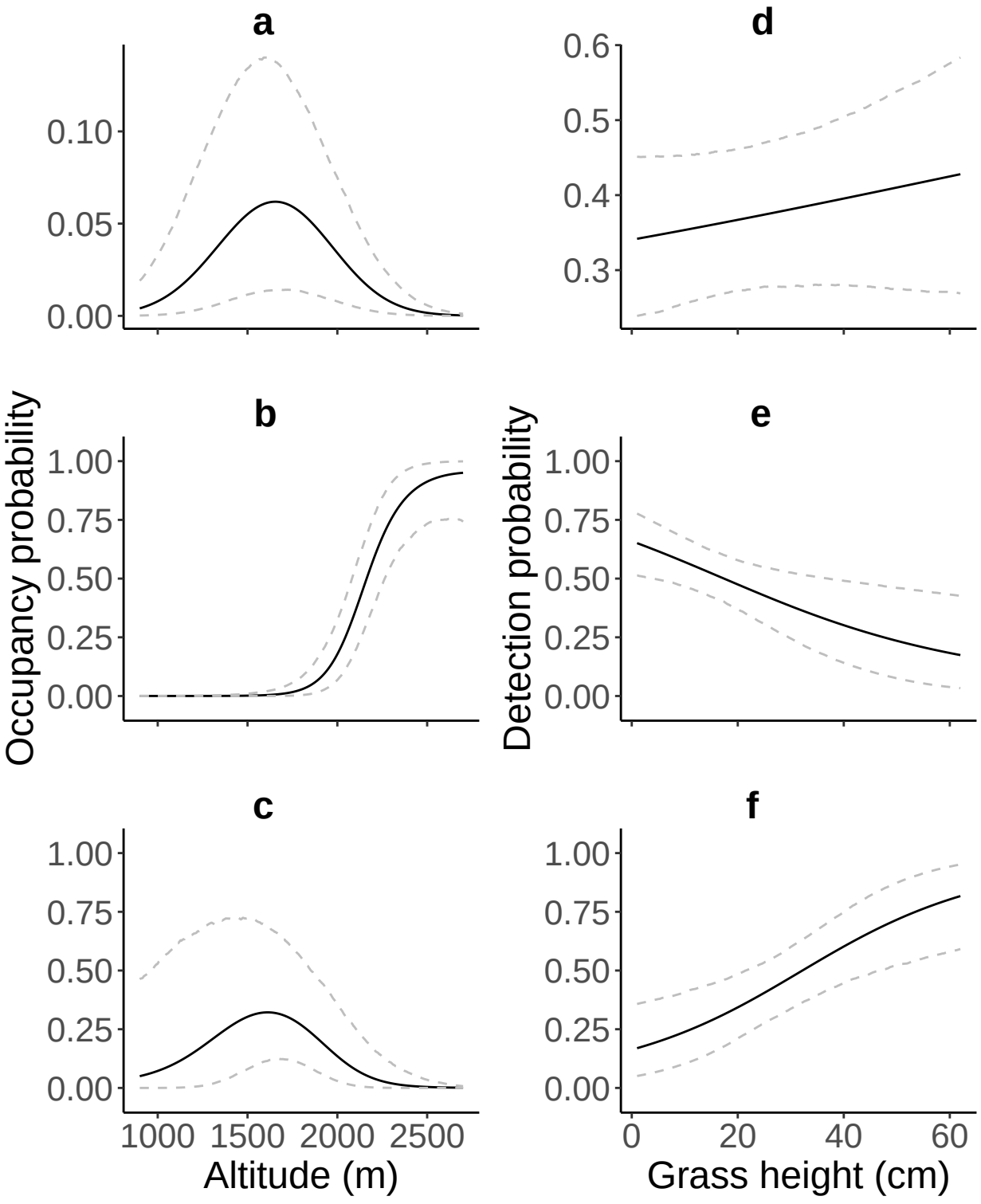

Figure 1: Effect of the altitude on the occupancy probability for (a) an average species at the community-level, (b) *Gomphocerus sibiricus sibiricus*, (c) *Antaxius pedestris*, and effect of the grass height on the probability of sighting (d) an average species at the community-level, (e) *Podisma dechambrei* and (f) *Euthystira brachyptera*. The solid lines represent the posterior mean, and the dashed lines correspond to the 95% credible interval.

3.2. Sampling effectiveness and optimization

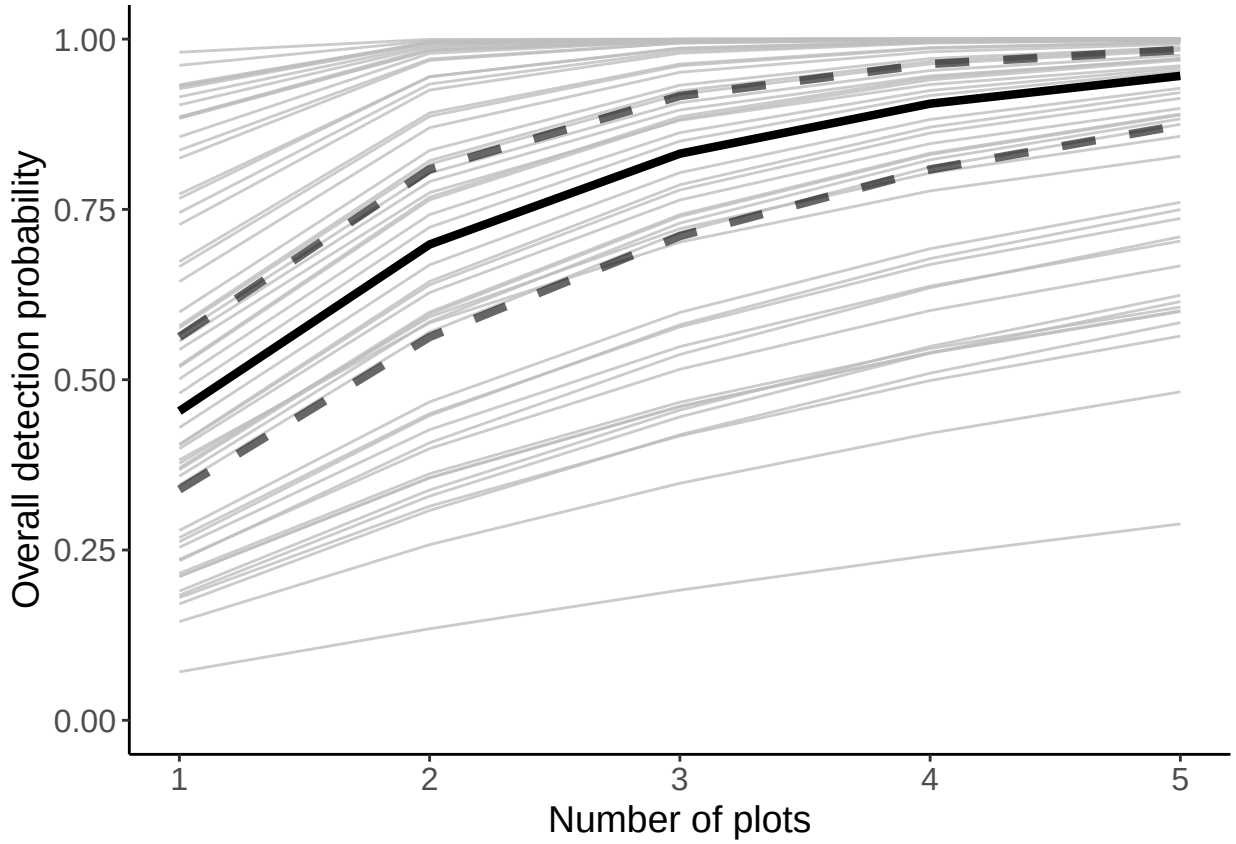

Figure 2: Mean overall detection probabilities at site-level according to the number of plots sampled for each grasshopper species (grey curves) and for an "average" species (black solid line). Black dashed lines correspond to the boundaries of the 95% credibility interval associated to the point estimate of the cumulative probability for an "average" species.

Considering all the detection techniques, the average number of species observed at site level was 9.54 ($CrI_{95\%}$ = [8.77, 10.32]), varying from 2 to 17 among sampling sites. Besides, the estimated site-level species richness obtained from the median of posterior distributions of each site was 10.31 ($CrI_{95\%}$ = [9.46, 11.16]) on average. The completeness (the ratio between the observed number of species and the estimated number of species) increased with the number of plots (Figure 3). With five plots, 76 sites (94%) had completeness superior to 80%, while considering four plots, this decreased to 70 sites (86%), and it further decreased to 41 sites (51%) when a survey in only three plots was simulated.

The overall detection probability of an 'average' species with all the detection techniques rose sharply when increasing from one to three plots, from 0.45 ($CrI_{95\%}$ = [c(2.5% = 0.34) ; c(97.5% = 0.56)]) to 0.83 ($CrI_{95\%}$ = [c(2.5% = 0.71) ; c(97.5% = 0.92)]), then grew slightly from 0.91 ($CrI_{95\%}$ = [c(2.5% = 0.81) ; c(97.5% = 0.96)]) to 0.95 ($CrI_{95\%}$ = [c(2.5% = 0.87) ; c(97.5% = 0.98)]) when adding a fourth and fifth plot respectively. The detection probabilities estimated at the site level for each detection technique highlighted the importance of sighting in overall detection

and the marginality of the two other techniques (see ESM 3 for details). The detection probability of an 'average' species at the site level (five plots) was 0.89 ($CrI_{95\%}$ = [0.79 ; 0.95]) with the sighting technique only. The average detection probabilities at the site level when considering only the listening or the sweep netting technique were much lower: 0.22 ($CrI_{95\%}$ = [0.1 ; 0.38]) and 0.4 ($CrI_{95\%}$ = [0.26 ; 0.54]) respectively. Hence, the detection probability of an 'average' species without the sweep netting step (0.91, $CrI_{95\%}$ = [0.82 ; 0.97]) was only slightly lower than the detection probability combining the three techniques.

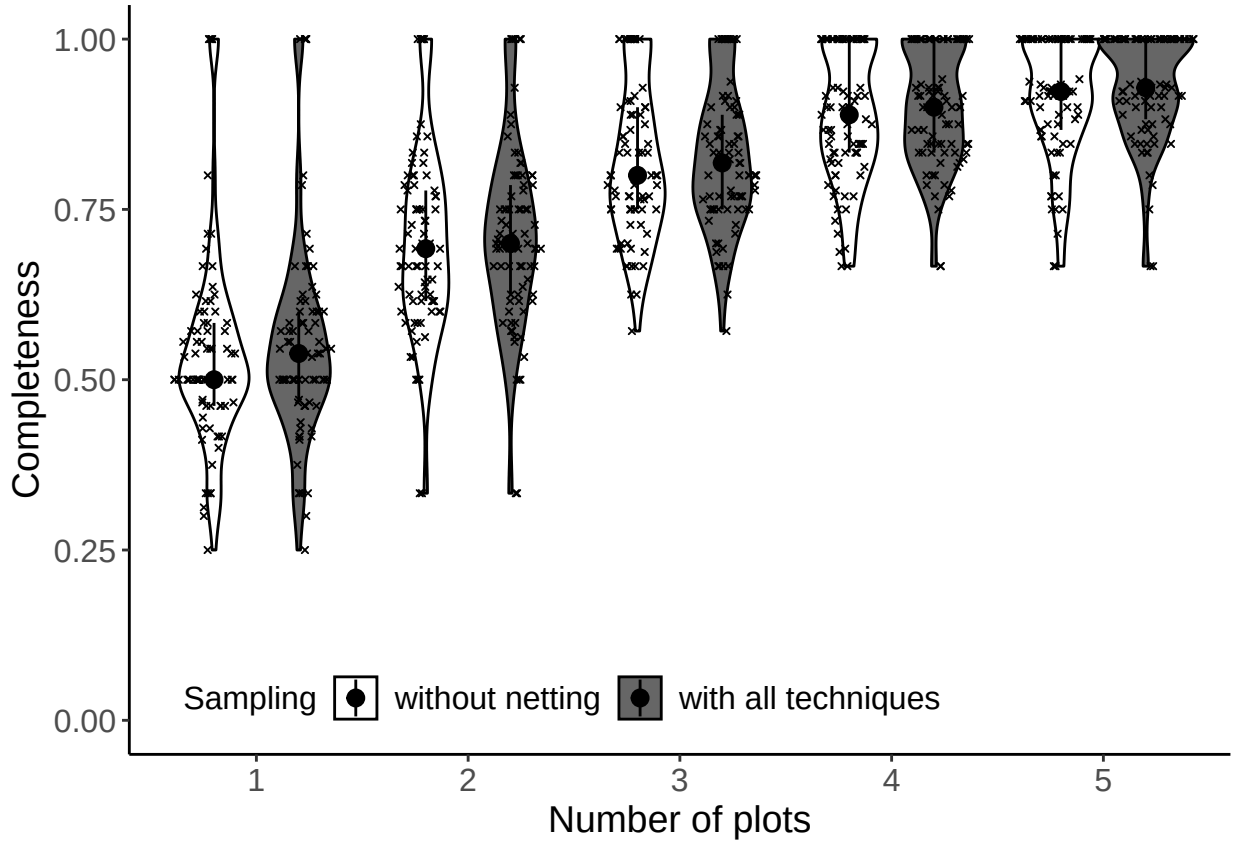

Figure 3: Completeness of inventories at site-level according to the number of plots sampled.

Detection probability varied consistently among species, ranging from 0.07 ($CrI_{95\%}$ = [0.01; 0.22]) to 0.98 ($CrI_{95\%}$ = [0.96; 1]) at the plot level and from 0.29 ($CrI_{95\%}$ = [0.07; 0.71]) to 1 at the site level when all the techniques were used (Figure 4; see shinyapp for details). Some species (38%) had a high probability of being detected after having surveyed a first plot ($P_{i,1}$ > 0.60). For these species, the overall detection probability followed an asymptotic curve approaching 1 after three or four plots. In contrast, there were species (27%) with low detection probability in one sampling plot ($P_{i,1}$ < 0.30) for which each additional plot sampled sharply increased the overall detection probability at the site level. For almost all species, the differences in detection probability with or without the netting step were not significant, except for *Oecanthus pellucens* (Scopoli, 1763) and *Leptophyes punctatissima* (Bosc, 1792). In these two species, the detection probability decreased from 0.95 ($CrI_{95\%}$ = [0.87;0.99]) to 0.66 ($CrI_{95\%}$ = [0.41;0.88]) for *O. pellucens*, and from 0.61

($CrI_{95\%}$ = [0.34;0.87]) to 0.45 ($CrI_{95\%}$ = [0.2;0.73]) for *L. punctatissima* when removing the sweep netting technique.

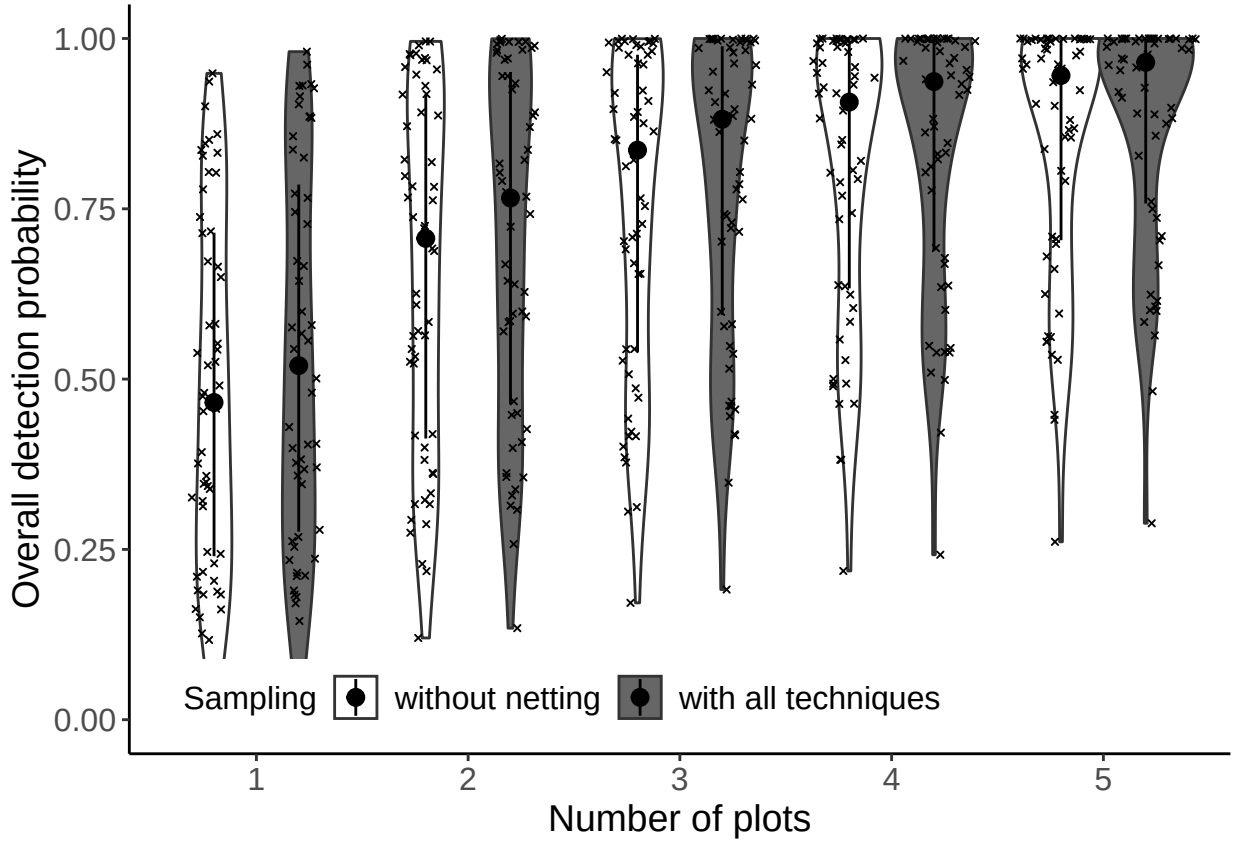

Figure 4: Distribution of the species-specific global detection probability at the site-level depending on the detection techniques used. White points represent the median and white segments the first and third quartiles.

4. Discussion

Using the data from our survey, we were able to estimate the occupancy probabilities of 56 Orthoptera species along an elevation gradient in the Mercantour National Park, while accounting for imperfect detection through a multi-species occupancy model. The species-specific detection probabilities varied widely between species, from 0.29 to 1 at the site level. This could be affected, positively or negatively, or unaffected by the grass height, depending on the species. The inventory completeness was more than 0.80 for 94% of the sites, and the overall detection probability at the community level was 0.95 ($CrI_{95\%}$ = [c(`2.5%` = 0.87) ; c(`97.5%` = 0.98)]) when using all of a site's five plots and the three sampling techniques. These values slightly decreased when we hypothetically reduced the sampling effort by omitting the netting step or by removing one plot, suggesting that the sampling effort could be reduced with minimal impact on estimate quality.

4.1. Reliability of MSOM estimates

Reliability of MSOMs estimates first relies on the respect of the two major assumptions implied by the model: site-closure and no false-presence. The closure assumption indicates that if a plot is occupied, all plots within the site are also occupied. Violating this assumption involves underestimation of detection probabilities and then a overestimation of occupancy probabilities (Kendall & White, 2009). In our study, as we sampled homogeneous grasslands, we were confident about the respect of this assumption. False-presences due to misidentification also induce overestimation of occupancy probabilities if not addressed in the model (Royle & Link, 2006). This assumption is likely to be violated in unexeperienced observer. Yet in our case, the observer is highly skilled. Hence, we were confident about the respect of the assumptions, but readers should remember those assumptions when planning to use MSOM.

As expected, the distribution of Orthoptera in our study was structured according to the elevation with (i) maximum in occupancy probabilities of thermophile species such as *Pezotettix giornae* (Rossi, 1794) estimated at low elevations, (ii) wide estimated distribution for generalist species such as *Stauroderus scalaris* (Fischer von Waldheim, 1846), and (iii) arctic-alpine species such as *Gomphocerus sibiricus sibiricus* (Linnaeus, 1767) having their estimated elevation optimum at high elevations. These results are consistent with what is known about Orthoptera species distribution in the Mercantour National Park area (Gueguen, 1990; Lemonnier, 1999); Braud, com. pers.). This reliability in the estimated distribution range confirms the relevance of MSOMs to model the effects of biotic or abiotic factors on Orthoptera communities.

MSOM estimates also proved to be useful to assess the inventory completeness reached at each site, referring in particular to the true number of species present, which is rarely known despite its importance in biological studies. However, in some cases, MSOMs may produce unreliable species richness estimates (Guillera-Arroita *et al.*, 2019), especially when detection and/or occupancy are low, inducing a lack of obervation and a large number of missing species. Aside from these cases, MSOMs seem to produce reliable estimates in most scenarios and often outperform commonly used estimators such as *iChao2* or *Jacknife* (Tingley *et al.*, 2020). In this study, the total species richness estimated by MSOM (75 species, $CrI_{95\%}$ = [61; 98]) is consistent with the 95 Orthoptera species known to be present above an altitude of 900 m in the Mercantour National Park in grasslands and ecotone habitats (Braud com. pers.). We chose to integrate ecotone species in the list of known species as we encoutered some of those during sampling, such as *Nemobius sylvestris* (Bosc, 1792) or *Pholidoptera griseoaptera* (De Geer, 1773). We are thus confident about the reliability of the MSOM estimates at the site level and of the derived inventory completeness.

4.2. Importance of detectability in orthopteran studies

Even with the inclusion of the full sampling process (all five plots per sampling site and all three sampling techniques), the overall detection probabilities we estimated were less than one for most species. In some cases it was quite low: for example 0.29 for *Eupholidoptera chabrieri* (Charpentier, 1825) or 0.48 for *Calliptamus italicus* (Linnaeus, 1758), confirming the importance of using methods that explicitly correct for imperfect detection. Our results also indicate that detection probability is affected by certain environmental covariates. For instance, we found that Orthoptera detectability by sight may vary with grass height, and that this relationship differs between species. Such a

correlation was expected as less mobile species or those living close to the ground, such as *Podisma dechambrei* Chopard, 1952, are likely to be less detectable with increasing grass height. In contrast, the abundance of some Orthoptera, such as *Euthystira brachyptera* (Ocskay, 1826), increases with grass height (Gardiner, 2018), which in turn may increase their detectability (McCarthy *et al.*, 2013). A positive relationship between grass height and species detectability could also be explained by an effect of the higher abundance expected at lower elevations, where grass is slightly higher. Such a correlation between detectability and habitat covariate is likely to generate strong bias when studying the distribution of a species and its relationship with habitat if detection is not modeled explicitly (Lahoz-Monfort *et al.*, 2014). When detection is affected by a habitat covariate, a model that does not include the effect of this covariate on the detection probability may incorrectly identify this habitat covariate as affecting the occupancy rate of the species (Lahoz-Monfort *et al.*, 2014). Such a bias could have huge repercussions in comparative approaches (Archaux *et al.*, 2012), which are commonly used for Orthoptera (e.g. Bomar, 2001; Marini *et al.*, 2009; Löffler *et al.*, 2019).

Some authors have questioned the benefits of modeling imperfect detection (Welsh *et al.*, 2013). They argue that in some cases, i.e. when occupancy is low and detectability is high, 'simple models' perform similarly or better than site-occupancy models. However, this is true only in limited scenarios and assumes high a priori knowledge on the detectability and occupancy of the studied species (Guillera-Arroita *et al.*, 2014). Little is known about the detectability of insects, as there are very few studies accounting for imperfect detection (Kellner & Swihart, 2014; Devarajan *et al.*, 2020). The results obtained studying the orthopteran community in the Mercantour National Park show that occupancy probability is not systematically low, detection probability is not systematically high, and detection probability is highly affected by habitat covariates. These results advocate for the systematic use of MSOMs when studying orthopteran distribution.

4.3. Sampling effectiveness and optimization

The proportion of species richness detected by a survey is a metric commonly used as an indicator of inventory completeness (Moreno & Halffter, 2000; Foggo *et al.*, 2003). Foggo *et al.* (2003) used a completeness threshold of 0.8 to indicate that an inventory is representative of the community composition in a given site. In our study, 94% of the sites exceeded this threshold with five plots sampled and with the three sampling techniques used. Hence, the composition of the Orthoptera community seems to be well described at the site scale with our sampling design. Our results also suggest that sampling effort may be reduced, notably by omitting one plot or by removing the sweep netting step, while still maintaining a completeness higher than 0.8 for 86% (omitting one plot) and 86% (removing sweep netting) of the sites.

Overall detection probability at the species scale may also be seen as an indicator of sampling efficiency (Moore *et al.*, 2014; Smart *et al.*, 2016). According to the usual detection probability threshold of 0.95 (see e.g. Moore *et al.* (2014); Smart *et al.* (2016)), the sampling design we used was effective for 31 of the 56 species observed (55%). The number of species above this threshold would decline by 3 species by omitting the sweep netting step or by 6 species by removing one plot. The overall detection probability of an 'average' species would also decrease, but slightly and not significantly, from 0.95 ($CrI_{95\%}$ = [c(`2.5%` = 0.87) ; c(`97.5%` = 0.98)]) with complete sampling, to 0.91 ($CrI_{95\%}$ = [0.82 ; 0.97]) without sweep netting, or 0.91 ($CrI_{95\%}$ = [c(`2.5%` = 0.81) ; c(`97.5%` = 0.96)]) with four plots. These results confirm that reducing the field effort is

possible with a weak impact on detectability.

The best way to optimize the sampling effort, either by removing a detection technique or by reducing the number of plots, may depend on the local species composition, the study objectives and the specific characteristics in the field. In our case, whether we chose to remove a plot or the sweep netting step, the loss in detection probability and in inventory completeness was almost the same. Moreover, in each sampling site the time required to perform five sweep netting steps is quite similar to that needed to fulfil a complete survey in a single plot, around 10 minutes each. Hence, in our case there is not really one choice that is better, especially since the costs are associated mainly with the travel time between sampling sites. However, this could be different in other studies. For example, if we had chosen temporal rather than spatial replicates, removing a plot would have been much more worthwhile than omitting the sweep netting. In the same way, while in our study missing some species was not problematic, as our aim was not to obtain an exhaustive inventory of the entire orthopteran community, other study aims may be different. It should be noted that we found that certain species such as *Oecanthus pellucens* may be missed without the sweep netting technique. It is also important to consider that efficiency of detection techniques depends on their execution order. For instance, walking accross the plot during the sighting step made individuals flushing away and so reduced the efficacity of sweep netting sessions. Hence, execution order for the different techniques should be chose accordingly with the behavior of the species of interest. Thus, we advocate for implementing a pilot study to help identify how sampling can be best optimized depending on the study objectives.

4.4. Conclusion

In this paper, we developed a multi-species occupancy model to demonstrate the need to account for imperfect detection in insect studies and highlight the potential use of MSOM to investigate the sampling efficiency and optimization. As our aim was not to explain ecologically Orthoptera distribution or detactability, we developed a relatively simple model including just few covariates in order to show MSOMs potential.

However, our model could be easily adapted, with a minimal knowledge in Bayesian computation, to other environmental gradients or pressures commonly investigated in entomological studies, such as management practices (Marini *et al.*, 2009), urbanization levels (Penone *et al.*, 2013), land use intensity (Weking *et al.*, 2016) or climate change (Löffler *et al.*, 2019). Similarly, the effects of other potential covariates on detection probability, such as meteorological (temperature, wind, irradiance, etc.) or phenological variables (date), can easily be implemented in MSOMs. Species traits known to influence detectability, such as mobility capacity, could also be incorporated in MSOMs, for example by grouping species *a priori* (Pacifici *et al.*, 2014). However, adding species traits in the model can be tricky when using 'data augmentation' approach to estimate species richness as the traits are unknown for species never detected. The unobserved species traits have to be integrated through the hierarchical structure of MSOMs as latent variables, which could however be complicated for non Bayesian experts. MSOM can also be extended in a dynamic approach to estimate the probability of the colonization or extinction of sites (Dorazio *et al.*, 2010) and thus to study temporal variations in occupancy probability at the community scale. Potential extensions of MSOMs are numerous. Existing model selection methods adapted to Bayesian hierarchical models (Hooten & Hobbs, 2015), thus MSOMs (Broms *et al.*, 2016), may help ecologists to chose the most

appropriate model. Bayesian computation can become time-consuming when dealing with large datasets or numerous covariates, and can be a constraint to perform model or variable selection. Yet, alternatives, such as indicator variable selection (Kuo & Mallick, 1998), exist to facilitate variable selection in Bayesian regression models (see: O'Hara & Sillanpää (2009), for a review), thus in MSOMs (Dorazio *et al.*, 2011). Model selection in MSOMs may be done using information criterion (Drouilly *et al.*, 2018), such as deviance information criterion (DIC; Spiegelhalter *et al.* (2002)) or Watanabe-Akaike information criterion (WAIC; Watanabe (2013)), or describing predictive performance, with cross-validation for example (Zipkin *et al.*, 2012). Current MSOMs are now highly flexible and offer an effective framework to model the state or dynamics of insect communities. They can also be used to optimize the efficiency of the sampling design, which could be of interest to many entomologists. We advocate for their systematic use in entomological studies.

References

Archaux, F., Henry, P.-Y. & Gimenez, O. (2012) When can we ignore the problem of imperfect detection in comparative studies? *Methods in Ecology and Evolution*, **3**, 188–194.

Badenhausser, I., Amouroux, P., Lerin, J. & Bretagnolle, V. (2009) Acridid (orthoptera: Acrididae) abundance in western european grasslands: Sampling methodology and temporal fluctuations. *Journal of Applied Entomology*, **133**, 720–732.

Bale, J.S., Masters, G.J., Hodkinson, I.D., Awmack, C., Bezemer, T.M., Brown, V.K., *et al.* (2002) Herbivory in global climate change research: Direct effects of rising temperature on insect herbivores. *Global Change Biology*, **8**, 1–16.

Bazelet, C.S. & Samways, M.J. (2011) Identifying grasshopper bioindicators for habitat quality assessment of ecological networks. *Ecological Indicators*, **11**, 1259–1269.

Bieringer, G. & Zulka, K.P. (2003) Shading out species richness: Edge effect of a pine plantation on the orthoptera (tettigoniidae and acrididae) assemblage of an adjacent dry grassland. *Biodiversity & Conservation*, **12**, 1481–1495.

Bomar, C.R. (2001) Comparison of grasshopper (orthoptera: Acrididae) communities on remnant and reconstructed prairies in western wisconsin. *Journal of Orthoptera Research*, **10**, 105–113.

Brodie, B.S., Popescu, V.D., Iosif, R., Ciocanea, C., Manolache, S., Vanau, G., *et al.* (2019) Non-lethal monitoring of longicorn beetle communities using generic pheromone lures and occupancy models. *Ecological Indicators*, **101**, 330–340.

Broms, K.M., Hooten, M.B. & Fitzpatrick, R.M. (2016) Model selection and assessment for multi-species occupancy models. *Ecology*, **97**, 1759–1770.

Coddington, J.A., Agnarsson, I., Miller, J.A., Kuntner, M. & Hormiga, G. (2009) Undersampling bias: The null hypothesis for singleton species in tropical arthropod surveys. *Journal of animal ecology*, **78**, 573–584.

Devarajan, K., Morelli, T.L. & Tenan, S. (2020) Multi-species occupancy models: Review, roadmap, and recommendations. *Ecography*, **n/a**.

Dorazio, R.M., Gotelli, N.J. & Ellison, A.M. (2011) In *Biodiversity loss in a changing planet*. InTech Rijeka, Croatia, pp. 277–302.

Dorazio, R.M., Kéry, M., Royle, J.A. & Plattner, M. (2010) Models for inference in dynamic metacommunity systems. *Ecology*, **91**, 2466–2475.

Dorazio, R.M. & Royle, J.A. (2005) Estimating size and composition of biological communities by

modeling the occurrence of species. *Journal of the American Statistical Association*, **100**, 389–398.
Dorazio, R.M., Royle, J.A., Söderström, B. & Glimskär, A. (2006) Estimating species richness and accumulation by modelling species occurrence and detectability. *Ecology*, **87**, 842–854.
Drouilly, M., Clark, A. & O'Riain, M.J. (2018) Multi-species occupancy modelling of mammal and ground bird communities in rangeland in the karoo: A case for dryland systems globally. *Biological Conservation*, **224**, 16–25.
Field, S.A., Tyre, A.J. & Possingham, H.P. (2005) Optimizing allocation of monitoring effort under economic and observational constraints. *The Journal of Wildlife Management*, **69**, 473–482.
Foggo, A., Rundle, S.D. & Bilton, D.T. (2003) The net result: Evaluating species richness extrapolation techniques for littoral pond invertebrates. *Freshwater Biology*, **48**, 1756–1764.
Gardiner, T. (2018) Grazing and orthoptera: A review. *Journal of Orthoptera Research*, **27**, 3–11.
Gardiner, T., Hill, J. & Chesmore, D. (2005) Review of the methods frequently used to estimate the abundance of orthoptera in grassland ecosystems. *Journal of Insect Conservation*, **9**, 151–173.
Gueguen, A. (1990) Impact du pâturage ovin sur la faune sauvage: Exemple des orthoptères.
Guillera-Arroita, G. (2017) Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. *Ecography*, **40**, 281–295.
Guillera-Arroita, G., Kéry, M. & Lahoz-Monfort, J.J. (2019) Inferring species richness using multi-species occupancy modeling: Estimation performance and interpretation. *Ecology and Evolution*, **9**, 780–792.
Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., *et al.* (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, **24**, 276–292.
Guillera-Arroita, G., Lahoz-Monfort, J.J., MacKenzie, D.I., Wintle, B.A. & McCarthy, M.A. (2014) Ignoring imperfect detection in biological surveys is dangerous: A response to 'fitting and interpreting occupancy models'. *PloS one*, **9**.
Guisan, A. & Thuiller, W. (2005) Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.
Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.
Hooten, M. & Hobbs, N. (2015) A guide to bayesian model selection for ecologists. *Ecological Monographs*, **85**, 3–28.
Kellner, K. (2018) JagsUI: A wrapper around "rjags" to streamline "JAGS" analyses. R package version 1.5.0.
Kellner, K.F. & Swihart, R.K. (2014) Accounting for imperfect detection in ecology: A quantitative review. *PloS one*, **9**.
Kendall, W.L. & White, G.C. (2009) A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. *Journal of Applied Ecology*, **46**, 1182–1188.
Kuo, L. & Mallick, B. (1998) Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, **60**, 65–81.
Lahoz-Monfort, J.J., Guillera-Arroita, G. & Wintle, B.A. (2014) Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, **23**, 504–515.
Lemonnier, M. (1999) Les peuplements d'orthoptères (insecta: Orthoptera) du parc national du mercantour (alpes maritimes, alpes-de-haute-provence). *Bulletin de la Société entomologique de France*, **104**, 149–166.
Löffler, F., Poniatowski, D. & Fartmann, T. (2019) Orthoptera community shifts in response to

land-use and climate change – lessons from a long-term study across different grassland habitats. *Biological Conservation*, **236**, 315–323.
MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, **84**, 2200–2207.
MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.
MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L. & Hines, J.E. (2006) *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence*. Elsevier.
Malinowska, A.H., Strien, A.J. van, Verboom, J., WallisdeVries, M.F. & Opdam, P. (2014) No evidence of the effect of extreme weather events on annual occurrence of four groups of ectothermic species. *PLOS ONE*, **9**, 1–10.
Marini, L., Fontana, P., Battisti, A. & Gaston, K.J. (2009) Response of orthopteran diversity to abandonment of semi-natural meadows. *Agriculture, Ecosystems & Environment*, **132**, 232–236.
Mata, L., Goula, M. & Hahs, A.K. (2014) Conserving insect assemblages in urban landscapes: Accounting for species-specific responses and imperfect detection. *Journal of Insect Conservation*, **18**, 885–894.
McCarthy, M.A., Moore, J.L., Morris, W.K., Parris, K.M., Garrard, G.E., Vesk, P.A., *et al.* (2013) The influence of abundance on detectability. *Oikos*, **122**, 717–726.
Moore, A.L., McCarthy, M.A., Parris, K.M. & Moore, J.L. (2014) The optimal number of surveys when detectability varies. *PLoS One*, **9**, e115345.
Moreno, C.E. & Halffter, G. (2000) Assessing the completeness of bat biodiversity inventories using species accumulation curves. *Journal of Applied Ecology*, **37**, 149–158.
Moritz, C., Patton, J.L., Conroy, C.J., Parra, J.L., White, G.C. & Beissinger, S.R. (2008) Impact of a century of climate change on small-mammal communities in yosemite national park, USA. *Science*, **322**, 261–264.
Nichols, J.D., Hines, J.E., Mackenzie, D.I., Seamans, M.E. & Gutiérrez, R.J. (2007) OCCUPANCY ESTIMATION AND MODELING WITH MULTIPLE STATES AND STATE UNCERTAINTY. *Ecology*, **88**, 1395–1400.
O'Hara, R.B. & Sillanpää, M.J. (2009) A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4**, 85–117.
Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: Hierarchical modeling of species communities. *Ecology*, **92**, 289–295.
Pacifici, K., Zipkin, E.F., Collazo, J.A., Irizarry, J.I. & DeWan, A. (2014) Guidelines for a priori grouping of species in hierarchical community models. *Ecology and Evolution*, **4**, 877–888.
Pecchi, M., Marchi, M., Burton, V., Giannetti, F., Moriondo, M., Bernetti, I., *et al.* (2019) Species distribution modelling to support forest management. A literature review. *Ecological Modelling*, **411**, 108817.
Penone, C., Le Viol, I., Pellissier, V., Julien, J.-F., Bas, Y. & Kerbiriou, C. (2013) Use of large-scale acoustic monitoring to assess anthropogenic pressures on orthoptera communities. *Conservation Biology*, **27**, 979–987.
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., *et al.* (2009) Sample selection bias and presence-only distribution models: Implications for background and

pseudo-absence data. *Ecological Applications*, **19**, 181–197.
Pierik, M.E., Gusmeroli, F., Marianna, G.D., Tamburini, A. & Bocchi, S. (2017) Meadows species composition, biodiversity and forage value in an alpine district: Relationships with environmental and dairy farm management variables. *Agriculture, Ecosystems & Environment*, **244**, 14–21.
Plummer, M. *et al.* (2003) JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria., p. 10.
R Core Team. (2018) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Roth, T., Allan, E., Pearman, P.B. & Amrhein, V. (2018) Functional ecology and imperfect detection of species. *Methods in Ecology and Evolution*, **9**, 917–928.
Royle, J.A., Dorazio, R.M. & Link, W.A. (2007) Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics*, **16**, 67–85.
Royle, J.A. & Link, W.A. (2006) Generalized occupancy models allowing false positive and false negative errors. *Ecology*, **87**, 835–841.
Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193–200.
Silva, D.P., Andrade, A.F., Oliveira, J.P., Morais, D.M., Vieira, J.E. & Engel, M.S. (2019) Current and future ranges of an elusive north american insect using species distribution models. *Journal of Insect Conservation*, **23**, 175–186.
Smart, A.S., Weeks, A.R., Rooyen, A.R. van, Moore, A., McCarthy, M.A. & Tingley, R. (2016) Assessing the cost-efficiency of environmental DNA sampling. *Methods in Ecology and Evolution*, **7**, 1291–1298.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.
Ter Braak, C.J. & Smilauer, P. (2002) *CANOCO reference manual and CanoDraw for windows user's guide: Software for canonical community ordination (version 4.5)*. www. canoco. com.
Tingley, M.W., Nadeau, C.P. & Sandor, M.E. (2020) Multi-species occupancy models as robust estimators of community richness. *Methods in Ecology and Evolution*, **11**, 633–642.
Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003) Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, **13**, 1790–1801.
Vaughan, I.P. & Ormerod, S.J. (2005) The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42**, 720–730.
Veran, S., Simpson, S.J., Sword, G.A., Deveson, E., Piry, S., Hines, J.E., *et al.* (2015) Modeling spatiotemporal dynamics of outbreaking species: Influence of environment and migration in a locust. *Ecology*, **96**, 737–748.
Walker, T.J. (1964) Cryptic species among sound-producing ensiferan orthoptera (gryllidae and tettigoniidae). *The Quarterly Review of Biology*, **39**, 345–355.
Watanabe, S. (2013) A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, **14**, 867–897.
Weking, S., Kämpf, I., Mathar, W. & Hölzel, N. (2016) Effects of land use and landscape patterns on orthoptera communities in the western siberian forest steppe. *Biodiversity and conservation*, **25**, 2341–2359.

Welsh, A.H., Lindenmayer, D.B. & Donnelly, C.F. (2013) Fitting and interpreting occupancy models. *PloS one*, **8**.
Wolda, H. (1988) Insect seasonality: why? *Annual review of ecology and systematics*, **19**, 1–18.
Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution*, **16**, 446–453.
Zipkin, E.F., DeWan, A. & Andrew Royle, J. (2009) Impacts of forest fragmentation on species richness: A hierarchical approach to community modelling. *Journal of Applied Ecology*, **46**, 815–822.
Zipkin, E.F., Grant, E.H.C. & Fagan, W.F. (2012) Evaluating the predictive abilities of community occupancy models using AUC while accounting for imperfect detection. *Ecological Applications*, **22**, 1962–1972.

ESM 1

Results comparison between model with grass height as occupancy covariate and model without this covariate

The aim of the study was to show MSOMs potential in entomological study. Hence, we developed a relatively simple MSOMs, with just few covariates. However, omitting potential important predictors of Orthoptera occupancy (or detectability) could bias the results. Therefore, we tried another model, similar in the detection model, but with the effect of grass height added in the occupancy model:

$$logit(\psi_{i,j}) = \alpha_{0_i} + \alpha_{1_i} \times \text{altitude}_j + \alpha_{2_i} \times \text{altitude}_j{}^2 + \alpha_{3_i} \times \text{height}_j$$

with α_{3_i} the linear effect of grass height on the occupancy probability of species i.

We compared the principal results found with this model to those presented in the manuscript. The relationship between the environmental parameters and the detection and occupancy probabilities did not change significantly between the two models (Fig S1). May be because the effect of grass height on the occupancy at the community-level and for most of the species was not significant (Fig S2). The estimated species richness was 77.6 ($IC_{95\%}$=[62.975, 102]), which is not signficantly different than the estimate of the simpliest model ($\hat{N}$=74.62, $IC_{95\%}$=[61, 98]). At the site level, the completeness estimates were very similar between the two models (Fig S3). With the simpliest model, 76 sites have completeness superior to 80%, while the second model estimated 77 sites above this threshold. The overall detection probabilities were also very similar (Fig S4). 31 species had an overall detection probability at the site level upper than 95% when considering that grass height influence both detection and occupancy, against 30 species when accounting only for grass height effect on detectability.
Results did not change meaningfully, neither did our inferences and interpretations. Hence, we kept the simpliest model to facilitate readers understanting.

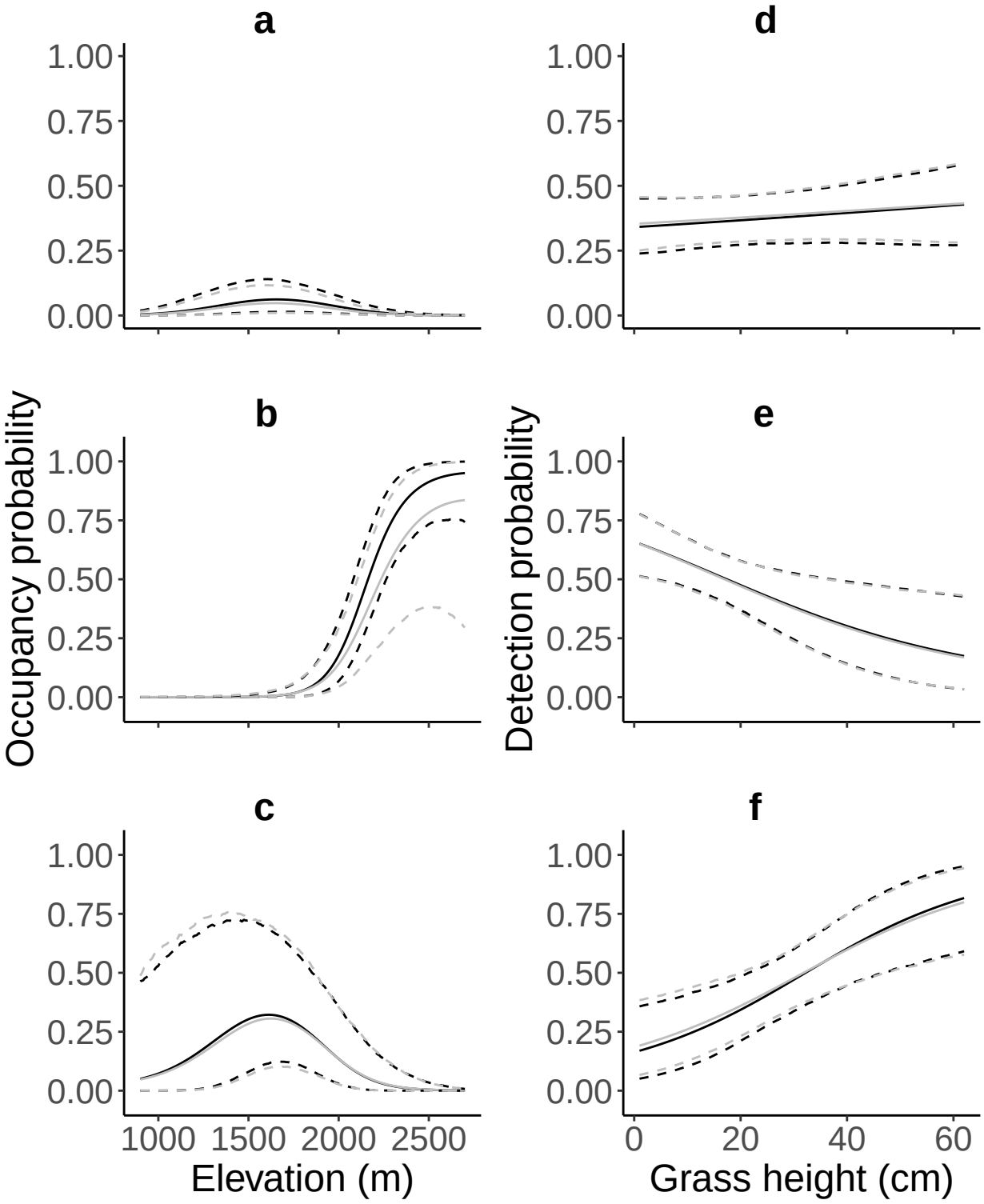

Figure 5: Effect of the altitude on the occupancy probability for (a) an average species at the community-level, (b) *Gomphocerus sibiricus sibiricus*, (c) *Antaxius pedestris*, and effect of the grass height on the probability of sighting (d) an average species at the community-level, (e) *Podisma dechambrei* and (f) *Euthystira brachyptera*. The colors correspond to model specification, with one model with grass height as occupancy covariate (grey lines) and one without (black lines). The solid lines represent the posterior mean, and the dashed lines correspond to the 95% credible interval.

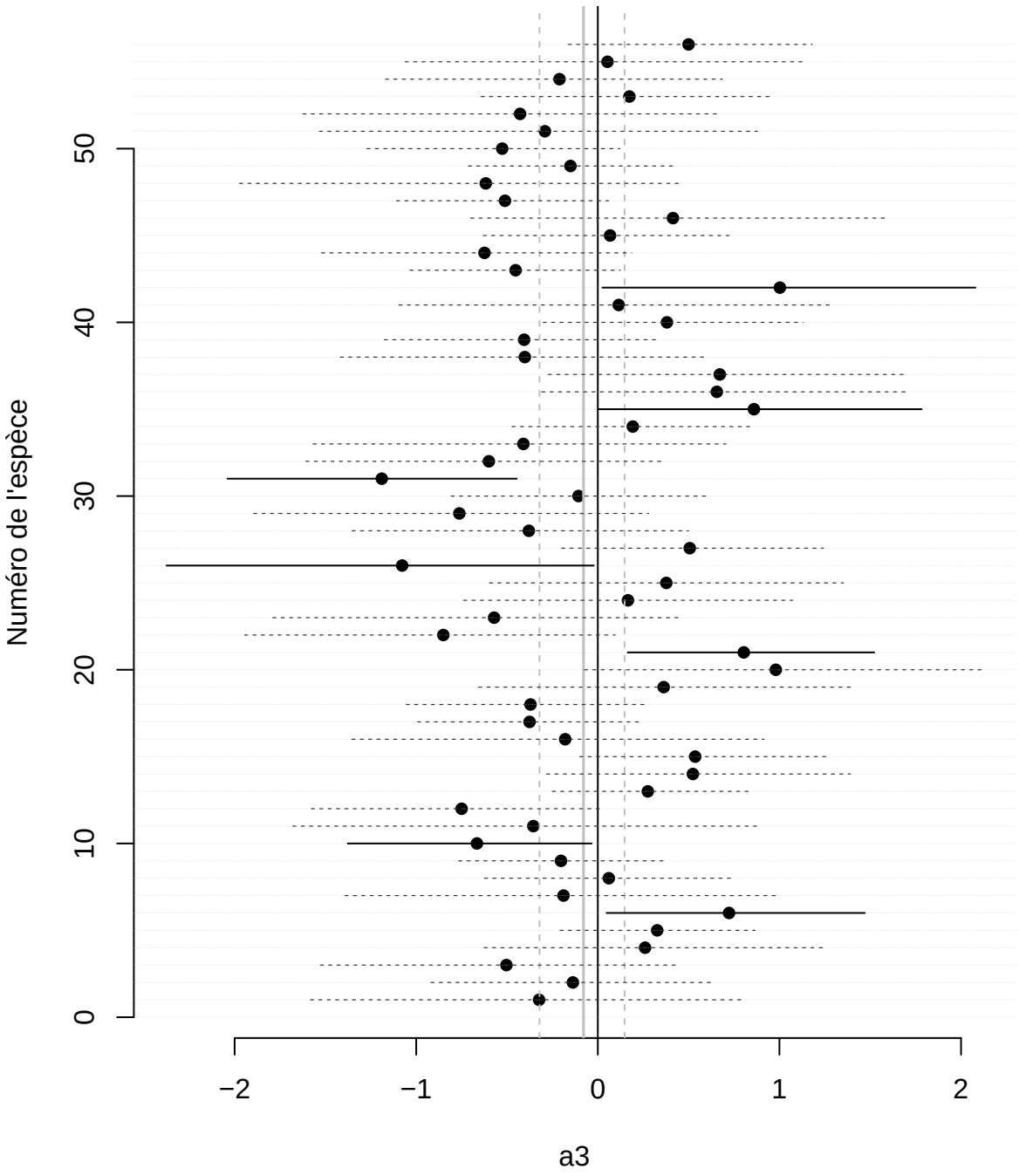

Figure 6: Estimates of the linear grass height effect on species-specific occupancy probabilities. Black points correspond to medians of posterior distributions and the segment to the credible interval associated. Segments with solid lines represent significant effects.

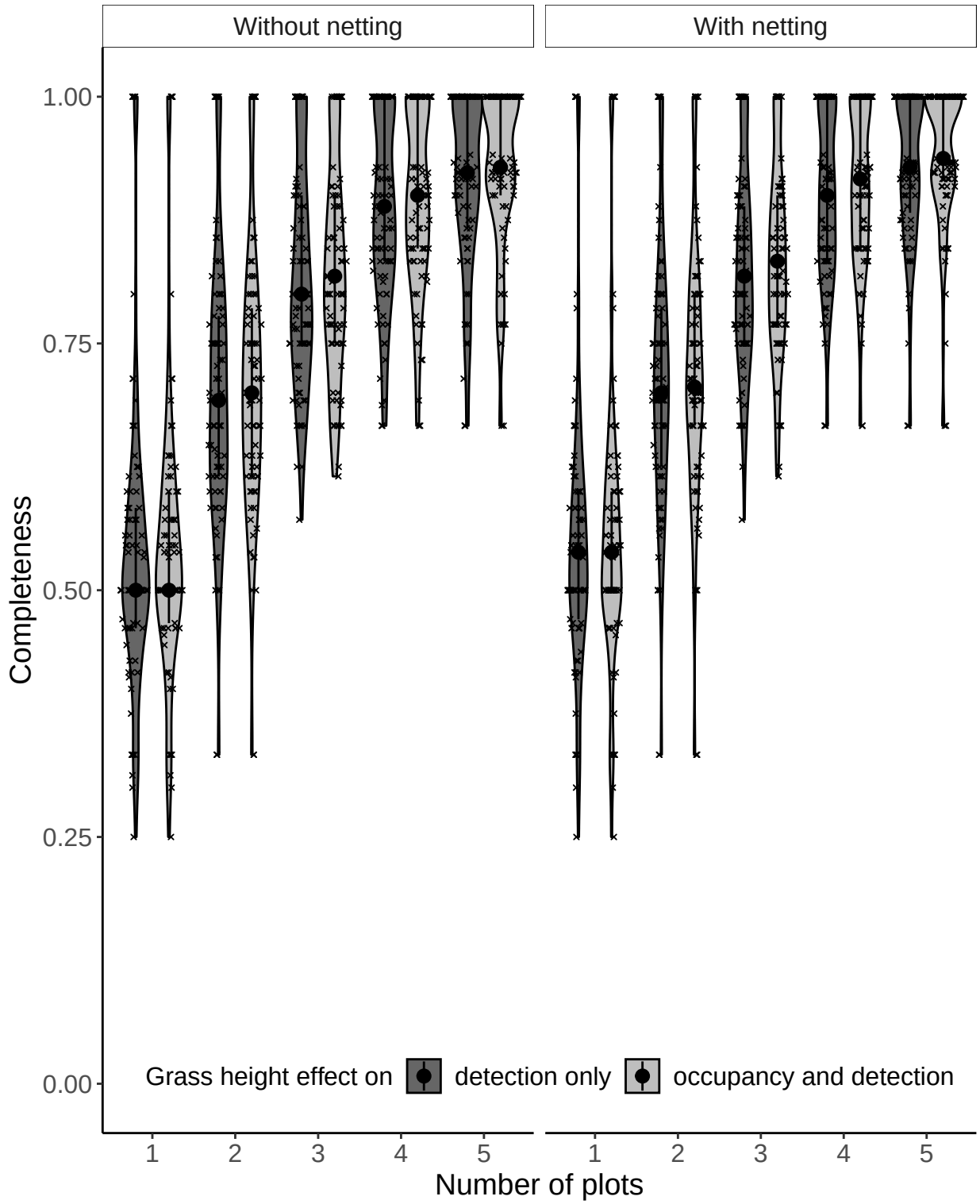

Figure 7: Inventory completeness at site level according to the model used, the number of plots sampled and the detection techniques used. Black dots represent the medians, and black segments represent the first and third quartiles.

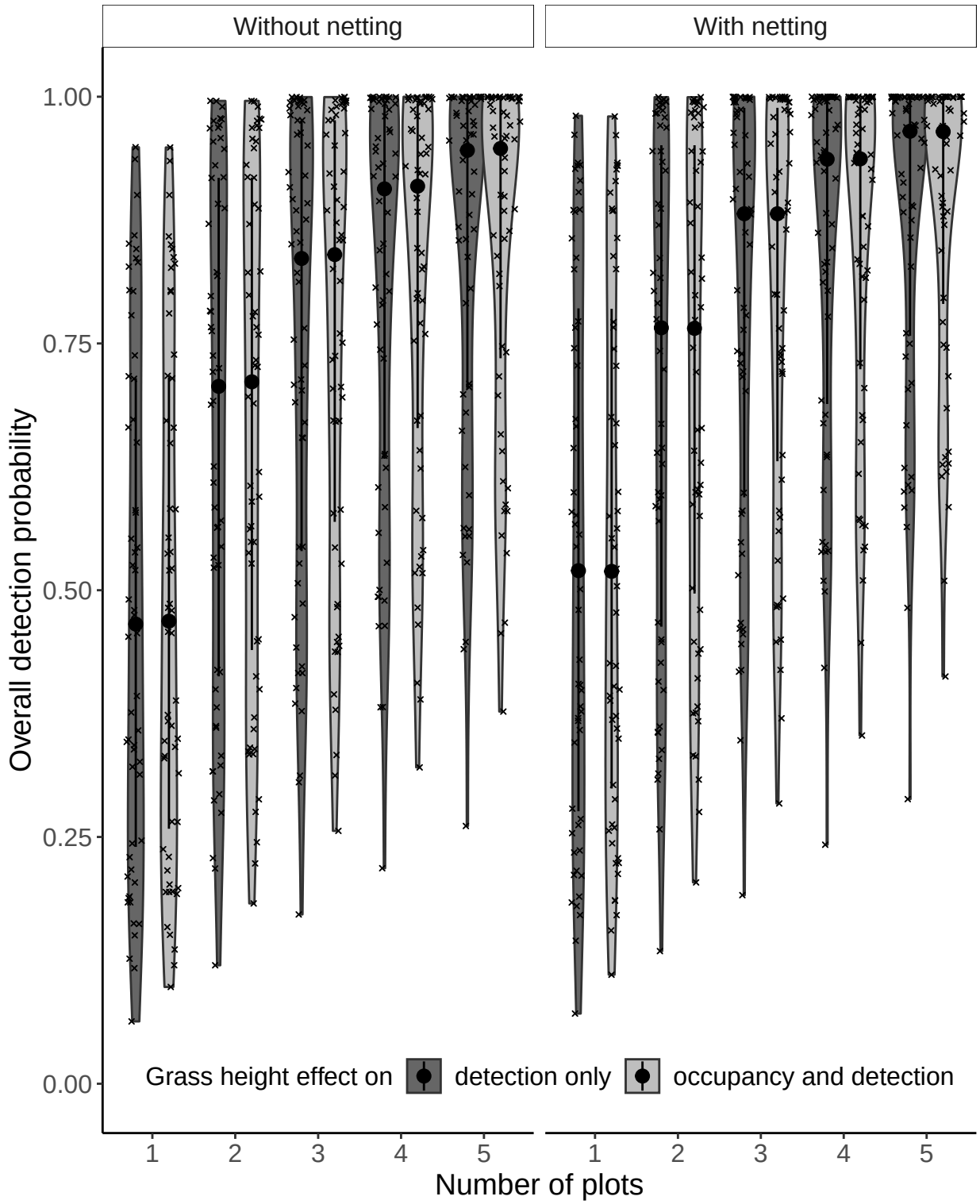

Figure 8: Distribution of the species-specific overall detection probabilities at the site level depending on the number of sampled plots, the model specification and the detection techniques used. Black dots represent the medians, and black segments represent the first and third quartiles.

ESM 2

```
print(out$model)
```

```
## $cluster1
## JAGS model:
##
##
##     model{
##
##     #Define prior distributions for community-level model parameters
##     omega ~ dunif(0,1)                                #probability of species
##
##     a0.mean ~ dunif(0,1)                              #species level effect c
##     mu.a0 <- log(a0.mean) - log(1-a0.mean)            #species level effect c
##     mu.a1 ~ dnorm(0, 0.001)                            #site covariable 1 ef
##     mu.a2 ~ dnorm(0, 0.001)
##     #mu.a3 ~ dnorm(0, 0.001)
##
##     tau.a0 ~ dgamma(0.1,0.1)                          #variability of species
##     tau.a1 ~ dgamma(0.1,0.1)                          #variability of site co
##     tau.a2 ~ dgamma(0.1,0.1)
##     #tau.a3 ~ dgamma(0.1,0.1)
##
##     mu.b1 ~ dnorm(0, 0.001)
##
##     mu.b0 ~ dnorm(0, 0.001)
##     mu.b2 ~ dnorm(0, 0.001)
##     mu.b3 ~ dnorm(0, 0.001)
##     #mu.b4 ~ dnorm(0, 0.001)
##
##     tau.b0 ~ dgamma(0.1,0.1)
##     tau.b2 ~ dgamma(0.1,0.1)
##     tau.b3 ~ dgamma(0.1,0.1)
##     #tau.b4 ~ dgamma(0.1,0.1)
##
##     tau.b1 ~ dgamma(0.1,0.1)
##
##
##     for (i in 1:(n+nzeroes)) {
##
##     #Create priors for species i from the community level prior distribut
##     w[i] ~ dbern(omega)                              #binary indicator: species
##
```

```
##     a0[i] ~ dnorm(mu.a0, tau.a0)                #species effect on occupa
##     a1[i] ~ dnorm(mu.a1, tau.a1)               #site covariable 1 effect
##     a2[i] ~ dnorm(mu.a2, tau.a2)
##     #a3[i] ~ dnorm(mu.a3, tau.a3)
##
##     b0[i] ~ dnorm(mu.b0, tau.b0)          #parameters for detection proba
##     b1[i] ~ dnorm(mu.b1, tau.b1)
##     b2[i] ~ dnorm(mu.b2, tau.b2)
##     b3[i] ~ dnorm(mu.b3, tau.b3)
##     #b4[i] ~ dnorm(mu.b4, tau.b4)
##
##
##     #Create a loop to estimate the Z matrix (true occurrence for species
##     #at point j.
##     for (j in 1:J) {
##     logit(psi[j,i]) <- a0[i] +
##             a1[i]*covSite1[j] + a2[i]*pow(covSite1[j],2) #+ #altitude ef
##             #a3[i]*covSite2[j]
##
##     mu.psi[j,i] <- psi[j,i]*w[i]          #site j could be occupied by sp
##     #if species i belongs to the community: mu.psi=psi, else: mu.psi=0
##     Z[j,i] ~ dbern(mu.psi[j,i])           #binary indicator: species i pr
##
##     #Create a loop to estimate detection for species i at point j during
##     #sampling period k.
##     for (k in 1:K[j]) {
##     logit(p[j,k,i]) <-  b0[i] + #view
##                         b1[i]*covDetection1[j,k] + #hearing
##                         b2[i]*covDetection2[j,k] + #sweep netting
##                         b3[i]*covDetection0[j,k]*covDetection3[j,k] #+ #s
##                         #b4[i]*covDetection2[j,k]*covDetection3[j,k] #net
##
##     mu.p[j,k,i] <- p[j,k,i]*Z[j,i]    #species i could be detected in sit
##     #if j is occupied: mu.p=p, else: mu.p=0
##     X[j,k,i] ~ dbern(mu.p[j,k,i])     #binary indicator observed: species
##
##     #Create simulated dataset to calculate the Bayesian p-value
##     Xnew[j,k,i] ~ dbern(mu.p[j,k,i])
##
##     #Pearson residuals
##     d[j,k,i]<-  abs(X[j,k,i] - mu.p[j,k,i])
##     dnew[j,k,i]<- abs(Xnew[j,k,i]- mu.p[j,k,i])
##     d2[j,k,i]<- pow(d[j,k,i],2)
##     dnew2[j,k,i]<- pow(dnew[j,k,i],2)
##
```

```
##     }
##
##     dsum[j,i]<- sum(d2[j,1:K[j],i])
##     dnewsum[j,i]<- sum(dnew2[j,1:K[j],i])
##
##     }
##     }
##
##
##     #Calculate the discrepancy measure
##     p.fit<-sum(dsum[1:J,1:(n)])
##     p.fitnew<-sum(dnewsum[1:J,1:(n)])
##
##     #Sum all species observed (n) and unobserved species (n0) to find the
##     #total estimated richness
##     n0 <- sum(w[(n+1):(n+nzeroes)])
##     N <- n + n0
##
##     #Create a loop to determine point level richness estimates for the
##     #whole community.
##     for(j in 1:J){
##       Nsite[j]<- sum(Z[j,1:(n+nzeroes)])
##     }
## }
## Fully observed variables:
##  J K X covDetection0 covDetection1 covDetection2 covDetection3 covSite1 r
##
## $cluster2
## JAGS model:
##
##
##     model{
##
##     #Define prior distributions for community-level model parameters
##     omega ~ dunif(0,1)                              #probability of species
##
##     a0.mean ~ dunif(0,1)                            #species level effect c
##     mu.a0 <- log(a0.mean) - log(1-a0.mean)          #species level effect c
##     mu.a1 ~ dnorm(0, 0.001)                          #site covariable 1 ef
##     mu.a2 ~ dnorm(0, 0.001)
##     #mu.a3 ~ dnorm(0, 0.001)
##
##     tau.a0 ~ dgamma(0.1,0.1)                        #variability of species
##     tau.a1 ~ dgamma(0.1,0.1)                        #variability of site co
##     tau.a2 ~ dgamma(0.1,0.1)
```

```
##      #tau.a3 ~ dgamma(0.1,0.1)
##
##      mu.b1 ~ dnorm(0, 0.001)
##
##      mu.b0 ~ dnorm(0, 0.001)
##      mu.b2 ~ dnorm(0, 0.001)
##      mu.b3 ~ dnorm(0, 0.001)
##      #mu.b4 ~ dnorm(0, 0.001)
##
##      tau.b0 ~ dgamma(0.1,0.1)
##      tau.b2 ~ dgamma(0.1,0.1)
##      tau.b3 ~ dgamma(0.1,0.1)
##      #tau.b4 ~ dgamma(0.1,0.1)
##
##      tau.b1 ~ dgamma(0.1,0.1)
##
##
##      for (i in 1:(n+nzeroes)) {
##
##      #Create priors for species i from the community level prior distribut
##      w[i] ~ dbern(omega)                          #binary indicator: species
##
##      a0[i] ~ dnorm(mu.a0, tau.a0)                  #species effect on occupa
##      a1[i] ~ dnorm(mu.a1, tau.a1)                 #site covariable 1 effect
##      a2[i] ~ dnorm(mu.a2, tau.a2)
##      #a3[i] ~ dnorm(mu.a3, tau.a3)
##
##      b0[i] ~ dnorm(mu.b0, tau.b0)           #parameters for detection proba
##      b1[i] ~ dnorm(mu.b1, tau.b1)
##      b2[i] ~ dnorm(mu.b2, tau.b2)
##      b3[i] ~ dnorm(mu.b3, tau.b3)
##      #b4[i] ~ dnorm(mu.b4, tau.b4)
##
##
##      #Create a loop to estimate the Z matrix (true occurrence for species
##      #at point j.
##      for (j in 1:J) {
##      logit(psi[j,i]) <- a0[i] +
##              a1[i]*covSite1[j] + a2[i]*pow(covSite1[j],2) #+ #altitude ef
##              #a3[i]*covSite2[j]
##
##      mu.psi[j,i] <- psi[j,i]*w[i]           #site j could be occupied by sp
##      #if species i belongs to the community: mu.psi=psi, else: mu.psi=0
##      Z[j,i] ~ dbern(mu.psi[j,i])            #binary indicator: species i pr
##
```

```
##      #Create a loop to estimate detection for species i at point j during
##      #sampling period k.
##      for (k in 1:K[j]) {
##      logit(p[j,k,i]) <-  b0[i] + #view
##                          b1[i]*covDetection1[j,k] + #hearing
##                          b2[i]*covDetection2[j,k] + #sweep netting
##                          b3[i]*covDetection0[j,k]*covDetection3[j,k] #+ #s
##                          #b4[i]*covDetection2[j,k]*covDetection3[j,k] #net
##
##      mu.p[j,k,i] <- p[j,k,i]*Z[j,i]    #species i could be detected in sit
##      #if j is occupied: mu.p=p, else: mu.p=0
##      X[j,k,i] ~ dbern(mu.p[j,k,i])     #binary indicator observed: species
##
##      #Create simulated dataset to calculate the Bayesian p-value
##      Xnew[j,k,i] ~ dbern(mu.p[j,k,i])
##
##      #Pearson residuals
##      d[j,k,i]<-  abs(X[j,k,i] - mu.p[j,k,i])
##      dnew[j,k,i]<- abs(Xnew[j,k,i]- mu.p[j,k,i])
##      d2[j,k,i]<- pow(d[j,k,i],2)
##      dnew2[j,k,i]<- pow(dnew[j,k,i],2)
##
##      }
##
##      dsum[j,i]<- sum(d2[j,1:K[j],i])
##      dnewsum[j,i]<- sum(dnew2[j,1:K[j],i])
##
##      }
##      }
##
##
##      #Calculate the discrepancy measure
##      p.fit<-sum(dsum[1:J,1:(n)])
##      p.fitnew<-sum(dnewsum[1:J,1:(n)])
##
##      #Sum all species observed (n) and unobserved species (n0) to find the
##      #total estimated richness
##      n0 <- sum(w[(n+1):(n+nzeroes)])
##      N <- n + n0
##
##      #Create a loop to determine point level richness estimates for the
##      #whole community.
##      for(j in 1:J){
##        Nsite[j]<- sum(Z[j,1:(n+nzeroes)])
##      }
```

```
## }
## Fully observed variables:
##  J K X covDetection0 covDetection1 covDetection2 covDetection3 covSite1 r
##
## $cluster3
## JAGS model:
##
##
##     model{
##
##     #Define prior distributions for community-level model parameters
##     omega ~ dunif(0,1)                                #probability of species
##
##     a0.mean ~ dunif(0,1)                              #species level effect c
##     mu.a0 <- log(a0.mean) - log(1-a0.mean)            #species level effect c
##     mu.a1 ~ dnorm(0, 0.001)                            #site covariable 1 ef
##     mu.a2 ~ dnorm(0, 0.001)
##     #mu.a3 ~ dnorm(0, 0.001)
##
##     tau.a0 ~ dgamma(0.1,0.1)                          #variability of species
##     tau.a1 ~ dgamma(0.1,0.1)                          #variability of site co
##     tau.a2 ~ dgamma(0.1,0.1)
##     #tau.a3 ~ dgamma(0.1,0.1)
##
##     mu.b1 ~ dnorm(0, 0.001)
##
##     mu.b0 ~ dnorm(0, 0.001)
##     mu.b2 ~ dnorm(0, 0.001)
##     mu.b3 ~ dnorm(0, 0.001)
##     #mu.b4 ~ dnorm(0, 0.001)
##
##     tau.b0 ~ dgamma(0.1,0.1)
##     tau.b2 ~ dgamma(0.1,0.1)
##     tau.b3 ~ dgamma(0.1,0.1)
##     #tau.b4 ~ dgamma(0.1,0.1)
##
##     tau.b1 ~ dgamma(0.1,0.1)
##
##
##     for (i in 1:(n+nzeroes)) {
##
##     #Create priors for species i from the community level prior distribut
##     w[i] ~ dbern(omega)                            #binary indicator: species
##
##     a0[i] ~ dnorm(mu.a0, tau.a0)                    #species effect on occupa
```

```
##     a1[i] ~ dnorm(mu.a1, tau.a1)                #site covariable 1 effect
##     a2[i] ~ dnorm(mu.a2, tau.a2)
##     #a3[i] ~ dnorm(mu.a3, tau.a3)
##
##     b0[i] ~ dnorm(mu.b0, tau.b0)          #parameters for detection proba
##     b1[i] ~ dnorm(mu.b1, tau.b1)
##     b2[i] ~ dnorm(mu.b2, tau.b2)
##     b3[i] ~ dnorm(mu.b3, tau.b3)
##     #b4[i] ~ dnorm(mu.b4, tau.b4)
##
##
##     #Create a loop to estimate the Z matrix (true occurrence for species
##     #at point j.
##     for (j in 1:J) {
##     logit(psi[j,i]) <- a0[i] +
##             a1[i]*covSite1[j] + a2[i]*pow(covSite1[j],2) #+ #altitude ef
##             #a3[i]*covSite2[j]
##
##     mu.psi[j,i] <- psi[j,i]*w[i]          #site j could be occupied by sp
##     #if species i belongs to the community: mu.psi=psi, else: mu.psi=0
##     Z[j,i] ~ dbern(mu.psi[j,i])           #binary indicator: species i pr
##
##     #Create a loop to estimate detection for species i at point j during
##     #sampling period k.
##     for (k in 1:K[j]) {
##     logit(p[j,k,i]) <-  b0[i] + #view
##                         b1[i]*covDetection1[j,k] + #hearing
##                         b2[i]*covDetection2[j,k] + #sweep netting
##                         b3[i]*covDetection0[j,k]*covDetection3[j,k] #+ #s
##                         #b4[i]*covDetection2[j,k]*covDetection3[j,k] #net
##
##     mu.p[j,k,i] <- p[j,k,i]*Z[j,i]    #species i could be detected in sit
##     #if j is occupied: mu.p=p, else: mu.p=0
##     X[j,k,i] ~ dbern(mu.p[j,k,i])     #binary indicator observed: species
##
##     #Create simulated dataset to calculate the Bayesian p-value
##     Xnew[j,k,i] ~ dbern(mu.p[j,k,i])
##
##     #Pearson residuals
##     d[j,k,i]<-  abs(X[j,k,i] - mu.p[j,k,i])
##     dnew[j,k,i]<- abs(Xnew[j,k,i]- mu.p[j,k,i])
##     d2[j,k,i]<- pow(d[j,k,i],2)
##     dnew2[j,k,i]<- pow(dnew[j,k,i],2)
##
##     }
```

```
##
##     dsum[j,i]<- sum(d2[j,1:K[j],i])
##     dnewsum[j,i]<- sum(dnew2[j,1:K[j],i])
##
##     }
##     }
##
##
##     #Calculate the discrepancy measure
##     p.fit<-sum(dsum[1:J,1:(n)])
##     p.fitnew<-sum(dnewsum[1:J,1:(n)])
##
##     #Sum all species observed (n) and unobserved species (n0) to find the
##     #total estimated richness
##     n0 <- sum(w[(n+1):(n+nzeroes)])
##     N <- n + n0
##
##     #Create a loop to determine point level richness estimates for the
##     #whole community.
##     for(j in 1:J){
##       Nsite[j]<- sum(Z[j,1:(n+nzeroes)])
##     }
## }
## Fully observed variables:
##  J K X covDetection0 covDetection1 covDetection2 covDetection3 covSite1 r
```

ESM 3

```
data.frame("taxon"=dimnames(X[,,1:n])[[3]],
           "occu"=plogis(out$mean$a0[1:n]),
           "occuI"=plogis(out$q2.5$a0[1:n]),
           "occuS"=plogis(out$q97.5$a0[1:n]),
           "pVu"=plogis(out$mean$b0[1:n]),
           "pVuI"=plogis(out$q2.5$b0[1:n]),
           "pVuS"=plogis(out$q97.5$b0[1:n]),
           "pE"=colMeans(plogis(out$sims.list$b0[,1:n]+out$sims.list$b1[,1:n
           t(apply(plogis(out$sims.list$b0[,1:n]+out$sims.list$b1[,1:n]),2,
           "pF"=colMeans(plogis(out$sims.list$b0[,1:n]+out$sims.list$b2[,1:n
           t(apply(plogis(out$sims.list$b0[,1:n]+out$sims.list$b2[,1:n]),2,
           probam_meth_sp[,10],
           probainf_meth_sp[,10],
           probasup_meth_sp[,10],
           probam_meth_sp[,9],
```

```
            probainf_meth_sp[,9],
            probasup_meth_sp[,9]) %>%
  mutate_at(-1,function(x)round(x,2)) -> tabS1

data.frame("taxon"=dimnames(X[,,1:n])[[3]],
            (out$mean$a1[1:n]),
            (out$q2.5$a1[1:n]),
            (out$q97.5$a1[1:n]),
            (out$mean$a2[1:n]),
            (out$q2.5$a2[1:n]),
            (out$q97.5$a2[1:n]),
            (out$mean$b3[1:n]),
            (out$q2.5$b3[1:n]),
            (out$q97.5$b3[1:n])) %>%
  mutate_at(-1,function(x)round(x,2)) -> tabS2

data.frame("Mean"=c(out$mean$mu.a0,out$mean$mu.a1,out$mean$mu.a2,
              out$mean$mu.b0,out$mean$mu.b1,out$mean$mu.b2,out$mean$mu.b3),
            "Q2.5"=c(out$q2.5$mu.a0,out$q2.5$mu.a1,out$q2.5$mu.a2,
              out$q2.5$mu.b0,out$q2.5$mu.b1,out$q2.5$mu.b2,out$q2.5$mu.b3),
            "Q97.5"=c(out$q97.5$mu.a0,out$q97.5$mu.a1,out$q97.5$mu.a2,
              out$q97.5$mu.b0,out$q97.5$mu.b1,out$q97.5$mu.b2,out$q97.5$mu.b
  mutate_all(function(x)round(x,2)) -> tabS3

write.csv(tabS1, file="tabS1.csv")
write.csv(tabS2, file="tabS2.csv")
write.csv(tabS3, file="tabS3.csv")
```