Supporting information for :

Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic occurrence data

Christophe Botella[1,2,3,4], Alexis Joly[1], Pierre Bonnet[3,5], François Munoz[6], and Pascal Monestiez[4]

[1]INRIA Sophia-Antipolis - ZENITH team, LIRMM - UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France.

[2]INRAE, UMR AMAP, F-34398 Montpellier, France.

[3]Univ Montpellier, UMR AMAP, Montpellier, France.

[4]INRAE, BioSP, Site Agroparc, 84914 Avignon, France.

[5]CIRAD, UMR AMAP, F-34398 Montpellier, France.

[6]Université Grenoble Alpes, 621 avenue Centrale, 38400 Saint-Martin-d'Hères, France.

1 Appendix A: Expected estimators and information matrix

Expected estimators. From the negative log-likelihood of the model expressed in equation **(3)** in the article, we derived an expression of the asymptotic density and intercept estimators in the system of equations 1. It shows that the density estimators minimize a weighted sum of Kullback-Leibler divergences from the true to estimated occurrence densities. We note in the following $n_i := |Z_i|$ and $\theta_i = (\alpha_i, \beta_i)$.

$$\begin{array}{lrcl} & \mathbb{E}((\hat{\gamma},\hat{\beta}_1,...,\hat{\beta}_N)) & = & \underset{\gamma,\beta_1,...,\beta_N}{\text{argmin}} \sum_{i=1}^{N}(\int_D s\lambda^i d\mu) D_{KL}^{D}(s\lambda^i || s_\gamma \lambda^i_{(0,\beta_i)}) \\ \forall i \in [|1,N|], & \mathbb{E}(\hat{\alpha}_i) & = & \log(\int_D s\lambda^i d\mu / \int_D s_{\mathbb{E}(\hat{\gamma})} \exp(\mathbb{E}(\hat{\beta}_i)^T x) d\mu) \end{array} \quad (1)$$

Proof:

$$
\begin{aligned}
&\mathbb{E}(\hat{\theta})\\
=\quad & \lim_{n_1,...,n_N \to \infty} \operatorname*{argmin}_{\theta} - \log(p(Z_1, ..., Z_n|\theta))\\
=\quad & \operatorname*{argmin}_{\theta} \lim_{n_1,...,n_N \to \infty} \sum_{i=1}^{N} n_i \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{n_i} - \frac{\sum_{k=1}^{n_i} \log(s_\gamma(z_i^k)\lambda^i_{\theta_i}(z_i^k))}{n_i} \right)\\
=\quad & \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} \lim_{n_i \to \infty} n_i \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{n_i} - \frac{\sum_{k=1}^{n_i} \log(s_\gamma(z_i^k)\lambda^i_{\theta_i}(z_i^k))}{n_i} \right)\\
=\quad & \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} \lim_{n_i \to \infty} n_i \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{n_i} - \int_D \frac{s(z)\lambda^i(z)}{\int_D s\lambda^i d\mu} \log(s_\gamma(z)\lambda^i_{\theta_i}(z))\mu(dz) \right) && \text{Large number law}\\
& && \text{and transfer theorem}\\
=\quad & \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} (\textstyle\int_D s\lambda^i d\mu) \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} + \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log(s\lambda^i) d\mu - \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log(s_\gamma \lambda^i_{\theta_i}) d\mu \right) && \text{Large number law}\\
& && \text{+ independent term}\\
=\quad & \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} (\textstyle\int_D s\lambda^i d\mu) \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} + \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{s\lambda^i}{s_\gamma \lambda^i_{\theta_i}} \right) d\mu \right)\\
=\quad & \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} (\textstyle\int_D s\lambda^i d\mu) \left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} - \log\left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} \right) + \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{s\lambda^i \int_D s_\gamma \lambda^i_{\theta_i} d\mu}{s_\gamma \lambda^i_{\theta_i} \int_D s\lambda^i d\mu} \right) d\mu \right)\\
=\quad & \operatorname*{argmin}_{\theta} \sum_{i=1}^{N} (\textstyle\int_D s\lambda^i d\mu) \left(\text{nlogL}(\alpha_i) + D^D_{KL}(s\lambda^i || s_\gamma \lambda^i_{\theta_i}) \right)
\end{aligned}
$$

Where $\text{nlogL}(\alpha_i) := \dfrac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} - \log\left(\dfrac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} \right) = -\log\left(\dfrac{\left(\frac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} \right)^1}{1!} \exp\left(-\dfrac{\int_D s_\gamma \lambda^i_{\theta_i} d\mu}{\int_D s\lambda^i d\mu} \right) \right)$ is the negative log-likelihood of a Poisson regression with a single count of value one. The likelihood is maximized when the Poisson parameter $\int_D s_\gamma \lambda^i_{\theta_i} d\mu / \int_D s\lambda^i d\mu = 1$, which then minimizes $\text{nlogL}(\alpha_i)$ with $\text{nlogL}(\alpha_i) = 0$, and translates into $\alpha_i = \log(\int_D s\lambda^i d\mu / \int_D s_\gamma \exp(\beta_i^T x) d\mu)$. In other words, we can choose α_i to minimize $\text{nlogL}(\alpha_i)$ whatever the values of $\gamma, \beta_1, ..., \beta_N, s, \lambda^1, ..., \lambda^N$. This means that the minimization of the whole sum with respect to $\gamma, \beta_1, ..., \beta_N$ is unaffected by the terms $(\int_D s\lambda^i d\mu)\text{nlogL}(\alpha_i)$ which can be removed in the expression of $\mathbb{E}(\hat{\gamma}, \hat{\beta_1}, ..., \hat{\beta_N})$, and gives us the first equation of system 1. The second equation of 1 is shown by remarking that, conversely, the term $D^D_{KL}(s\lambda^i || s_\gamma \lambda^i_{\theta_i})$ is totally independent of α_i. Indeed, when replacing α_i by $\alpha_i + \delta$ we have:

$$
\begin{aligned}
D^D_{KL}(s\lambda^i || s_\gamma \exp(\alpha_i + \delta + \beta_i^T x)) &= \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{s\lambda^i \int_D s_\gamma \exp(\alpha_i + \delta + \beta_i^T x) d\mu}{s_\gamma \exp(\alpha_i + \delta + \beta_i^T x) \int_D s\lambda^i d\mu} \right) d\mu\\
&= \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{e^\delta s\lambda^i \int_D s_\gamma \exp(\alpha_i + \beta_i^T x) d\mu}{e^\delta s_\gamma \exp(\alpha_i + \beta_i^T x) \int_D s\lambda^i d\mu} \right) d\mu\\
&= \int_D \frac{s\lambda^i}{\int_D s\lambda^i d\mu} \log\left(\frac{s\lambda^i \int_D s_\gamma \exp(\alpha_i + \beta_i^T x) d\mu}{s_\gamma \exp(\alpha_i + \beta_i^T x) \int_D s\lambda^i d\mu} \right) d\mu\\
&= D^D_{KL}(s\lambda^i || s_\gamma \exp(\alpha_i + \beta_i^T x))
\end{aligned}
$$

Finally, the computation of the expected estimators can be separated as follows. First, the density parameter estimates $\gamma, \beta_1, ..., \beta_N$ are given by resolving the first equation of the system 1, and then the intercept parameter estimates $\alpha_1, ..., \alpha_N$ are given by resolving the other equations.

Fisher information matrix of the model. Here we describe $I(\theta)$, the global Fisher information matrix of our model parameters, and show its particular structure. Note that the Fisher information matrix is also the Hessian, or curvature, matrix of the negative log-likelihood. Indeed, $I(\theta)$ includes the second and cross derivatives of the negative log-likelihood described in **equation (3)** of section **2.2 - Inference** of the article (see also Bickel and Doksum [2015], section 6.2.2 , p.386, for more details on the Fisher information matrix).

Because of our model structure, $I(\theta)$ has many 0. We compute its non-null submatrices as follows. To simplify the notations, we consider here that all species densities are functions of the same vector of environmental features x, such that $\forall z \in D, x(z) \in \mathbb{R}^p$.

$\beta_i \in \mathbb{R}^p$ is the vector of parameters that model species i density in the environmental space for any $i \in [|1, N|]$. The Fisher information matrix for this parameter is derived from the second and cross derivatives of the negative log-likelihood, in equation **equation (3)** of the article, with respect to the components of β_i. That is:

$I(\beta_i) = \int_D xx^T s\lambda^i_{\theta_i} d\mu$

$\alpha_i \in \mathbb{R}$ is the intercept parameter of species i that is directly linked to the global abundance and detection/reporting probability of the species. It equals the total expected occurrence count of species i:

$I(\alpha_i) = \int_D s\lambda^i_{\theta_i} d\mu = \mathbb{E}(n_i)$

$\gamma_j \in \mathbb{R}$ is the parameter of the sampling effort in cell j. The cross information between cell j and j' is null when $j \neq j'$ cells form a partition of D and do not intersect. It equals the total expected occurrence count of cell j:

$$\begin{aligned} I(\gamma_j) &= \sum_{i=1}^{N} \int_D s\lambda^i_{\theta_i} d\mu \\ &= e^{\gamma_j} \sum_{i=1}^{N} \int_{c_j} \lambda^i_{\theta_i} d\mu \\ &= \mathbb{E}(n^j) \end{aligned} \tag{2}$$

The cross information of γ_j and β_i is written:

$I(\gamma_j, \beta_i) = \int_{c_j} xe^{\gamma_j} \lambda^i_{\theta_i} d\mu$

The cross information of γ_j and α_i equals the expected occurrence count of species i in cell j:

$I(\gamma_j, \alpha_i) = \int_{c_j} e^{\gamma_j} \lambda^i_{\theta_i} d\mu = \mathbb{E}(n_i^j)$

The cross information of β_i and α_i is written:

$I(\beta_i, \alpha_i) = \int_D xs\lambda^i_{\theta_i} d\mu$

The remaining information matrix is null. In particular we have:

$$I(\gamma) = \begin{pmatrix} I(\gamma_2) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I(\gamma_Q) \end{pmatrix}$$

Thus, we exhibit the structure of $I(\theta)$ as follows:

$$I(\theta) = \begin{pmatrix} I(\gamma) & I(\gamma,\alpha_1)^T & I(\gamma,\beta_1)^T & \dots & I(\gamma,\alpha_N)^T & I(\gamma,\beta_N)^T \\ I(\gamma,\alpha_1) & I(\alpha_1) & I(\beta_1,\alpha_1)^T & 0 & 0 & 0 \\ I(\gamma,\beta_1) & I(\beta_1,\alpha_1) & I(\beta_1) & 0 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 & 0 \\ I(\gamma,\alpha_N) & 0 & 0 & 0 & I(\alpha_N) & I(\beta_N,\alpha_N)^T \\ I(\gamma,\beta_N) & 0 & 0 & 0 & I(\beta_N,\alpha_N) & I(\beta_N) \end{pmatrix} \quad (3)$$

2 Appendix B: Model identifiability and robustness

2.1 Necessary and sufficient conditions for structural identifiability.

The structural identifiability of a model means that, for any set of true parameters, there are two equivalent properties: (i) the parameter estimates converge to the true parameters for any infinite sample, (ii) the estimates are unbiased, i.e. they are exact in expectation. Our model is structurally identifiable (for all sets of parameters) in the multi-species case if it is structurally identifiable in the single-species case. The single-species case is a Poisson process whose log-linear intensity function may be noted $z \to \theta^T v(z)$ where $\forall z \in D,\ v(z) = (1, 1_{z\in c_2}, ..., 1_{z\in c_Q}, x_1(z), ..., x_p(z))$, containing the intercept, the indicator functions of the cells c_j, and the environmental features vector. Then, according to the CNS identifiability condition shown for log-linear Poisson processes in Rathbun and Cressie [1994], the model is identifiable if and only if the matrix $\int_D v(z)v(z)^T dz$ is of full rank, i.e. of rank $1+p+Q-1$.

This condition means that there must be no linear condition of the non-constant functions of v that is constant. This condition is fulfilled if there is no linear combination of the environmental features that is constant across all sampling cells. For a single environmental feature, this would mean that this feature must vary inside at least one sampling cell. In the multivariate case, a simple interpretable identifiability condition is hard to provide. Fulfilling the condition above is sufficient to ensure unicity and convergence of the estimator for any dataset. However, for a finite number of occurrences, being close to non-identifiability is often a synonym of facing numerical approximation problems in the likelihood optimization, or getting high correlations between distinct parameter estimators. We need stronger conditions to ensure good estimability ([Jacquez and Greif, 1985]) of the model parameters. We thus advise the user, after having fit the model, to check the condition number of the inverse observed Fisher information matrix. This matrix may be computed by replacing parameters of the information matrix in equation 3 with their estimates. The closer the condition number is to 1, the lower the global covariance between pairs of distinct parameter estimators.

Another option for the user, before fitting the model, is to numerically compute the condition number of the matrix $\int_D v(z)v(z)^T dz$ when designing the sampling mesh. Then, the user may choose a sampling mesh that has a condition number inferior to 10^6 (in our experience) while keeping in mind the other conditions provided in the article. This may directly eliminate some designs and is much faster than fitting the model and computing the condition number for the whole information matrix, even though the latter is a more accurate indice of estimability as it accounts for the data point distribution.

2.2 Remarks on model robustness.

The structural identifiability of the model means that we expect good separation of the sampling effort density and the species density in our estimates, but this is on the restrictive condition that the model is well specified. The sampling effort and species density model representation must be able to exactly fit their true values. In general, this does not happen in reality, as it is not realistic to assume that sampling effort is constant per sampling cell. The ability of a statistical model to converge to estimates that are close to the true values even though the model specification is wrong is called its robustness. The simulation study described in the article shows that our estimates are robust as long as the sampling effort variation within cells is reasonable. In the following section, we provide more detail on the conditions that induce bias in the model. First, we describe two examples where such bias appeared, then, we provide theoretical arguments to explain what type of model misspecification causes bias.

Lack of robustness: Two examples. In profile (3) of the simulation experiment in **Appendix F**, the sampling model does not allow the estimate to converge exactly with the true sampling model, which decreases continuously as the environmental feature increases. As the sampling cells are segments along the environmental gradient, the sampling effort actually decreases as the environmental variable increases in every cell. In this setting, we observe a significant deviation between the sampling effort estimate and the species density. As can be seen in Figure 3, the species density modes both deviate on the left of the environmental range, compensating the underestimation of the sampling effort in this range. This indicates that the error on the parameters of both species have the same sign. This bias thus coincides with a trend of monotonic variation in the true sampling within the model sampling cells.

Bias also appears in case **x:alti H:-20** of the simulation experiment described in the article. The environmental variable here is the elevation gradient, a variable that negatively impacts the sampling effort and that has a much finer resolution than the sampling cells and varies strongly inside certain cells. This bias does not appear as much in the case of the precipitation variable (**x:**$chbio_{12}$). This is probably because, even though precipitation is linked with sampling effort in the same way as elevation, it varies much less within sampling cells.

Theoretical arguments. Here we clarify the robustness problem and then provide some mathematical arguments that corroborate the previous empirical observations. In the single species case, we derive from equation 1 the following estimator expectation:

$$\mathbb{E}(\hat{\gamma}, \hat{beta}) = \underset{\gamma,\beta}{\text{argmin}}\ D_{KL}(s\lambda \circ x || s_\gamma \lambda_{0,\beta} \circ x)$$

The model is optimized so that the variation of the fitted occurrence density $s_\gamma \lambda_\beta \circ x$ across space fits the variation of observed occurrence density $s\lambda \circ x$. When the model is misspecified for the sampling effort, i.e. $s \notin \{s_\gamma, \gamma \in \mathbb{R}^{Q-1}\}$, then the best approximation of $s\lambda \circ x$ is not necessarily the product of $\lambda \circ x$ and $s_{\gamma_{BCCA}} := \underset{\gamma}{\text{argmin}} D_{KL}(s\lambda || s_\gamma \lambda)$, the best cell-wise constant approximation (BCCA) of s for λ. We note that bias due to a lack of robustness appears if there is a parameterization of the sampling effort $\gamma * \in \mathbb{R}^{Q-1}$ that maximizes the likelihood $\mathbb{E}(\hat{\gamma}) = \gamma*$ but is not the BCCA $\gamma* \neq \gamma_{BCCA}$. This happens if $D_{KL}(s\lambda || s_{\gamma*}\lambda_{\mathbb{E}(\hat{\beta})}) < D_{KL}(s\lambda || s_{\gamma_{BCCA}}\lambda)$. In this case, the estimator of the species density $\lambda_{\mathbb{E}(\hat{\beta})}$ will be necessarily biased ($\lambda_{\mathbb{E}(\hat{\beta})} \neq \lambda$) because, by definition, the BCCA is the solution that maximizes the likelihood if the estimator of species density is unbiased. Thus, a bias due to lack of robustness results in a deviation of both the sampling effort and the species density estimators from the values that we want to obtain.

Secondly, we propose an explanation regarding the properties of s that cause a lack of robustness in our model. We can characterize this phenomenon more accurately in the multi-species case with a re-expression and analysis of the asymptotic model negative log-likelihood given in equation (1) of **Appendix A**. By re-expressing the equation with a single environmental variable $x \in \mathrm{Im}(x)$, we obtain the equation 4. For large samples, fitting the model is equivalent to minimizing the right term of equation 4, where the terms $\mathrm{Err}^{W_j}_{s,\lambda^i}(s, s_\gamma)$ and $\mathrm{Err}^{W_j}_{s,\lambda^i}(\lambda^i, \lambda^i_{\beta_i})$ can be seen as logarithmic density errors over the range of environment W_j for the sampling effort and the species i density, respectively. Those errors are spatially weighted by the occurrence density of species i, $s\ \lambda^i \circ x$, and its number of occurrences n_i. If sampling effort s is badly approximated by the sampling mesh, i.e. by the BCCA, and if s shows a strong and monotonic co-variation with x within cells, then $\mathrm{Err}^{W}_{s,\lambda^i}(s, s_\gamma)$ can show monotonic variation along the environmental gradient. The effect can be counterbalanced by an opposite variation profile in the error terms of the species densities, which can be achieved by adjusting their parameters to minimize the overall error. Such lack of robustness of the sampling mesh to environmentally structured variations within cells is a consequence of the latent lack of identifiability of the model. In contrast, if the sampling effort variation within cells is independent from that of the environmental variables, no bias is caused, whatever the strength of sampling effort variation. This problem is related to the problem of spatial confounding in spatial statistics Hodges and Reich [2010], or to interlinked biases between covariates and purely spatial effects in generalized linear mixed models.

$$\begin{array}{rcl}
\{\hat{\gamma}, \hat{\beta}_1, ..., \hat{\beta_N}\} & = & \underset{\gamma,\beta_1,\ldots,\beta_N}{\mathrm{argmin}} \sum_{j=1}^{B} \sum_{i=1}^{N} n_i \left(\mathrm{Err}^{W_j}_{s,\lambda^i_{\beta^*_i}}[s, s_\gamma] + \mathrm{Err}^{W_j}_{s,\lambda^i_{\beta^*_i}}[\lambda^i_{\beta^*_i}, \lambda^i_{\beta_i}] \right) \mu(x^{-1}(W_j)) \\
\text{Where} & & (W_j)_{j\in[|1,B|]} \text{ is a partition of } \mathrm{Im}(x) \text{ into small intervals} \\
\text{and} & & \forall f, g \in \mathbb{R}+^{D} \text{ densities over } D \\
\mathrm{Err}^{W}_{s,\lambda}[f, g] & := & \frac{\int_{x^{-1}(W)} s(z)\lambda\circ x(z)(\log(f)-\log(g))dz}{\mu(x^{-1}(W))}
\end{array} \quad (4)$$

Note that in equation 4, we consider that all densities integrate to 1 over D.

3 Appendix C: Estimation variance analysis

Our model is in the canonical exponential family, and thus the vector or parameter estimators $\hat{\theta} := (\hat{\gamma}, \hat{\alpha_1}, \hat{\beta_1}, ..., \hat{\alpha_N}, \hat{\beta_N})$ asymptotically follow a multivariate Gaussian distribution (see Bickel and Doksum [2015], section 5.3.3, p.322-323). In this case of one realization from a Poisson process, the variance-covariance matrix is simply the inverse of the Fisher information matrix, introduced in equation 3 of Appendix A.

$\Sigma(\hat{\theta}) \ = \ I(\theta)^{-1}$.

Effect of occurrence rate. We used this formula and equation 3 in the R script `Variance_Script.R` (downloadable from the article Github repository: `https://github.com/ChrisBotella/SamplingEffort`) to efficiently compute the model parameters variance-covariance matrix for a given scenario: a spatial domain D, sampling effort s, species number N and intensity $\lambda_1, ...\lambda_N$ (defined from their density and expected occurrence $n_1, ..., n_N$) and the model sampling cells. We computed the variance for profile 2 of the complementary simulation setting (see **Appendix F**). We set the number of occurrences for species 1 to 100 while varying the number of occurrences for the other species, conversely. Figure 1 shows, in the upper panel (resp. lower panel), how species 1 (resp. 2) parameter variance decreases when increasing the number of occurrences of a species 1 (resp. 2) through the curve in blue (resp. curve in red). The upper panel (resp. lower panel) also shows, through the curve in red (resp. in blue), that the variance of the focal species 1 (resp. 2) parameter decreases when increasing the occurrence rate of the other species 2 (resp. 1) while the occurrence rate of the focal species is kept constant. Indeed, increasing the occurrences of any species enables the model to better estimate the sampling effort, which makes the estimation of every other species parameter easier . In equation 2, we see that the information gained on the sampling effort in cell j is the expectation of the total number of occurrences in this cell $\mathbb{E}(n^j)$ of all species so that each species contributes proportionally to its number of occurrences in the cell to improve the estimation of γ_j. Still, as shown by Figure 2, the indirect variance reduction mechanism from one species to another is slower than increasing the occurrence rate of the focal species itself.

Effect of removing the parameter. As proposed in the **'Model design guidelines'** paragraph of section 2.1 of the article, we can drastically reduce the estimation variance in all species parameters by excluding an environmental variable from the model of one species (say species i) while keeping it in the model training data. This is a special case of conditional estimation (see next paragraph) where we condition on $\beta_i = 0$. It means that we assume a priori that species i is indifferent to variation in the environmental variable across the study domain D. In this case, the model knows that the species intensity is constant along this environmental variable (all others are kept constant) and can then use the variation in occurrence concentration along this gradient to better estimate the variation in sampling effort. We show this in the same theoretical context as in the previous paragraph, which corresponds to the sampling effort profile 2 of the simulation experiment. We now compute the asymptotic parameter variance of species 1 (β_1) given that we know the exact niche parameters of species 2 (β_2) along the environmental variable x. This variance is simply obtained by removing the columns and lines of the information matrix $I(\theta)$ (see equation 3 in Appendix A) that are associated with β_2, obtaining $I(\theta_{-\beta_2})$, and numerically inverting $I(\theta_{-\beta_2})$ to get the new estimators variance-covariance matrix $\Sigma(\hat{\theta}_{-\beta_2})$. In the upper panel of Figure 1 we represent the estimation variance on density parameters of species 1 extracted from $\Sigma(\hat{\theta}_{-\beta_2})$ with a growing occurrence rate for species 1 (purple curve) or species 2 (green curve). We can see that (i) the variance is always lower or equal compared to the cases where β_2 is estimated (green *le* red, purple *le* blue), (ii) it is lower for a small sample size (for 100 occurrences, green is well below red, and purple is well below blue), (iii) it enhances the indirect variance reduction effect by increasing the occurrence rate on another species (green is well below red for all occurrence rates). To lighten the graph, we did not add to the lower panel the effect of removing parameters β_1 on estimation of β_2, but it works in the same way.

Variance reduction with conditional estimation, the general case. The previous paragraph showed that when setting the parameters β_i of species i to 0, estimation variance is reduced on all other species parameters. We show this for a specific simulation scenario that is only a particular case of conditional estimation, i.e. estimating some parameters when the value of others is given, which can be used more broadly with our method. We show here mathematically that (i) the variance reduction is not specifically due to the chosen simulation scenario but appears in any case, and (ii) it appears whatever the parameters θ_i over which we condition. We first recall that when we have many occurrences for all species, we have the below (see Bickel and Doksum [2015], section 5.3.3, p.322-323):

$$\lim_{n_1,\ldots,n_N \to \infty} \mathcal{L}(\hat{\theta}) = \mathcal{N}(\theta, \Sigma(\theta))$$

Here we re-order the parameter estimation vector $\hat{\theta} = (\hat{\gamma}, \hat{\theta_1}, ..., \widehat{\theta_{i-1}}, \hat{\theta}_{i+1}, ..., \hat{\theta}_N, \hat{\theta}_i)$ and decompose its variance-covariance matrix as follows:

$$\Sigma(\theta) = \begin{pmatrix} \Sigma_{-\theta_i} & \Sigma_c^T \\ \Sigma_c & \Sigma_{\theta_i} \end{pmatrix}$$

We also note $\hat{\theta}_{-i} := (\hat{\gamma}, \hat{\theta_1}, ..., \widehat{\theta_{i-1}}, \hat{\theta}_{i+1}, ..., \hat{\theta}_N)$. The Gaussian conditioning theorem states that the conditional law $\hat{\theta}_{-i}|\hat{\theta}_i$ is a multivariate Gaussian distribution with variance-covariance matrix $\Sigma(\theta_{-i}) = \Sigma_{-\theta_i} - \Sigma_c^T \Sigma_{\theta_i}^{-1} \Sigma_c$. The individual variances of all parameters are the diagonal elements of the latter matrix. We can now easily show that they are all smaller than the original variances, i.e. the diagonal elements of $\Sigma_{-\theta_i}$, because the diagonal elements in the matrix $\Sigma_c^T \Sigma_{\theta_i}^{-1} \Sigma_c$ are all strictly positive. Indeed, $\Sigma_{\theta_i}^{-1}$ is positive definite as the inverse of Σ_{θ_i}, which is positive definite as a variance-covariance matrix. Then, the jth diagonal element of $\Sigma_c^T \Sigma_{\theta_i}^{-1} \Sigma_c$ is of the form $a_j^T \Sigma_{\theta_i}^{-1} a_j > 0$ (where a_j is jth column of Σ_c) by definition of positive definite matrices. In summary, the variance reduction of the estimator conditionally to the parameters of species i is strict whatever the value of θ_i.

Effect of the number of sampling cells. With the same setting, we evaluate the effect of the number of modeled sampling cells, evenly spaced along the longitude of the square domain. In Figure 2, we plot the asymptotic estimation variance on species parameters, computed numerically through the inversion of the information matrix, as a function of the number of cells. All estimator variance increases with the number of cells, but not at an equal speed for all types of parameters. More precisely, we see that the variances on $\beta_{1,1}$ and $\beta_{2,1}$, which both control the optimum of the species Gaussian density along the environmental gradient x, explode very quickly, whereas the parameters controlling the niche breadth remain reasonable even for 20 cells. Above 20 cells, the model shows a weak numerical identifiability, checked through the high condition number of the information matrix. When including too many cells, we decrease the ability of the model to separate the effect of the environmental variable, which varies less within each cell, from the cell effect. However, the identifiability may not concern all parameters simultaneously: the species niche breadth parameters do not seem very sensitive to the increased number of cells. However, the sampling effort approximation error increases as we decrease the number of cells, and this effect is not taken into account in the estimation variance. Thus, determining the best size of cells should be based on cross-validation using a density evaluation metric

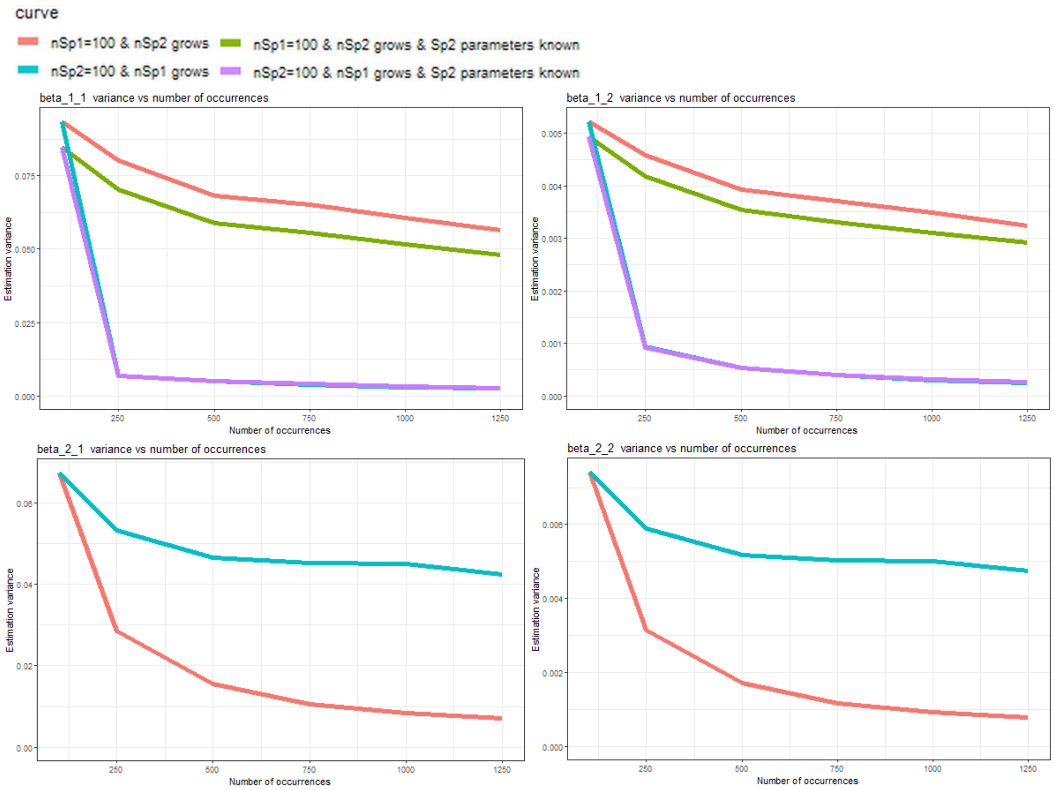

Figure 1: Asymptotic species density parameters estimation variance as a function of the number of each species occurrence for the simulation setting of profile 2 described in section 2.4 of the article. $\beta_{1,1}$ and $\beta_{1,2}$ (resp. $\beta_{2,1}$ and $\beta_{2,2}$) are respectively the first and second parameters modeling the Gaussian density of species 1 (resp. species 2) along the environmental gradient x.

(Tsybakov [2009]). For a K-fold cross-validation, we recommend building the folds so that each one contains a proportion of approximately $1/K$ of the occurrences of every individual cell, as no sampling cell should be empty or scarce for training.

4 Appendix D: Inference and implementation details

For a given mesh across which a cell-wise constant sampling effort is defined, we fit log-linear Poisson processes for multiple species with a shared term in their linear predictor, i.e. the log-sampling effort. We here present a maximum-likelihood fitting procedure. We use an approximation of the Poisson process likelihood by a Poisson regression likelihood using background points, as described in Berman and Turner [1992] and Warton et al. [2010], which we extend to the joint likelihood of a marked Poisson process.

We consider the set of observed occurrences for any species $i \in [1, N]$ $Z_i = \{(z_1^i, i, 1), ..., (z_{n_i}^i, i, 1)\}$, i.e. a set of points marked with the species label i and the state 1. We have to maximize the joint likelihood of $Z_1, ..., Z_N$

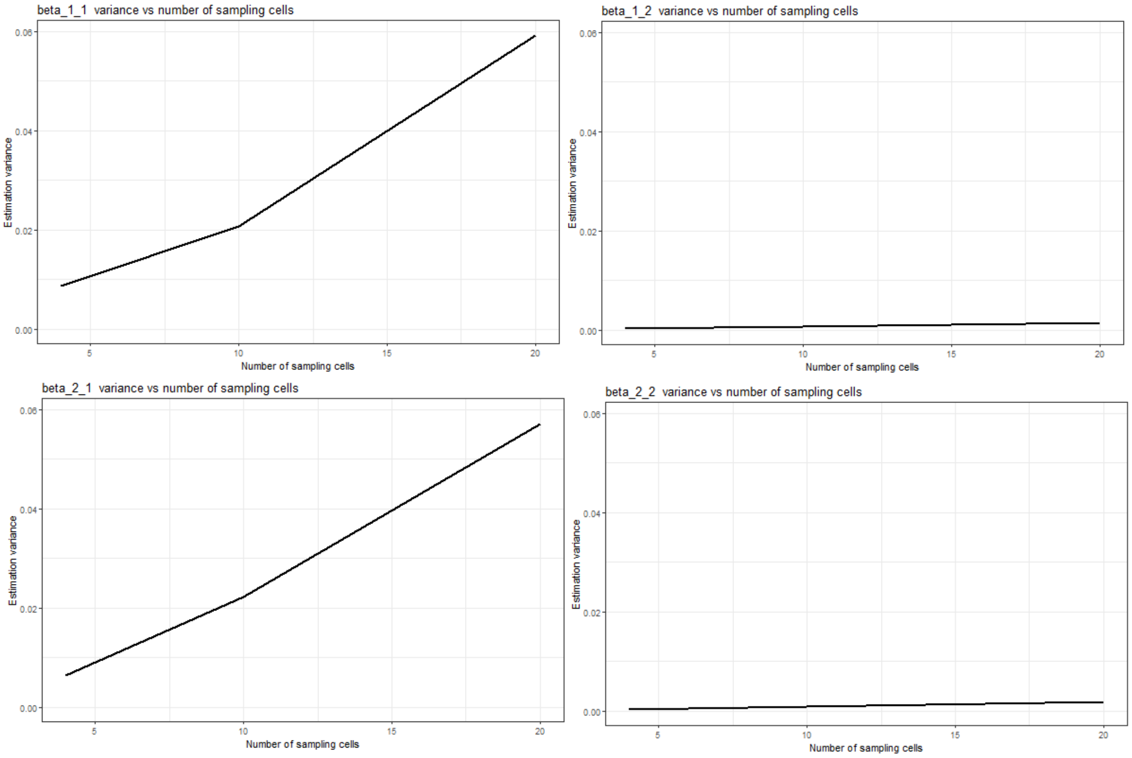

Figure 2: Asymptotic species density parameters estimation variance as a function of the number of modeled sampling cells (regularly spaced along the longitude of the domain) in the simulation setting of profile 2 described in section 2.4 of the article. $\beta_{1,1}$ and $\beta_{1,2}$ (resp. $\beta_{2,1}$ and $\beta_{2,2}$) are respectively the first and second parameters modeling the Gaussian density of species 1 (resp. species 2) along the environmental gradient x. Above 20 cells, we began to diagnose weak numerical identifiability (through the condition number of $I(\theta)$) of the model, making the variance-covariance matrix unreliable.

with respect to all model parameters introduced in the previous section $\theta := (\alpha_1, ..., \alpha_N, \beta^1, ..., \beta^N, \gamma_1, ..., \gamma_C)$:

$$\begin{array}{lrcl} & p(Z_1, ..., Z_N|\theta) & = & \prod_{i=1}^{N} \left[\frac{(\int_D s(z)\lambda_i(z)dz)^{n_i}}{!n_i} \exp\left(- \int_D s(z)\lambda_i(z)dz\right) \prod_{k=1}^{n_i} \frac{s(z_k^i)\lambda_i(z_k^i)}{\int_D s(z)\lambda_i(z)dz} \right] \\ \Leftrightarrow & p(Z_1, ..., Z_N|\theta) & \propto & \prod_{i=1}^{N} \left[\exp\left(- \int_D s(z)\lambda_i(z)dz\right) \prod_{k=1}^{n_i} s(z_k^i)\lambda_i(z_k^i) \right] \\ \Leftrightarrow & log(p(Z_1, ..., Z_N|\theta)) & = & \sum_{i=1}^{N} \left[\sum_{k=1}^{n_i} log(s(z_k^i)\lambda_i(z_k^i)) - \int_D s(z)\lambda_i(z)dz \right] \end{array} \tag{5}$$

The likelihood is factorized over species as we assume that their processes are independent given the environment.

The integral terms are often very costly to compute exactly when dealing with multiple high resolution rasters of environmental variables. Instead we use a numerical approximation. Each integral is replaced by a weighted sum of $s\lambda_i$ computed at some quadrature points $Z_i^q = \{(z_1^q, i, 0), ..., (z_Q^q, i, 0)\}$ marked with their species label i and state 0 indicating it is a background point, associated with weights $w_1^i, ..., w_Q^i$, selected such that $\int_D s(z)\lambda_i(z)dz \approx \sum_{k=1}^{Q} w_k s(z_k^q)\lambda_i(z_k^q)$. Background points are also called quadrature points, or pseudo-absences in the Poisson process SDM literature (Warton et al. [2010]).

Numerical quadrature strategy and background points. We chose to draw uniformly background points to achieve the approximation of the integral through the unbiased Monte Carlo estimator. More precisely, Berman and Turner [1992] re-expressed the likelihood by including the points of Z_i among the quadrature points Z^q, and by defining adapted weights. We note $w(z, i, e)$ the weight associated with the marked point (z, i, e).

$$\begin{array}{rcl} \log(p(Z_1, ..., Z_N|\theta)) & \approx & \sum_{i=1}^{N} \sum_{(z,i,e)\in Z_i \cup Z_i^q} 1_{e=1} \log(s(z)\lambda_i(z)) - w(z,i,e)s(z)\lambda_i(z) \\ & = & \sum_{(z,k,e)\in \cup_i (Z_i \cup Z_i^q)} w(z,k,e) \left[\, y(z,k,e) \log(s(z)\lambda_i(z)) - s(z)\lambda_i(z) \,\right] \end{array} \tag{6}$$

Where the $y(z, k, e) := 1_{e=1}/w(z, k, e)$ are the Poisson regression pseudo-counts (non-integers), and we recall that by design in our model $s(z)\lambda_i(z) = \exp(\sum_{j=1}^{C} \gamma_j 1_{z\in c_j} + \alpha_i + \beta^{iT} x_i(z))$. We end up with a Poisson regression log-likelihood that satisfactorily approximates our initial log-likelihood when there are enough properly selected quadrature points. We use the same quadrature points and associated weights for all species. Now, we need to explain how those points are selected and their weights computed $w(z, i, e)$. The Monte Carlo method is an unbiased way to approximate the integral: we use the average of $s\lambda_i$ over uniformly sampled background points on D to approximate the integral $\int_D s(z)\lambda_i(z)dz$. However, occurrences in Z_i's are not uniformly distributed over D, and we need to ensure that they will not bias our approximation. For this purpose, the sum of weights of occurrences is negligible compared to the sum of weights of quadrature points and the total sum:

$$\forall(z,i,e) \in \cup_i(Z_i \cup Z_i^q) w(z,i,e) = \begin{cases} \frac{|D|}{100n_i} & \text{if } e = 1 \\ \frac{99|D|}{100Q} & \text{if } e = 0 \end{cases}$$

This yields the following expression for the approximation of integral term $\int_D s(z)\lambda_i(z)dz$:

$$\begin{aligned} \int_D s(z)\lambda_i(z)dz &\approx \sum_{z\in Z_i \cup Z_i^q} w(z)s(z)\lambda_i(z) \\ &= \frac{1}{100}\sum_{z\in Z_i} \frac{|D|}{n_i} s(z)\lambda_i(z) + \frac{99}{100}\sum_{z\in Z_i^q} \frac{|D|}{Q} s(z)\lambda_i(z) \end{aligned}$$

With this setting, all weights sum to $|D|$ (area of D), while weights of species occurrences alone represent only 1%, which we note is enough not to bias the approximation in our experience.

Application to the real dataset. For the real dataset of occurrences, we used an alternative strategy to ensure that all the sampling cells had background points and that they captured the environmental variability of each cell. We uniformly drew a fixed number (6) of background points uniformly in each sampling cell. As each sampling cell had the same size in this case, we could keep the same weighting scheme as previously, and the procedure weighted sum also converged to the target integral. We can show this by decomposing the integral into a sum of integrals over each sampling cell multiplied by the inverse of the total number of cells and then using the Monte Carlo (because points are uniformly drawn inside cells).

Implementation details. The inference was performed using software for generalized linear models penalized with L1 (with R package `glmnet`) to estimate parameter values that maximize the penalized version of the likelihood, for given y_j, $Z_1, ..., Z_N$ and w.

The `R` code used for fitting the model can be found on the following Github repository: `https://github.com/ChrisBotella/SamplingEffort`. Equation 7 gives the R formula for building the model design matrix passed to `glmnet`.

$$\begin{aligned} \mathtt{y} \sim\ & 1 + \mathtt{SamplingCell} + \mathtt{species1} : (x_1^1 + ... + x_{p_1}^1) + \mathtt{species2} : (1 + x_1^2 + ... + x_{p_2}^2) \\ & ... + \mathtt{speciesN} : (1 + x_1^N + ... + x_{p_N}^N) \end{aligned} \tag{7}$$

The categorical effect of a point `SamplingCell` is the effect of its cell. There are $C-1$ parameters for the sampling effort because it is impossible to identify the global intercept and the parameters of all sampling cells. Thus, we needed to choose a way to constrain the effects of the C cells with $C-1$ parameters, or in other words, to define contrasts. We chose the `SamplingCell` contrasts as `contr.sum`, $\sum_{j=1}^{C} \gamma_j = 0$. This way the L1 penalty induces a shrinkage of all sampling cell parameters toward zero, rather than a shrinkage toward a reference cell as the `contr.treat` contrasts would have done. Concerning the species niche parameters, there are $p_i + 1$ parameters for species i and different species may depend on different environmental predictors. Note that the intercept of species 1 is grouped with the global intercept, again for identifiability reasons. This explains why we can only estimate the species intensity and the sampling effort up to a constant factor. Using `glmnet` allows handling sparse matrices and performing our model with a large number of sampling cells, environmental features, background points, and occurrences, as explained in the real data illustration section.

5 Appendix E: Environmental variables tables

Name	Description	Values	Resolution (m)
CHBIO_1	Annual mean temperature	[-10.6,18.4]	1000
CHBIO_5	Max temperature of warmest month	[36.4,6.2]	1000
CHBIO_12	Annual precipitation	[318,2543]	1000
etp	Potential evapotranspiration	[133,1176]	1000
alti	Elevation	[-188,4672]	90
slope	Absolute elevation gradient	[0,13457]	90
awc_top	Topsoil available water capacity	$\{0, 120, 165, 210\}$	1000
bs_top	Base saturation of the topsoil	$\{35, 62, 85\}$	1000
spht	Aggregated land cover	{culti.,for.,past.,urb.,other}	100

Table 1: Table of environmental variables used in this study.

CLC category description	spht category name	Raster code
Non-irrigated arable land	cultivated	12
Permanently irrigated land	cultivated	13
Vineyards	cultivated	15
Fruit trees and berry plantations	cultivated	16
Complex cultivation patterns	cultivated	20
Land principally occupied by agriculture, with significant areas of natural vegetation	cultivated	21
Agro-forestry areas	cultivated	22
Pastures	grasslands	18
Natural grasslands	grasslands	26
Moors and heathland	grasslands	27
Sclerophyllous vegetation	grasslands	28
Broad-leaved forest	forest	23
Coniferous forest	forest	24
Mixed forest	forest	25
Transitional woodland-shrub	forest	29
Continuous urban fabric	urban	1
Discontinuous urban fabric	urban	2
Industrial or commercial units	urban	3
Road and rail networks and associated land	urban	4
Airports	urban	6
Green urban areas	urban	10
Sport and leisure facilities	urban	11
Port areas	other	5
Mineral extraction sites	other	7
Dump sites	other	8
Construction sites	other	9
Rice fields	other	14
Olive groves	other	17
Annual crops associated with permanent crops	other	19
Beaches, dunes, sands	other	30
Bare rocks	other	31
Sparsely vegetated areas	other	32
Burned areas	other	33
Glaciers and perpetual snow	other	34
Inland marshes	other	35
Peat bogs	other	36
Salt marshes	other	37
Salines	other	38
Intertidal flats	other	39
Water courses	other	40
Water bodies	other	41
Coastal lagoons	other	42
Estuaries	other	43
Sea and ocean	other	44
No data	other	48
Unclassified land surface	other	49
Unclassified water bodies	other	50

Table 2: spht (Aggregated land cover) categories correspondence with Corine Land Cover 2012.

6 Appendix F: Complementary simulation study, a closer look at the density estimates

6.1 Methodology

We designed the following simulation study to examine more closely whether our approach allows a reliable inference of sampling effort density and species density from observed occurrences of two virtual species with heterogeneous sampling effort. Note that we did not use intercepts in the simulation because, as explained in section **2.1**, we cannot estimate absolute intensity across space but only relative intensity. We evaluated the estimation quality as the ability to recover the density over the environmental gradient, because it is the space over which both the species intensity and the sampling effort are defined by our design. This space is one-dimensional to enable visualization. To reproduce this experiment, one must run the script called `Simu_and_graphs.R` on the article Github repository: `https://github.com/ChrisBotella/SamplingEffort`.

Spatial domain and species variable. We considered a square spatial domain $D = [0, 10]^2$ where the only environmental variable x was a linear gradient from west to east, such that $x(z) = z - 5$.

Virtual species. The environmental intensity of virtual species was modeled as a Gaussian function over the gradient x, i.e. $\forall z \in D,\ \lambda_i(z) \propto \exp((x(z) - \mu_i)^2/(2\sigma_i^2))$. This means that the expected x of a given species individual is μ_i (optimum constraint), and the variance of x over many individuals is σ_i^2 (niche breadth constraint), and λ_i is maximum entropy. We used the following re-parameterization of species density:

$$\begin{aligned} \forall z \in D,\ \lambda_i(z) &\propto \exp\left(-\frac{(x(z)-\mu_i)^2}{2\sigma_i^2}\right) \\ &\propto \exp\left(\beta_1^i x(z) + \beta_2^i x(z)^2\right) \end{aligned}$$

With $\left\{\begin{array}{lcl} \beta_1^i & = & \frac{\mu_i}{\sigma_i^2} \\ \beta_2^i & = & -\frac{1}{2\sigma_i^2} \end{array}\right. \Leftrightarrow \left\{\begin{array}{lcl} \mu_i & = & -\frac{\beta_1^i}{2\beta_2^i} \\ \sigma_i & = & \frac{1}{\sqrt{-2\beta_2^i}} \end{array}\right.$

β_2^i being strictly negative. This re-expression will be useful as the method implementation gives us estimates of β_1^i, β_2^i for each i (see Inference section). In our simulation study we had two virtual species $i \in \{1, 2\}$ and we chose the optima to be $\mu_1 = -2.5$, $\mu_2 = 2.5$. The standard deviation of their intensities are $\sigma_1 = \sigma_2 = 1.6$.

Types of sampling effort. We designed a case where the relative sampling effort strongly depended on the environment x, which made it harder to separate sampling effort from species intensity. The relative sampling effort is a step function over D depending on the longitude only (like the feature x), and not the latitude. We designed three profiles for relative sampling effort:

1. $s(z) = 1_{x(z)<0}$. This profile has a constant non-null effort on the western half of the domain, and no sampling on the eastern half.
2. $s(z) = 1 + 5\ 1_{x(z)\in[-4.5,-2.5[\cup[-0.5,1.5[\cup[2.5,4.5[}$. This profile has sharp variation within the sampling cells of the model design.
3. $s(z) = 9 * \dfrac{\exp(-5x(z))}{1 + exp(-5x(z))} + 1$. This profile is a decreasing sigmoïdal function. It has also sharp variations within sampling cells, plus they are continuous and monotonic across the domain.

The fitted sampling model was well specified for type (1). Indeed, the point of discontinuity of the simulated sampling effort was the boundary between the sampling cells. Thus, we expected to get exact estimates of species niches and sampling effort density. In our test case, the method recovered the species niches with only a partial sampling of the environmental range. However, for type (2), the simulated sampling effort varied in the middle of some modeled sampling cells, making it impossible to get a perfect estimation. If the method is robust, we would expect the sampling effort estimate to approximate the average of the target in every sampling cell. The estimation was not perfect for type (3) either. Here, the sampling effort co-varies strongly and monotonically with the environmental variable, so it is expected to be the most problematic profile for use with this method.

Simulating species observed points. We drew $200,000$ occurrences for both species in each of the 3 sampling effort scenarios. For a defined relative sampling effort s and species intensity λ, we drew points according to a conditional Poisson process of intensity function $s\lambda$ over D. This was done using the following acceptance-rejection algorithm:

- Initialization: Determine an upper bound B of $s\lambda$ on D.
- Repeat:

1. Draw a point $z \sim U(D)$.

2. Draw a variable $y \sim U([0, B])$

3. We accept z if $y <= s(z)\lambda(z)$.

4. If $200,000$ points are accepted, finish the procedure, otherwise go back to 1).

We chose $200,000$ points as this is enough for a satisfying convergence of the sampling effort and species intensity estimates, as shown by the standard deviation bounding curves of Fig. 3.

Background points. For each experiment, $50,000$ background points were uniformly drawn over D, which is enough for likelihood convergence in this simple setting.

6.2 Results

We analyze here the reliability of our joint estimation method for two simulated species with three scenarios of sampling effort. Fig. 3 shows the mean and standard deviations of estimated relative sampling effort.

Unbiased niches and sampling effort estimates under good model specifications. Our simulation results first show that estimation of the relative sampling effort and of relative species intensity are unbiased under the observation scenario (1), i.e. when the species and sampling model is well designed. In scenario (1), there was no sampling in the eastern part of the domain, and constant sampling in the western part. The left graph of box A on Fig. 3 shows that the model perfectly captures the non-sampled area, and the estimate for the western part is almost exact. Center and right graphs of box A show that species intensity is also well recovered. The model uses the variation in species points occurrences in the western part to fit the whole species intensity model and is then able to make a good prediction on the eastern part. Blue curves in Fig. 3 represent the observed standard deviation, which approximately indicate the 95% confidence interval (mean +/- 2 times the standard deviation) of the estimate over the 20 repetitions of the simulation. We note a small bias likely due to numerical approximation in the fitting algorithm. It is not due to the regularization path, as we had a bias of similar order with the implementation `glmn`.

Approximation bias under bad sampling model design. Secondly, the graphs of box B illustrate the results of scenario (2). It shows that even though the sampling effort model neglects actual variation within sampling cells, the method provides a reasonably good approximation, as the estimate is often close to the

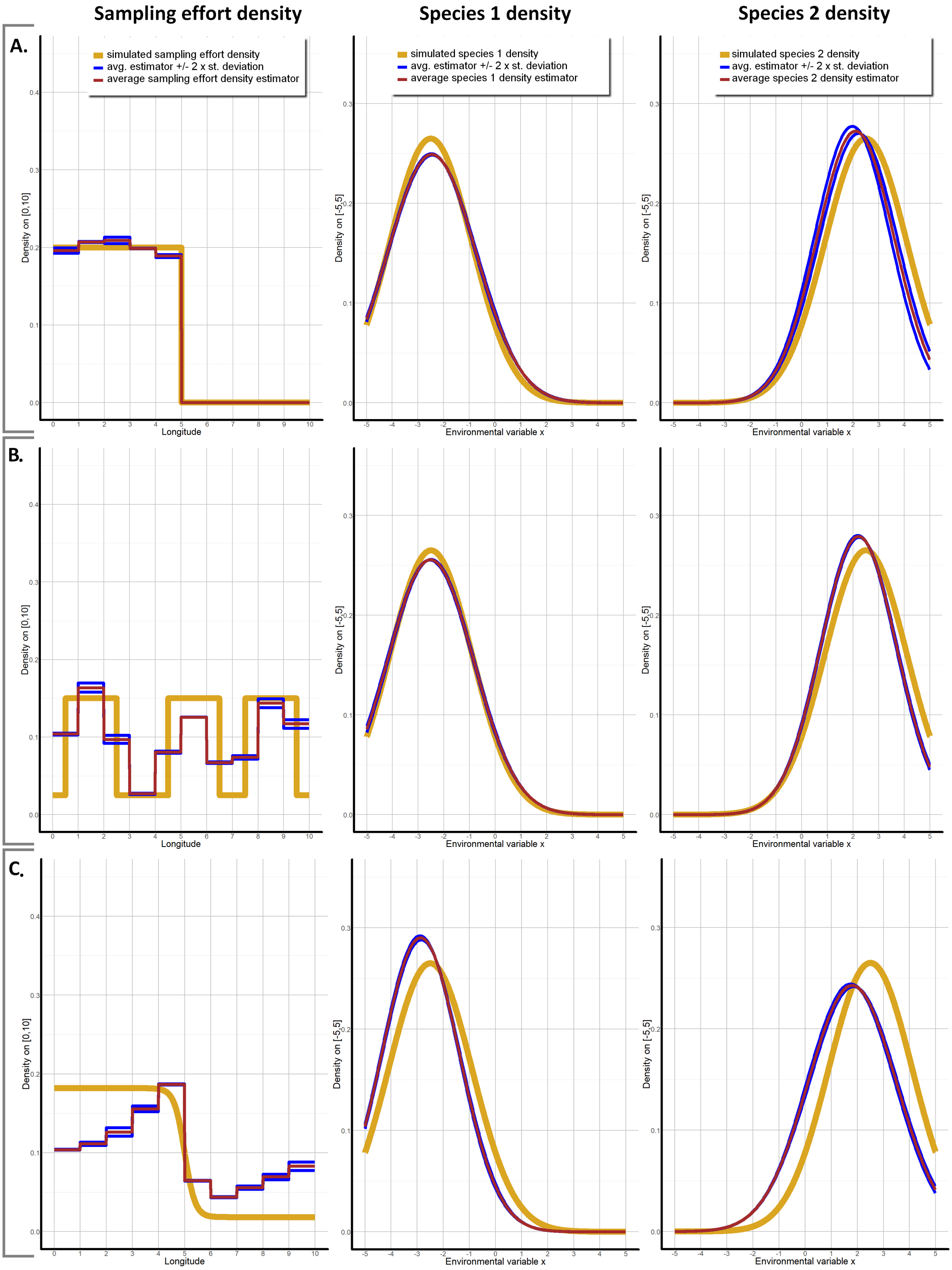

Figure 3: Sampling effort and the two species estimated densities for the three profiles of simulated sampling effort in the simulation experiment. A. type (1); B. type (2); C. type (3); see the paragraph 'Types of sampling effort'. Red curves are the mean estimates over 20 repetitions of the simulation scenario, with the blue curves indicating the approximate 95% confidence interval. Yellow curves are the targets. Sampling density (graphs on the left) is plotted against longitude, while species density (graphs in the center and right) is plotted against x values (which are in bijection). The vertical gray lines on the graphs represent the longitudinal limits of sampling effort square cells.

average of the true sampling effort in each cell. The species intensity estimates, in the center and right graphs of box B, are slightly more biased than in case (1). For scenario (3), illustrated by the densities of box C, we see bias in both the estimation of species density and the sampling effort. The species density deviates on the left, associated with an underestimation of the sampling effort for low x values and an overestimation for high x values.

7 Appendix G: Assumptions on detection probability and data selection

Several assumptions regarding detection probability in the proposed model may deviate from reality.

1. **Detection probability varies similarly across space for all species.** Sampling effort was assumed to be identical across species. While our model can allow detection probability to vary across species (R_is), this is not distinguishable from overall species abundance. We thus assumed detection probability density to vary similarly across space for all species, which is not specific to our method (see Fithian et al. [2015]). Bias can appear if species detection probability varies differently in space from one species to another. For instance, some species might be looked for only in specific areas and such sampling peculiarity can induce bias in the estimation of species density.

2. **Homogeneous detection and identification skills across observers.** We also made the assumption that for each modeled species, the detection and identification probability was identical across observers. This may be problematic in citizen science programs, in which identification skills are heterogeneous. Thus, it is preferable to include only species that are well identified by most observers. In Pl@ntNet data, this is possible thanks to the automatic identification system.

3. **No saturation of interest.** Lastly, we assumed the expected number of occurrences to be proportional to the local intensity (expected abundance) of the species and the sampling effort, which means that there was no saturation of interest. If for instance, observers report a maximum of only one individual from the local population, there is saturation of reporting interest, and this may impact the estimation of our model. Saturation of interest in observers' reports is not always problematic. If the number of observers is high (everywhere) and their probability of detection of specimens is generally low, then estimates provided by our model should not change drastically. However, if the number of observers

is low everywhere and their probability of detection is high, then we could expect that our model's estimation of the environmental density will be shrunken toward the uniform density. This assumption seems consistent with the citizen science context, but otherwise, occurrence thinning strategies may be useful to avoid bias (Boria et al. [2014], Fourcade et al. [2014], Varela et al. [2013]).

References

Berman, M. and Turner, T. R. (1992). Approximating point process likelihoods with glim. *Applied Statistics*, pages 31–38.

Bickel, P. J. and Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press.

Boria, R. A., Olson, L. E., Goodman, S. M., and Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275:73–77.

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438.

Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (2014). Mapping species distributions with maxent using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PloS one*, 9(5):e97122.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334.

Jacquez, J. A. and Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, 77(1-2):201–227.

Rathbun, S. L. and Cressie, N. (1994). Asymptotic properties of estimators for the parameters of spatial inhomogeneous poisson point processes. *Advances in Applied Probability*, 26(1):122–154.

Tsybakov, A. (2009). Introduction to nonparametric estimation. In *Springer Series in Statistics, ISBN 978-0-387-79051-0*. Springer-Verlag New York.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2013). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37:1084–1091.

Warton, D. I., Shepherd, L. C., et al. (2010). Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. *The Annals of Applied Statistics*, 4(3):1383–1402.