Received: 18 May 2019 | Accepted: 19 January 2021

DOI: 10.1111/2041-210X.13565

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data

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

[1]INRIA Sophia-Antipolis - ZENITH team, LIRMM - UMR, Montpellier, France; [2]INRAE, UMR AMAP, Montpellier, France; [3]AMAP, University of Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France; [4]INRAE, BioSP, Site Agroparc, Avignon, France; [5]CIRAD, UMR AMAP, Montpellier, France and [6]Laboratoire Interdisciplinaire de Physique, Université Grenoble Alpes, Saint-Martin-d'Hères, France

Correspondence
Christophe Botella
Email: christophe.botella@gmail.com

Handling Editor: Robert B. O'Hara

Abstract

1. Building reliable species distribution models (SDMs) from presence-only information requires a good understanding of the spatial variation in the sampling effort. However, in most cases, the sampling effort is unknown, leading to biases in SDMs. This study proposes a method to jointly estimate the parameters of sampling effort and species densities to avoid such biases. The method is particularly suited to the analysis of massive but highly heterogeneous presence-only data.
2. The proposed method is based on estimating the variation in sampling effort over units of a spatial mesh in parallel with the environmental density of multiple species using a marked Poisson process model. Based on simulations with realistic settings, we examined the performance and robustness of parameter estimations. We also analysed a large-scale citizen science dataset with highly heterogeneous sampling (Pl@ntNet), including around 300,000 occurrences of 150 plant species.
3. We found that sampling effort was correctly estimated when the true sampling effort was constant within the cells of a spatial mesh. Estimation bias arose when sampling effort and environmental drivers strongly covaried within cells. Otherwise, the inference was correct and robust to sampling variation within cells. Running the model on real occurrences of 150 plant species provided an estimated map of relative sampling effort for 15% of French territory. We also found that the density estimated for an exotic invasive plant was consistent with prior data.
4. This is the first method jointly estimating species densities depending on environment, and sampling effort as an explicit spatial function, from occurrence data of multiple species. An asset of the method is that a few frequently observed species greatly contribute to correctly estimate sampling effort, thereby improving density estimation of all other species. This approach can thus provide reliable SDM for

Alexis Joly and Pierre Bonnet contributed equally to the paper.

François Munoz and Pascal Monestiez contributed equally to the paper.

large opportunistic presence-only datasets, with broad spatial variation in sampling effort but also many species, such as datasets from citizen science programmes.

KEYWORDS
citizen science, marked Poisson point process, multi-species data, presence-only data, sampling effort, species distribution model, unbiased estimation

1 | INTRODUCTION

Understanding biodiversity dynamics and designing conservation strategies require characterizing and analysing the distribution of species in space and time. Today, international citizen science projects and naturalist networks provide massive geolocated occurrence data for multiple species around the world. Yet, the observed distribution of species occurrences depends not only on the actual species abundance but also on the sampling effort of observers. To correctly estimate environmental effects in species distribution models (SDMs, Elith & Leathwick, 2009), it is crucial to design a statistical approach that can separate these two intertwined signals in the data.

Until recently, digitized geolocated presence of species, or species occurrences, were extracted from expert collections, mainly naturalist field surveys and natural history museums (Soberón & Peterson, 2004). Today, species occurrence data have become widely available from worldwide citizen science programmes or naturalist community platforms (e.g. iNaturalist, e-Bird, Pl@ntNet, Naturgucker; see Chandler et al., 2017), in part thanks to new digital tools and smartphone applications (Teacher et al., 2013). For example, eBird has collected around 500 million valid geolocated occurrences of bird species worldwide, which are accessible online on the GBIF website.[1] Moreover, automatic identification of images or sound (Joly et al., 2018) and the collaborative review of observations have enhanced the quality of species identification by non-professional observers. However, contributors do not follow a planned sampling protocol and submit observations of specimens that are remarkable, atypical or new to them. Such 'opportunistic' sampling (Kery et al., 2010) reflects the specific behaviour and reporting choices of contributors. The sampling effort is then neither spatially uniform nor balanced between species. The objective of the present study is to propose a joint estimation of spatial sampling effort and species ecological niches, to alleviate biases in SDMs due to heterogeneous sampling.

Sampling effort, or 'observation effort' (Calenge et al., 2015), is defined as an intensity function measuring the number of visits during which observers can report a specimen occurrence at a given point. Here, we assume that sampling effort does not depend on species detectability or reporting interest (Fithian et al., 2015; Giraud et al., 2016) so that sampling effort represents a common function influencing in the same way the observation of multiple species.

Estimating spatial variation in sampling effort in a set of species occurrences is crucial for several purposes. An unknown spatial variation in sampling effort can be correlated to an environmental factor and result in bias in SDM results (Beck et al., 2014; Boakes et al., 2010; Botella et al., 2020; Bystriakova et al., 2012; Costa et al., 2010). Therefore, it is crucial to take sampling effort into account in SDMs. Several approaches have been proposed to tackle this problem. Sampling effort can be approximated from external information about the sampling protocol when available. Calenge et al. (2015) used the number of driven kilometres reported by agents as an approximation of the relative sampling effort when reporting the occurrence of dead animals on roadsides. Solan et al. (2019) provided another approach to estimate sampling effort from multi-species occurrence based on an a priori model of sampling effort. Alternatively, when no external knowledge is available, the background points used for inference of the environmental density of a species (Warton et al., 2010) can approximate the heterogeneous sampling effort. Phillips et al. (2009) proposed the target-group background (TGB) procedure, in which sites with at least one observation of a target group of species are integrated as background points to provide a proxy of sampling effort. Bradter et al. (2018) proposed using information about the prospecting behaviour and detection skills of very active reporters to infer the true absence of a species, and then integrating it into a presence–absence modelling framework. Finally, Warton et al. (2013) proposed to jointly model sampling effort and species density, under the assumption that sampling effort depends on specific variables. Those variables reflect prior knowledge on what influences observers behaviour. Yet, the bias correction efficiency of the existing approaches is conditional on external information or specific assumptions about sampling effort.

In this study, we propose a new SDM method for multi-species presence-only data, which requires less prior knowledge about the sampling process. It models sampling effort as a common component across multiple species, and as a step function with constant values within cells of a spatial mesh. Therefore, our sampling effort model only assumption concerns its spatial scale of variation. This model is related to those spatial statistics models, where bases of spatially smooth functions, called smoothers, are often used to estimate response surfaces in a computationally efficient way when the number of samples is large (Johannesson & Cressie, 2004). The response surface typically represents unobserved spatially smooth predictors. Using realistic simulated data, we show that the method can alleviate bias in sampling effort and species niche estimates while allowing computational efficiency for large occurrence datasets. We further examined the method's robustness to the approximation of constant sampling within cells by varying the amplitude of spatial variation and the curvature of the sampling effort within cells. We also analysed a real dataset including opportunistic occurrence data of plant species stemming

[1]https://www.gbif.org/

from automatic identification and sourced from the citizen science observatory Pl@ntNet. We present the results obtained for an exotic invasive plant species in France.

2 | MATERIALS AND METHODS

2.1 | A spatial model for sampling effort

We jointly modelled the occurrence of multiple species as independent marked Poisson point processes. In this model, the density of each species occurrence process is the product of the sampling effort and of the given species density. Species density represents a spatial variation in relative abundance in function of environmental variables. Figure 1 illustrates the principle and the components of the statistical model.

2.1.1 | Species occurrence processes and density functions

Let D be a two-dimensional geographical domain where occurrences were collected for N species. We assume that individuals of any species i are distributed over D according to a Poisson process depending on an intensity function λ_i. λ_i is assumed to be a log-linear function of environmental variables defined across D. $x^i(z) = \left(x^i_1(z), \ldots, x^i_{p_i}(z)\right)$ denote the environmental features of species i at point z, where p_i is the number of features. A feature can be any function of an environmental variable. Different features can be derived from the same variable: for instance, if we include the identity and quadratic features, we model a Gaussian density response to the environmental variable. $\beta^i = \left(\beta^i_1, \ldots, \beta^i_{p_i}\right)$ denote the parameters associated with the features so that species intensity is calculated as $\lambda_i(z) = \exp\left(\alpha_i + \sum_{k=1}^{p_i} \beta^i_k x^i_k(z)\right)$. The model can only estimate species density across space, not its absolute intensity, so we assume that $\alpha_1 = 0$ by convention. In addition, the model does not estimate the overall intensity of one species relative to others, as this cannot be differentiated from the probability of detection/reporting in presence-only data. This species density model belongs to the family of species distribution models based on point processes (Chakraborty et al., 2011; Dorazio, 2014; Fithian et al., 2015; Koshkina et al., 2017; Phillips et al., 2006; Renner et al., 2015; Warton et al., 2013).

2.1.2 | Assumption on the sampling effort process

The model defines the sampling effort as a spatial function representing a cumulated number of visits of all observers at a particular point over a time period. This function is likely to vary at a high spatial resolution, but it makes sense to model it by a random function with some smooth spatial intensity. In addition, we assume that reporting probability for a given species is constant in space, time and across observers. The sampling effort at point $z \in D$, noted $s(z)$, represents the probability of observing a spatial point z. If a specimen is present at z, it is then detected and reported with probability R_i, which implies that the specimen is sampled with probability $R_i\, s(z) \in [0, 1]$. Although the probability of sampling species i varies proportionally to s across space, it can be more or less detected than other species overall. The distribution of observed species occurrence follows a thinned Poisson process, that is, a Poisson process of intensity $z \to R_i s(z)\, \lambda_i(z)$ (Chiu et al., 2013). A discussion of the assumptions about the observation process is provided in Appendix G.

2.1.3 | Spatial variation in sampling effort

We model s as a cell-wise constant function in a spatial mesh defined over D. This assumption makes sense if the sampling effort is known to vary reasonably slowly across space, at the scale of mesh cells. In subsequent analyses, we chose a mesh with rectangular cells for simplicity, but any other form of partition of D could be considered. The sampling effort is a factor in the intensity function as shown in Equation 2 of Figure 1. We set, at any point $z \in D$, $s(z) = \exp\left(\sum_{j \in [|1,C|]} \gamma_j \mathbf{1}_{z \in c_j}\right)$, where $(c_j)_{j \in [|1.C|]}$ are the cells of the mesh verifying $\cup_{j \in [|1,C|]} c_j = D$, and $\cap_{j \in [|1,C|]} c_j = \varnothing$, and $\gamma = \left(\gamma_1, \ldots \gamma_C\right)$ are the unknown sampling effort parameters. A parameter is defined in $\mathbb{R}$ for each unit of the spatial mesh. In other words, the sampling effort model associates a parameter to each cell indicator function, which equals 1 in the cell and 0 elsewhere. We can only estimate the relative sampling effort across space, and thus we assume by convention that $\gamma_1 = 0$.

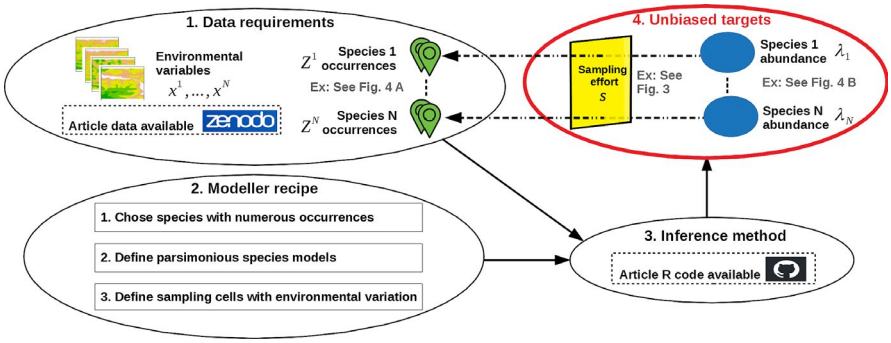
2.1.4 | Related methods

Our method can be seen as a multi-species extension of the model of Warton et al. (2013), apart from the sampling effort design. It is also a particular case of the model in Fithian et al. (2015), in which we removed the presence–absence term (equation 10 in Fithian et al., 2015) from the joint log-likelihood. Although we focus here on presence-only data, data integration (Miller et al., 2019) is probably the best strategy to correct sampling bias when more standardized data are also available (see Dorazio, 2014; Fithian et al., 2015; Giraud et al., 2016; Koshkina et al., 2017). A recent integrated data model especially used a related random spatial surface model for the sampling effort based on a log-Gaussian Cox process (Simmonds et al., 2020).

2.2 | Model identifiability and estimability

2.2.1 | Intensity is estimated up to an unknown factor

It is impossible to identify absolute values of R_i, the sampling effort s and the species density λ_i from presence-only data. We can



• **Species distribution model**

The environmental intensity of species i is

$$\begin{array}{rcccl} \lambda_i & : & \mathbb{R}^{p_i} & \to & \mathbb{R}^+ \\ & & x^i & \to & \exp(\alpha_i + \sum_{k=1}^{p_i} \beta_k^i x_k^i), \end{array} \tag{1}$$

where $x^i : D \to \mathbb{R}^{p_i}$ are the environmental variables considered for species i. The intensity represents the expected species abundance in an environment. We assume that the locations of species i specimens Z_i are distributed according to the inhomogeneous Poisson process: $Z_i \sim IPP(\lambda_i \, o x)$.

• **Sampling effort model**

The sampling effort $s : D \to [0, 1]$ represents the probability that a specific location is observed, which does not depend on the species (it does not include the detection probability or reporting interest). We approximate it by a step function over a mesh (i.e. a partition) of D: $s(z) = \exp(\sum_{j \in [|1,C|]} \gamma_j 1_{z \in c_j})$, where $(c_j)_{j \in [|1,C|]}$ are the cells of the mesh. We chose a regular mesh in the test cases for simplicity, but this is not a requirement.

• **Full model**

A specimen located at $z \in D$ is reported with probability $R_i \, s(z)$, where $R_i \in [0, 1]$ is the constant probability of detecting and reporting i. Thus, with the thinning property of the Poisson process, the joint probability distribution of the model is:

$$(Z_1, ..., Z_N) \sim \otimes_{i=1}^{N} IPP(s \; R_i \; \lambda_i). \tag{2}$$

• **Main assumptions**

1. Individual locations are independent given the environment.
2. The probability of detection and reporting of any species is constant in space, time and across observers.
3. The reporting of two individuals at distinct point locations are independent random variables.
4. The proportion of sampled individuals is small everywhere.
5. The sampling effort is constant per sampling cell.

FIGURE 1 Method workflow summary and statistical model

only estimate sampling effort and species density up to a factor that is constant across space (see Fithian & Hastie, 2013; Hastie & Fithian, 2013). For this reason, our approach allows estimating density variation across space, or relative intensity, but not absolute intensity.

2.2.2 | Disentangling sampling effort and species intensity parameters

Our method separates two spatial densities from a single distribution of points, and it is important to ensure that the parameters

of these densities are in fact estimable (Jacquez & Greif, 1985). Estimability relies on the orthogonality of the spatial covariates on which the density components depend. As shown in Appendix B, if the basis of spatial functions composed of environmental features and sampling cell indicator functions is close to multicollinearity, the true sampling effort and species density will be mixed together in the estimates. Such potential issue can be detected by assessing multicollinearity with the condition number, that is, the ratio of the highest over the lowest eigenvalues of the observed variance–covariance matrix of all model parameters. It is a common measure for this purpose (see e.g. Dorazio, 2014). If the condition number is high, the variance–covariance matrix exhibits multicollinearity and high covariance between the parameter estimates. More precisely, given that there is no collinearity between the chosen environmental features, and given that, by design, sampling cell indicators have no multicollinearity, a high condition number necessarily means that there is collinearity between environmental features and sampling cell indicators. This can be solved by increasing the size of sampling cells until the condition number (always ≥ 1) becomes reasonably small. In our experience, a condition number inferior to 10^6 was still reasonable for fitting the model. For a more detailed discussion of the issue, see Appendix B. The observed Fisher information matrix is provided in Appendix A to compute the condition number. An implementation in R language is provided in file VARIANCE _ SCRIPT.R of the R package accompanying this article (Botella, 2020).

2.3 | Parameter inference

We summarize here the procedure for inferring parameter values from multi-species occurrence data. Appendix D further includes more explicit and detailed description of the procedure. A log-linear Poisson process is fitted over multiple species with a shared term in their linear predictor, that is, the log-sampling effort. The procedure minimizes the global negative log-likelihood in Equation 3, with respect to the parameters. It is the sum of negative log-likelihoods over species Poisson processes.

$$\log(p(Z_1, \ldots, Z_N|\theta)) = \sum_{i=1}^{N}\left[\sum_{k=1}^{n_i}\log\left(s\left(z_k^i\right)\lambda_i\left(z_k^i\right)\right) - \int_D s(z)\lambda_i(z)dz\right]. \quad (3)$$

This objective function is similar to the one in Fithian et al. (2015), yet without a presence–absence term (see Equation 14). To approximate the integrals, we sum over uniformly distributed background points (Warton et al., 2010), and use the re-expression as a Poisson regression likelihood (Berman & Turner, 1992) to optimize it with standard generalized linear model software. Our implementation is based on the glmnet library, in R language. It is similar to the implementation in Renner et al. (2015), except that it is extended to a multi-species case and with a cell-wise constant sampling effort. glmnet handles sparse matrices and is very efficient in terms of memory and computational load, given the structure of the model design matrix. The R code for reproducing the results and fitting the model is provided in a publicly available Github repository (Botella, 2020).

2.4 | Simulation study

We simulated occurrence data and tested the reliability of inferences with our method. The R code to reproduce this simulation study, that is, to generate sampling effort rasters, to simulate species occurrences, to fit the model and to run the analysis over all scenarios, is provided in the article code repository (Botella, 2020).

2.4.1 | Geographical area

The French Mediterranean region was used as a reference spatial domain D for simulation of species occurrences (over the longitude/latitude extent $[1.5, 8] \times [41,45]$).

2.4.2 | Simulated species density

We simulated $n_o = (n_1, \ldots, n_{50})$ occurrences of 50 virtual species over D. The n_is were chosen from real occurrence data (i.e. the 50 most represented plant species in the Pl@ntNet queries dataset; Botella et al., 2019). This resulted in the following statistics: $\min(n_o) = 1502$, $\max(n_o) = 5002$ and $\sum_i n_i/50 \approx 2206$. All the virtual species densities (λ_i for species i in our model) were defined as Gaussian functions of the same single environmental variable (two cases considered: elevation/**alti** or annual precipitation/**chbio_12**, see Appendix E). The expectation of the Gaussian density was drawn uniformly inside the quantiles 0.1 and 0.9 of the environmental variable range of values while the standard deviation was drawn according to a gamma distribution of shape parameter 3 and scale parameter 50. The two environmental variables were selected because they are both strongly linked to the simulated sampling effort, and thus could challenge joint estimation of sampling effort and species densities. In addition, the resolution of the **alti** variable (around 90 m) was much finer than that of **chbio_12** (around 1 km), and **alti** varied more within sampling cells of our model. Therefore, estimation bias was more likely to arise with **alti** than with **chbio_12**.

2.4.3 | Sampling effort

The spatial density of sampling effort, s_H was parameterized by a bandwidth parameter $H > 0$, which controlled the level of spatial smoothness. It was a continuous approximation of the density of real occurrences, obtained by filtering the Pl@ntNet dataset (Botella et al., 2019) over D. More precisely, s_H was an

exponential quadratic kernel density estimator (KDE) function applied to the counts of those occurrences per very small square cells (resolution = 0.002 in longitude and latitude) over D. The test case included four values for the bandwidth parameter $H = \{20, 50, 80, 100\}$ in cell units, which corresponded to 3.2, 8, 12.8, 16 km in longitude, or 4.4, 11, 17.6, 22 km in latitude. The value of s_H at point $z \in D$ was a weighted average of the occurrence counts of surrounding cells. The weight of a count of a cell at Euclidean distance d was proportional to $\exp(-n^2/H)$. For instance, for $H = 20$, the weight decreased by 80% at 3.8 km in longitude. For the highest bandwidth $H = 100$, only large-scale demographic and coastline effects were visible in the simulated sampling effort. For the lowest bandwidth $H = 20$, fine-grain effects such as the influence of rivers or roads connecting cities were visible. In addition to these KDE-based sampling efforts, we also considered a simulated sampling effort assumed to be constant within the cells of the mesh. This profile, called **H=+Inf**, was used as a reference and enabled an evaluation of the performance of the method under the best model specification to characterize the error due to estimation variance. The sampling effort sharply decreased, on average, when the values of **alti** and **chbio_12** increased. This strong covariation in sampling effort and environmental variables would lead to bias in a naive SDM model (Ref?), which argues for the use of sampling bias correction. For computational details, the reader can refer to the script virtual_species_and_bias_final. R in the article repository (Botella, 2020).

2.4.4 | Simulated species occurrences

For a given species i with spatial intensity $\lambda_i \circ x$, and for a given sampling effort surface s, we independently simulated n_i occurrences according to the conditional Poisson process of intensity $s\lambda_i \circ x: D \to \mathbb{R}^+$, using an acceptation–rejection algorithm (Devroye, 1986). To do this, the maximum M of $s\lambda_i \circ x$ over D was determined. Then, it was iterated until n_i points were obtained: uniformly drawing a point $z \in D$, drawing a random variable $X \sim U([0, M])$, accepting z if $X \leq s(z)\,\lambda_i \circ x(z)$ or rejecting it otherwise. This procedure was consistent with our distribution and observation model as described in Figure 1.

2.4.5 | Model fitting

We fitted the model for the 50 species with a spatial mesh of rectangular cells with (0.1, 0.1) dimensions in (longitude, latitude), or approximately (8, 11) in kilometres. Thus, except for the case where the simulated sampling effort was constant cell-wise, the fitted model was deliberately misspecified. Indeed, the simulated sampling effort varied strongly within cells for the lowest bandwidth $H = 20$, and much more weakly for the highest $H = 100$. After defining the mesh, only cells with at least 50 occurrences were used to fit the model. We drew background points uniformly across cells as explained in Appendix D. We drew these points until there were at least 10 points per sampling cell. We fitted the model on data with different combinations of the environmental variables (elevation or precipitation) and sampling effort profiles (4 based on KDE with varying smoothness and the one constant by cell).

2.4.6 | Performance evaluation

We used two metrics to evaluate the estimation performance of the sampling effort:

1. The coefficient of determination between the simulated sampling effort and its estimation over the points of a fine regular spatial grid across D (approximately 200-m resolution).
2. The coefficient of determination between the simulated sampling effort averaged per sampling cell and its estimation over the same points. In other words, this metric computes the correlation with the best possible approximation of the true sampling effort and is necessarily superior to the first.

We also evaluated the estimation performance of species i density parameters as the coefficient of determination between λ_i and its estimate $\widehat{\lambda_i}$ across uniformly distributed values of x in the range $\left[\min\{x(z), z \in D\}, \max\{x(z), z \in D\}\right]$. We computed the metric over the environmental gradient x rather than over the geographical space D, to avoid biasing the evaluation towards the most represented environmental values.

2.5 | Application to a real dataset

We also fitted the model to real occurrences recorded in the Pl@ntNet query dataset (Botella et al. 2019). The occurrences in this dataset were collected by citizens using the Pl@ntNet mobile application (Joly et al., 2016). They were automatically identified by the Pl@ntNet AI engine. Details on the identification system and the database infrastructure of Pl@ntNet are provided in Affouard et al. (2017). The dataset is publicly available on the open-access repository Zenodo (http://doi.org/10.5281/zenodo.2634137). The code for extracting occurrence and environmental data and fitting the model is provided in the code repository accompanying this article (Botella, 2020).

2.5.1 | Species occurrences

We selected species occurrence records in France from the beginning of 2017 to October 2018. The process involved a user of the Pl@ntNet mobile application taking one or several pictures of parts of a plant specimen (e.g. leaf, flower, fruit, bark, etc.). The pictures were then sent to the Pl@ntNet server to carry out automatic

identification of the species and produce a probability distribution across species. The highest of these probabilities was then the identification confidence score. We only kept species occurrences whose confidence score (field FirstResPLv2Score) was above 0.85. We also removed all occurrences with missing values for the selected environmental variables (described below). In the last step, we kept only the **150** species with the highest number of occurrences. The list of species is provided in the table speciesTable.csv on the article code repository (Botella, 2020). The mesh used for our model was defined as a regular spatial grid of 8-km-wide squares over France, including Corsica, which we restricted to squares whose centre was inside the territory or closer than 4 km to the border or coast. Only squares with more than 30 occurrences were used to fit the model: occurrences within other squares were excluded. This resulted in a set of 302,961 occurrences, distributed over 2,869 spatial squares covering around 15% of the French territory. These squares are coloured on the map in Figure 3. To illustrate the method output for species density, we compared the fitted density of *Phytolacca americana* L., an exotic invasive plant species in France, to externally available distribution data. For this comparison, we referred to the occurrences recorded by the Federation of National Botanical Conservatories (FCBN), geographically summarized at http://siflore.fcbn.fr/?cd_ref=&r=metro, and to occurrences listed in Dumas (2011) and Pl@ntNet.

2.5.2 | Environmental data

A set of nine environmental variables was used to model the environmental density of species. These were selected carefully to model the macroecological niche of plant species, following the recommendations in Mod et al. (2016). The set included mean and annual temperature variation, annual precipitation, potential evapotranspiration, available soil water capacity and a soil pH proxy. The variables are presented in Table 1 of Appendix E. We got the environmental data from multiple sources (Karger et al., 2016; Panagos, 2006; Panagos et al., 2012; Van Liedekerke et al., 2006; Zomer et al., 2007, 2008). We extracted the values at occurrence points from the geographical rasters described and downloadable at Botella (2019).[2]

2.5.3 | Species density model

We modelled the distribution of species along continuous environmental gradients with a Gaussian density function. We combined annual rainfall chbio_12 and potential evapotranspiration etp into chbio_12-etp, known as the water balance, which is commonly used in plant SDM (Mod et al., 2016). We included pedologic variables representing categories of physicochemical properties. To summarize, Equation (4) shows the R formula of the linear predictor of any species density, with 12 feature terms computed from the environmental variables of Table 1 of Appendix E. This resulted in 13 parameters for each species density, including the intercept, plus $2,869 - 1$ observation parameters in sampling cells, resulting in $4,817$ parameters in total, for $302,961$ occurrences.

$$\begin{aligned}\sim 1 + \text{etp} + \text{I}(\text{etp}^2) + \text{I}(\text{chbio\_12} - \text{etp}) + \text{I}((\text{chbio\_12} - \text{etp})^2) \\ + \text{chbio\_1} + \text{I}(\text{chbio\_1}^2) + \text{chbio\_5} + \text{I}(\text{chbio\_5}^2) + \text{awc\_top} \\ + \text{I}(\text{awc\_top}^2) + \text{bs\_top} + \text{I}(\text{bs\_top}^2). \end{aligned} \quad (4)$$

2.5.4 | Background points

We uniformly drew a fixed number of points per sampling cell as described in Appendix D. This avoided the problems of total uniform sampling, that is, cells with no background points. We drew 15 points per sampling cell to account for environmental heterogeneity within cells, which resulted in around 43,000 background points duplicated for each species, or 6,450,000 background points in total. The dimensions of the model design matrix were then (6,752,961; 4,817). A standard R numerical matrix with these dimensions would require around 231 GBytes of RAM memory. However, as our design matrix is sparse, with only $2 * (p_i + 1) + 1 = 27$ non-null values per row, its storage cost was divided by a factor of around 180 with the R sparse matrix format (see library Matrix). Consequently, we could fit this model on a laptop with R-glmnet (it requires about 20 Gbytes of RAM overall).

3 | RESULTS

The simulation study allowed us to evaluate the reliability of our joint model estimation method (see Section 2.4). The two performance metrics obtained for the 10 simulation scenarios (2 environmental variables and 5 sampling effort profiles) are summarized in Figure 2. We were also able to illustrate the effectiveness of the method on the Pl@ntNet queries dataset as described in Section 2.5. The estimated sampling effort is displayed in Figure 3. The estimation results for an exotic invasive plant, *Phytolacca americana* L, are provided in Figure 4.

3.1 | Simulation: Very good fit when the simulated sampling effort was cell-wise constant

The estimated sampling effort had an R^2 of 0.97 (for both **alti** and **chbio_12**) compared to the simulated sampling effort when the latter was constant within cells (columns **x:alti H=+Inf** and **x:*chbio*$_{12}$ H=+Inf** of Figure 2a). This means that the estimate was almost collinear with the true sampling effort. Regarding species density estimates, the average R^2 over all species was 0.95 for **alti** and 0.88 for **chbio_12** (columns **x:alti H=+Inf** and **x:*chbio*$_{12}$ H=+Inf** of Figure 2b). This shows that the method recovers unbiased niches and sampling

[2]http://doi.org/10.5281/zenodo.2635501

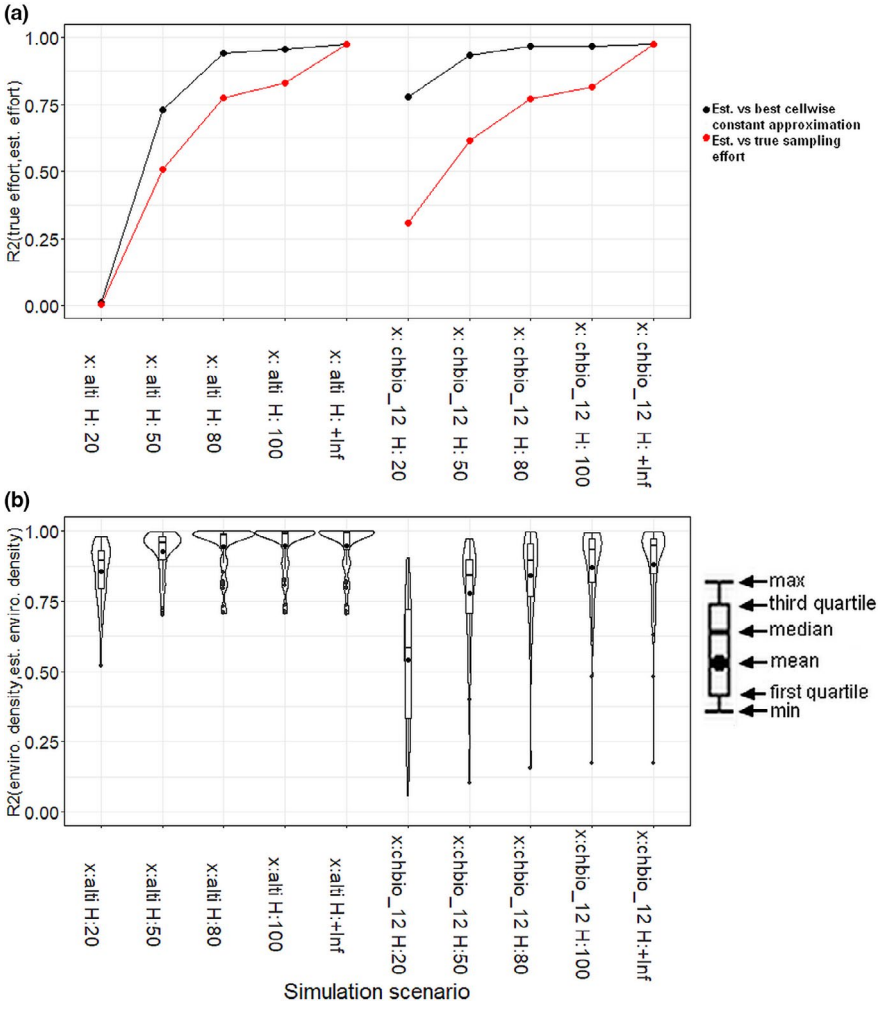


FIGURE 2 R^2 between generative and estimated model components in the 10 simulation scenarios for the sampling effort (a) and the species environmental density (b). In (a), the R^2 was computed between the simulated sampling effort density (raw in red or averaged per estimation cell in black) and the estimated density over the geographical space. Regarding the evaluation of the species density estimates, the same metric was computed between the true and the estimated density across the environmental gradient and for the 50 species, for each scenario. In (b), the 50 species metrics values are summarized through boxplots overlaid on a density plot

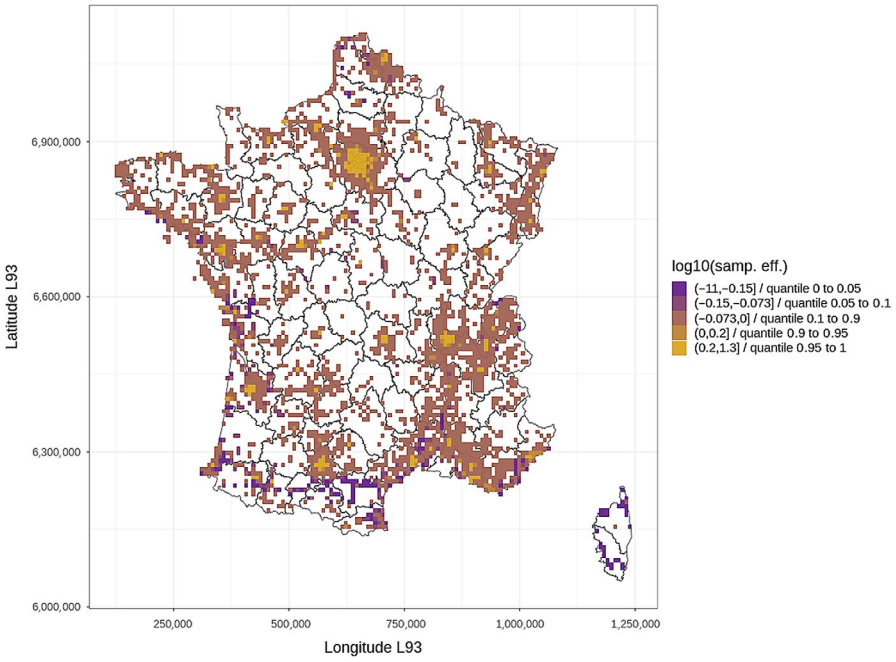


FIGURE 3 Relative sampling effort estimated from occurrences recorded in Pl@ntNet in France. The model was fitted on 302,961 occurrences of 150 plant species in France reported between 2017 and 2018 using the Pl@ntNet application. We represent the estimated sampling effort in the logarithm with base 10 to more clearly shows the orders of variation. The white cells are those with too few occurrences to be integrated in the model

effort estimates under good model specifications (sampling effort constant within cells) and that it is almost unaffected by estimation variance with this realistic simulated sample size and parameterization. However, in reality, sampling effort is not constant within cells, so the effect of violating this assumption also needed to be assessed, as shown below.

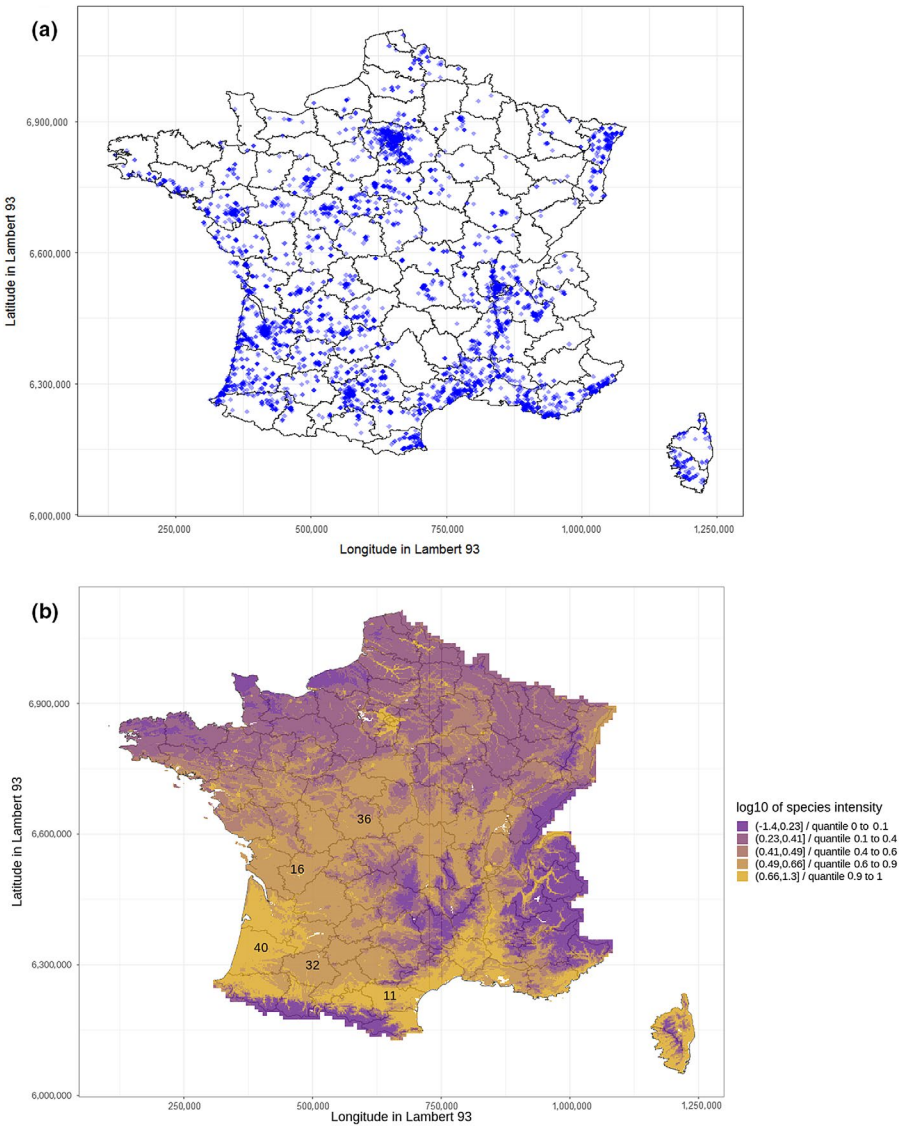


FIGURE 4 Raw occurrence and estimated density of *Phytolacca americana* L. from Pl@ntNet data. (a) 4,640 occurrences of *Phytolacca americana* L. recorded by Pl@ntNet users with automatic identification over the 2017–2018 period. (b) Decimal logarithm of predicted relative density of *Phytolacca americana* L. across France estimated from the occurrences with the proposed study method. The discrete gradient of colours represents quantile interval ranges

3.2 | Simulation: Smoother is better

The red and black curves of Figure 2a show that the approximation of the sampling effort was better when the sampling effort was smoother, for both environmental variables. While the red curve represents the fit of the raw sampling effort, the black curve represents the fit of the sampling effort averaged per cell (i.e. the best cell-wise constant approximation (BCCA) of the true sampling effort that can be estimated by the model in the ideal case) and is always above the red curve. As H increased, the sampling effort variation within cells became smoother, that is, the curvature of the spatial function decreased, and was thus closer to constant within cells. This tended to reduce the gap between the red and black curves when H increased and the model converged towards the BCCA. However, it is surprising that for **x:alti H:20** the gap between the red and the black curve was much smaller than for **x:alti H:50** ($R^2 = 0.0044$ for the true sampling effort and $R^2 = 0.01$ for the BCCA).

3.3 | Bias under joint variation in sampling effort and environmental variables within cells

The high error of **x:alti H:20** cannot be due to estimation variance, as the fit was almost perfect for the cell-wise constant effort. The error was most likely due to an estimation bias when the model of sampling effort cannot fit the variation in occurrence density within cells. To explain it in the simplified context of a single species case, the model is optimized so that variation in occurrence intensity $s_\gamma(z)\lambda_\beta \circ x(z)$ (product of the sampling effort and the species density estimates) fits the variation in observed occurrence density $s(z)\lambda \circ x(z)$ across space. However, the best approximation of this product of densities is not necessarily the product of the best approximations per density, namely the BCCA of s and $\lambda \circ x$ itself. More precisely, bias may appear if sampling effort strongly and monotonically varies with the environmental feature x within cells. We visualize and describe such bias in sampling effort profile (3) in the complementary simulation experiment

in Appendix F, with a joint visualization of species and sampling effort density estimates. The examination of a re-expression of the asymptotic model likelihood (Equation 3, second subsection of Appendix B) suggested that, if bias happens, the N species density parameters $\beta^1, \ldots, \beta^N$ controlling the log-linear response of the species densities $\lambda^1, \ldots, \lambda^N$ to x are all likely to be biased. Their errors should have the same sign, to compensate for the increase or decrease in sampling effort along the environmental feature within cells. This bias is related to the problem of spatial confounding in spatial statistics (Hodges & Reich, 2010).

3.4 | Simulation: Estimation of species density improved with smoother sampling effort

Figure 2b shows that species responses using the model were, on average, well estimated in most scenarios, even when sampling effort estimation was worst. In the scenario **x:alti H:20**, the average R^2 of the 50 species densities was around 0.85. In fact, as shown by the asymmetry of density plots in all scenarios, most species had a good fit with similar performance while a few had a significantly worse fit. As for the estimation of sampling effort, quality notably increased with H. This indicates that the robustness issue with sampling effort variation within cells translated into bias in species estimates. In addition, some species were consistently badly estimated, with R^2 below 0.50 even for $H = +\text{Inf}$. This could be the consequence of a simulated niche optimum being in a scarcely sampled area and/or a lack of occurrences. Species density estimation was, overall, worse for **chbio_12** than **alti**, even with good model specification (0.88 for **x:chbio_12 H:+Inf** on average compared to 0.95 for **x:alti H:+Inf** on average, see Figure 2b), whereas the estimation of sampling effort density was almost perfect. This implies that lower performance was not due to estimation bias, but to estimation variance, due to responses that were harder to estimate given the sampling effort and occurrences. The lower estimation quality with chbio_12 was thus not intrinsically due to the variable itself, but a consequence of species niches (which were randomly defined) that are harder to estimate. It also highlights that even when species estimation is unbiased, its precision necessarily depends on the overall intensity of sampling, that is, a sufficient number of points are required everywhere (all species included) in environmental space to ensure homogeneity in the estimation quality across species, as highlighted in section **model design guidelines**.

3.5 | Application: *Phytolacca americana* L. distribution

Using the Pl@ntNet queries dataset, we fit the model to provide species density estimates for 150 plant species. Figure 4b displays the decimal logarithm density estimation of *Phytolacca americana* L. The estimation provided by the model is consistent with the knowledge of the *Phytolacca* habitat as described in Dumas (2011). This species is cultivated as an ornamental shrub all over France—one of the reasons for its introduction—and often becomes established on disturbed soils in surrounding areas. In rural areas, it prefers managed forests with acidic, sandy soils. It is also found along rivers bordered with trees, as predicted by the model along the Rhone and the Garonne. Northern France is not favourable to this species. The model identified true hotspots even in scarcely sampled areas. It also predicted that the species is abundant in several relatively unsampled departments, such as the Indre, Aude, Charente and Gers. Indeed, Figure 3, representing the fitted log-relative sampling effort, shows that most cells in those regions had too few occurrences to be included in the model, and the ones that were included had a relatively low sampling effort. The FCBN records from 2000, which can be seen at http://siflore.fcbn.fr/?cd_ref=113418 r = metro, confirm that the species is indeed widely present in Indre. Conversely, there are very few reports in the National Inventory of Natural Assets (INPN) data for Aude, Charente or Gers, although presence-only records exist (Dumas, 2011 and Pl@ntNet). Those regions have been undersurveyed by conservatory experts in the last 20 years. Thus, the current estimated abundance of *Phytolacca americana* has either stayed undetected by sampling or is the result of a recent invasion.

4 | DISCUSSION

We found that our method to jointly estimate densities of multiple species, with a spatial function representing a common sampling effort, provides unbiased estimation of species relative density and sampling effort if the latter is constant within the cells of a spatial mesh. This allows the flexible estimation of the sampling effort, with no other prior knowledge than the grain of its spatial variation. Although the condition of constant sampling effort within cells is crucial to disentangle species and sampling densities, the method is robust for reasonable variation of sampling effort within cells, and even to stronger variation unrelated to environmental drivers of species density. We also found that the information gain on sampling effort from the most observed species helps to better estimate the niche of less observed species (Appendix C). Our method is devised for analysing large volumes of occurrences. Nevertheless, the simulation experiments and complementary results (Appendix F) showed that the artefactual influence of an environmental feature on species density can bias estimates when the sampling effort model is misspecified. More precisely, such bias appears along an environmental feature gradient if the true sampling effort strongly and monotonically covaries with this feature within cells (Appendix B). Removing this variable from the model of a species known to not respond to it should eliminate this bias for all species.

In the context of the model's application to Pl@ntNet data, we fit the model on a total of 302,961 citizen science occurrences of 150 plant species, covering 15% of France partitioned into 2,869 sampling cells. The estimated density of *Phytolacca americana* L. suggests

potential invaded areas yet undetected in published data, especially in scarcely sampled regions. Nevertheless, predictions out of the training area must be carefully examined, as they may present different environmental conditions and be subject to extrapolation errors.

4.1 | Method use guidelines

This method should be useful for large datasets of opportunistic occurrences in which some species are highly observed: for instance, with data from large citizen science or naturalist programmes. The recommendations on model design and sample size below indicate the conditions under which the method is most potentially useful:

1. Include at least several tens of occurrences (all species included) per sampling cell. Otherwise discard the cells and their occurrences, and do not include any background points over these cells. Alternatively, the user can increase the size of cells or include more species, widely distributed and with many occurrences, to meet the condition. Indeed, the information gain on the sampling effort parameter in a cell is equal to the total number of occurrences in this cell (see Appendix A). Scarce cells are a useless computational burden, as they need background points, and a potential source of variance. As the sampling effort in those cells is very uncertain, they consume degrees of liberty but do not contribute to reducing the variance in the species parameters. The method is not suited to contexts where the concentration of occurrences per sampling cell is too low: for example, herbarium datasets with few samples collected over large areas with very heterogeneous sampling effort. In such cases, the FactorBiasOut/TGB method (Phillips et al., 2009) should be more reliable because it does not require many degrees of freedom to model sampling effort. For example, our test case had an average of 105 occurrences per cell.
2. There should be at least several tens of occurrences for each environmental feature for each species. This is because the information gained on the parameters of a species comes only from its intensity of occurrence $s\lambda^i \circ x$ as can be seen in the expression of $I(\beta^i)$ given in Appendix A, because $\mathbb{E}(n_i) = \int_D$.
3. For each environmental feature, the standard deviation of this feature over all occurrences divided by the standard deviation over background points should not be too small (at least 1/3 in practice). This is a proxy of the spread of the overall occurrence intensity along the feature gradient. The estimation of this parameter with a certain confidence will require more occurrences if this indicator is low. Indeed, $I(\beta^i_k) = \int_D s(z)x_k(z)^2\lambda^i(x(z))\,\mathrm{d}z$ thus, if x is centred, the information on the species parameter is proportional to a (spatially weighted) variance in the corresponding environmental feature across space.
4. Regarding the choice of cell size, an optimal compromise should exist, but we have no definite procedure to reach it yet. Three main limitations can prevent good estimation when the sampling mesh reaches a resolution that is too high: the estimation variance (see the first point above), the identifiability (discussed in Section 2.2) and the memory limitation (number of background points required). Conversely, designing cells that are too large results in more variation in sampling effort within cells, which tends to favour estimation bias (see Section 3 and Appendix B, paragraph 2). In practice, a cross-validation scheme should be run for each tested cell area. Decreasing the size of cells can very quickly increase the estimation variance in the species parameters, as shown for a simulation example in paragraph 4 of Appendix C.
5. It is important to include some species with many occurrences in the model if available, especially if they are generalist respective to the environmental features. As shown in Appendix C, an increase in the number of occurrences of a single species reduces the estimation variance in sampling effort, which, in turn, reduces estimation variance in all other species parameters. Moreover, species with many occurrences contribute more to estimation variance as they are widely distributed in the environmental space.
6. An environmental variable should be removed from the model of a species if the species is known to be generalist along this gradient. This (a) reduces the estimation variance for all other species density parameters associated with this gradient, as shown in paragraphs 2 and 3 of Appendix C and (b) drastically reduces the estimation bias. Indeed, generalist species clarified in this way in the model provide a reference for sampling effort along the environmental gradient for the model.

4.2 | Scalability

This method can handle datasets that include a massive number of total occurrences over large geographical areas with many sampling cells, as shown by our test case with the Pl@ntNet data across France. This is favoured by the cell-wise constant sampling effort model and the use of a sparse design matrix. The memory load increases sublinearly with the number of sampling cells, but is roughly proportional to the inverse cell area of the highest resolution environmental raster. While there must be background points in all sampling cells, their total number just needs to enable a good screening of the overall environmental variability. The number of species may also be limiting as this linearly increases the memory load, independently of their number of occurrences. There is a room for improvement in reducing the need for memory in the fitting procedure: for instance, by optimizing the selection of background points or using a batch gradient descent algorithm.

5 | CONCLUSIONS

Our results demonstrate that our method can estimate sampling effort from presence-only data in a geographical space with considerably fewer prior assumptions than previous methods. The method can be extended to allow its use in a broad range of situations. It is

especially suited to analyse massive occurrences of multiple species at a large spatial scale, and should decrease bias in species distribution estimates. We thus think the approach will be useful to recover information about sampling effort from purely opportunistic occurrence data, enabling post-analysis of sampling effort variation in citizen science programmes and guiding strategies for further data collection. Insofar as citizen science data can provide time series over a long enough period, our method should allow monitoring of remarkable or noxious species such as exotic invasive species and help to guide conservation and management strategies (Botella et al., 2018).

ACKNOWLEDGEMENTS

The authors would like to thank the consortium Floris'Tic, the GDR CNRS 3645 'Ecologie Statistique', the CiSStats network—Statistics for Citizen Sciences—and the ANR project EcoNet who all supported this work. The authors also thank the anonymous associate editor for its investment and deep remarks all along the reviewing process. Finally, the authors would like to thank Gilles Le Moguedec for his contribution to this work and general support during the Phd of Christophe BOTELLA.

AUTHORS' CONTRIBUTIONS

All authors conceived and designed the methodology; A.J. and P.B. provided the Pl@ntNet data; C.B. compiled and published all the data, carried out the mathematical analysis, developed the code, performed the data analysis and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

Following the FAIR principles, the datasets and source code used in this manuscript are provided at the URLS below:

- Species occurrence data may be freely downloaded at http://doi.org/10.5281/zenodo.2634137 (Botella et al., 2019).
- Environmental rasters may be freely downloaded at http://doi.org/10.5281/zenodo.2635501 (Botella, 2019).
- The R code for running simulations and real data illustration, as well as the list of modeled species are provided on the manuscript dedicated Github repository https://doi.org/10.5281/zenodo.4455857 (Botella, 2020).

ORCID

Christophe Botella https://orcid.org/0000-0002-5249-911X
Alexis Joly https://orcid.org/0000-0002-2161-9940
Pierre Bonnet https://orcid.org/0000-0002-2828-4389
François Munoz https://orcid.org/0000-0001-8776-4705
Pascal Monestiez https://orcid.org/0000-0001-5851-2699

REFERENCES

Affouard, A., Goëau, H., Bonnet, P., Lombardo, J. C., & Joly, A. (2017). Toulon: s.n., 6 p. Pl@ntNet app in the era of deep learning. ICLR: International Conference on Learning Representations. 5, Toulon, France, 24 Avril 2017/26 Avril 2017. Retrieved from https://hal.archives-ouvertes.fr/hal-01629195

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the gbif database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15.

Berman, M., & Turner, T. R. (1992). Approximating point process likelihoods with glim. *Applied Statistics*, 31–38. https://doi.org/10.2307/2347614

Boakes, E. H., McGowan, P. J., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, *8*(6), e1000385.

Botella, C. (2019). A compilation of environmental geographic rasters for sdm covering france (version 1). *Zenodo*, https://doi.org/10.5281/zenodo.2635501

Botella, C. (2020). R code repository for 'jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic occurrence data'. https://doi.org/10.5281/zenodo.4455857

Botella, C., Bonnet, P., Joly, A., Lombardo, J.-C., & Affouard, A. (2019). Pl@ntnet queries 2017–2018 in france. *Zenodo*, https://doi.org/10.5281/zenodo.2634137

Botella, C., Joly, A., Bonnet, P., Monestiez, P., & Munoz, F. (2018). Species distribution modeling based on the automated identification of citizen observations. *Applications in Plant Sciences*, *6*(2), e1029.

Botella, C., Joly, A., Monestiez, P., Bonnet, P., & Munoz, F. (2020). Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection. *PLoS One*, *15*(5), e0232078.

Bradter, U., Mair, L., Jönsson, M., Knape, J., Singer, A., & Snäll, T. (2018). Can opportunistically collected citizen science data fill a data gap for habitat suitability models of less common species? *Methods in Ecology and Evolution*, *9*(7), 1667–1678. https://doi.org/10.1111/2041-210X.13012

Bystriakova, N., Peregrym, M., Erkens, R. H., Bezsmertna, O., & Schneider, H. (2012). Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. *Systematics and Biodiversity*, *10*(3), 305–315.

Calenge, C., Chadoeuf, J., Giraud, C., Huet, S., Julliard, R., Monestiez, P., Piffady, J., Pinaud, D., & Ruette, S. (2015). The spatial distribution of mustelidae in France. *PLoS One*, *10*(3), e0121689.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., & Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *60*(5), 757–776.

Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., & Rosemartin, A. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294.

Chiu, S. N., Stoyan, D., Kendall, W. S., & Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

Costa, G. C., Nogueira, C., Machado, R. B., & Colli, G. R. (2010). Sampling bias and the use of ecological niche modeling in conservation planning: A field evaluation in a biodiversity hotspot. *Biodiversity and Conservation*, *19*(3), 883–899.

De Solan, T., Renner, I., Cheylan, M., Geniez, P., & Barnagaud, J.-Y. (2019). Opportunistic records reveal mediterranean reptiles' scale-dependent responses to anthropogenic land use. *Ecography*, *42*(3), 608–620.

Devroye, L. (1986). Sample-based non-uniform random variate generation. In *WSC '86: Proceedings of the 18th conference on Winter simulation*, December 1986, pp. 260–265. https://doi.org/10.1145/318242.318443

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484.

Dumas, Y. (2011). Que savons-nous du raisin d'amérique (*Phytolacca americana*), espèce exotique envahissante? synthèse bibliographique. *Rendez-vous Techniques ONF*, *2011*, 48–57.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697.

Fithian, W., Elith, J., Hastie, T., & 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.

Fithian, W., & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, *7*(4), 1917.

Giraud, C., Calenge, C., Coron, C., & Julliard, R. (2016). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, *72*(2), 649–658.

Hastie, T., & Fithian, W. (2013). Inference from presence-only data; The ongoing controversy. *Ecography*, *36*(8), 864–867.

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

Johannesson, G., & Cressie, N. (2004). Finding large-scale spatial trends in massive, global, environmental datasets. *Environmetrics: The Official Journal of the International Environmetrics Society*, *15*(1), 1–44.

Joly, A., Bonnet, P., Goëau, H., Barbe, J., Selmi, S., Champ, J., Dufour-Kowalski, S., Affouard, A., Carré, J., Molino, J.-F., & Boujemaa, N. (2016). A look inside the pl@ ntnet experience. *Multimedia Systems*, *22*(6), 751–766.

Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Vellinga, W. P., Planqué, R., & Müller, H. (2018). Overview of lifeclef 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of ai. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. September, 2018 (pp. 247–266). Springer. ISBN: 978-3-319-98932-7. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-98932-7_24

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N., Linder, H. P., & Kessler, M. (2016). Climatologies at high resolution for the earth's land surface areas. *arXiv Preprint arXiv:1607.00217*.

Kery, M., Royle, J. A., Schmid, H., Schaub, M., Volet, B., Haefliger, G., & Zbinden, N. (2010). Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology*, *24*(5), 1388–1397.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., & Stone, L. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, *8*(4), 420–430.

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*, 22–37.

Mod, H. K., Scherrer, D., Luoto, M., & Guisan, A. (2016). What we use is not what we know: Environmental predictors in plant distribution models. *Journal of Vegetation Science*, *27*(6), 1308–1322.

Panagos, P. (2006). The European soil database. *GEO: Connexion*, *5*(7), 32–33.

Panagos, P., Van Liedekerke, M., Jones, A., & Montanarella, L. (2012). European soil data centre: Response to european policy support and public data requirements. *Land Use Policy*, *29*(2), 329–338.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3), 231–259.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), 366–379.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*(10), 1413–1422.

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 689–698.

Teacher, A. G., Griffiths, D. J., Hodgson, D. J., & Inger, R. (2013). Smartphones in ecology and evolution: A guide for the apprehensive. *Ecology and Evolution*, *3*(16), 5268–5278.

Van Liedekerke, M., Jones, A., & Panagos, P. (2006). ESDBv2 Raster Library - a set of rasters derived from the European Soil Database distribution v2.0 (published by the European Commission and the European Soil Bureau Network, CD-ROM, EUR 19945 EN). Retrieved from https://esdac.jrc.ec.europa.eu/content/european-soil-database-v2-raster-library-1kmx1km

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS One*, *8*(11), e79168.

Warton, D., & Shepherd, L. (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. https://doi.org/10.2307/29765559

Zomer, R. J., Bossio, D. A., Trabucco, A., Yuanjie, L., Gupta, D. C., & Singh, V. P. (2007). *Trees and water: Smallholder agroforestry on irrigated lands in Northern India* (Vol. 122). IWMI.

Zomer, R. J., Trabucco, A., Bossio, D. A., & Verchot, L. V. (2008). Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, Ecosystems & Environment*, *126*(1), 67–80.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Botella C, Joly A, Bonnet P, Munoz F, Monestiez P. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods Ecol Evol*. 2021;12:933–945. https://doi.org/10.1111/2041-210X.13565