Bayesian estimation of species relative abundances and habitat preferences using opportunistic data

Camille Coron[1] · **Clément Calenge**[2] · **Christophe Giraud**[1] · **Romain Julliard**[3]

Received: 14 April 2017 / Revised: 2 January 2018 / Published online: 1 February 2018

Abstract We develop a new statistical procedure to monitor relative species abundances and their respective preferences for different habitat types, using opportunistic data. Following Giraud et al. (Biometrics 72(2):649–658, 2015), we combine the opportunistic data with some standardized data in order to correct the bias inherent to the opportunistic data collection. Species observations are modeled by Poisson distributions whose parameters quantify species abundances and habitat preferences, and are estimated using Bayesian computations. Our main contributions are (i) to tackle the bias induced by habitat selection behaviors, (ii) to handle data where the habitat type associated to each observation is unknown, (iii) to estimate probabilities of selection of habitat for the species. As an illustration, we estimate common bird species habitat preferences and abundances in the region of Aquitaine (France).

Keywords Citizen science · Estimation of species relative abundances · Habitat selection · Opportunistic data · Resource selection function

Handling Editor: Pierre Dutilleul.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10651-018-0398-2) contains supplementary material, which is available to authorized users.

✉ Camille Coron
camille.coron@math.u-psud.fr

[1] Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France

[2] Office National de la Chasse et de la Faune Sauvage, Saint Benoist, BP 20, 78612 Le Perray en Yvelines, France

[3] CESCO, UMR CNRS 7204, Muséum National d'Histoire Naturelle, 55 rue Buffon, 75005 Paris, France

1 Introduction

Citizen science programs have been increasingly developed for biodiversity monitoring during the last 20 years. These programs usually enroll a large number of volunteers to work on a given scientific issue. For example, breeding bird surveys aim at estimating population trends of bird species in a given area (Link and Sauer 1998); the bird observations by the volunteers of the program populate a database describing the number of individuals of every focus species observed at a given time and place. Since the observational effort is usually much larger in citizen science programs than in "professional" scientific programs, citizen science programs usually gather much more observations than classical programs.

The issues tackled by citizen science programs can be very diverse, including the estimation of the spatial distribution of a set of species at different spatial scales (Royle et al. 2005; Fithian et al. 2014; Giraud et al. 2015), the study of certain ecological behaviors such as habitat selection (e.g., Biggs and Olden 2011), or the monitoring of population trends of endangered species (Link and Sauer 1998). Although some citizen science programs rely on data collected with standardized protocols and sampling designs (e.g. the North American Breeding Bird Survey; Link and Sauer 1998), many others rely on the opportunistic collection of observations by the volunteers, with unknown observation intensity. In the following, we will refer to this sort of uncontrolled data collection by "opportunistic data collection". In this paper, we focus on the estimation of some relative abundances based on such opportunistic data.

The opportunistic nature of these data raises important statistical issues, as shown in Dickinson et al. (2010) and in the review Isaac et al. (2014). A major issue is due to the non-uniform observation intensity: the collected data cannot be considered as an unbiased sample of the individuals present on this area. Any statistical approach relying on such data must tackle this data collection bias in some way. Several approach to this problem have been developed recently, notably based on data filtering (Roy et al. 2012), on defining a proxy for observation intensity and its variations through space and time (Telfer et al. 2002; Ball et al. 2011; Roy et al. 2012), and on the use of replicated visits leading to occupancy–detection models (van Strien et al. 2013). Some recent papers (Giraud et al. 2015; Fithian et al. 2014) proposed to handle this bias by combining this biased opportunistic dataset with a (possibly much smaller) dataset collected in the same area by a more classical program with a known observational effort (hereafter called "standardized dataset"). Under some restrictive assumptions (discussed below), such a combination provides some unbiased estimates of the relative abundance for the species monitored in at least one of the two programs. An attractive feature of these estimation schemes is to provide relative abundance estimation for species monitored in the opportunistic dataset, but not in the standardized dataset. This allows, in principle, to use opportunistic data collection to monitor some rare species that would be much more costly to monitor with a classical standardized program.

The approach proposed in Giraud et al. (2015) provides some relative abundance estimates, for a set of species at a collection of sites. The statistical modeling accounts for unequal and unknown detectability and reporting rates for the monitored species, both in the opportunistic and the standardized dataset, and for the unequal and unknown observational intensity in the opportunistic dataset. Yet, a crucial hypothesis is that the animals are distributed uniformly within each site. When a site gathers several areas with different habitat, and if the proportions of these different habitats differ among sites, this assumption is likely to be violated due to habitat selection behavior. Similarly the observational effort in opportunistic data is not equally distributed across the different habitat types due to observer preference for some habitats (Tulloch and Szabo 2012). This lack of homogeneity induces some important bias in the estimation, as shown in Bellamy et al. (1998), Mason and Macdonald (2004), Fuller et al. (2005) and Fithian et al. (2014). For example, if the volunteers participating to a bird monitoring program are mostly interested in waterbirds, they will strongly select for humid habitat within each site. If humid habitat is rare yet present within a site, most of the observations in this site will be performed in this rare habitat, and the resulting waterbird abundance in this site will be strongly overestimated.

The aim of our paper is to extend the approach of Giraud et al. (2015) by handling (unknown) habitat preferences that might influence both observers and observed animal behaviors. The whole monitored area is described by several habitat categories for which both observers and animals have different preferences. The habitat type associated to each observation is *not* assumed to be known exactly (e.g. the exact location of the observation is only known approximately, or an observed species observation may not be attributed unambiguously to a surrounding habitat). It can be seen as a hidden variable. Preferences of the observers and of each species for each habitat types are also unknown. Our approach provides estimation for all these quantities. By taking the habitat stratification into account, compared to what is obtained in Giraud et al. (2015), we produce (i) some more accurate relative abundance estimates; in particular, for given site, it allows to decompose a species relative abundance in a habitat-specific component (e.g., forest birds are relatively more abundant because forests are over-represented in that site) and an additive site-specific component, (ii) relative abundance maps at a finer spatial scale, and (iii) some estimates of the resource selection functions of the species (Manly et al. 2002), which has major implications for biological conservation. To sum-up, our main contributions are:

- To incorporate habitat type preferences in the statistical modeling of Giraud et al. (2015);
- To handle data where the habitat type associated to each observation is unknown (which allows to gather data at different spatial scales);
- To estimate the relative probabilities of selection of habitat for the monitored species.

We develop our statistical modeling in Sect. 2. In this new model, the respective habitat selection behaviors of observers and animals are modeled using hidden variables. The spatial distribution of observers in the sites, as well as the habitat selection within the sites is modeled differently for the two datasets (opportunistic and standardized). Animals are assumed to distribute according to their preferences for different habitat types within a site. We illustrate our approach using simulated data to demonstrate that it recovers the values of model parameters that were used to simulate the data. Finally, using a real dataset concerning birds in the Aquitaine region (France), we assess the performance of the model for estimating species relative abundances as well as their habitat selection parameters.

2 Model and parameters

In this section, we introduce our statistical modeling of the available data. These data are the outcome of some ecological features (species abundances) and some observational bias (detectability, partial reporting, heterogeneous observational effort, etc). Both the ecological features and the observational bias are affected by some ecological variables (for example habitat type, population and/or road density, altitude, as presented in Mair and Ruete 2016), which will be called habitats, from now on. Our modeling takes into account this double source of bias induced by the habitat. We first describe the ecological ingredients, which are independent from the considered datasets, and then the observational ingredients which are dataset dependent.

Species abundances and habitat selection probability The space-time is divided into units, we call henceforth *sites*, which correspond to the scale at which we will predict the relative abundances. More concretely, each site refers to the couple of a spatial domain and a time interval (although note that for real data applications we will focus only on spatial relative abundances variability, so a site will be simply a spatial region). We index the sites by $j \in [\![1, J]\!]$. The species we focus on, are indexed by $i \in [\![1, I]\!]$, and we denote by N_{ij} the number of individuals of species i in the site j. Our aim is to estimate the relative abundances N_{ij}/N_{i1} for all i and j, so that we can plot a relative abundance map and/or its temporal dynamics for each species. The choice of the reference site 1 (among sites that are visited by all datasets) is arbitrary.

The habitat types of a given site j are not homogeneous. Each site j is composed with several spatial domains, each presenting a specific habitat type. We index by $h \in [\![1, H]\!]$ the habitat types. Species are not uniformly distributed in the spatial domain of j: each species prefers some habitat types to some others and hence is more or less frequent in the different habitat types. In order to avoid biases in our estimation, we must take this heterogeneity into account. Our modeling assumes that the fraction of the animals of the species i present in the habitat type h inside the site j is proportional to the area V_{hj} of the habitat type h inside the site j (which is assumed to be known) weighted by a number $S_{ih} \in [0, 1]$ representing the preference toward the habitat type h for the species i. More precisely, we assume that the density of the species i at location x in the site j is given by

$$\frac{N_{ij} S_{ih(x)}}{\sum_{h'} S_{ih'} V_{h'j}},$$

with $h(x)$ the habitat type at location x. Following the concept definitions clarified in Lele et al. (2013), the parameters S_{ih} can be interpreted as the probability of selection of habitat h by species i. As an example, if S_{ih} is twice the value of $S_{ih'}$ then the probability that an individual with species i is present at a given point with habitat h is twice the value of this probability for habitat h'. These probabilities of selection of habitat are *unknown* and we will estimate them.

Observations and reporting As in Giraud et al. (2015), our relative abundance estimation is based on two datasets : (i) a standardized dataset, labeled by $k = 0$, collected under a program with a known sampling effort and (ii) an opportunistic dataset, labeled by $k = 1$, characterized by a completely unknown sampling effort. The datasets gather counts of animals for all sites j. We emphasize that each site j must be surveyed by both datasets, and each species i must be surveyed by at least one of the two datasets (at least one species must be surveyed in both datasets).

We assume that we have informations about the locations of the observations at a finer scale than the site j. Each site j is divided into several (possibly many) cells indexed by c and for each observation, we know in which cell c the observation occurred. We emphasize that the cell paving can completely differ between the two datasets. In each dataset, only a (possibly very small) fraction of the cells have been visited at least once by the observers, so we do not have counts for all cells c, but only for a (possibly very small) fraction of them. For a cell c visited in the dataset k, we denote by X_{ick} the corresponding count for the species i. This count X_{ick} is not homogeneously proportional to the abundance of the species i in c. Actually, the counts are biased by the inhomogeneous observational effort (total amount of observation time, number of observers, number or density of traps, etc) and the unequal probability of reporting of the species i (varying detectability, partial reporting, etc). Following Giraud et al. (2015), we denote by E_{ck} the observation intensity (or effort) in the cell c for the dataset k, and by P_{ik} the probability of detection/reporting of the species i in the dataset k. When the species i are not monitored in the dataset k, the probability of detection/reporting P_{ik} is set to 0. Within a cell c, the observers do not scan the space uniformly. Actually, they have some preferences for some habitat types (which are not the same for the two datasets). These preferences induce some specific biases, which must be properly addressed. Similarly as for the probability of selection of habitat, the preference of the observers of the dataset k for the habitat h is represented by a real number $q_{hk} \in [0, 1]$. For the dataset k, we model the observation intensity at location x within the cell c by

$$\frac{q_{h(x)k} E_{ck}}{\sum_{h'} q_{h'k} V_{h'c}},$$

where V_{hc} is the known area of cell c covered by habitat h. Note that in our model we assume that observers all have the same preference for each habitat. Writing $\mathcal{A}_c$ for the spatial domain of the cell c and taking into account both the probabilities of selection of habitat and the observers habitat preferences, we obtain the modeling for the count of the species i in the cell c for the dataset k

$$\begin{aligned} X_{ick} &\sim \mathcal{P}\text{oisson}\left(\int_{\mathcal{A}_c} N_{ij} \frac{S_{ih(x)}}{\sum_{h'} S_{ih'} V_{h'j}} \times E_{ck} \frac{q_{h(x)k}}{\sum_{h'} q_{h'k} V_{h'c}} \times P_{ik} \, dx\right) \\ &= \mathcal{P}\text{oisson}\left(N_{ij} E_{ck} P_{ik} \sum_h \frac{q_{hk}}{\sum_{h'} q_{h'k} V_{h'c}} \times \frac{S_{ih}}{\sum_{h'} S_{ih'} V_{h'j}} V_{hc}\right). \end{aligned} \tag{1}$$

In the above model, recall that the volumes V_{hj} and V_{hc} are known. For the standardized dataset, the observation intensities E_{c0} are assumed to be known (up to a common multiplicative constant), and we assume that (i) either the habitat type associated to each observation X_{jc0} is known, (ii) or the ratios q_{h0}/q_{10} are known for all h (generally equal to 1). All the other parameters are unknown. Their identifiability and the implementation of model (1) are detailed in Appendix A.2.

We point out that, here, we do not take into account a dependence of the detectability with habitat types (due notably to different levels of visibility in different types of habitat). We refer to Sect. A.5.2 for an extension of this model, integrating a dependence of detection probability with habitat types.

Note finally that when neglecting differences in habitat selection probabilities both for observers and observed individuals, the total number X_{ick} of observations of individuals of species i in cell c of domain j for the dataset k follows the following model:

$$X_{ick} \sim \mathcal{P}\text{oisson}(N_{ij} E_{ck} P_{ik}/V_j), \tag{2}$$

which is the model introduced and studied in Giraud et al. (2015), with an appropriate scaling of the observation intensities E_{ck}.

3 Numerical results

We test our modeling framework both with some simulated data and with some real data. The likelihood of (1) cannot be maximized easily, so we opt for a non-informative Bayesian estimation approach as implemented in the *JAGS* software (Plummer 2003). In addition, it is straightforward to model overdispersion in a Bayesian context, for example by supposing a random normal variation in habitat preferences. This program is called within *R* (R Core Team 2014) using the *rjags* package (Plummer 2014). We choose uninformative priors (Gamma distributions with very large variance) for the unknown parameters and the *JAGS* sampler provides samples distributed according to the posterior distributions for these parameters. The details about the implementation of the estimation procedures are given in Appendix A.2.

3.1 Illustration with simulated data

We illustrate the ability of our estimation procedure to recover the actual parameter values with some simulated datasets, and we compare the results of our procedure to those computed using Model (2) (studied in Giraud et al. (2015)). More precisely, Model (2) is a particular case of our Model (1), and involves a lower number of unknown parameters. When taking into account habitat preferences, we therefore expect a gain in accuracy but a loss in precision of our estimation of relatives abundances, which is illustrated now.

We simulate two datasets according to Model (1): A standardized one ($k = 0$) with known relative effort intensities E_{j0}/E_{10} and an opportunistic one ($k = 1$) with unknown relative effort intensities. For this simulated dataset we consider 20 different species at 30 different sites that are covered by 2 types of habitat. For standardized (resp. opportunistic) data, 10 (resp. 30) cells are visited in each site. The other parameters, such as species abundances, detection probabilities, habitat selection probabilities, or efforts in the opportunistic dataset are sampled according to uniform distributions.

These two datasets are simulated only once, and the results described further do not show an important dependence on the simulated datasets.

In order to illustrate the impact of the habitat modeling and the gain of using opportunistic data, we compare the three following estimation frameworks:

[Opp+Stand with hab] Our model (1) with unknown habitat selection probabilities and using both opportunistic and standardized data, denoted below by [Opp+Stand with hab];

[Stand only with hab] Our model (1) with unknown habitat selection probabilities and using *only* standardized data (which corresponds to Equation (1), with $k = 0$ only). It is denoted by [Stand only with hab];

[Opp+Stand no hab] Model (2) (introduced in Giraud et al. 2015) which neglects differences in selection probabilities, using both opportunistic and standardized data. This model is denoted by [Opp+Stand no hab].

In Fig. 1, we plot the posterior distributions of relative species abundances obtained for these three frameworks, and the reference relative species abundance values that we estimate are given in red. This figure shows, as proved in Giraud et al. (2015), the improvements brought by opportunistic data, since the estimation obtained by combining the two datasets is both more precise and more accurate. It also illustrates that neglecting habitat preferences can lead to biased estimation of species relative abundances. Figure 5 in Appendix A.4 also gives the posterior distributions of habitat selection probabilities, that give good approximations of the real values given in red. Figure 2 gives the boxplots of the relative differences between the estimated and real relative abundances, for each of the three models. Here, the estimated relative abundances is defined as the mean of the associated posterior distribution, and similar results are obtained using the median of these distributions. Again, we observe that the estimation combining both datasets and taking habitat types into account produces better results; and ignoring habitat types induces a significant bias. Finally, when simulating 50 standardized and 50 opportunistic datasets (indexed by n) according to Model (1) we obtain that the empirical L^2 distance $\frac{1}{50*I*J}\sum_{n=1}^{50}\sum_{i,j}\left[\hat{N}_{ij}^{[model],(n)} - N_{ij}^{(n)}\right]^2$ between our estimation and real relative abundances is equal to 0.07 in the [Opp+Stand with hab] framework, to 201.25 in the [Stand only with hab] framework, and to 1.70 in the [Opp+Stand no hab] framework, which confirms improvement brought both by taking habitat into account and by combining standardized and opportunistic data.

3.2 Real data

3.2.1 Datasets and habitats

Datasets To investigate whether taking the habitat types into account improves the estimation of real data, we consider the same datasets as in Giraud et al. (2015). These are two different datasets of common birds observations in Aquitaine (south-western French region): standardized data are provided by the French National Hunting and

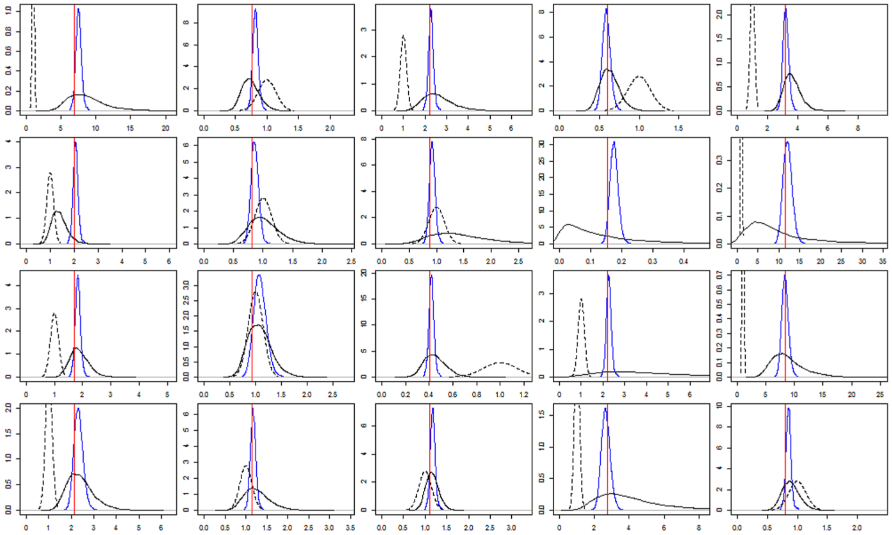

Fig. 1 Posterior distributions of the relative abundances N_{i2}/N_{i1} for all i, estimated with [Stand only with hab] (in black), [Opp+Stand with hab] (in blue) and [Opp+Stand no hab] (dotted line). The reference values are given in red

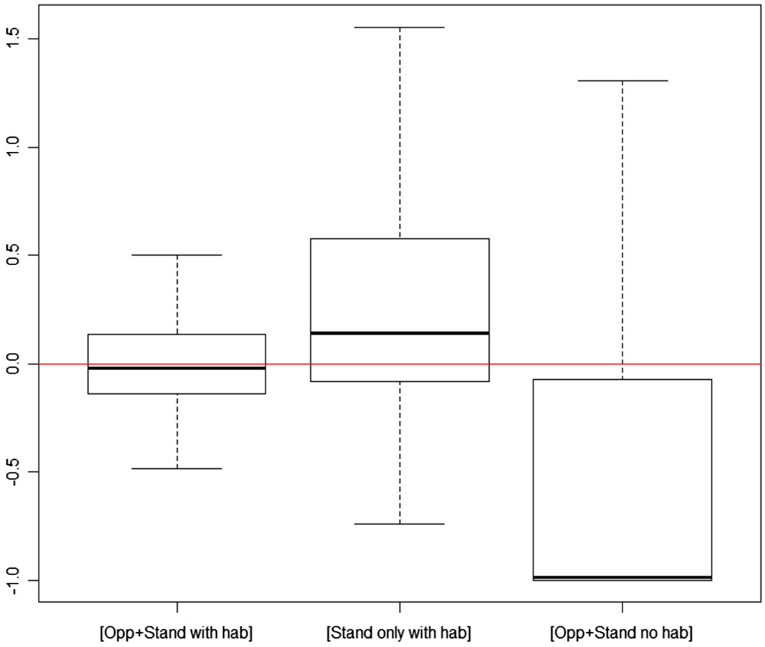

Fig. 2 Boxplots of the relative differences $\frac{(\hat{N}_{ij}^{[model]}/\hat{N}_{i1}^{[model]})-(N_{ij}/N_{i1})}{N_{ij}/N_{i1}}$ between the estimated and “real” relative abundances, using data and models [Opp+Stand with hab], [Stand only with hab], and [Opp+Stand no hab]

Wildlife Agency (ONCFS, Office National de la Chasse et de la Faune Sauvage), by the French National Hunters' Association (FNC, Fédération Nationale des Chasseurs) and by the French Departemental Hunters' Associations (FDC, Fédérations Départementales des Chasseurs), while opportunistic data are provided by the French program Faune-Aquitaine, managed by the protection association Ligue pour la Protection des Oiseaux (LPO). Our estimation is assessed using a validation dataset, produced by the French Museum of Natural History. These datasets provide us 15535 standardized and 693581 opportunistic observations of 34 bird species, which have been made on 66 quadrats. These data are used in the rest of the article, to estimate 6 habitat preferences parameters per species, as well as 65 relative abundances per species. More details concerning datasets are given in Appendix A.1. Note in particular that we gather observations which were made during 4 years and that our estimations were made under the assumption that relative abundances are constant from one year to the other (although they are modeled by stochastic variables that could account for some temporal variation in the Bayesian framework). However we recall that in Model (1), sites j can be seen as the couple of a spatial domain and a time interval, so temporal variation could also be studied using this type of data (preferably on a longer time period).

Habitats Land use (habitat) types were based on corine landcover typologies that were grouped in 7 categories in order to reduce complexity and ensure identifiability (see Appendix A.2): urbanized area, intensive agriculture with homogenous landscape (arable land or permanent crop), open natural landscape (natural or pasture), farmland with heterogenous landscape, mixed forest, deciduous forest, and coniferous forest.

3.2.2 Assessing estimation performances

In this section, we compare the estimation performances of the same three statistical models as in the Sect. 3.1: [Opp+Stand with hab], [Stand only with hab] and [Opp+Stand no hab]. We obtain estimation for the main parameters of interest of the model, namely the relative abundances N_{ij}/N_{i1} for all i, j, and the habitat selection probabilities S_{ih}/S_{i1} for all i, h and q_{h2}/q_{12} for all h. Our goal here is to compare the three estimation schemes, so we do not to discuss the ecological aspects of our estimates. Yet, in Appendix A.5, we provide some abundances maps and habitat selection probabilities for some species of interest.

To assess the performances of the three models, we investigate their ability to predict the STOC observations X_{ij}^{STOC}, that give the total number of observations of an individual of species i in each quadrat j surveyed in the STOC dataset. Since some species (21 among 34, see Appendix A.1 for the exact list) are not surveyed in the ACT dataset, we consider separately the results for the species surveyed in ACT and the results for the others. For each species i and each of the three models, we compute a predictor $\widehat{X}_{ij}^{model}$ of X_{ij}^{STOC} (defined in Appendix A.3 as the expectation of X_{ij}^{STOC} knowing the estimates of the unknown parameters of the model). Then for each species i and each model we compute the Pearson correlation between the vector $(\widehat{X}_{ij}^{model})_j$ and the vector $(X_{ij}^{STOC})_j$. The medians (as well as the first and third

Table 1 Medians of Pearson correlation coefficients (as well as first and third quartiles) between the STOC observations and the estimates of species relative abundances computed with the models [Opp+Stand with hab], [Stand only with hab], and [Opp+Stand no hab]

Data and model	Correlations (in ACT)	Correlations (not in ACT)
[Opp+Stand with hab]	0.49 (0.30, 0.54)	0.39 (0.12, 0.54)
[Stand only with hab]	0.29 (0.03, 0.46)	–
[Opp+Stand no hab]	0.44 (0.32, 0.68)	0.31 (0.19, 0.42)

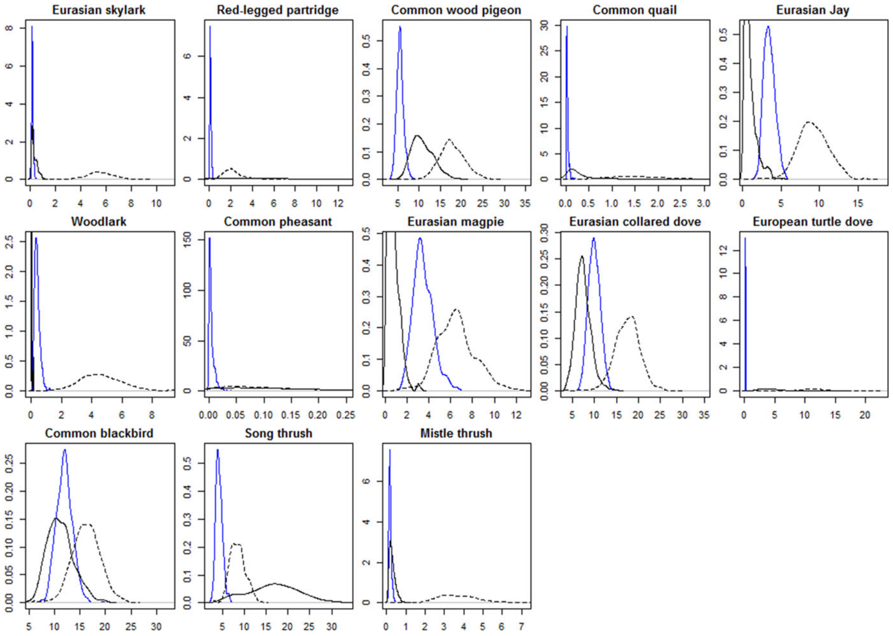

Fig. 3 Posterior of the relative abundances $\hat{N}_{ij}^{[model]}/\hat{N}_{i1}^{[model]}$ of species i monitored in the ACT program and $j = 62$, and using data and models [Opp+Stand with hab] (blue), [Stand only with hab] (black), and [Opp+Stand no hab] (dotted line)

quartiles) of these correlations (calculated for each species i) are given in Table 1. We notice that the results for the Model (2) slightly differ from the results obtained in Giraud et al. (2015), this can be explained by the fact that we use non-informative Bayesian estimation performed with JAGS instead of maximum likelihood estimation.

Figure 3 shows the posterior distributions of the relative abundances $N_{ij}^{[model]}/N_{i1}^{[model]}$ obtained for the three considered situations.

We observe an improvement of the predictions when we take the habitat into account. Actually, when we move from the Model (2) without habitat modeling (introduced in Giraud et al. 2015), to the Model (1) with habitat modeling, the median of the correlations between the vectors $(\widehat{X}_{ij}^{model})_j$ and $(X_{ij}^{STOC})_j$ increases from 0.44 to 0.49 for the species surveyed in the standardized dataset (ACT) and from 0.31 to

0.39 for the other species. Both the bias and variance contribute to the fit as measures by the correlation between observations and predictions, so the reduction of the bias obtained by modeling the habitat is stronger than the increase of the variance induced by the inflation of the number of parameters. This feature is confirmed by the difference between the BIC value for the model [Opp+Stand with hab] and the model [Opp+Stand no hab]:

$$\Delta BIC = BIC(\text{Opp} + \text{Stand with hab}) - BIC(\text{Opp} + \text{Stand no hab}) = -8867.$$

This improvement can be explained by the strong differences in habitat selection parameters S_{ih} that we have estimated (see Fig. 7), confirming that habitat biases cannot be ignored. This feature is also supported by the shape of the posterior distribution of the model with habitat [Opp+Stand with hab] which is much more spiked than the shape of the posterior distribution of the model without habitat [Opp+Stand no hab].

In addition, as in Giraud et al. (2015), we observe that adding the opportunist data in our estimation improves the estimation. In particular, we observe that for our dataset, the improvement brought from the use of opportunistic data is stronger than the improvement brought from habitat modeling. In the next paragraph, we investigate the importance of the habitat structure in the spatial variation of the abundance.

3.2.3 Models of spatial repartition

In our model (1), we incorporate two sources of abundance variation. Part of the variation in abundance is explained by the habitat structure. This part is driven by the habitat selection probability S_{ih}. The remaining of the abundance variation comes from some other factors, acting differentially on the different sites j. We investigate below whether the spatial repartition of individuals is mainly explained by habitat structure, or if some other factors have a major role. In the first case, most of the spatial variation would be explained by the variations of the ratio $S_{ih(x)} / \left(\sum_h S_{ih} V_{hj}\right)$; while in the second case, a major part of the spatial variation would be explained by the variations of the N_{ij} with j. In order to compare the relative importance of both effects, we compare the models derived from (1) by first neglecting the variation in the habitat selection probabilities S_{ih} (setting them all to 1); second by neglecting the variation in j of the N_{ij}, replacing all of them by a single value N_i. As explained before, when all the habitat selection probabilities S_{ih} are equal, the model (1) reduces to the model (2) of Giraud et al. (2015). So to investigate the relative importance of the habitat types with respect to the other factors, we compare the model [Opp+Stand with hab] and [Opp+Stand no hab] with the [One Quadrat with hab] model where

$$\text{[One Quadrat with hab]} \quad X_{ick} \sim \mathcal{P}\text{oisson}\left(N_i E_{ck} P_{ik} \sum_h \frac{q_{hk}}{\sum_{h'} q_{h'k} V_{h'c}} \frac{S_{ih}}{\sum_{h'} S_{ih'} V_{h'}} V_{hc} \right),$$

where V_h is the total area of habitat h in the considered space.

For each model, we compute as previously the Pearson correlations between the observations of the STOC dataset and the estimation of these observations using each of the three models. The median and the quartiles for all the species are given in Table 2.

Table 2 Medians (as well as first and third quartiles) of Pearson correlation coefficients between estimation of species relative abundances using different models of species spatial repartition

Data and model	Correlations
[Opp+Stand with hab]	0.45 (0.23, 0.54)
[One Quadrat with hab]	0.15 (–0.03, 0.36)
[Opp+Stand no hab]	0.38 (0.26, 0.52)

We observe that, in our case, the overall fit for the [one quadrat] model is poorer. This point is confirmed by the BIC values

$$\Delta BIC = BIC([\text{Opp} + \text{Stand no hab}]) - BIC([\text{One Quadrat with hab}]) = -813853.$$

The habitat type therefore does not seem to be the main driver of the spatial variation for most of the species considered in the specific spatial domain (though this ecological result can be different when considering other species on an other space, notably with more diverse types of habitats). The study of the impact of habitat structure on the estimates for each species relative abundance is presented in Sect. 4.

4 Discussion

Main results. We have developed a new statistical approach relying on the joint use of two datasets collected respectively by an opportunistic data collection program and by a classical standardized monitoring program with a known (and ideally controlled) observation intensity. By combining these two datasets, our approach estimates the relative abundance of a set of species in a set of sites, while accounting for the different detectability of the species in the two programs, variable habitat preferences by both the species and the observers, and unknown observation intensity in the opportunistic data collection program. The use of opportunistic data in this approach results in a considerably increased precision in comparison to the estimation that would be based only on the standardized data. Note also that our approach allows the estimation of the relative abundances of some species monitored only with the opportunistic data collection program, as long as there are at least several other species monitored by the two programs. Our approach extends the statistical modeling developed in Giraud et al. (2015), by taking into account the variable preferences of habitat types by both the species and the observers. We show that by accounting for habitat preferences by both the species and the observers in the citizen science program, our approach results in a lower bias and a better prediction of standardized observations using the relative abundances estimates. This is illustrated by a simulated dataset, as well as a practical case study.

A useful byproduct of this approach is the estimation of the relative preferences of each species for each habitat type: more precisely, the estimated value of the habitat selection parameters S_{ih} corresponds to the relative probability of selection of habitat h, which is exactly the definition of a resource selection function (RSF, Lele et al. 2013), a tool widely used in biological conservation and wildlife management to identify important habitats for a given species on a study area (Boyce and McDonald 1999). Existing statistical approaches for the fit of RSF rely on the comparison of an unbiased sample of the habitat used by the focus species, and an unbiased sample of either the unused habitat or the habitat available to the species (see a list of possible statistical approaches in Manly et al. (2002), especially chapter 4 for the case where habitat is defined by several categories). The collection of such data can be expensive, and when the study area is large and/or the focus species is rare it can become prohibitive (see the conclusion of MacKenzie (2005)). However, endangered rare species are precisely those for which information on selected places is the most crucial. The situation is generally worsened when several rare and endangered species are under study. Citizen science programs relying on opportunistic data collection are a very attractive alternative in this context because of the large observation effort carried out, but are often notoriously flawed by an unequal and unknown observation effort, which make their use in such studies difficult (Phillips et al. 2009). If at least a part of the species monitored in the opportunistic data collection program are also monitored in a more classical standardized program, our approach provides a way to correct for the biases caused by the unequal observation effort in opportunistic data collection programs, and therefore to benefit of the large observation effort for the RSF estimation. Our approach therefore allows the batch estimation of the RSF for all species in the opportunistic data collection program.

Limitations Our approach relies on the hypothesis that the preferences of a given species for a given habitat type does not vary into space and time. Several authors have shown that this might not always be the case: animals sometimes show a functional response of habitat selection, i.e. an habitat selection pattern that depends on habitat availability (Mysterud and Ims 1998); an habitat type can therefore be preferred by a species in a context and avoided in another (e.g. Calenge et al. 2005). Similarly, our approach supposes that the observers in the citizen science program show a constant preference in space and time. However, the observers preferences can also be characterized by functional responses. For example, in an opportunistic data collection program focusing on birds, observers may be more interested by waterbirds in humid regions and therefore prefer to spend their time close to lakes and ponds in such regions, as this is where they are more likely to observer the species of interest. On the other hand, in a mountainous region, observers might be more interested into raptors and avoid lakes and ponds. Such functional response of the observers can bias the resulting estimates, and should be seriously considered when fitting this model.

Possible extensions So far, we assumed that the detectability P_{ik} of species i in dataset k does not depend on the habitat type, which might be unrealistic since, in particular, the range of vision of an observer can be different from one habitat type to the other. If so, our estimation of habitat selection parameters S_{ih} can be biased. Due to identifiability constraints, our model cannot include an unknown list of parameters α_h taking into account the dependence of detectability on the habitat (since they will be undistinguishable from the species habitat selection probabilities S_{ih}). However, it is possible to include these parameters in the model and define an informative prior distribution for these detectabilities, if information is available elsewhere. Another solution is to use additional data concerning the detectability associated to each considered habitat. In Appendix A.5.2 we demonstrate how to implement this approach with the dataset used in this paper, by using additional data that give the respective numbers of observations made with different distance ranges for different kinds of habitats. Based on an idea similar to the statistical approach underlying the abundance estimation based on distance sampling (Buckland et al. 1993), we demonstrate how to estimate and account for the variable detectability between habitat types when estimating the species relative abundance in each site.

Issues when implementing the model Several issues must be carefully handled when implementing our model for specific datasets. A key step in the implementation of our model lies in the choice of the habitat types and their number. This choice, which is dataset dependent, must be handled with care. First, the choice of the habitat types must be meaningful for the monitored species. For example, assume that some of the monitored species have a very different selection probability for two given habitat types, say “deciduous forest” and “coniferous forest”. If the proportion of “deciduous forest” and “coniferous forest” varies from one site to the other, then the merging of these two habitat types into a single habitat type “forest” would induce a significant bias in the estimation. To avoid such biases, we may be tempted to select a very large number of habitat types, ensuring a strong homogeneity of each type. Yet, the multiplication of the habitat types is limited by the number of available observation points in each habitat type. Actually, in order to avoid a detrimental increase of the variance, we need to have enough observation points in each habitat type. These observations can be both in the opportunistic or in the standardized dataset. So, when choosing the habitat types (and their number), one must find a good balance between defining meaningful habitat types for the monitored species and having enough observations in each habitat type.

Another major degree of freedom in the implementation of the model, is the choice of the number of species. On the one hand, increasing the number of species helps the estimation, since the ratio between the number of observations and the number of parameters then decreases. On the other hand, including some rare species, or including some species which require to add some new habitat types, can harm the estimation by increasing the variance.

For a fair comparison of the statistical models with and without habitat modeling, we have implemented our model with very flat priors on all the parameters. In practice, we may have access to some existing estimates of some parameters. For example, for some species, we can have some estimates of some habitat selection probabilities (Manly et al. 2002). In this case, it is worth to incorporate this knowledge by designing some more informative priors on the habitat selection probabilities.

Acknowledgements The authors would like to thank Benjamin Auder for his contribution to the computer code for data treatment. We sincerely thank Denis Roux and all observers from the ACT network. We also thank the managers of both programs «Faune d’Aquitaine» and STOC-EPS, as well as the observers participating to these programs. This work was partially funded by public grants as part of the “Investissement d’avenir” project, reference ANR-11-LABX-0056-LMH, LabEx LMH, and reference ANR-10-CAMP-0151-02, Fondation Mathématiques Jacques Hadamard, by the Chair “Modélisation Mathématique et Biodiversité” of VEOLIA-Ecole Polytechnique-MNHN-F.X, and by the Mission for Interdisciplinarity at CNRS.

A Appendix

A.1 Datasets

For the first (standardized) dataset we used the ACT (Alaudidae, Columbidae, Turdidae) monitoring survey (see Boutin et al. (2003) for more details concerning this dataset and its protocole) in which 13 species of birds are monitored (see Table 3). The observers are professionals from the technical staff of the participating organisms. The Aquitaine region was discretized into 66 quadrats, in which a 4-km-long route was randomly placed in non-urban habitat (see Fig. 4). Each route was traveled twice between April and mid-June and included 5 points separated by exactly 1 km: at each travel, each point was visited for exactly 10 min. The species of every bird heard or seen was recorded, and for each point and each species, we have access to the maximum of the counts from the two visits (in order to take advantage of the maximum detectability and to avoid effects due to migration, as explained in Pollock 1982). This protocol was repeated for several years and we use data from 2008 to 2011, which finally leads to 239 visits of quadrats (some of the quadrats were not visited each year), therefore leading to $13 * 5 * 239 = 15535$ data, corresponding to the reporting of 7899 birds observations (some species are not always detected).

For opportunistic data, we used the dataset collected by the website www.fauneaquitaine.org (handled by the LPO), on which anyone can register and report the species, number of detected individuals, date, and location associated to any bird observations made in Aquitaine. The level of precision of the location is variable: exact location, locality indication, or municipality indication. For numerical analyses, to deal with this inhomogeneity in location information, we will use the municipality in which each observation was made, which is always given. As previously, we selected all such records between April and mid-June for the years 2008–2011. This led to 693,581 birds observations in 1622 municipalities (see Fig. 4), monitoring 34 species. Note that observers can go anywhere, for an unknown amount of time, and that they report their observations with an unknown probability (that might depend on the observed species); therefore these data do not provide any information concerning observation effort.

For the validation dataset used to assess the predictive power of our approach, we used the data from the STOC program (*Suivi temporel des oiseaux communs*), which is a French breeding bird survey carried out by the French Museum of Natural History (MNHN, Museum National d'Histoire Naturelle). The protocole of this survey (see Jiguet et al. 2012 for more details) is the following: each observer is assigned a 2×2 km square whose position is uniformly randomly chosen within 10 km of his/her house. The observer then distributes on the considered square, 10 observation points that have to be representative of the different habitats areas on the square, and each point is visited twice between April and mid-June, during 5 min. Every observation of each species (hearing or seeing) is reported and the maximum count among the two visits is kept, as for the ACT program. As previously, we use all such records for the years 2008–2011. This leads to 86526 birds observations in 38 squares (see Fig. 4), monitoring 34 species (the same than for the LPO dataset).

Table 3 provides the list of the 34 bird species under study.

Table 3 List of the 34 bird species under study

Latin name	Species	Latin name	Species
Aegithalos caudatus	Long-tailed tit	*Alauda arvensis**	Eurasian skylark
*Alectoris rufa**	Red-legged partridge	*Carduelis carduelis*	European goldfinch
Carduelis chloris	European greenfinch	*Certhia brachydactyla*	Short-toed treecreeper
*Columba palumbus**	Common wood pigeon	*Coturnix coturnix**	Common quail
Cuculus canorus	Common cuckoo	*Cyanistes caeruleus*	Eurasian blue tit
Dendrocopos major	Great spotted woodpecker	*Erithacus rubecula*	European robin
Fringilla coelebs	Common chaffinch	*Garrulus glandarius**	Eurasian jay
Hippolais polyglotta	Melodious warbler	*Lullula arborea**	Woodlark
Luscinia megarhynchos	Common nightingale	*Milvus migrans*	Black kite
Parus major	Great tit	*Passer domesticus*	House sparrow
*Phasianus colchicus**	Common pheasant	*Phoenicurus ochruros*	Black redstar
Phylloscopus collybita	Common chiffchaff	*Pica pica**	Eurasian magpie
Pica viridis	Eurasian green woodpecker	*Sitta europaea*	Eurasian nuthatch
*Streptopelia decaocto**	Eurasian collared dove	*Streptopelia turtur**	European turtle dove
Sylvia atricapilla	Eurasian blackcap	*Troglodytes troglodytes*	Eurasian wren
*Turdus merula**	Common blackbird	*Turdus philomelos**	Song thrush
*Turdus viscivorus**	Mistle thrush	*Upupa epops*	Eurasian hoopoe

The 13 species that are monitored by the ACT survey are indicated by an asterisk

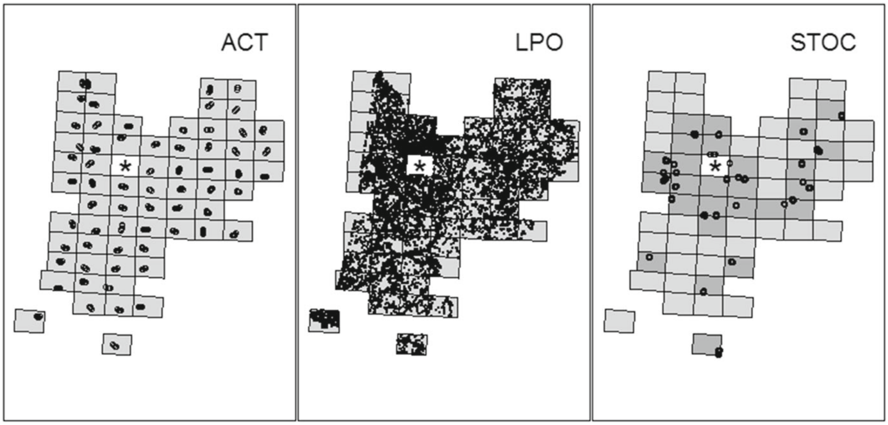

Fig. 4 Positions of data collection

A.2 Identifiability and models implementation

A.2.1 Reparametrization of the model

Let us recall our model for the observations, where c is a cell of the site j:

$$\begin{aligned} X_{ick} &\sim \mathcal{P}\text{oisson}\left(\int_{\mathcal{A}_c} N_{ij}\frac{S_{ih(x)}}{\sum_{h'} S_{ih'}V_{h'j}} \times E_{ck}\frac{q_{h(x)k}}{\sum_{h'} q_{h'k}V_{h'c}} \times P_{ik}\, dx\right) \\ &= \mathcal{P}\text{oisson}\left(N_{ij}E_{ck}P_{ik}\sum_h \frac{q_{hk}}{\sum_{h'} q_{h'k}V_{h'c}}\frac{S_{ih}}{\sum_{h'} S_{ih'}V_{h'j}}V_{hc}\right). \end{aligned}$$

For standardized data, we can assume either that q_{h0}/q_{10} is known for all h (generally equal to 1), or that each cell for the standardized dataset is small enough to be composed with only one habitat. In addition, we assume that E_{c0}/E_{10} is known for all c for the standardized dataset. To implement our model while ensuring identifiability of the parameters, we use the following change of variables

$$\begin{aligned} &\tilde{N}_{ij} = \frac{N_{ij}P_{i0}E_{10}}{\sum_{h'} \frac{S_{ih'}}{S_{i1}}V_{h'j}}, \quad \tilde{P}_{ik} = \frac{P_{ik}P_{10}}{P_{i0}P_{1k}}, \quad \tilde{E}_{ck} = \frac{E_{ck}P_{1k}}{P_{10}E_{10}}\frac{V_c}{\sum_{h'} \frac{q_{h'k}}{q_{1k}}V_{h'c}}, \\ &\tilde{q}_{hk} = \frac{q_{hk}}{q_{1k}}, \quad \tilde{S}_{ih} = \frac{S_{ih}}{S_{i1}} \end{aligned}$$

where $V_c = \sum_h V_{hc}$. Using this change of variables, we get that for all i, c and k,

$$X_{ick} \sim \mathcal{P}\left(\tilde{N}_{ij}\tilde{E}_{ck}\tilde{P}_{ik}\sum_h \tilde{q}_{hk}\tilde{S}_{ih}V_{hc}\right)$$

with

$$\frac{\tilde{N}_{ij}}{\tilde{N}_{i1}} = \frac{N_{ij}}{N_{i1}} \frac{\sum_{h'} \tilde{S}_{ih'} V_{h'1}}{\sum_{h'} \tilde{S}_{ih'} V_{h'j}}, \quad \tilde{P}_{i0} = 1, \quad \tilde{P}_{11} = 1, \quad \tilde{q_{h0}} = 1, \quad \tilde{q_{11}} = 1, \quad \tilde{S_{i1}} = 1$$

for all i, c, k, and $\tilde{E}_{c0}$ is known for all c.

In particular for standardized data, for which we can assume that the habitat associated to each cell c is known (denoted by $h(c)$), we get:

$$X_{ic0} \sim \mathcal{P}\left(\tilde{N}_{ij} \tilde{E}_{c0} \tilde{S}_{ih(c)}\right),$$

where $\tilde{E}_{c0}$ is known for each cell c. This is a generalized linear model with $IJ + I(H - 1)$ unknown parameters (the quantities $\tilde{N}_{ij}$ as well as habitat selection parameters $\tilde{S}_{ih}/S_{i1}$ for $h > 1$). These parameters are identifiable if and only if the matrix Y with size $C \times (J + H - 1)$ giving for each cell c visited by the STOC dataset, the site and habitat associated to this cell (when this habitat is not the first habitat), has rank $J + H - 1$. More precisely, the matrix Y is such that for all $c \in \llbracket 1, C \rrbracket$, $Y_{cj(c)} = 1$, $Y_{cJ+(h(c)-1)} = 1$ if $h(c) > 1$, and $Y_{cl} = 0$ elsewhere.

A.2.2 Implementation with JAGS

The computer code associated to the Sect. 3.1 is given in the numerical Additional File SimulatedData.Rnw. This program calls three models that are written in separate files: one for our model (Additional file ModelSimulatedData.txt), one for the model in which we use only standardized data (Additional file ModelStandardized-SimulatedData.txt), and one for the model in which differences in habitat preferences are neglected (Additional file ModelWithoutHabitatSimulatedData.txt). To make our estimations we used 10000 iterations ($n.iter$), with a thinning value to 10, 1 chain and 1000 iterations for adaptation ($n.adapt$). The computation time is about one hour per generated dataset for the three models, using a 2 cores Intel i5 processor.

The computer code associated to the Sect. 2 is given in the numerical Additional File RealData.Rnw. This program calls four models that are written in separate files: one for our model (Additional file ModelWithHabitat.txt), one for the model in which we use only standardized data (Additional file ModelStandardizedOnly.txt), one for the model in which differences in habitat preferences are neglected (Additional file ModelWithoutHabitat.txt), and one for the model in space is considered as one single quadrat (Additional file ModelOneQuadrat.txt). To make our estimations we used 10000 iterations ($n.iter$), with a thinning value to 10, 1 chain and 1000 iterations for adaptation ($n.adapt$). The estimations is about one hour per model, using a 2 cores Intel i5 processor.

A.3 Some details on the numerics: the prediction of the STOC data

Let $\mathcal{C}_j^{STOC}$ denote the set of all the observation points c in the quadrat j surveyed in the STOC dataset. The STOC counts for the species i in the quadrat j are

$$X_{ij}^{STOC} = \sum_{c \in \mathcal{C}_j^{STOC}} X_{ic}^{STOC}.$$

Let us denote by $h(c)$ the habitat type of the observation point c. In our model (1), the average number of individuals of the species i in the square $c \in \mathcal{C}_j^{STOC}$ is given by

$$\int_{\mathcal{A}_c} \frac{N_{ij} S_{ih(c)}}{\sum_{h'} S_{ih'} V_{h'j}} dx = \frac{N_{ij} S_{ih(c)} V_c}{\sum_{h'} S_{ih'} V_{h'j}}.$$

Taking into account a variable observational effort E_c^{STOC} on each observation point c, we then predict X_{ij}^{STOC} from the estimation based on our Model (1) by

$$\widehat{X}_{ij}^{model} = \hat{N}_{ij}^{model} \sum_{c \in \mathcal{C}_j^{STOC}} E_c^{STOC} \frac{\hat{S}_{ih(c)}^{model} V_c}{\sum_{h'} \hat{S}_{ih'}^{model} V_{h'j}},$$

where the observational effort E_c^{STOC} is given by the number of years of observation at the observation point c.

For the one quadrat model with habitat [One Quadrat with hab] displayed in Table 2, the prediction is given by

$$\widehat{X}_{ij}^{model} = \hat{N}_{i}^{model} \sum_{c \in \mathcal{C}_j^{STOC}} E_c^{STOC} \frac{\hat{S}_{ih(c)}^{model} V_c}{\sum_{h'} \hat{S}_{ih'}^{model} V_{h'}},$$

with $V_{h'}$ the area of the habitat type h' in the whole quadrat.

When the Model (2) is used for estimation, then the predictions are given by

$$\widehat{X}_{ij}^{model} = \hat{N}_{ij} \sum_{c \in \mathcal{C}_j^{STOC}} E_c^{STOC} \frac{V_c}{V_j}.$$

A.4 Additional results on simulated data

In this section we provide additional results to the ones presented in Sect. 3.1 for simulated data. Figure 5, as a complement to Fig. 1, provides the posterior distributions of the habitat selection probabilities S_{i2} for all i, showing that these posterior are a good approximations to the reference values that we wish to estimate.

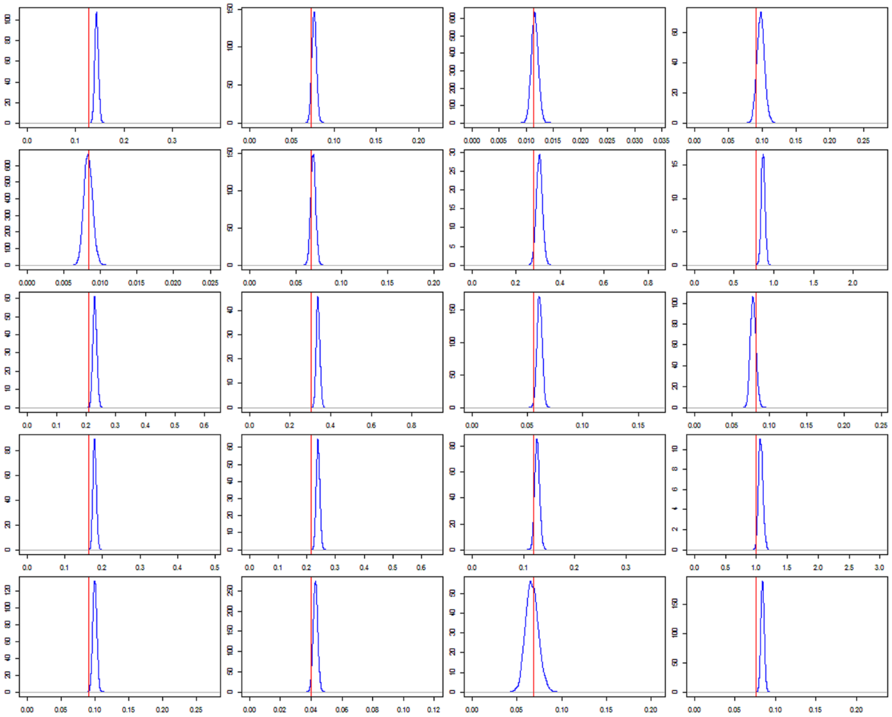

Fig. 5 Posterior distributions of the habitat selection probabilities S_{i2} for all i ($S_{i1} = 1$ for all i, each graph corresponds to a different value of i). The reference values are given in red

A.5 Additional results on real data

A.5.1 Some ecological results

In this section we provide additional results answering to natural ecological motivations. In Fig. 6 we give maps of the estimated densities of the Eurasian nuthatch, with and without habitat structure. The important differences between these two maps highlight the necessity of taking into account habitat. In Fig. 7 we give the mean preferences of all considered species, for each habitat type. The values of these preferences present important differences (the highest being 10 times larger than the lowest), which is crucial to take into account when predicting reaction of species to environmental change for instance.

A.5.2 Taking into account habitat dependent detectability

We so far assumed that the detectability P_{ik} of species i in dataset k does not depend on the habitat, which might be unrealistic since, in particular, the range of vision (or hearing) can be different from one habitat to the other. If so, our estimation of habitat selection parameters S_{ih} but also of species relative abundances can be biased. Due

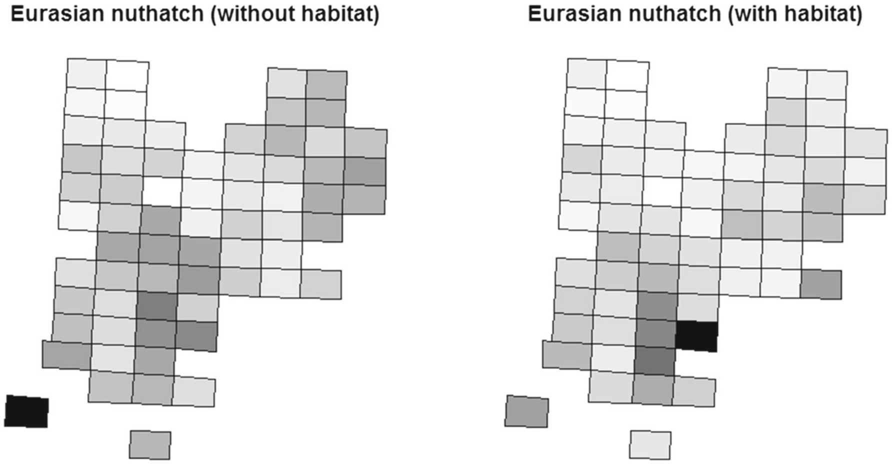
Eurasian nuthatch (without habitat)
Eurasian nuthatch (with habitat)

Fig. 6 Relative density maps of the European nuthatch, without (left), and with (right) taking into account habitat structure. For each quadrat the gray level indicates the relative density $\frac{\hat{N}_{ij}^{model} V_1}{\hat{N}_{i1}^{model} V_j}$ where V_l is the area of quadrat l

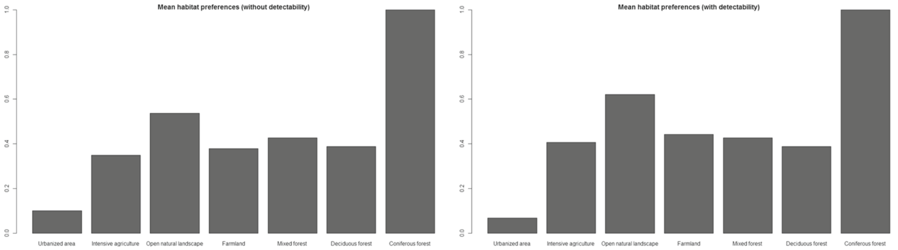

Fig. 7 Mean species estimated habitat selection probabilities, without (left) and with (right) habitat dependent detectability

to identifiability constraints, we cannot add and estimate an unknown list of parameters α_h taking into account the dependence of detectability on the habitat (since they will be undistinguishable from the species habitat selection probabilities S_{ih}). Our proposition is to use an auxiliary dataset that can provide informations concerning the detectability associated to each considered habitat. We test this idea using a dataset provided by VigieNature that gives for different kinds of habitats the respective numbers of observations made in different distance ranges. The program associated to this section is given in the file alpha.R.

For each habitat h, we can assume that detection probability is equal to 1 when observed individuals are "close enough" to the observer, since we only want to quantify the loss in detectability in each habitat due to the limitation in the range of vision (or hearing) in this habitat. More precisely, we assume that the detection probability is equal to 1 when the observed individual is less than 25 m far from the observer. Then if we denote by Y_h the number of observed individuals in habitat h and by Y_{1h} the number of observed individuals in habitat h, at distance less than 25 m from the observer, we can quantify the detectability in habitat h by the quantity

Table 4 The detectability associated to each habitat, taking account differences in ranges of vision

Corine land cover habitat	Detectability α_h
Urbanized area	1
Intensive agriculture	1.72
Open natural landscape	1.71
Farmland with heterogenous landscape	1.72
Mixed forest	1.47
Deciduous forest	1.47
Coniferous forest	1.47

$$\alpha_h = \frac{Y_h/Y_{1h}}{Y_1/Y_{11}}.$$

The result of these calculations is given in Table 4. As expected, the detectability is lower in urbanized area and forest than in open and agricultural landscapes. The impact of taking habitat detectability into account is illustrated in Fig. 7. This figure shows that difference between the highest and the lowest mean habitat selection probabilities is even higher than predicted without taking account habitat detectability (the former being about 15 times larger than the latter).

References

Ball S, Morris R, Rotheray G, Watt K (2011) Atlas of the hoverflies of great britain (diptera, syrphidae). Centre for Ecology and Hydrology, Wallingford

Bellamy PE, Brown NJ, Enoksson B, Firbank LG, Fuller RJ, Hinsley SA, Schotman AGM (1998) The influences of habitat, landscape structure and climate on local distribution patterns of the nuthatch (*Sitta europaea* L.). Oecologia 115(1–2):127–136

Biggs CR, Olden JD (2011) Multi-scale habitat occupancy of invasive lionfish (*Pterois volitans*) in coral reef environments of roatan, honduras. Aquat Invasions 6:347–353

Boutin J, Roux D, Eraud C (2003) Breeding bird monitoring in France: the act survey. Ornis Hung 12(13):1–2

Boyce M, McDonald L (1999) Relating populations to habitats using resource selection functions. Trends Ecol Evol 14:268–272

Buckland S, Anderson D, Burnham K, Laake J (1993) Distance sampling: estimating abundance of biological populations. Chapman & Hall, New York

Calenge C, Dufour A, Maillard D (2005) K-select analysis: a new method to analyse habitat selection in radio-tracking studies. Ecol Model 186:143–153

Dickinson JL, Zuckerberg B, Bonter DN (2010) Citizen science as an ecological research tool: challenges and benefits. Annu Rev Ecol Evol Syst 41(1):149–172

Fithian W, Elith J, Hastie T, Keith D (2014) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol Evol 6:424–438

Fuller RM, Devereux BJ, Gillings S, Amable GS, Hill RA (2005) Indices of bird-habitat preference from field surveys of birds and remote sensing of land cover: a study of south-eastern England with wider implications for conservation and biodiversity assessment. Glob Ecol Biogeogr 14:223–239

Giraud C, Calenge C, Coron C, Julliard R (2015) Capitalizing on opportunistic data for monitoring species relative abundances. Biometrics 72(2):649–658

Isaac NJB, van Strien AJ, August TA, de Zeeuw MP, Roy DB (2014) Statistics for citizen science: extracting signals of change from noisy ecological data. Methods Ecol Evol 5:1052–1060

Jiguet F, Devictor V, Julliard R, Couvet D (2012) French citizens monitoring ordinary birds provide tools for conservation and ecological sciences. Acta Oecol 44:58–66

Lele SR, Merrill EH, Keim J, Boyce MS (2013) Selection, use, choice and occupancy: clarifying concepts in resource selection studies. J Anim Ecol 82:1183–1191

Link WA, Sauer JR (1998) Estimating population change from count data: application to the North American breeding bird survey. Ecol Appl 8:258–268

MacKenzie D (2005) What are the issues with presence–absence data for wildlife managers? J Wildl Manag 69:849–860

Mair L, Ruete A (2016) Explaining spatial variation in the recording effort of citizen science data across multiple taxa. PLoS ONE 11(1):1–13

Manly B, McDonald L, Thomas D, MacDonald T, Erickson W (2002) Resource selection by animals. Statistical design and analysis for field studies. Kluwer Academic Publisher, London

Mason CF, Macdonald SM (2004) Distribution of foraging rooks, corvus frugilegus, and rookeries in a landscape in Eastern England dominated by winter cereals. Folia Zool 53(2):179–188

Mysterud A, Ims R (1998) Functional responses in habitat use: availability influences relative use in trade-off situations. Ecology 79:1435–1441

Phillips S, Dudík M, Elith J, Graham C, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19:181–197

Plummer M (2003) Jags: a program for analysis of bayesian graphical models using Gibbs sampling. In:3rd International workshop on distributed statistical computing (DSC 2003), vol 124. Vienna, Austria

Plummer M (2014) Rjags: Bayesian graphical models using MCMC. R package version, pp. 3–13

Pollock KH (1982) A capture recapture design robust to unequal probability of capture. J Wildl Manag 46:752–757

Core Team R (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Roy H, Adriaens T, Isaac N, Kenis M, Martin G, Brown PEA (2012) Invasive alien predator causes rapid declines of native European ladybirds. Divers Distrib 18:717–725

Royle JA, Nichols JD, Kéry M (2005) Modelling occurrence and abundance of species when detection is imperfect. Oikos 110(2):353–359

Telfer M, Preston C, Rothery P (2002) A general method for measuring relative change in range size from biological atlas data. Biol Conserv 107:99–109

Tulloch A, Szabo J (2012) A behavioural ecology approach to understand volunteer surveying for citizen science datasets. Emu 112:313–325

van Strien A, van Swaay C, Termaat T (2013) Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. J Appl Ecol 50:1450–1458

Camille Coron is an assistant professor in probability and statistics applied to ecology. She works in the Laboratoire de Mathématiques d'Orsay at Université Paris Sud (Orsay, FRANCE).

Clément Calenge is a biometrician and an engineer. He works in the Office national de la chasse et de la faune sauvage (Le Perray en Yvelines, FRANCE).

Christophe Giraud is a professor in statistics. He works in the Laboratoire de Mathématiques d'Orsay at Université Paris Sud (Orsay, FRANCE).

Romain Julliard is a professor in conservation biology. He works in the laboratory "Conservation des espèces, Restauration et Suivi des Populations" in the Muséum National d'Histoire Naturelle (Paris, FRANCE).