Modeling Earth Systems and Environment
https://doi.org/10.1007/s40808-022-01417-3

ORIGINAL ARTICLE

Inhomogeneous Poisson point process for species distribution modelling: relative performance of methods accounting for sampling bias and imperfect detection

Yannick Mugumaarhahama[1,2] · Adandé Belarmain Fandohan[1,3] · Arsène Ciza Mushagalusa[1,2] · Idelphonse Akoeugnigan Sode[1] · Romain L. Glèlè Kakaï[1]

Received: 9 November 2021 / Accepted: 8 April 2022

Abstract
Species distribution models (SDMs) have become tools of great importance in ecology, as advanced knowledge of suitable species habitat is required for the process of global biodiversity conservation. Presence-only data are the more abundant and readily available data widely used in SDM applications. These data should be treated as a thinned Poisson process to account for detection errors related to sampling bias and imperfect detection that arise in them. Failure to do so could be detrimental to SDM's predictions. This study assesses the effects of the species abundance, the variation in detection probability, and the number of sites visited in planned surveys on the performance of SDMs accounting for detection errors using simulated data. The results show that the accuracy and precision of estimates differ depending on models and species abundance. Their main difference lies in their ability to estimate β_0, the model intercept. The lower the species abundance, the higher the bias and variance of $\hat{\beta}_0$. Furthermore, the lower the detection probability, the higher the bias and variance of $\hat{\beta}_0$. However, β_1, the slope parameter, is estimated with almost high accuracy and precision for all models. This study demonstrates the low efficiency of accounting for sampling bias and imperfect detection based on presence-only data alone. Analysing presence-only data in conjunction with point-count outperformed the other approaches, whatever the species abundance, as long as the detection probability is at least 0.25 with average values of detectability covariates. The acceptable accuracy and precision, the minimum number of sites to consider vary depending on species abundance. At least 200 sites are required for the rare species, whereas 50 sites can suffice for the abundant species. Since collecting high-quality data are very expensive, this study emphasizes the need to promote initiatives such as citizen science programs that aim to collect species occurrence data with as little bias as possible.

Keywords Integrated Species distribution models · Hierarchical models · Maximum likelihood estimates · Data quality · Presence-only data · Point-count · Site-occupancy

✉ Romain L. Glèlè Kakaï
romain.glelekakai@fsa.uac.bj

Yannick Mugumaarhahama
lesmas2020@gmail.com

Adandé Belarmain Fandohan
bfandohan@gmail.com

Arsène Ciza Mushagalusa
shaga.ciza@gmail.com

Idelphonse Akoeugnigan Sode
sdidelphonse@gmail.com

[1] Laboratoire de Biomathématiques et d'Estimations Forestières, Faculty of Agronomic Sciences, University of Abomey-Calavi, 01, PO. Box: 526 Cotonou, Benin

[2] Unit of Applied Biostatistics, Faculty of Agriculture and Environmental Sciences, Université Evangélique en Afrique, PO. Box: 3323 Bukavu, Democratic Republic of Congo

[3] Unité de Recherche en Foresterie et Conservation des Bioressources, Ecole de Foresterie Tropicale, Université Nationale d'Agriculture, PO. Box: 45 Kétou, Benin

Published online: 11 May 2022

Introduction

In the context of climate change, species distribution models (SDMs) have become crucial tools in ecology for guiding conservation initiatives. They are used to find suitable habitats for vulnerable, invasive, endangered, or special-interest species, as well as to analyze the influence of climate change or land use on their future distribution (Fuller et al. 2008; Kremen et al. 2008; De Siqueira et al. 2009; Kearney et al. 2010; Fei et al. 2012; Crall et al. 2013; Li and Wang 2013; Guillera-Arroita et al. 2014; Hefley et al. 2017; Koshkina et al. 2017). The primary objective of SDMs is to identify areas where conservation activities such as habitat restoration, biomonitoring, or reserve networks should be strengthened. Therefore, accurate predictive SDMs are particularly important for efficient biodiversity management and conservation. Hence, several modelling approaches have been developed.

In the context of absence data being unavailable or inadequate, SDMs estimating species distribution using presence-only (PO) data have been developed (Brotons et al. 2004; Guisan and Thuiller 2005; Peterson et al. 2011). The use of PO data in the SDM is driven by the fact that high-quality data (presence–absence or abundance data), in addition to being scarce, are very expensive to collect. In contrast, PO data are abundant and readily available and accessible. However, the way they are not known for most them, leading to uncertainty in the predictions of SDMs (Elith et al. 2002; Barry and Elith 2006). Moreover, despite numerous attempts by ecologists, it is not feasible to accurately predict the true spatial distribution of species of interest based merely on PO data of doubtful quality (Fithian and Hastie 2013; Hastie and Fithian 2013). The uncertainty related to ecological data (mainly collected online) is a great challenge for SDMs. It has to be taken into account when the analysis results have to be interpreted appropriately or when they serve as a basis for decision-making process (Elith et al. 2002; Barry and Elith 2006).

SDM accuracy is affected by a number of factors, including sample bias and imperfect detection. Sampling bias is the process of gathering occurrence records of the species of interest in such a way that some sites are less likely to be visited than others. The failure to detect an individual of that species in a specific location, even if it is present, is known as imperfect detection. The accuracy of PO SDMs can be improved by handling sample bias effects; otherwise, the model may reflect sampling effort rather than the "true" species distribution (Phillips et al. 2009). Furthermore, imperfect detection can dramatically impair maximum likelihood estimates of presence-only SDM parameters (Dorazio 2012). The violation of the assumption of perfect detection makes parameter estimate more difficult and statistical inference less reliable. As a result, this situation could lead to a misunderstanding of the issue of interest, and consequently, conclusions for policy-making could be erroneous (Kellner and Swihart 2014). Yet, it is unclear to what extent incorporating sampling bias and poor detection into the model specification will increase the accuracy and precision of PO SDMs (Koshkina et al. 2017).

To address the challenge related to sampling bias and imperfect detection, integrated SDMs that analyses presence-only data in conjunction with replicated point-count (PC data) (Dorazio 2014) or with repeated surveys site-occupancy records (SO data) (Koshkina et al. 2017) were introduced. These integrated models are extensions of the Poisson point process models proposed by Warton and Shepherd (2010). They simultaneously account for sampling bias and imperfect detection using a thinned Poisson point-process model for presence-only data and incorporate either an N-mixture model for PC data (Royle 2004) or an extension of the conventional site-occupancy model for SO data (MacKenzie et al. 2002). Yet, species abundance may also affect SDMs performances (i.e., rare versus abundant). The number of occurrences is likely to vary with species abundance and thus affect the model performance. Furthermore, the effects of sampling bias and imperfect detection on model performance are also expected to vary with species abundance. Therefore, it is crucial to assess the performance of these models in this respect. For instance, how will SDMs performance be affected depending on species abundance?

It is only recently that the effect of species rarity on model performance has become a matter of interest in the SDM literature. While sample bias and imperfect detection are accounted for, the effects of species rarity on PO SDMs is investigated in this work. To the best of our knowledge, no previous research has assessed the performance of these newly introduced SDMs in the situation of rare vs. abundant species.

Methods

Simulation design

This study was conducted using simulated data. Three types of data were simulated: (i) presence-only data assumed to be collected opportunistically and subject to different levels of detection probability; (ii) point-count data and (iii) site-occupancy data assumed to be collected in spatial sample units (quadrats) using standard survey protocols.

The simulation approach for the data generation process was similar to that reported in Dorazio (2014) and Koshkina et al. (2017). Let us designate our virtual area B, which is

a square divided into 1000 x 1000 grid cells and represents the study area where our virtual species lives. Let us also refer to $x(s)$ and $w(s)$, two environmental predictors (covariates) generated using spatially varying bivariate distributions that are assumed to be independent of one another. $x(s)$ and $w(s)$ were generated in such a way that they were defined at every location, s on B's 2D grid. In other words, $x(s)$ and $w(s)$ are rasters that contain values at each location s within B, a subset of $\mathbb{R}^2$ (see Dorazio 2014; Koshkina et al. 2017).

Simulation of presence-only data

Let us consider n individuals of a particular species residing within B. PO data are a set $s = s_1, s_2, s_3, \ldots, s_n$ of point locations in B, where individuals of that species are recorded. It has been shown that these point locations can be modelled as a realization of Poisson point process. Hence, it is hypothesised that they represent the activity centres of the observed individuals. The the intensity function $\lambda(s)$ define the species distribution (Dorazio 2014). The process that characterizes the PO data is inhomogeneous because the intensity $\lambda(s)$ varies with location. Consequently, the expected number of occurrences in region B is a Poisson random variable with mean:

$$\mu(B) = \int_B \lambda(s)ds. \tag{1}$$

Since the intensity $\lambda(s)$ varies in space, it is hypothesised that this variation is a function of location-specific environmental covariates at each location s. The most standard formulation of the relationship between intensity $\lambda(s)$ and environmental covariates is the log-linear function (Warton and Shepherd 2010; Dorazio 2014; Fithian et al. 2014; Koshkina et al. 2017). In this study, we simulated the species intensity $\lambda(s)$ using the log-linear function that depends on the single covariate $x(s)$ (see, Dorazio 2014; Koshkina et al. 2017) :

$$\log(\lambda(s)) = \beta_0 + \beta_1 x(s). \tag{2}$$

In this study, two scenarios of species abundance were considered in the simulation process. We considered $\beta_1 = 0.5$ and $\beta_0 = \log(8000) \approx 8.9872$ for what we considered as an abundant species, while we considered $\beta_1 = 0.5$ and $\beta_0 = \log(200) \approx 5.2983$ for the rare species. In the context of this study, fitting the model involves estimating both the unknown intercept β_0 and the unknown slope β_0. Consequently, accurate estimates of β_0 and β_1 can be used to infer the abundance or occurrence of the species of interest for any location or sub-region within B.

As species occurrence data are prone to sampling bias and imperfect detection, only a subset of m individuals is observed among the n individuals. Each individual species present at locations s belonging to B is detected following $b(s)$, the detection probability (Dorazio 2014; Koshkina et al. 2017). In this study, the detection probability $b(s)$ includes both the imperfect detection (that is defined as the ability of observers to detect individuals) and the geographic sampling bias (that is defined as the probability that a location is sampled). Imperfect detection and sampling bias are analogous if they both depend on environmental covariates (Guillera-Arroita et al. 2014; Dorazio 2014; Koshkina et al. 2017). In this study, the function $b(s)$, subsequently referred to as 'detectability' or 'probability of detection', was considered to be determined by a single covariate $w(s)$. As in Dorazio (2014) and Koshkina et al. (2017), we used the logit function to simulate $b(s)$ as follows:

$$\text{logit}(b(s)) = \alpha_0 + \alpha_1 w(s). \tag{3}$$

As PO data are prone to sampling bias and imperfect detection, the m locations, $s = s_1, s_2, s_3, \ldots, s_m$ $(m < n)$, where the individuals have been observed opportunistically, can be modelled as a thinned Poisson point process. At each location s, the intensity $\nu(s)$ is equal to the product of $\lambda(s)$ and $b(s)$ (Dorazio 2014). Consequently, the expected number of detected occurrences in region B is defined as follows:

$$\nu(B) = \int_B \lambda(s)b(s)ds. \tag{4}$$

For both species abundance scenarios considered in this study, the point patterns representing the true presences were simulated with the intensity function $\lambda(s)$ but thinned by the detection probability $b(s)$. In the simulation process, we considered $\alpha_1 = -1$ and α_0 was assigned values ranging from -5 to 5 so that detection probabilities at the average value of $w(s)$ ranged from very low values ($b(s) \approx 0$) to very high values ($b(s) \approx 1$). In addition, the data were assumed to be subject to an extra sampling bias that cannot be accounted for by covariate $w(s)$. To achieve this, we did as Koshkina et al. (2017), randomly selected 30% of the total number of detected occurrences and considered them as the observed occurrences. Thus, the number of occurrence records for abundant species slightly exceeded 10000, assuming perfect detection and decreased as the probability of detection decreased. For rare species, the number of occurrence records did not exceed 300, assuming perfect detection, and decreased as the detection probability decreased until the minimum of 5 occurrence records.

Fitting the SDM assuming $b(s) = 1$ gives estimates of the intensity of $\lambda(s)b(s)$ rather than the intensity of the species $\lambda(s)$. Treating these estimates as estimates of species distribution can be dramatically misleading (Fithian et al. 2014). It is therefore mandatory to estimate $\lambda(s)$ accounting for $b(s)$. Different approaches have been proposed to estimate both unknown β parameters (β_0 and β_1 in this study) and unknown parameter α (α_0 and α_1 in this study). Careful attention was

paid to whether or not it was feasible to get reliable estimates of abundance or probability of occurrence based on PO data as a sole basis (Fithian et al. 2014). One option to address the limited information in the PO data is to analyse them in conjunction with PC or SO data (Dorazio 2014; Koshkina et al. 2017). In this study, we assessed the performance of the Poisson point process SDMs that analyse PO data alone and integrated SDMs that analyse PO data in conjunction with PC data or SO data. Their ability to predict species distribution could differ depending on the accuracy and precision they estimate β and α. The performance in estimating β is dependent on the accuracy and precision of α estimates. Additional PC data and SO data should increase the accuracy and precision of the integrated SDMs' estimates. See Dorazio (2014), and Koshkina et al. (2017) for a more in-depth understanding of the log-likelihood functions of these SDMs.

Simulation of point-count and site-occupancy data

Point-count and site-occupancy surveys provide high-quality information on a species' abundance and occupancy in a given area. In contrast, these data consist of a small number of observations made at a small number of sampling sites. Moreover, they usually cover a limited area and are not indicative of a species' geographic range. In this study, we divided the study region B into 100×100 square quadrats of equal size, to simulate point-count and site-occupancy data. Each quadrat contained 100 grid cells of B. As the parameter estimations of the integrated SDMs may be affected by the number of sample locations. The performance of integrated SDMs was assessed as a function of the number of sample sites in this study. Consequently, Z = 50, 100, 200, 400, or 800 were randomly selected throughout region B. By aggregating (summing) the intensities corresponding to grid cells that fall in the considered quadrat, the corresponding values of intensity $\lambda(s)$ or a true number of individuals present in it were estimated for each quadrat. The corresponding values of covariates $x(s)$ and $w(s)$ for each quadrat were determined using the aggregated areas' mean values of $x(s)$ and $w(s)$. By doing J = 4 independent binomial draws from individuals in each quadrat, simulated point-count and presence-absence (site-occupancy status) were obtained.

The detection probability in the PO data ($b(s)$) includes both sampling bias and imperfect detection (i.e., failure to detect the species when it is present). In contrast, the detection probability during planned surveys denoted $p(s)$ account for imperfect detection only because sites are chosen by design (Koshkina et al. 2017). In this study, the covariate of detectability in the PO data, $w(s)$, was also considered as the driver of detection probability, $p(s)$, in the planned surveys. The probability of detection ($p(s)$) at any site s was considered to be the same for all four repeated surveys but depending only on the single covariate $w(s)$, as follows :

$$\text{logit}(p_j(s)) = \gamma_0 + \gamma_1 w(s). \tag{5}$$

Here j represents the jth survey. For our simulations, we considered $\gamma_0 = 0$ and $\gamma_1 = -1.5$. With thinned intensity $\lambda(s)p(s)$, in each quadrat, observed individuals were counted during the j^{th} survey. These PC data were analysed in conjunction with PO data as proposed by Dorazio (2014).

To obtain SO datasets, PC data were converted into presence–absence data. If count ≥ 1, the species is present in the quadrat. Otherwise, the species is absent. In other words, a quadrat was considered as occupied by the virtual species, if at least one individual is observed in it (occupancy = 1). Otherwise, it is deemed to be unoccupied (occupancy = 0). These SO data were analysed in conjunction with PO data as proposed by Koshkina et al. (2017).

Data analysis

All of the experiments produced 2000 replications of the simulated data, which included presence-only, site-occupancy, and point-count data. Maximum likelihood estimates of the predictor's coefficient β_1 and the intercept β_0 were computed for each simulated dataset by fitting the presence-only model (PO model), the integrated model which combines PO data and SO data (PBSO model), and the integrated model which combines PO data and PC data (PBPC model). All the analyses were conducted in R (version 3.5.1) (R Core Team 2018), and the maximum likelihood approach was used to fit all of these models.

We estimated parameters β_0 and β_1 based on the data sets containing records of points selected randomly over the study area B and the associated covariate. Hence, we obtained $\hat{\beta}_0$ and $\hat{\beta}_1$ for each model that was compared to the 'true' values (β_0 and β_1) that were used to simulate the virtual species distribution $\lambda(s)$ in Eq. 2.

The models were compared by testing their performance in estimating β_0 and β_1 while accounting for sampling bias and imperfect detection under varying detection probability $b(s)$ and the number of sites in planned surveys. The performance of SDMs was assessed based on the operating characteristics of the estimators (bias and variance), the mean squared error (MSE) of estimates, and the relative efficiency.

In this study, $N = 2000$ replications of the simulated data were generated and $\beta_k = \{\beta_0, \beta_1\}$. For each parameter β_k, the operating characteristics and efficiency measures were calculated as follows:

Mean bias

$$\text{Bias}(\hat{\beta}_k)_{absolute} = \frac{1}{N}\sum_{i=1}^{N}\left(\beta_k - \hat{\beta}_{k,i}\right) \quad (6)$$

Mean relative bias

$$\text{Bias}(\hat{\beta}_k)_{relative} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\beta_k - \hat{\beta}_{k,i}}{\beta_k} \times 100\right) \quad (7)$$

Variance

$$\text{Var}(\hat{\beta}_k) = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{\beta}_k - \hat{\beta}_{k,i}\right)^2 \quad (8)$$

Mean squared error

$$\text{MSE}(\hat{\beta}_k) = \frac{1}{N}\sum_{i=1}^{N}\left(\beta_k - \hat{\beta}_{k,i}\right)^2 \quad (9)$$

Relative efficiency

For pairwise comparison of $\hat{\beta}_k$ of pairs of SDMs denoted M_1 and M_2, we calculated relative efficiency using the concept of the mean squared error (MSE).

$$\text{RE}(\hat{\beta}_{k,M_1}, \hat{\beta}_{k,M_2}) = \frac{\text{MSE}(\hat{\beta}_{k,M_1})}{\text{MSE}(\hat{\beta}_{k,M_2})} \quad (10)$$

with $\hat{\beta}_{k,M_1}$ and $\hat{\beta}_{k,M_2}$ estimates of parameter β_k obtained with model M_1 and model M_2 respectively.

Results

Comparison of β_0 and β_1 estimates

Table 1 shows the results of the test of conformity of $\hat{\beta}$ to their 'true' values (β) and the results of the equality of variances of the estimates as a function of the abundance of the species. The Student t test was used to test the conformity of the $\hat{\beta}$ while the Levene test was used to compare the variance of $\hat{\beta}$ as a function of species abundance.

The results of Table 1 showed that the higher the species abundance, the lower the variance of estimates of parameters β_0 and β_1. In addition, $\hat{\beta}_0$ are highly underestimated for the PO and PBSO model for the rare species as well as the abundant species. The best $\hat{\beta}_0$ were obtained with the PBPC model. However, almost all $\hat{\beta}_0$ are equal to β_0 for all studied SDMs. These results suggest that the combination of PO data and SO data does not significantly increase the accuracy of SDMs under any species abundance scenario. The PBPC model has shown the best performance in the estimation of parameters β_0 and β_1 for both rare and abundant species when compared to other models.

Table 1 Operating characteristics of maximum-likelihood estimates of parameters β_0 and β_1 obtained by fitting different SDMs

Parameter	Model	Species abundance	Real	Estimate	SD	t test p value
β_0	PO	Rare	5.298	4.842	1.866	0.000
		Abundant	8.987	7.718	1.193	0.000
		Levene test p value			0.000	
	PBSO	Rare	5.298	0.434	3.13	0.000
		Abundant	8.987	7.471	1.66	0.000
		Levene test p value			0.000	
	PBPC	Rare	5.298	5.286	0.153	0.111
		Abundant	8.987	8.987	0.027	0.93
		Levene test p value			0.000	
β_1	PO	Rare	0.5	0.521	0.222	0.025
		Abundant	0.5	0.498	0.051	0.06
		Levene test p value			0.000	
	PBSO	Rare	0.5	0.508	0.126	0.361
		Abundant	0.5	0.552	0.217	0.347
		Levene test p value			0.09368	
	PBPC	Rare	0.5	0.504	0.106	0.342
		Abundant	0.5	0.5	0.017	0.401
		Levene test p value			0.000	

For illustration, we considered $\alpha_0 = 1$ and 200 sites for high-quality data

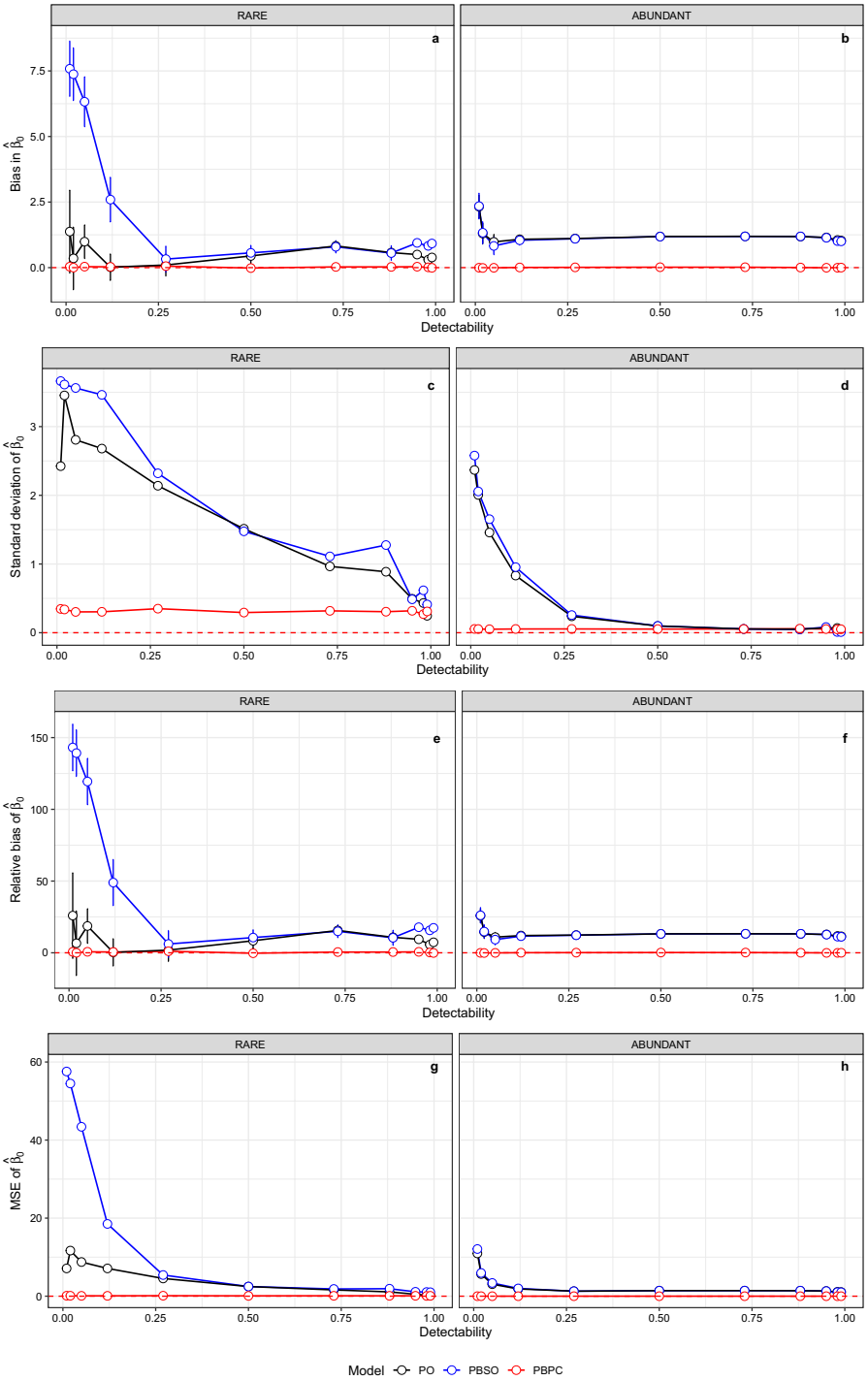

Fig. 1 Effects of variation detection probability ($b(s)$) on maximum-likelihood estimates of β_0. The red dashed line indicates bias equal zero. Error bars indicate 95% confidence intervals

It appears that accounting for sampling bias and imperfect detection using the PBPC model is a promising alternative.

Effects of detection probability on maximum likelihood estimates of β_0 and β_1

Figure 1 illustrates how $\hat{\beta}_0$ vary depending on the detection probability $be(s)$ while Fig. 2 shows how $\hat{\beta}_1$ is affected by the variation in the detection probability $be(s)$.

The results of $\hat{\beta}_0$ of the PO model and PBSO model are biased and sensitive to the variation in the detection

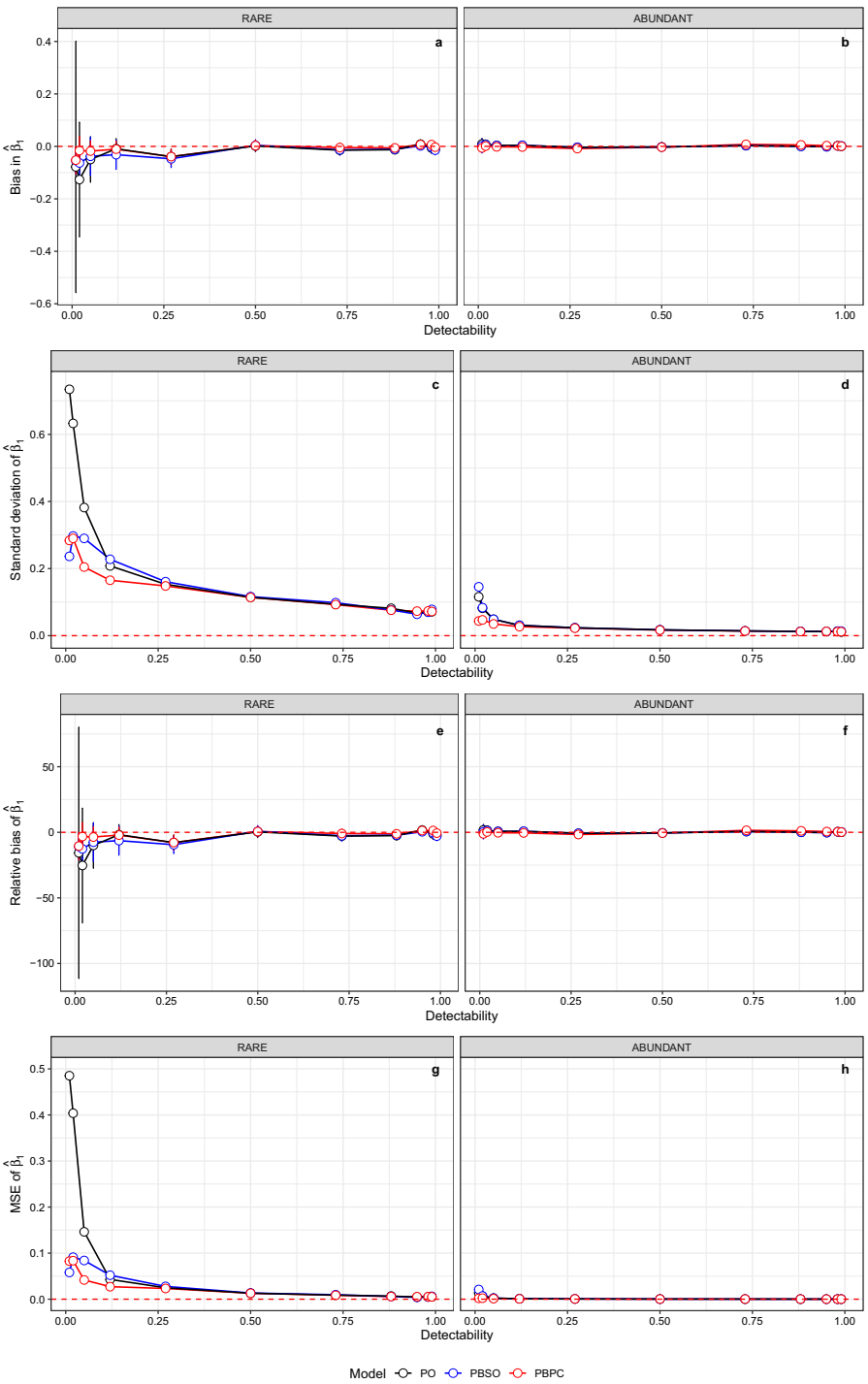

Fig. 2 Effects of variation detection probability ($b(s)$) on maximum-likelihood estimates of β_1 The red dashed line indicates bias equal zero. Error bars indicate 95% confidence intervals

probability (Fig. 1). For these SDMs, as $b(s)$ decreases, the bias in $\hat{\beta}_0$, mean squared error and variance of $\hat{\beta}_0$ increase. This trend is observed for both abundant and rare species. However, in both scenarios of species abundance, the PBPC model differs from PO and PBSO models. The PBPC model has shown to be less sensitive to changes in $b(s)$. For this model, the bias in $\hat{\beta}_0$ and the mean squared error of $\hat{\beta}_0$ are almost zero. Even the variance of $\hat{\beta}_0$ is very close to 0, but only its variance is slightly altered if the species is rare, regardless of the detection probability (Fig. 1). Figure 1 also illustrates that biases in $\hat{\beta}_0$ become extremely high when $b(s) < 0.25$.

Apart from the fact that species rarity induces high variability in $\hat{\beta}_1$, it was noted that the variation in $b(s)$ does not affect significantly the quality of $\hat{\beta}_1$ for all the SDMs. For both abundant and rare species, as long as $b(s) > 0.25$, the

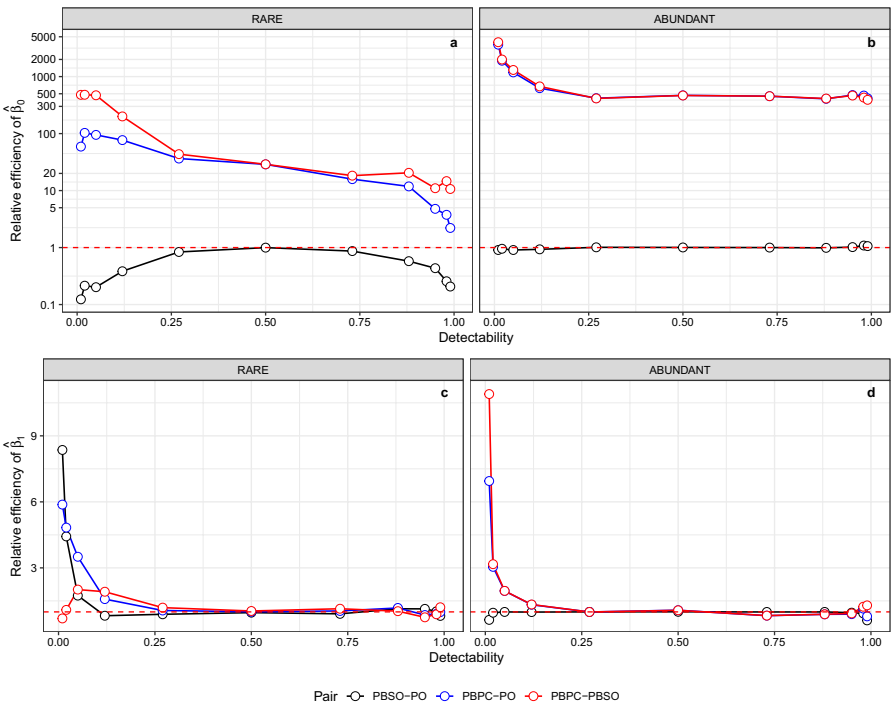

Fig. 3 Relative efficiency of SDMs as a function of detection probability $b(s)$. The red dashed line indicates the threshold of 1

effects of $b(s)$ on $\hat{\beta}_1$ are negligible. With $b(s) < 0.25$, the quality of $\hat{\beta}_1$ estimates is strongly affected (Fig. 2).

To compare the performance of the SDMs as a function of $b(s)$, the relative efficiency was calculated for each pair of SDMs (see Fig. 3).

The PBPC model outperformed the other SDMs (Fig. 3). For β_0 and β_1, estimates obtained with the PBPC model are either more efficient or equivalent to those of the other SDMs, whatever the detection probability. These results also show that $\hat{\beta}_0$ of the PO model are more efficient or equivalent to $\hat{\beta}_0$ of the PBSO model. However, for $\hat{\beta}_1$, the opposite trend was observed.

Effects of the number of sampled sites on maximum likelihood estimates of β_0 and β_1

Integrated SDMs are an alternative to overcome sampling bias and imperfect detection. However, it is important to know how many sites are required for a good performance of these models. Figures 4, 5 and 6 present the results with respect to this.

The first observation that emerges from the results in Fig. 4 is the poor performance of the PBSO model compared to the PBPC model in estimating β_0 regardless of the number of sites considered for high-quality data. These results also show the increase in the variance of $\hat{\beta}_0$ resulting from the species rarity. For the PBPC model, high performance in estimating β_0 is obtained from 50 sites for abundant species, while for rare species, good performance is obtained with at least 200 sites.

One significant observation emerges from the results in Fig. 5: varying the number of sites has almost no effect on $\hat{\beta}_1$. $\hat{\beta}_1$ are very close to their 'true' values for all integrated SDMs and all abundance scenarios.

As for the results presented in Fig. 3, the results of Fig. 6 show that the PBPC model outperformed the PO and PBSO models. $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained with the PBPC model are more efficient compared to those of the other SDMs, whatever the number of sites sampled during repeated planned surveys. $\hat{\beta}_0$ and $\hat{\beta}_1$ of the PO model are almost equivalent to $\hat{\beta}_0$ and $\hat{\beta}_1$ of the PBSO model, respectively.

Discussion

Performance in accounting for imperfect detection and sampling bias

PO models assume $b(s) \approx 1$. However, as PO data are prone to sampling bias and imperfect detection (Dorazio 2014), they should be treated as a thinned Poisson process (Fithian and Hastie 2013; Hefley et al. 2013; Warton et al. 2013) because estimating $\lambda(s)$ without accounting for $b(s)$ could lead to estimating $\lambda(s)b(s)$ instead of $\lambda(s)$. If interpreted as estimates of $\lambda(s)$, these estimations could be extremely wrong, potentially misdirecting species conservation efforts (Fithian et al. 2014). $\lambda(s)b(s)$ simply captures sampling effort and does not reflect ecological mechanisms that determine

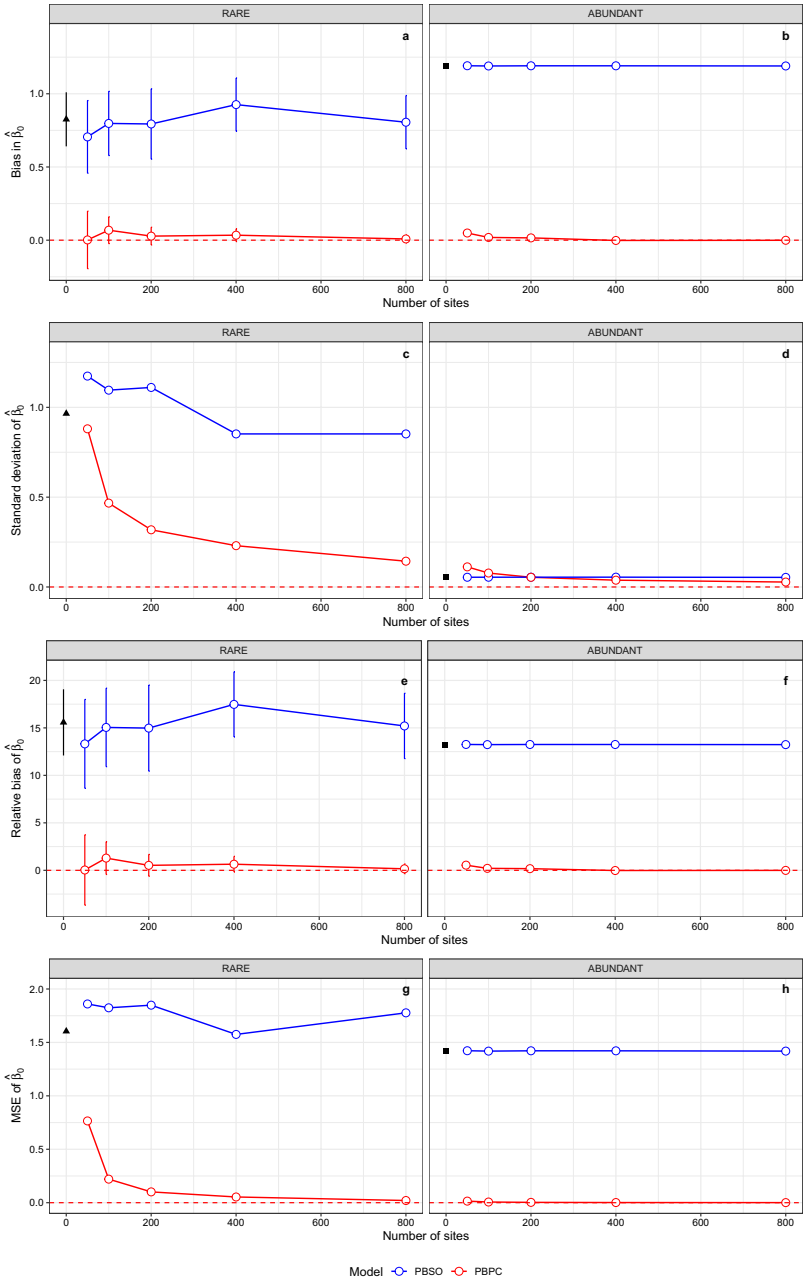

Fig. 4 Effects of the number of sites sampled in planned surveys on maximum-likelihood estimates of β_0. The red dashed line indicates bias equal zero. Black square and triangle are the results of the PO model. Error bars indicate 95% confidence intervals

the species' spatial distribution (Lahoz-Monfort et al. 2014; Guillera-Arroita 2017). Therefore, sampling bias and imperfect detection must be accounted for if predictions are the basis of a decision-making process (Elith et al. 2002; Barry and Elith 2006).

This study assessed the benefit of analysing PO data in conjunction with PC data or SO data, as proposed recently by Dorazio (2014) and Koshkina et al. (2017). Particularly, the aim was to assess the effects of the species abundance, the variation in detection probability, and the number of sites on the performance of these SDMs to inform ecologists about the efficiency of using such additional data. Dorazio (2007) demonstrated that the PC and SO models' predictions of species site occupancy status and abundance are technically identical. Moreover, MacKenzie et al. (2002) and Mackenzie and Royle (2005) also showed that the SO model may correct for imperfect detection. Furthermore, the PC model is well-known for producing reliable abundance and detectability estimations (Royle 2004). Then, we could expect that PBSO and PBPC models would display almost the same performance. Surprisingly, the results of this study show that analysing PO data alone or in conjunction with SO data was unable to estimate accurately β_0. The two models seem to behave almost in the same way. Then, it is clear that

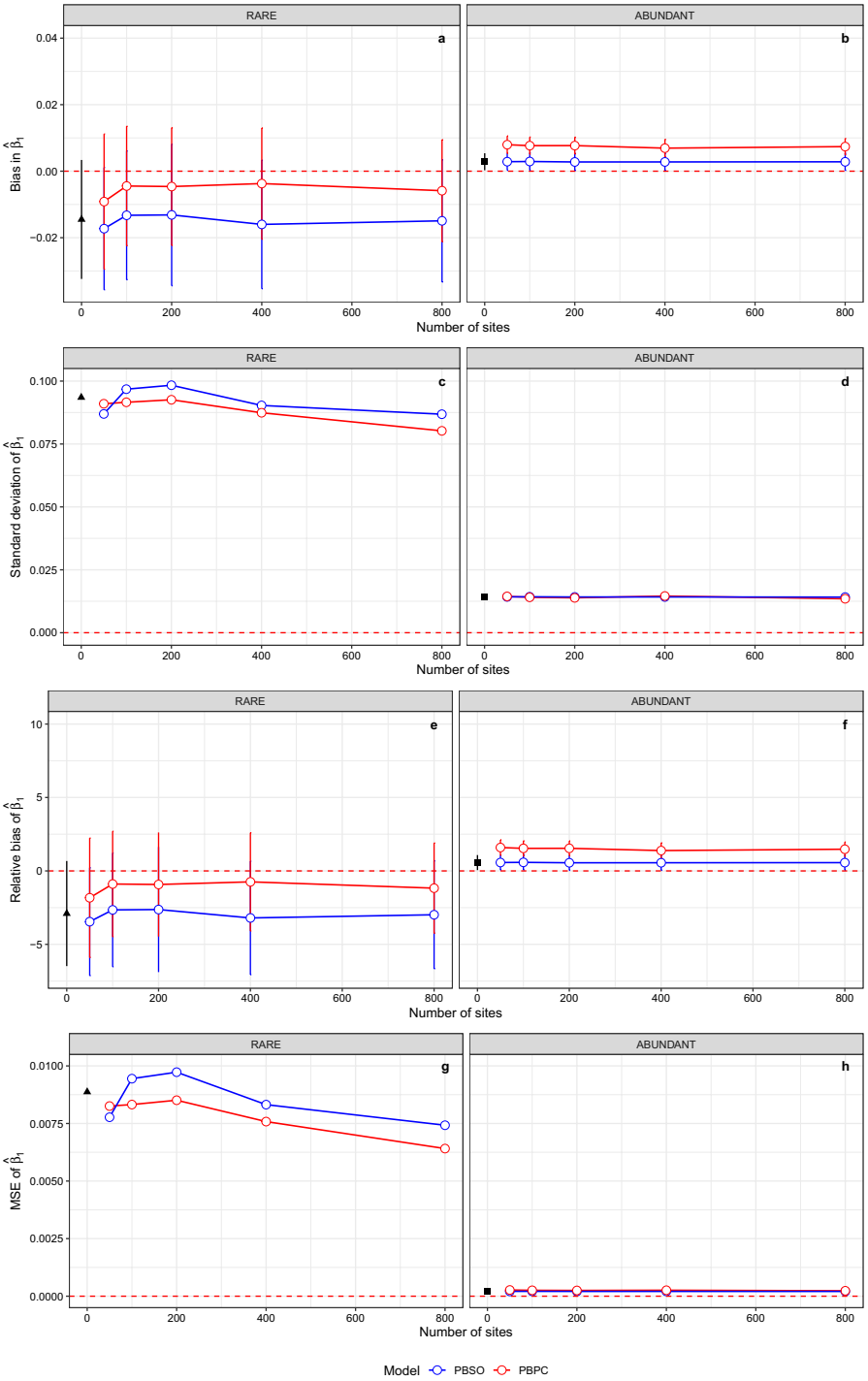

Fig. 5 Effects of the number of sites sampled in planned surveys on maximum-likelihood estimates of β_1. The red dashed line indicates bias equal zero. Black square and triangle are the results of the PO model. Error bars indicate 95% confidence intervals

SO data do not bring any gain in and accuracy and precision of $\hat{\beta}_0$ and $\hat{\beta}_1$. These findings do not corroborate those of Koshkina et al. (2017), while the simulation plan was almost identical to ours.

Comte and Grenouillet (2013) demonstrated that SO data can be used to estimate the distribution of species that are hard to detect. If the abundance within sample units is very low, the SO data can be utilised to lessen or eliminate the detrimental effect of detection errors, according to Dorazio (2014). With this statement, the PBSO model would be expected to give accurate estimates $\lambda(s)$ in the case of rare species. Our results show that, even with the rare species, the PBSO model did not offer any particular gain in terms of model performance. We presume that the SO data's main flaw is that they are not very informative on species abundance. Other empirical research, such as Pearce et al. (2001);

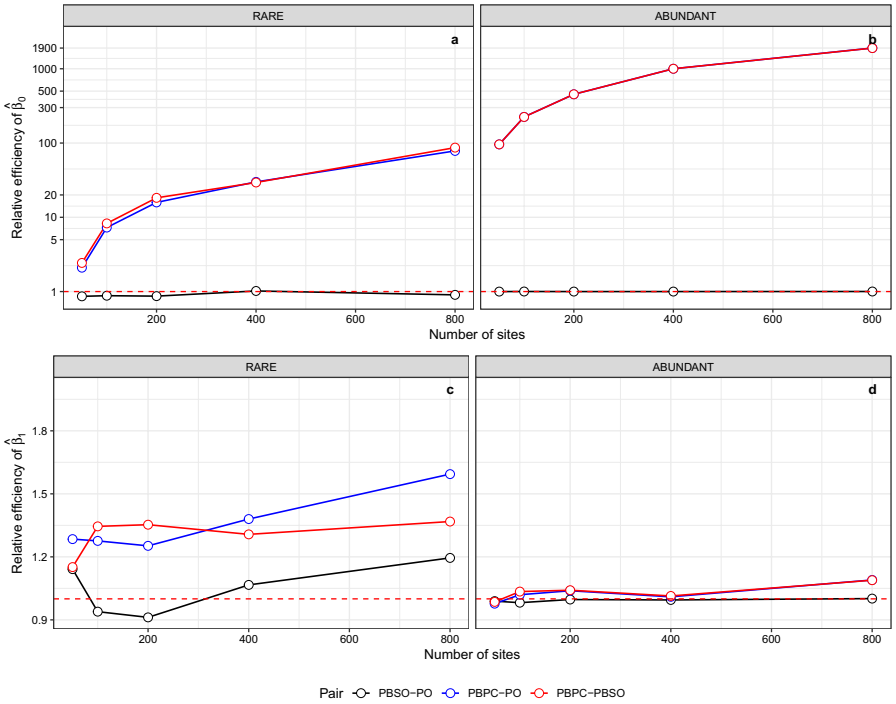

Fig. 6 Relative efficiency of SDMs as a function of the number of sites sampled during repeated planned surveys. The red dashed line indicates the threshold of 1

Nielsen et al. (2005); Jiménez-Valverde et al. (2009), have shown that putting a presence-absence design on a point-count design results in information loss, unless each geographical unit has only one observation. When each spatial unit has at least one observation, using SO data eliminates the potential to explore the impact of environmental factors on the distribution of species (Aarts and Fieberg 2012). Thus, because PC data potentially contain more information (related to species abundance) than SO data, the PBPC model may produce more accurate estimates than those produced by the PBSO model. Thus, we suspect that this may be among the reasons why the PBSO model did not perform well. Comte and Grenouillet (2013) showed that using SO data does not always result in a significant improvement above conventional model predictions.

Our findings are in line with those of Dorazio (2014). Indeed, the PBPC model has shown good performance by producing the most accurate and precise estimates of both β_0 and β_1. Then, PC data are highly valuable in increasing the accuracy and precision of $\hat{\beta}_0$ and $\hat{\beta}_1$. PO data are abundant but prone to sampling bias and imperfect detection. However, PC data are very sparse but allow for better accuracy and precision (Fithian et al. 2014). Using these two data types separately results in less accurate and precise estimates. This study demonstrates the benefit of using both types of data for better performance. Planned surveys are costly, and biologists interested in modeling species distribution are not usually fortunate enough to gather PC data during planned surveys. The PBPC model has the advantage of not necessitating a large number of PC observations to be collected during repeated planned surveys.

Effects of variation in detection probability and species rarity on the models' performance

Readers should not be confused with the results of this study. All the models studied in this work represent a particular approach to account for sampling bias and imperfect detection. The results of this study highlight the effects of the species rarity and variation of the detection probability on the performance of SDMs, accounting for sampling bias and imperfect detection. Our results have shown that the species being rare tends to increase the bias and imprecision of the $\hat{\beta}$. This could be due to the number of PO records used to fit these SDMs. Species rarity leads to small numbers of PO records and hence alters the ability of models to fully capture the species–environment relationship (Gábor et al. 2020). The results show that with the PO model, even $b(s)$ close to 1, the biases in the estimates are not zero. These biases and the imprecision of estimates increase as $b(s)$ decrease. We, therefore, deduce that not accounting for sampling and imperfect detection would imply much more biased estimates. For values of $b(s)$ greater than 0.25, the losses in terms of performance are non-zero, but remain small. With values of $b(s)$ lower than 0.25, we notice substantial losses in accuracy and precision depending on the SDMs. The PO and PBSO models are sensitive to the variation in detection probability during opportunistic surveys. In contrast, the PBPC model

is less sensitive. According to Welsh et al. (2013), poor detectability may lead to convergence issues of detectability models. That could be why SDMs produced poor estimates with $b(s)$ lower than 0.25. Another fact that would have arisen in the simulations is the relationship between the number of occurrences, species abundance, and detection probability. Low values of b(s) corresponded to low numbers of occurrences. With few data and of poor quality, it is therefore not surprising that there were such significant losses in performance.

Model limitations and suggestions for further works

For simplicity in simulations, we assumed that the detection probability at each surveyed site remained constant during repeated planned surveys, and we did not assess the effect of variation in detection probability during planned surveys to see how the PBPC and PBSO models performed in that situation. However, this assumption is not always met. For example, detection likelihood varies greatly among individuals within a species. Individual variances in levels of activity or movement, crypsis, and possibly even body size could all be signs of intrinsic heterogeneity (Cutrera et al. 2006; Karlsson et al. 2008; Karpestam et al. 2014). Studies of the performance of PC models on real data have also revealed that variation in individual detection probabilities can impair model performance. Then we can question if the models studied behave the same way in situations of heterogeneous detection. Veech et al. (2016) and Royle (2004) found that in the case of homogeneity in individual detection or constant detection probabilities of individuals, the Point-count model generally performs well with relatively little bias. Furthermore, Veech et al. (2016) showed that even when detection probability is heterogeneous, models perform well as long as the detection probability is reasonably high on average. Since the PBPC model is a combination of the PO model and PC model, one would expect this model to perform well even in the case of heterogeneity in detecting individuals during planned surveys. However, this should be verified in further studies.

Furthermore, PO data have the disadvantage of containing the spatial auto-correlation issue (non-independence). We cannot afford to disregard this bias in species distribution modeling. Environmental layers utilised as hypothetical predictive factors and coupled with the geographical records of species do not show problems of spatial auto-correlation, as assumed by species distribution models (Segurado et al. 2006; Cruz-Cárdenas et al. 2014). One will wind up with low accuracy in model coefficients if one ignore and do not avoid spatial auto-correlation, which will inflate type I errors (Dormann 2007; Cruz-Cárdenas et al. 2014). As a consequence, model predictions will be inaccurate where biological or population processes induce substantial auto-correlation in species distribution and this is not modelled. Consequently, the models we investigated should be evaluated for their predictive performance when the PO data are prone to spatial auto-correlation. Another unexplored aspect of this research is the advantage of including interaction terms in the models. More research is needed to investigate it.

Conclusion

This study assessed the performance of SDMs accounting for sampling bias and imperfect detection with respect to species abundance, variation in detection probability, and the number of sites visited in planned surveys. The results show that analysis of presence data alone cannot accurately predict species distribution. Other data are needed for better accuracy and precision. Analysis of presence-only data in conjunction with point-count data outperformed other approaches regardless of species abundance, as long as the probability of detection is at least 0.25 with mean values of detectability covariates. The minimum number of sites required for better performance varied with species abundance. At least 200 sites are needed for rare species, while 50 sites may be sufficient for abundant species. Because of the high cost of collecting these data, this study highlights the need to promote initiatives to collect species occurrence data with as little bias as possible.

Acknowledgements Support for this research was made possible through a capacity-building competitive grant Training Next Generation of Scientists (Grant #RU/2020/GTA/DRG/036) provided by Carnegie Cooperation of New York through the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM). The Université Evangélique en Afrique (UEA) is acknowledged for manifold support to this work through the University project on improving research, and teaching quality funded by Pain pour le Monde (Project A-COD-2018-0383). Authors greatly appreciated technical support of the team of the Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF). A. Belarmain Fandohan was supported by Georg Forster-Hermes Research Fellowship programs of the Alexander von Humboldt Foundation: postdoctoral fellowship number 3.4-BEN/1155509 STP, return fellowship no. 3.4 - RKS - BEN/1155509 and equipment subsidy no 3.6-BEN/1155509.

Authors' contributions Conceptualization: YM, ABF, and ACM;Methodology: YM, ABF, and ACM; Writing-original draft preparation: YM,ABF, ACM, and IAS; Writing-review and editing: YM, ABF, ACM, IAS, andRGK; Supervision: ABF and RGK. All authors read and approved the finalmanuscript.

Funding Not applicable.

Availability of data and materials The datasets used and/or analyzed-during the current study are available from the corresponding author onreasonable request.

Declaration

Conflict of interest The authors declare no competing interests.

References

Aarts G, Fieberg J (2012) Comparative interpretation of count, presence-absence and point methods for species distribution models. Methods Ecol Evol 3(1):177–187. https://doi.org/10.1111/j.2041-210X.2011.00141.x

Barry S, Elith J (2006) Error and uncertainty in habitat models. J Appl Ecol 43(3):413–423. https://doi.org/10.1111/j.1365-2664.2006.01136.x

Brotons L, Thuiller W, Araújo MB et al (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27(4):437–448. https://doi.org/10.1111/j.0906-7590.2004.03764.x

Comte L, Grenouillet G (2013) Species distribution modelling and imperfect detection: comparing occupancy versus consensus methods. Diversity Distributions 19(8):996–1007. https://doi.org/10.1111/ddi.12078

Crall AW, Jarnevich CS, Panke B et al (2013) Using habitat suitability models to target invasive plant species surveys. Ecol Appl 23(1):60–72. https://doi.org/10.1890/12-0465.1

Cruz-Cárdenas G, López-Mata L, Villaseñor JL et al (2014) Potential species distribution modeling and the use of principal component analysis as predictor variables. Revista Mexicana de Biodiversidad 85(1):189–199. https://doi.org/10.7550/rmb.36723

Cutrera AP, Antinuchi CD, Mora MS et al (2006) Home-range and activity patterns of the south american subterranean rodent ctenomys talarum. J Mammal 87(6):1183–1191. https://doi.org/10.1644/05-MAMM-A-386R1.1

De Siqueira MF, Durigan G, de Marco Junior P et al (2009) Something from nothing: using landscape similarity and ecological niche modeling to find rare plant species. J Nat Conserv 17(1):25–32. https://doi.org/10.1016/j.jnc.2008.11.001

Dorazio RM (2007) On the choice of statistical models for estimating occurrence and extinction from animal surveys. Ecology 88(11):2773–2782. https://doi.org/10.1890/07-0006.1

Dorazio RM (2012) Predicting the geographic distribution of a species from presence-only data subject to detection errors. Biometrics 68(4):1303–1312. https://doi.org/10.1111/j.1541-0420.2012.01779.x

Dorazio RM (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecol Biogeography 23(12):1472–1484. https://doi.org/10.1111/geb.12216

Dormann CF (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecol Biogeography 16(2):129–138. https://doi.org/10.1111/j.1466-8238.2006.00279.x

Elith J, Burgman MA, Regan HM (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. Ecol Modell 157(2–3):313–329. https://doi.org/10.1016/S0304-3800(02)00202-8

Fei S, Liang L, Paillet FL et al (2012) Modelling chestnut biogeography for american chestnut restoration. Diversity Distribut 18(8):754–768. https://doi.org/10.1111/j.1472-4642.2012.00886.x

Fithian W, Hastie T (2013) Finite-sample equivalence in statistical models for presence-only data. Ann Appl Stat 7(4):1917–1939. https://doi.org/10.1214/13-AOAS667

Fithian W, Elith J, Hastie T et al (2014) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol Evol 6(4):424–438. https://doi.org/10.1111/2041-210X.12242

Fuller T, Morton DP, Sarkar S (2008) Incorporating uncertainty about species' potential distributions under climate change into the selection of conservation areas with a case study from the arctic coastal plain of alaska. Biol Conserv 141(6):1547–1559. https://doi.org/10.1016/j.biocon.2008.03.021

Guillera-Arroita G (2017) Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography 40(2):281–295. https://doi.org/10.1111/ecog.02445

Guillera-Arroita G, Lahoz-Monfort JJ, MacKenzie DI et al (2014) Ignoring imperfect detection in biological surveys is dangerous: A response to "fitting and interpreting occupancy models." PLoS ONE 9(7):1–14. https://doi.org/10.1371/journal.pone.0099571

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8(9):993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Gábor L, Moudrý V, Barták V et al (2020) How do species and data characteristics affect species distribution models and when to use environmental filtering? Int J Geographical Inform Sci 34(8):1567–1584. https://doi.org/10.1080/13658816.2019.1615070

Hastie T, Fithian W (2013) Inference from presence-only data; the ongoing controversy. Ecography 36(8):864–867. https://doi.org/10.1111/j.1600-0587.2013.00321.x

Hefley T, Tyre A, Baasch D et al (2013) Nondetection sampling bias in marked presence-only data. Ecol Evol 3(16):5225–5236. https://doi.org/10.1002/ece3.887

Hefley TJ, Brost BM, Hooten MB (2017) Bias correction of bounded location errors in presence-only data. Methods Ecol Evol 8(11):1566–1573. https://doi.org/10.1111/2041-210X.12793

Jiménez-Valverde A, Diniz F, Eduardo BdA et al (2009) Species distribution models do not account for abundance: The case of arthropods on terceira island. Annales Zoologici Fennici 46(6):451–464. https://doi.org/10.5735/086.046.0606

Karlsson M, Caesar S, Ahnesjö J et al (2008) Dynamics of colour polymorphism in a changing environment: Fire melanism and then what? Oecologia 154(4):715–724. https://doi.org/10.1007/s00442-007-0876-y

Karpestam E, Merilaita S, Forsman A (2014) Body size influences differently the detectabilities of colour morphs of cryptic prey. Biol J Linnean Soc 113(1):112–122. https://doi.org/10.1111/bij.12291

Kearney MR, Wintle BA, Porter WP (2010) Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. Conservation Lett 3(3):203–213. https://doi.org/10.1111/j.1755-263X.2010.00097.x

Kellner KF, Swihart RK (2014) Accounting for imperfect detection in ecology: A quantitative review. PLoS One 9(10):1–8. https://doi.org/10.1371/journal.pone.0111436

Koshkina V, Wang Y, Gordon A et al (2017) Integrated species distribution models: combining presence-background data and site-occupany data with imperfect detection. Methods Ecol Evol 8(4):420–430. https://doi.org/10.1111/2041-210X.12738

Kremen C, Cameron A, Moilanen A et al (2008) Aligning conservation priorities across taxa in madagascar with high-resolution planning tools. Science 320(5873):222–226. https://doi.org/10.1126/science.1155193

Lahoz-Monfort JJ, Guillera-Arroita G, Wintle BA (2014) Imperfect detection impacts the performance of species distribution models.

Global Ecol Biogeography 23(4):504–515. https://doi.org/10.1111/geb.12138
Li X, Wang Y (2013) Applying various algorithms for species distribution modelling. Integrative Zool 8(2):124–135. https://doi.org/10.1111/1749-4877.12000
Mackenzie DI, Royle JA (2005) Designing occupancy studies: general advice and allocating survey effort. J Appl Ecol 42(6):1105–1114. https://doi.org/10.1111/j.1365-2664.2005.01098.x
MacKenzie DI, Nichols JD, Lachman GB et al (2002) Estimating site occupancy rates when detection probabilities are less than one. Ecology 83(8):2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2
Nielsen SE, Johnson CJ, Heard DC et al (2005) Can models of presence-absence be used to scale abundance? two case studies considering extremes in life history. Ecography 28(2):197–208. https://doi.org/10.1111/j.0906-7590.2005.04002.x
Pearce J, Ferrier S, Scotts D (2001) An evaluation of the predictive performance of distributional models for flora and fauna in north-east new south wales. J Environ Manag 62(2):171–184. https://doi.org/10.1006/jema.2001.0425
Peterson AT, Soberón J, Pearson RG et al (2011) Ecological Niches and Geographic Distributions, Monographs in Population Biology (MPB-49). Princeton University Press, Princeton,. https://doi.org/10.1515/9781400840670
Phillips SJ, Dudik M, Elith J, et al (2009) Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. Ecological Applications 19(1):181–197. 0.1890/07-2153.1
R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/
Royle JA (2004) N-mixture models for estimating population size from spatially replicated counts. Biometrics 60(1):108–115. https://doi.org/10.1111/j.0006-341X.2004.00142.x
Segurado P, Araújo MB, Kunin WE (2006) Consequences of spatial autocorrelation for niche-based models. J Appl Ecol 43(3):433–444. https://doi.org/10.1111/j.1365-2664.2006.01162.x
Veech JA, Ott JR, Troy JR (2016) Intrinsic heterogeneity in detection probability and its effect on n-mixture models. Methods Ecol Evol 7(9):1019–1028. https://doi.org/10.1111/2041-210X.12566
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/10.1214/10-AOAS331
Warton D, Renner I, Ramp D (2013) Model-based control of observer bias for the analysis of presence-only data in ecology. PLoS ONE 8(11):e79,168. https://doi.org/10.1371/journal.pone.0079168
Welsh AH, Lindenmayer DB, Donnelly CF (2013) Fitting and interpreting occupancy models. PLoS ONE 8(1):1–21. https://doi.org/10.1371/journal.pone.0052015