Ecological Informatics 57 (2020) 101086

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Correcting the effect of sampling bias in species distribution modeling – A new method in the case of a low number of presence data

Yi Moua[a,b,*], Emmanuel Roux[b], Frédérique Seyler[b], Sébastien Briolant[c,d]

[a] ESPACE-DEV, UMR 228 (IRD, UM, UR, UA, UG), Université de Guyane, Cayenne, French Guiana
[b] ESPACE-DEV (IRD, Univ Montpellier, Univ La Réunion, Univ Guyane, Univ des Antilles), Institut de Recherche pour le Développement (IRD), Montpellier, France
[c] Unité de Parasitologie et Entomologie, Département des Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, 19-21 Boulevard Jean Moulin, 13005 Marseille, France
[d] IRD, AP-HM, SSA, UMR Vecteurs – Infections Tropicales et Méditerranéennes (VITROME), IHU - Méditerranée Infection, Aix Marseille Université, 19-21 bd Jean Moulin, 13385 Marseille Cedex 5, France

ARTICLE INFO

Keywords:
Sampling bias
Data scarcity
Species distribution models
Maxent

ABSTRACT

Species distribution models that only require presence data provide potentially inaccurate results due to sampling bias and presence data scarcity. Methods have been proposed in the literature to minimize the effects of sampling bias, but without explicitly considering the issue of sample size.

A new method developed to better take into account environmental biases in a context of data scarcity is proposed here. It is compared to other sampling bias correction methods primarily used in the literature by analyzing their absolute and relative impacts on model performances.

Results showed that the number of presence sites is critical for selecting the applicable method. The method proposed was regularly placed in the first or second rank and tends to be more proficient than other methods in the context of presence site scarcity (<100). It tends to improve results regarding environment-based performance indexes. Eventually, its parametrization, requiring background knowledge on species bio-ecology, appears to be more robust and convenient to perform than those based on geographical criteria.

1. Background

Species distribution models (SDMs) are widely used in ecology to predict species habitat distribution in space and time from the geographical coordinates of species occurrences and environmental features that characterize species habitats (Pearson et al., 2007). Each environmental feature corresponds, in practice, to a raster layer, i.e., a grid of cells (or *pixels*), and each cell being associated with a numerical or categorical value. Species distribution models can predict future changes in species distributions regarding climate or habitat change scenarios (Alimi et al., 2015).

Numerous SDM approaches are proposed in the literature, e.g., generalized linear model (GLM, Guisan et al., 2002), generalized additive model (GAM, Guisan et al., 2002), artificial neural networks (Pearson et al., 2002), support vector machine (Guo et al., 2005), HABITAT (Walker and Cocks, 1991), genetic algorithms for rule-set production (GARP, Stockwell, 1999), and Maxent (Phillips et al., 2006). Some of them require both presence and absence data, whereas others – presence-only, presence-background, and presence-pseudoabsence models – require only species presence information and offer a significant advantage because of the difficulty to obtain reliable absence information (Hirzel et al., 2002; Peterson et al., 2011).

Among models exploiting only presence data, it has been shown that presence-background models were more discriminant (Peterson et al., 2011). Background sites are supposed to reflect all the environmental conditions present in the study area and are generally randomly selected. Maxent is a popular and widely used presence-background SDM (Elith et al., 2006). However, similar to all presence models, Maxent is very sensitive to sampling bias, as a result of different sampling efforts from one environmental context to another (Elith et al., 2011; Phillips et al., 2009).

Sampling bias can be due to heterogeneous geographical sampling by, notably, favoring easily accessible areas, and can induce a significant environmental bias. Biased input data can lead to incorrect model outputs. Indeed, a model built with biased data corresponds more to a model of the survey effort than a model of the actual species habitat distribution (Phillips et al., 2009). Theoretically, sampling bias is minimized when both the presence and the background datasets are

* Corresponding author.
E-mail addresses: moua.yi@yahoo.fr (Y. Moua), emmanuel.roux@ird.fr (E. Roux), frederique.seyler@ird.fr (F. Seyler).

https://doi.org/10.1016/j.ecoinf.2020.101086
Received 28 September 2019; Received in revised form 20 March 2020; Accepted 21 March 2020
Available online 01 April 2020

equally biased (Phillips et al., 2009). This is achieved by transforming the model input data using four approaches which can be categorized according to two criteria: i) the dataset concerned with the transformation (presence sites or background sites); and ii) the implementation spaces of the transformation (geographical or environmental space). When the presence dataset is considered, the authors aim at modifying its distribution to tend toward a uniform distribution, and the background set is obtained via a uniform random selection. The implementation space of such an approach is either the geographical space (see for example Boria et al., 2014; Kramer-Schadt et al., 2013) or environmental feature space (see Fourcade et al., 2014; Varela et al., 2014). This approach seems to be effective, but requires many presence sites (Fourcade et al., 2014). When the background set is considered, it is built according to a selection bias that reflects the same sampling bias than that existing in the set of presence sites. This approach does not require many presence sites, but it requires the estimation of the sampling bias (Phillips et al., 2009), which is not trivial. In the same way, such an approach can be considered in either the geographical space (Elith et al., 2011; Fourcade et al., 2014) or environmental feature space (Hill and Terblanche, 2014; Moua et al., 2017).

Moreover, the number of presence sites used to build the SDM is often limited because of the scarcity of the species or the inadequate number of species records. Thus, many datasets suffer not only from sampling bias, but also from a low number of presence sites.

Some authors have performed comparative studies of the methods intended to correct the sampling bias effects (Fourcade et al., 2014, Kramer-Schadt et al., 2013, Varela et al., 2014). The study by Fourcade et al. (2014) compared and evaluated correction method performances with different bias types and intensities. However, these studies have not compared presence sites filtering and biased background approaches in the case of a small number of presence sites. As mentioned above, a background-based correction method appears to be adapted to a small number of presence sites more. Theoretically, the background-based correction method implemented in the environmental space, as proposed by Moua et al. (2017), handles environmental bias and is less sensitive to a low number of presence sites. This study aims to empirically compare this proposed correction method with other existing ones – covering the four main categories previously described – in the specific context of presence site scarcity.

2. Materials and methods

2.1. Correction methods

Four correction methods were considered to represent the four main approaches described in the introduction. Only the method proposed by Moua et al. (2017) is fully described hereafter since it has not yet been described in detail in the literature. The three other methods are briefly described; the reader is invited to consult related references for more details.

Two methods consisting of estimating the sampling effort to adapt the background set were implemented.

(1) The method initially proposed by Moua et al. (2017) (referred to as ***BGenv*** hereafter) is based on the estimation of the sampling effort within the environmental space. The principle, the implementation of which is detailed below, is as follow: the sampling effort associated with a given pixel i of the study area corresponds to the ratio of the number of sampled pixels over the total number of pixels, within the environmental neighborhood of i defined by using a Gaussian-like membership function.

Firstly, all pixels of the study area are represented in the environmental variable space. This is accomplished by performing a Principal Component Analysis (PCA), a Multiple Correspondence Analysis (MCA) or a Factor Analysis of Mixed Data (FAMD) (Pages, 2004) depending on the fact that the set of environmental variables is, respectively, exclusively numerical, exclusively categorical or a mix of both variables. The factor analysis allows the representation of the pixels within a Euclidean, orthonormal space defined from the whole set of environmental variables.

The membership degree of a pixel j to the neighborhood of pixel i, denoted w_{ij}, is defined by a Gaussian-like membership function:

$$w_{ij} = 0.5\text{^}(d_{ij}/D_{min})^2 \quad (1)$$

where d_{ij} is the Euclidean distance between i and j in the factorial space, and D_{min} the threshold distance over which j does not significantly belong to the environmental neighborhood of i, i.e., over which $w_{ij} < 0.5$. The membership degree w_{ij} has the following properties:

- $w_{ij} \in]0, 1]$;
- $w_{ij} = 1$ if $d_{ij} = 0$;
- $w_{ij} < 0.5$ if $d_{ij} > D_{min}$

The parameter D_{min} is set from a priori knowledge of the species bio-ecology. Particularly, given P, the set of pixels where the species was observed and U, the set of pixels where the species is known to be absent, D_{min} can be defined as follows:

$$D_{min} = \min_{p\in P, u\in U}(d_{pu}) \quad (2)$$

The general concepts of *environmental space* and *environmental neighborhood* are schematically represented in Fig. 1.

Given X, the set of pixels of the study area, and S the set of sampled pixels, the related sampling effort for pixel i, denoted z_i, is defined as:

$$z_i = \text{sum}_{j\in S} w_{ij} / \text{sum}_{k\in X} w_{ik} \quad (3)$$

When using a species target group (Phillips et al., 2009), the sets P and S will be different. However, P and S will be the same if only the target species is used to estimate the sampling effort. The relative sampling effort is computed for each pixel of the study area and corresponds to the sampling bias within the environmental space. The resulting map is then used to bias the random selection of background points. The higher the sampling effort of a cell, the greater is the likelihood to select a background site in it. The algorithm for the construction of the biased background set is shown in pseudo-code in a supplementary file.

Hereafter, the set of background sites selected with this method is denoted as bg_C_{env}.

(2) For each pixel, Elith et al. (2011) proposed to estimate the sampling effort, for any pixel i, by the ratio of the number of presence sites and the number of terrestrial cells, in the *geographical* neighborhood of i defined by a Gaussian-like membership function (method denoted as ***BGgeo***). This sampling estimate based on geographical criteria assumes that the habitat characteristics are similar within the geographical neighborhood. Background sites are selected by using a weighted random selection, and the weight distribution is defined by the estimated sampling effort.

Hereafter, the resulting background set is denoted as bg_C_{geo}.

The two methods chosen to implement the approach consisting of manipulating presence sites, in order to make uniform the presence data distribution in the study area, are:

(3) the method applied by Boria et al. (2014) (denoted ***Fgeo*** hereafter), based on a *geographical* filter ensuring a more uniform distribution of presence points within the geographical space, by reducing the spatial aggregation of the samples. It consists of removing capture sites located below a given distance from the others. This distance is defined according to the home range of the species (Boria et al., 2014).

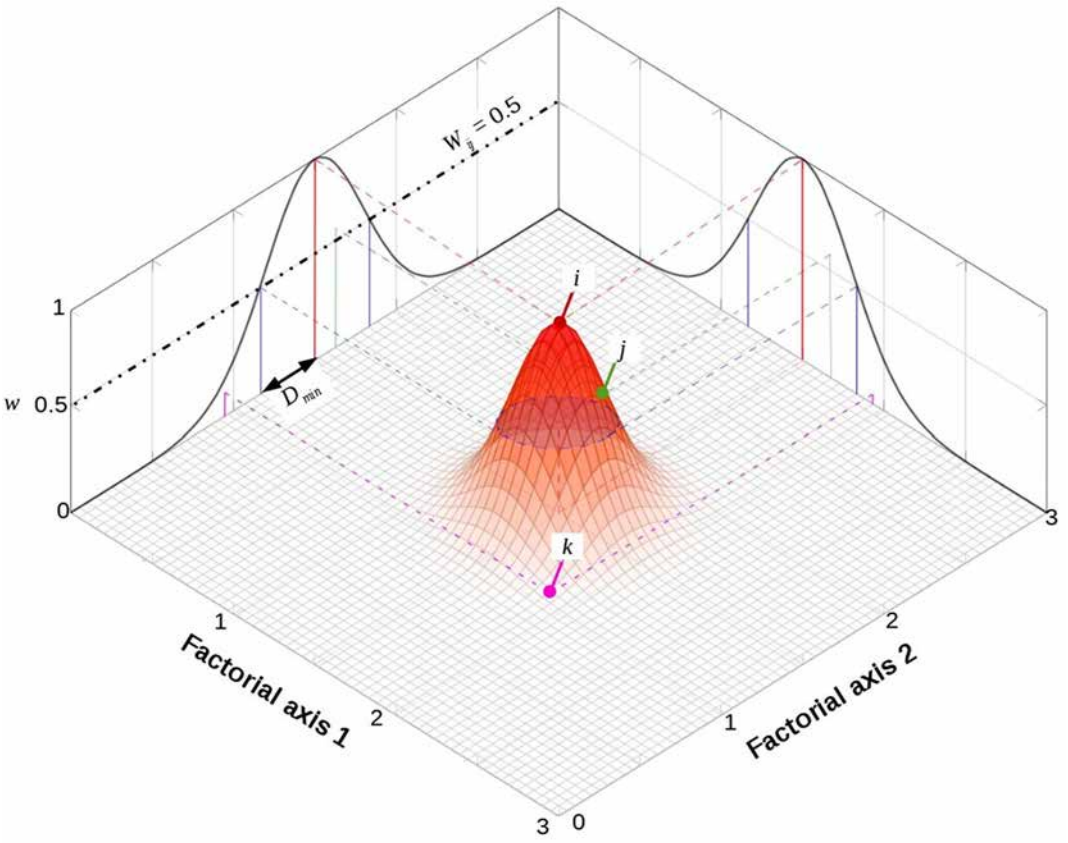

Fig. 1. Neighborhood of a pixel *i* in the environmental space represented by the first and second axes of the factorial analysis. The environmental neighborhood of point *i* is represented by the Gaussian function. The blue lines define the limit of the neighborhood of *i*. Only point *j* is located above these lines. Thus *j* is in the neighborhood of *i* in the first factorial plane (Adapted from Moua et al., 2017). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

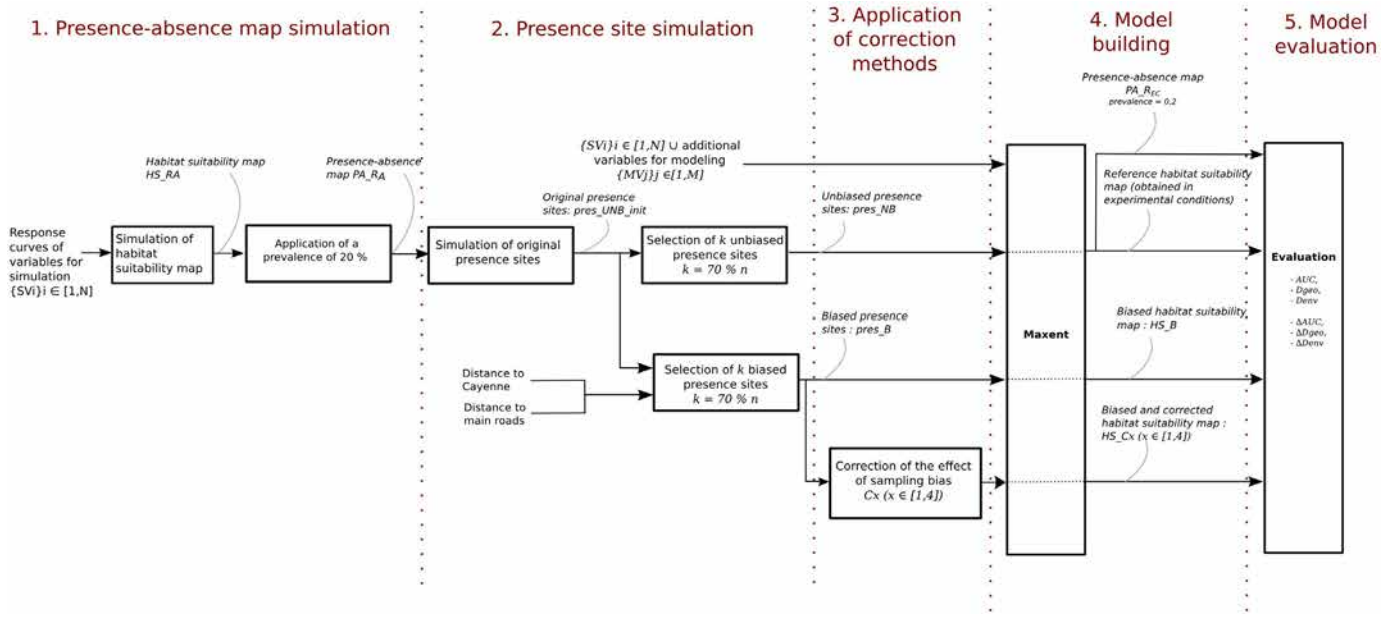

Fig. 2. Overall processing chain for data simulation, bias application, and the subsequent application, evaluation, and comparison of the correction methods in the framework of Maxent modeling.

Hereafter, the set of the remaining presence sites, defining the model input dataset, is referred to as $pres\_C_{fgeo}$.

(4) The method proposed by Fourcade et al. (2014) (referred as to ***Fenv***), corresponding to a filter method based on *environmental* criteria. This involves performing a principal component analysis (PCA) on the environmental data associated with presence sites. Subsequently, a cluster analysis is performed within the resulting

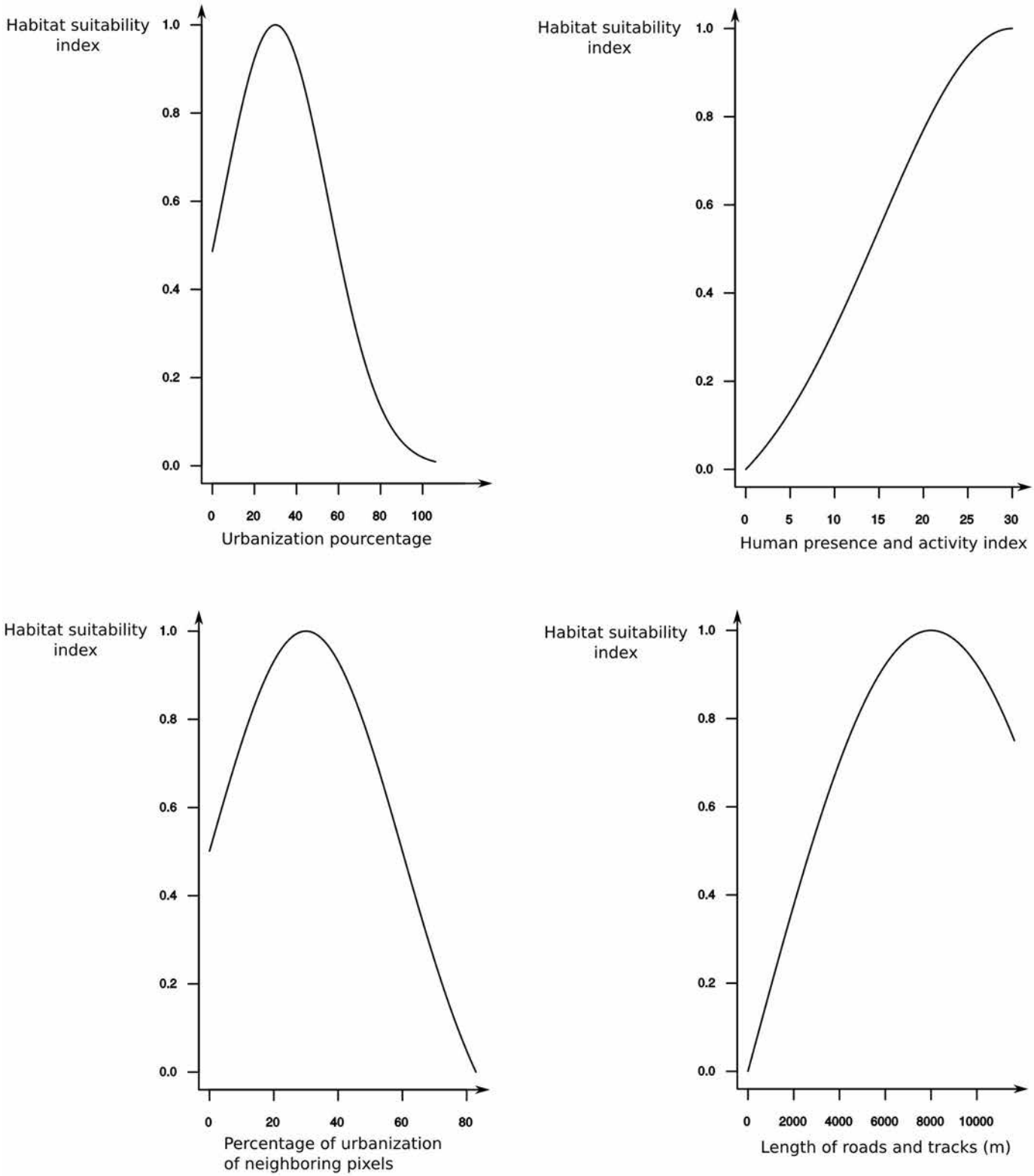

Fig. 3. Simulated response curves of the habitat descriptors of *Anopheles darlingi*.

PCA space, with the number of clusters set to half of the presence site number. Subsequently, only one occurrence is randomly chosen per cluster.

The set of selected presence sites is denoted as $pres\_C_{fenv}$ hereafter. Table S1 (Supplementary file) summarizes the four approaches previously described.

2.2. Simulation scheme for method evaluation

Virtual species data were simulated using the R software environment (R Development Core Team, 2018) and the R package *virtualspecies* (Leroy et al., 2015) to obtain baseline data and evaluate the different methods in both absolute and relative ways. The simulated datasets were generated building on: i) knowledge on the bioecology of

an actual species: *Anopheles* (Nyssorhynchus) *darlingi* Root (Diptera: Culicidae), the primary malaria vector in the Amazonian region; ii) the environmental data related to French Guiana, a French overseas region located in South America between Suriname and Brazil, where malaria is endemic-epidemic. Knowledge and data used to produce the simulated datasets were retrieved mostly from the study by Moua et al. (2017). They are briefly described here, but readers are invited to refer to Moua et al. (2017) for details on how they were generated.

The general simulation methodology was inspired by Fourcade et al. (2014). It comprised five steps that can be summarized as follows (please see Fig. 2 and following paragraphs for implementation details):

1. A simulated presence-absence map was produced;
2. *n* presence sites were simulated through a uniform random selection within the presence area of the presence-absence map. Among them, we randomly selected, on the one hand, *k* presence sites called *unbiased* presence sites and, on the other hand, *k biased* presence sites, where $k \in \{20, 50, 100, 150, 200\}$;
3. The four correction methods described above were applied to each set of biased presence sites. The methods were parameterized as per the current knowledge of the bioecology and behavior of the species of interest, mainly in order to be placed in the application conditions met by any user of these correction methods, and also because the simulation of the effects of the parameterization would have made the results too complex to be interpreted;
4. Maxent was applied to the unbiased data, biased data, and biased-corrected data, and allowed building, respectively, an unbiased model, a biased model, and a corrected model, the outputs of which were habitat suitability maps. A reference presence-absence map was also produced from the unbiased model. Thus, all items compared in the evaluation phase are outputs of the Maxent algorithm, ensuring that only the effects of sampling bias correction methods – and not those specific to the Maxent algorithm – are evaluated;
5. Produced maps from the biased and biased-corrected data were compared with maps obtained from the unbiased data.

This method was replicated 100 times for each value of *k*.

Thus, 500 unbiased models, 500 biased models, and 2000 biased-corrected models (four correction methods × 500 biased models) were built.

2.3. Simulated presence-absence map

The most contributive habitat descriptors to the habitat suitability of *An. darlingi* in Moua et al. (2017), and those whose influence has been demonstrated in the literature were chosen to build a presence-absence map of the species at a 1 km spatial resolution. Selected habitat descriptors were the urbanization percentage in the neighboring pixels (derived from Human Footprint, de Thoisy et al., 2010), the density of roads in m/km^2 (derived from the BD TOPO®, a database on roads and tracks of the French National Institute of Geographic and Forestry Information, IGN), human presence and activities that non-permanently alter natural environment (derived from Human Footprint, de Thoisy et al., 2010), and the land use classes *woodland savanna*, *swamps*, *mixed high and open forest,* and *open forest* from the land use map (Gond et al., 2011). The response curves associated with these habitat descriptors were rebuilt (see Fig. 3).

A normalized habitat suitability index map was obtained by combining these response curves with an additive function to obtain a less restrictive definition than a multiplicative one and obtain a habitat suitability map as similar as possible to the one obtained by Moua et al. (2017) (see Fig. 4). A logistic function deriving the presence probability from the habitat suitability was used (see Leroy et al., 2015) to convert this map into a presence-absence map, and thereafter, a prevalence rate of 0.2 (20%) was applied (see Fig. 4).

2.4. Simulated unbiased and biased presence sites

Different steps were required to obtain unbiased and biased sets of presence sites named *pres*_UNB and *pres_B*, respectively.

First, a set of *n* unbiased presence sites named *pres_UNB_init* was generated by uniformly and randomly selecting the sites in the species presence area provided by the presence-absence map. Subsequently, *k* unbiased and *k* biased capture sites (constituting, respectively, the sets *pres_UNB* and *pres_B*) were selected from the set *pres_UNB_init*. *n* was chosen for each *k* so that *pres_UNB* and *pres_B* have a significant number of sites in common. In practice, *k* was set to 0.7*n,* allowing *pres_UNB* and *pres_B* to have at least 40% of the initial *pres_UNB_init* set in common.

pres_UNB was selected randomly and uniformly among *pres_UNB_init*, whereas the random selection of the *pres_B* sites was weighted by a simulated sampling bias inversely proportional to the distance to the administrative capital (Cayenne) and the main road of French Guiana (see Fig. 4). In fact, in French Guiana, the sampling was generally performed near urban areas, and in villages where access was facilitated by proximity to the main road located along the coastline (Moua et al., 2017). An example of *pres_UNB* and *pres_B* sites, for $k = 20$, were mapped in Fig. 4.

One hundred replicates were generated for each set *pres_UNB* and *pres_B* and for each value of *k*.

2.5. Parameterization of the correction methods

The knowledge found in the literature and provided by experts of the ecology of the *An. darlingi* species were used to parameterize the correction methods described in paragraph 2.1 (Table S1 summarizes the key aspects of such parameterizations).

2.5.1. BGenv

In this study, as in Moua et al. (2017), D_{min} was calculated with respect to the knowledge that highly urbanized areas are known to be not suitable for *An. darlingi* (see, for example, De Castro et al., 2006; Stefani et al., 2013). Consequently, we stated that a pixel associated with *An. darlingi* presence cannot belong to the environmental neighborhood of a highly urbanized pixel and, reciprocally, that a pixel considered to be highly urbanized cannot belong to the environmental neighborhood of a pixel where *An. darlingi* was observed. The distance D_{min} was consequently set as the minimum distance between the pixels where *An. darlingi* was observed and the highly urbanized pixels, within the environmental feature space. A pixel was considered to be highly urbanized if it belonged to the Landscape type *Urban* and if its eight neighboring pixels showed a mean urbanization percentage higher than or equal to 50% (see hereafter and Moua et al., 2017, for details on the environmental layers and their construction).

It is worth noting that in Moua et al. (2017), a species target group (Phillips et al., 2009) was considered to define the set of sampled pixels *S* and estimate the relative sampling effort with Eq. (3). Here, only the (simulated) presence sites of the species of interest (*P*) were used to estimate the relative sampling effort.

2.5.2. BGgeo

In this study, the standard deviation of the Gaussian function defining the geographical neighborhood was set at 7000 m, which is the maximum distance moved of *An. darlingi* found in the literature (Charlwood and Alectrim, 1989).

2.5.3. Fgeo

The distance *r*, used to filter the presence sites and thus make the presence site distribution uniform in the geographical space, was set at 7000 *m* for the same reason as for the previous method.

2.5.4. Fenv

In this study, a factorial analysis of mixed data (FAMD) (Pages,

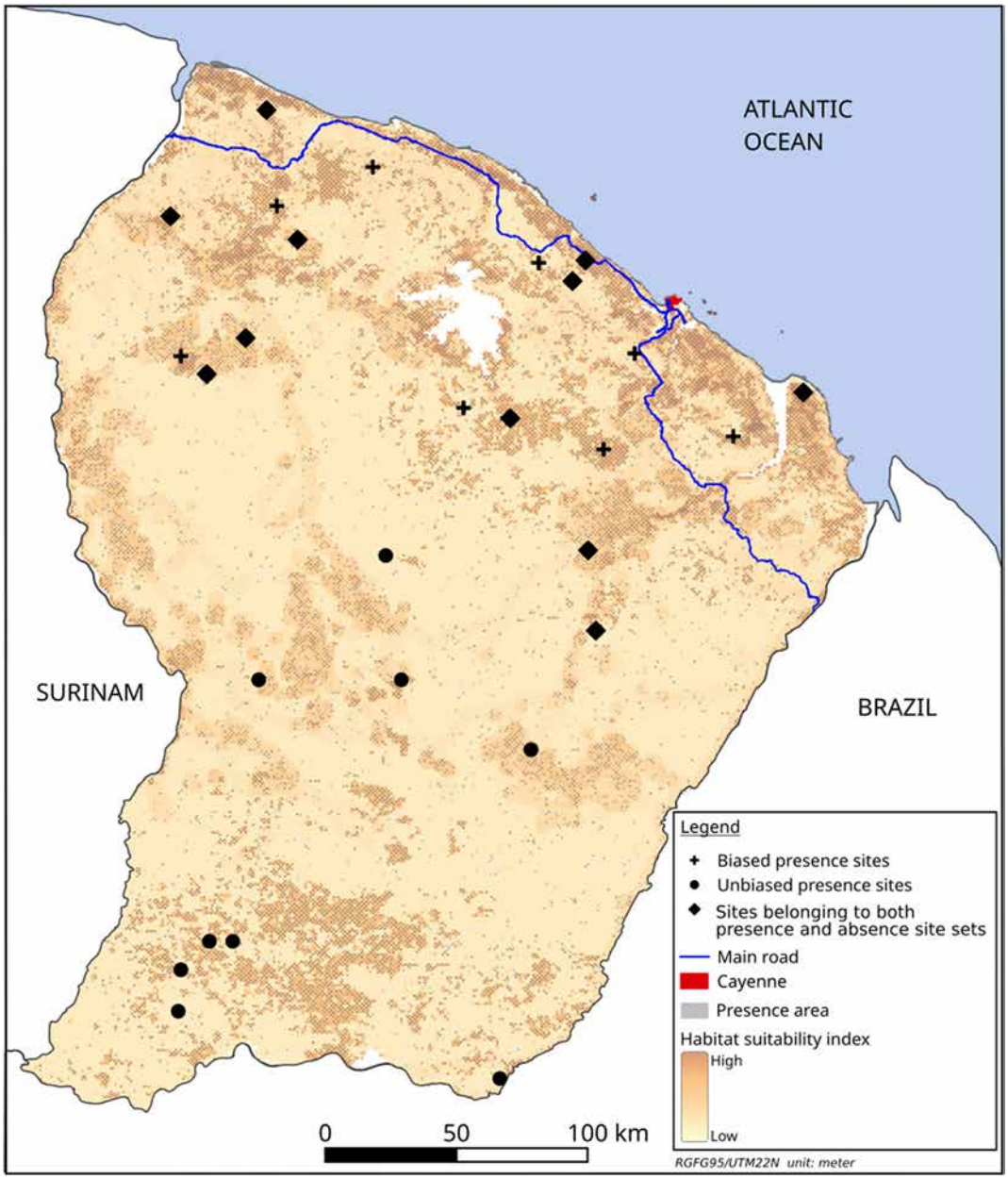

Fig. 4. Map of unbiased and biased presence sites, for $k = 20$. Cayenne city and the main road, which were used to weight biased presence sites selection, are represented.

2004) was applied, instead of the PCA applied by Fourcade et al. (2014), because of the presence of categorical and continuous environmental predictors. A hierarchical ascendant classification was used, and the number of selected clusters was set to half of the presence site number as in Fourcade et al. (2014). Subsequently, one capture site was randomly selected from each cluster.

2.6. Environmental predictors and species distribution modeling

The seven environmental predictors used in Moua et al., 2017 were considered to build models with Maxent (see complementary details on the environmental variables in Supplementary file, Table S2):

- Geomorphological landscapes (Guitet et al., 2013);
- Geomorphological landforms (Guitet et al., 2013);
- Landscape types (Gond et al., 2011);
- Minimum altitude (from the STRM, NASA);
- Length of roads and tracks outside of urban areas (from the topographic database BD TOPO®, IGN);
- Percentage of urbanization of neighboring pixels (derived from Human Footprint, de Thoisy et al., 2010);
- Human activities that non-permanently alter the natural environment (derived from Human Footprint, de Thoisy et al., 2010).

All raster layers had a resolution of 1 km^2.

Maxent aims to estimate the habitat suitability of a given species by determining the probability distribution of maximum entropy while adhering to the constraints derived from occurrences data (see Phillips et al., 2006, Elith et al., 2011, and Moua et al., 2017 for details). Maxent version 3.3.3 k was used within the R environment.

Unbiased models were built from unbiased data (*pres_UNB*). These models allowed to obtain the "*reference*" habitat suitability maps in *experimental conditions*, HS_R_{EC}. Reference presence-absence maps (PA_R_{EC}) were obtained by defining a prevalence rate of 0.2 from these

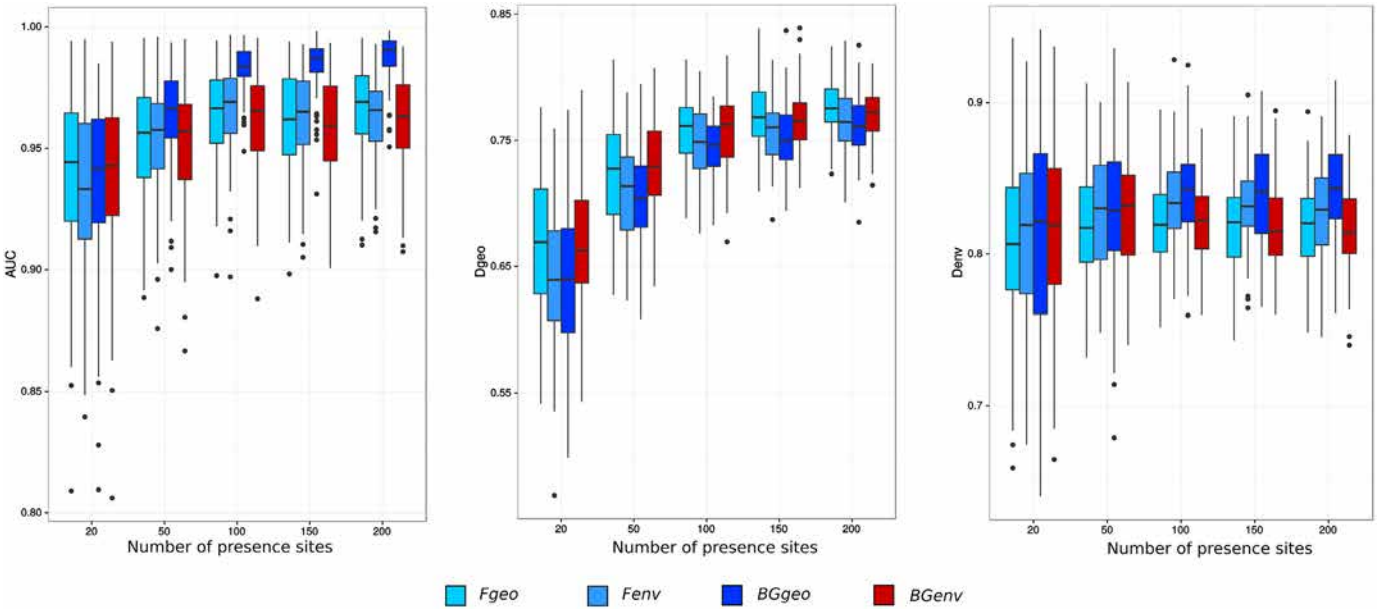

Fig. 5. Boxplots of *AUC*, *Dgeo*, and *Denv* results as a function of the correction method and the number of presence sites.

maps.

Biased models were built from biased data (*pres_B*). These models achieved the "*biased*" habitat suitability maps named *HS_B*.

Corrected models were built from biased data (*pres_B*) corrected with the four correction methods. These models allowed to produce the "*corrected*" habitat suitability maps, named *HS_Cx,* where *x*∈ *{Fgeo, Fenv, Bggeo, Bgenv}*.

2.7. Metrics for evaluation of sampling bias correction methods

Three metrics were used to evaluate the ability of each model to correct the effect of sampling bias:

- The area under the receiver operating characteristic curve (AUC). It represents the overall probability of correctly separating the presence and absence sites. The AUC values were computed by comparing the resulting habitat suitability maps, to which a set of thresholds was applied to obtain presence-absence maps, with the presence-absence map PA_R_{EC} used as a reference. Thus, in this study, the AUC is the capability of the model to distinguish presence from absence and not presence from random as implemented in Maxent (Phillips et al., 2006).
- The Schoener's index *Dgeo*, which was initially developed to quantify the overlap of two species habitats in the geographical space (Schoener, 1968). The index values range between 0 and 1, corresponding to no and complete overlap between the two species, respectively. *Dgeo* was defined as:

$$Dgeo\ (p_Y, p_Z) = 1 - ½ \cdot \mathrm{sum}_{i \in X} |p_{Y,i} - p_{Z,i}| \quad (4)$$

where p_Y and p_Z are, respectively, the probability distributions of the species *Y* and *Z*, obtained from the output of the model; $p_{Y,i}$ and $p_{Z,i}$ are, respectively, the probabilities of the presence of the two species assigned to the pixel *i*. In this study, $p_{Y,i}$ and $p_{Z,i}$ are related to the same species and are the probabilities of presence at *i* obtained, respectively, using unbiased and corrected or biased data;

- The Schoener's index *Denv* initially developed for measuring the overlap of habitat suitabilities of two species *Y* and *Z*, in the environmental space (Broennimann et al., 2012). Like *Dgeo*, *Denv* ranges between 0 (no overlap) and 1 (complete overlap).

In this study, the environmental feature space was built by using a FAMD. Subsequently, the first factorial plane was divided into $m \times m$ cells, where *m* was arbitrarily set to 100. Each cell corresponds to a specific and single environment v_{ij}, where *i* and *j* refer to the position of the cell.

Denv was defined as:

$$Denv\ (z_Y, z_Z) = 1 - ½ \cdot \mathrm{sum}_{ij \in X} |z_{Y,ij} - z_{Z,ij}| \quad (5)$$

where $z_{Y,ij}$, and $z_{Z,ij}$ were, respectively, the densities of presence sites of species *Y* and *Z* in the environment v_{ij}. In this study, $z_{Y,ij}$, and $z_{Z,ij}$ are the densities of presence sites obtained at the environment v_{ij}, respectively, with unbiased and corrected or biased data.

For a given experimental condition (unbiased, biased, or corrected), the estimate of presence density was obtained by randomly selecting the presence sites, with the corresponding habitat suitability map as selection weighting. The number of present sites was arbitrarily set to 500 (see Broennimann et al., 2012 for details).

Three indexes were computed to quantify the relative impact of the bias correction methods on model performances, as follows:

$$\Delta AUC_{Cx} = (AUC_{Cx} - AUC_B)/(1 - AUC_B) \quad (6)$$

$$\Delta Denv_{Cx} = (Denv_{Cx} - Denv_B)/(1 - Denv_B) \quad (7)$$

$$\Delta Dgeo_{Cx} = (Dgeo_{Cx} - Dgeo_B)/(1 - Dgeo_B) \quad (8)$$

where *x*∈*{Fgeo, Fenv, Bggeo, Bgenv}*.

In the previous three equations, the value 1 stands for AUC_{UNB}, $Denv_{UNB}$ and $Dgeo_{UNB}$, i.e. the evaluation of the unbiased model output compared with itself. In that sense, these indexes vary from $-\inf$ to 1. A negative value indicates that the biased model outperformed the corrected one, whereas a positive value indicates that the corrected model outperformed the biased one; 1 shows that the corrected model was perfectly corrected and was comparable to the unbiased one (Fourcade et al., 2014).

3. Results

The distributions of *AUC*, *Denv*, and *Dgeo* values for the 100 replicates are shown by boxplots in Fig. 5.

AUC ranged from 0.806 to 0.998, *Dgeo* from 0.504 to 0.829, and *Denv* from 0.639 to 0.948.

The *AUC* and *Dgeo* results show that, for $k \leq 100$, the higher the

Table 1
Percentages of strictly positive values of ΔAUC, ΔDgeo, and ΔDenv.

	Number of presence sites	*Fgeo*	*Fenv*	*BGgeo*	*BGenv*
Δ*AUC*	20	10 (71)	38	**40**	35
	50	45 (11)	52	**71**	43
	100	70	65	**91**	38
	150	77	68	**94**	46
	200	86	52	**98**	38
Δ*Dgeo*	20	13 (71)	23	35	**45**
	50	31 (11)	27	28	**46**
	100	**59**	29	31	47
	150	**77**	35	30	57
	200	**78**	37	41	53
Δ*Denv*	20	53	61	52	**68**
	50	58	66	58	**69**
	100	67	**82**	71	67
	150	69	**84**	70	67
	200	66	**79**	72	65

A positive value shows that the corrected model performs better than the biased one. The maximum of each evaluation index, for each value of k, is shown in bold. The numbers in brackets represent the percentages of null values, i.e., the percentage of cases where the corrected models and biased models were strictly equivalent.

number of presence sites, the higher the *AUC* and *Dgeo* values. The results were comparable for $k > 100$. No significant trend of *Denv* values as a function of the number of sites was observed, regardless of the correction method used. When considering the *AUC*, the method *BGgeo* provided significantly better results than the other methods for $k \geq 100$.

Table 1 lists the percentages of positive values of the different relative evaluation indexes as a function of the correction methods and the number of presence sites. These values show the number of times the corrected models yielded better results than the biased ones for each experimental condition. These values show that, when the number of presence sites was only 20 or 50, *BGenv*, on average, had achieved better results than the other methods.

The mean relative rank of each method for each performance index is shown in Fig. 6. The ranking varied by correction type and the number of presence sites. Concerning Δ*Denv*, *Fenv* was the best method, followed by *BGenv* when the amount of presences sites was 20. When $k = 50$, *BGenv* appeared to provide the best results in terms of Δ*Denv* and Δ*Dgeo*, even if this result does not appear to be significant considering the dispersed values.

When the number of presence sites increased, *BGgeo* was ranked the best in terms of Δ*AUC*, *BGgeo*, and *Fenv* in terms of Δ*Denv*, and *Fgeo* in terms of Δ*Dgeo*.

4. Discussion

This study aimed to present a new sampling bias effect correction method in the framework of species distribution modeling and to compare it with other representative approaches of the literature.

4.1. Absolute evaluation

The metrics *AUC*, *Denv,* and *Dgeo* allow the evaluation of corrected models by comparing them with unbiased ones. Fig. 4 shows that, for *AUC* and *Dgeo*, the higher the number of presence sites, the better the performance of each corrected and biased model, and the lower the variability of results.

Although Wisz et al. (2008) showed that Maxent is less sensitive to the number of presence sites in comparison with other presence-background methods, this result shows that the performances were altered when the number of presence sites was below 100. However, given the variability of the results, it is difficult to declare the best correction method using such an absolute evaluation approach.

4.2. Relative evaluation

Table 1 shows that, when the number of presence sites was 20 and 50, respectively 71% and 11% of models corrected with *Fgeo* provided strictly identical results as the biased models, according to Δ*AUC* and Δ*Dgeo*. This is explained by the fact that all the presence sites were located at a distance greater than *r* from one another, and that, consequently, filtering had no effect. When the number of presence sites was small ($k < 100$), the percentages of Δ*Dgeo* and Δ*Denv* strictly positive values were higher for *BGenv*. Negatives indexes indicate that the corrected models do not provide better results than the biased ones. It is worth noting that Fourcade et al. (2014) also achieved results with negative indexes, whereas they considered different biasing schemes

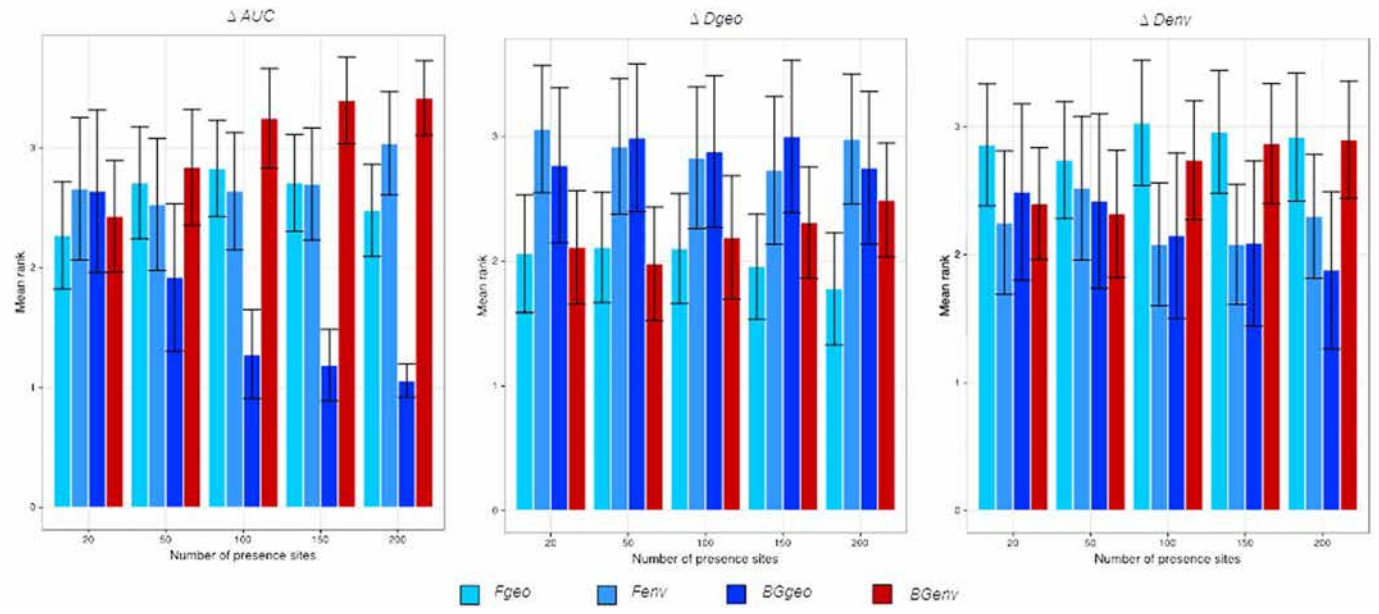

Fig. 6. Mean relative rank of the method according to the number of presence sites. The method that achieves the best results is ranked 1, and the method that achieves the worst result is ranked 4. The black lines represent the standard deviation.

and a high number of presence sites (2000). In this study, this may be due to the simulation process of the sets of biased and unbiased data (*pres_B* and *pres_UNB*). Indeed, these sets had a significant number of presence sites in common with *pres_UNB_init*. The proportion of *pres_UNB_init* selected to build *pres_B* and *pres_UNB* is 70%, which indicates that at least 40% of *pres_UNB_init* were in both *pres_B* and *pres_UNB* datasets, and that consequently, these two datasets share 60% of sites (for example, for $k = 20$, *pres_UNB* and *pres_B* shared at least 11 presence sites). This proportion may have been too high, resulting in a distribution of unbiased presence sites (*pres_UNB*) similar to the biased ones (*pres_B*), and minimizing the effect of the different correction methods. However, when very few sites are considered, it also appears crucial to control the random component in the definition of *pres_UNB* and *pres_B*, to ensure that the differences in performances between the unbiased model and the corrected one arise more likely from the correction methods than from the random selection of the sites. Thus, the chosen parameterization appeared the most appropriate by achieving a good compromise.

The ranking results of the different methods showed a variation of relative evaluation indexes according to the number of presence sites, and no correction method appears to be the best one. However, when we focus on a small number of presence sites, *BGenv* regularly achieved the first or second rank regardless of the evaluation index used.

In this study, the rank difference as per the evaluation index could be due to the random component in the selection of presence sites. Indeed, for the assessment of *AUC*, the presence-absence map PA_R_{EC} was used as a reference, whereas the assessment of *Denv* was based on a weighted and random selection of 500 presence sites from the reference habitat suitability map. Despite the weighting with the habitat suitability, the random effect was significant. The assessment of *Dgeo* was based on the probability distribution over the whole study area. Notably, *BGenv* was ranked first or second when $k \leq 50$, after *Fgeo* according to $\Delta Dgeo$ and after *Fenv* according to $\Delta Denv$. This indicates that *BGenv* allows a correction of the effect of sampling bias both in the environmental and geographical spaces.

Thus, it is challenging to declare the best correction method regardless of the number of presence sites, but the results show that *BGenv* achieves significant correction effects when the number of presence sites is small.

4.3. Parameterization of correction methods

The methods *Fgeo* and *BGgeo* were parameterized by referring to knowledge on the maximum distance moved of the species found in the literature. The absolute evaluation shows that *BGgeo* achieved better results than other methods when the number of presence sites increases. This could be due to the relatively large distance (7000 m) chosen in this study, supposed to be the maximum distance moved of *An. darlingi*. Indeed, Barve et al. (2011) showed that the greater the surface of background selection, the greater is the *AUC*.

However, the definition of *r* is not trivial, because the maximum distance moved depends on the land use and land cover, variation of environmental conditions in the area of study, and availability and distribution of resources (resting and breeding sites, mammals – particularly humans for *An. darlingi*, a highly anthropophilic species – for blood meals). Moreover, the flight range and habitat suitability are not theoretically related. The assumption is that, if the species is present at a specific location, according to its displacement distance *r*, it might as well originate from any point within a radius of *r*. However, such a definition is not directly based on assumptions on species habitat. A more appropriate way to parameterize methods in the geographical space would have been to consider the spatial autocorrelation of predictors (Naimi et al., 2011) in order to estimate the distance that makes occurrences spatially independent, whatever the species. This approach has not been implemented since we believe that it has not been applied with such an objective in the literature, but also because it would partially transpose in the geographical space, with an additional constraint of geographical contiguity, the notion of environmental proximity that has been implemented in this study through the *Bgenv* method.

The parameterization of *BGenv* was based on the a priori knowledge that *An. darlingi* cannot be present in a highly urban area. When referring to bioecological assumptions, the definition of *Dmin* in the environmental space is more robust than the definition of the geographical distance. For instance, in this study, it was based on unambiguous and scientifically proven knowledge on the bioecology of the species of interest (the fact that *An. darlingi* is not present in highly urbanized areas), whereas the assumptions leading to the size of the geographical buffer were highly questionable.

5. Conclusion

Since the 90's, SDM has been being increasingly used for various objectives including invasive species expansion control, endangered species track and reintroduction support, as well as health risk assessment in the framework of zoonotic and vector-borne diseases, by mapping the habitats of vectors and/or reservoirs of pathogens. Consequently, important issues, ranging from safeguarding biodiversity to public health, arise from SDM outputs and the capacity of correcting sampling bias.

This study stresses that it is challenging to select the best correction method, regardless of the number of presence data. However, when the amount of presence sites is small (<100), it shows that, besides producing relevant results validated by ecologists in a real study case (Moua et al., 2017), the definition of a biased background defined in the environmental feature space appears to be the approach the most likely to improve the model results, in comparison with other methods. Even if this method was applied with the Maxent algorithm only, it could be used in association with other presence-background modeling approaches, such as ecological niche factor analysis (ENFA) or GLM and GAM. Eventually, it can contribute to the improvement of SDM reliability in contexts of presence data scarcity.

Ethics approval and consent to participate

Not applicable (the study involves simulated species data).

Consent to publish

Not applicable (the study does not include details, images, or videos relating to any individual).

Availability of data and materials

The simulated datasets generated during the current study are available from the corresponding author upon reasonable request.

Funding

This study was funded by the Fond Social Européen (FSE), the Centre National d'Études Spatiales (CNES), and the Collectivité Territoriale de Guyane. Financial support was partially provided by the "Investissement d'Avenir" grants managed by the Agence Nationale de la Recherche (Center for the study of Biodiversity in Amazonia, ANR-10-LABX-0025), the GAPAM-Sentinela project of the Franco-Brazilian scientific and academic cooperation program Guyamazon (funds: IRD, CIRAD, French Embassy in Brazil, Territorial Collectivity of French Guiana, Brazilian State-level research agencies of Amapá, Amazonas and Maranhão) and the TélePal project of the TOSCA program (CNES). It is part of the research activities of the ODYSSEA project (H2020 MSCA RISE program funded by the European Commission) and the International Mixed Laboratory (LMI) Sentinela, funded by the French

National Research Institute for Sustainable Development (IRD).

Authors' contributions

YM and ER participated in the research design, data collection and simulation, program development, analysis and interpretation of the results, and prepared the manuscript. FS and SB participated in the research design, analysis and interpretation, and reviewed the manuscript. All authors read and approved the final manuscript.

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgments

The authors are grateful to the Pasteur Institute of French Guiana (Institut Pasteur de la Guyane), in particular, Benoît de Thoisy, Antoine Adde, Romain Girod, and Isabelle Dusfour, for the help they provided in carrying out this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2020.101086.

References

Alimi, T.O., Fuller, D.O., Qualls, W.A., et al., 2015. Predicting potential ranges of primary malaria vectors and malaria in northern South America based on projected changes in climate, land cover and human population. Parasit. Vectors 8 (431).

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J., Villalobos, F., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecol. Model. 222 (11), 1810–1819.

Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275, 73–77.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.-J., Randin, C., Zimmermann, N.E., Graham, C.H., Guisan, A., 2012. Measuring ecological niche overlap from occurrence and spatial environmental data: measuring niche overlap. Glob. Ecol. Biogeogr. 21 (4), 481–497.

Charlwood, J.D., Alecrim, W.A., 1989. Capture-recapture studies with the South American malaria vector Anopheles darlingi. Root. Ann. Trop. Med. Parasitol. 83 (6), 569–576.

De Castro, M.C., Monte-Mór, R.L., Sawyer, D.O., Singer, B.H., 2006. Malaria risk on the Amazon frontier. Proc. Natl. Acad. Sci. U. S. A. 103 (7), 2452–2457.

de Thoisy, B., Richard-Hansen, C., Goguillon, B., Joubert, P., Obstancias, J., Winterton, P., Brosse, S., 2010. Rapid evaluation of threats to biodiversity: human footprint score and large vertebrate species responses in French Guiana. Biodivers. Conserv. 19 (6), 1567–1584.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., et al., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography. 29, 129–151.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists: statistical explanation of MaxEnt. Divers. Distrib. 17 (1), 43–47.

Fourcade, Y., Engler, J.O., Rödder, D., Secondi, J., 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. Edited by J. F. Valentine. PLoS ONE 9 (5) e97122.

Gond, V., Freycon, V., Molino, J.-F., Brunaux, O., Ingrassia, F., Joubert, P., Pekel, J.-F., Prévost, M.-F., Thierron, V., Trombe, P.-J., Sabatier, D., 2011. Broad-scale spatial pattern of forest landscape types in the Guiana shield. Int. J. Appl. Earth Obs. Geoinf. 13 (3), 357–367.

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157 (2), 89–100.

Guitet, S., Cornu, J.-F., Brunaux, O., Betbeder, J., Carozza, J.-M., Richard-Hansen, C., 2013. Landform and landscape mapping, French Guiana (South America). J. Maps. 9 (3), 325–335.

Guo, Q., Kelly, M., Graham, C.H., 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecol. Model. 182 (1), 75–90.

Hill, M.P., Terblanche, J.S., 2014. Niche overlap of congeneric invaders supports a single-species hypothesis and provides insight into future invasion risk: implications for global Management of the Bactrocera dorsalis complex. Edited by J. Pinto. PLoS ONE 9 (2), e90121.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecol. 83 (7), 2027–2036.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., et al., 2013. The importance of correcting for sampling Bias in MaxEnt species distribution models. Divers. Distrib. 19 (11), 1366–1379.

Leroy, B., Meynard, C.N., Bellard, C., Courchamp, F., 2015. Virtualspecies, an R package to generate virtual species distributions. Ecography. 39 (6), 599–607.

Moua, Y., Roux, E., Girod, R., Dusfour, I., de Thoisy, B., Seyler, F., Briolant, S., 2017. Distribution of the habitat suitability of the main malaria vector in French Guiana using maximum entropy modeling. J. Med. Entomol. 54 (3), 606–621. https://doi.org/10.1093/jme/tjw199.

Naimi, B., Skidmore, A.K., Groen, T.A., Hamm, N.A.S., 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling: spatial autocorrelation and positional uncertainty. J. Biogeogr. 38 (8), 1497–1509.

Pages, J., 2004. Analyse factorielle de données mixtes. Rev. Stat. Appl. 52 (4), 93–111.

Pearson, R.G., Dawson, T.P., Berry, P.M., Harrison, P.A., 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. Ecol. Model. 154 (3), 289–300.

Pearson, R.G., Raxworthy, C.J., Nakamura, M., Peterson, A.T., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar: predicting species distributions with low sample sizes. J. Biogeogr. 34 (1), 102–117.

Peterson, A.T., Soberón, J., Pearson, R.G., Martinez-Meyer, E., Nakamura, M., Araújo, M.B., 2011. Ecological Niches and Geographic Distributions. Princeton University Press, Princeton, NJ, USA.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190 (3–4), 231–259.

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

R Development Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schoener, T.W., 1968. The Anolis lizards of Bimini: resource partitioning in a complex fauna. Ecol. 49 (4), 704–726.

Stefani, A., Dusfour, I., Corrêa, A.P.S.A., Cruz, M.C.B., Dessay, N., Galardo, A.K.R., Galardo, C.D., Girod, R., Gomes, M.S.M., Gurgel, H., Lima, A.C.F., Moreno, E.S., Musset, L., Nacher, M., Soares, A.C.S., Carme, B., Roux, E., 2013. Land cover, land use and malaria in the Amazon: a systematic literature review of studies using remotely sensed data. Malar. J. (12), 1–8. https://doi.org/10.1186/1475-2875-12-192.

Stockwell, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Sci. 13 (2), 143–158.

Varela, S., Anderson, R.P., García-Valdés, R., Fernández-González, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography. 37 (11), 1084–1091.

Walker, P.A., Cocks, K.D., 1991. HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. Glob. Ecol. Biogeogr. 1 (4), 108–118.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., 2008. NCEAS predicting species distributions working group. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14 (5), 763–773.