Spatial Distribution Modeling (SDM) for endangered species using the Random Forest (RF) algorithm

Modèle de distribution spatiale pour les espèces en danger utilisant les forêts d'arbres décisionnels

Tidjani FOUSSENI SALAMI CISSE
Anaïs SPIRE

Université Côte d'Azur – 2023

Number of words (from page 3 to 14, without titles and references): 4462

Abstract

The conservation of endangered species has become a major challenge nowadays, as a consequence of huge anthropic impacts on biodiversity (climate change, pollution, deforestation, …) and the awareness that biodiversity is necessary for a healthy earth. The red-bellied monkey (Cercopithecus Erythrogaster), an endemic primate of sub-Saharan regions, is one of these species: the monkey has been recently classified as “endangered” by the IUCN Red List in 2016 (IUCN). To work towards its conservation, many approaches need to be combined and quantitative tools can be of great help. Species distribution models (SDMs) are the most widely used modelling framework in quantitative ecology, aiming at calculating the ecological niche of a species, both on the current timeline and on the future timeline (considering climate change, deforestation, …). SDM are particularly convenient because they use very low data and storage requirements, with no use of DNA sequences. SDMs constitute a big family and there are many different models and algorithms that have been developed, including algorithms using machine learning methods. We did our study with one of the latter: the Random Forest (RF) algorithm, using the R software. The project is available in the following GitHub repository: git or on UCA OneDrive: Drive.

Key words

conservation biology, Spatial Distribution Modeling, endangered species, red-bellied monkey, random forest

Résumé

La conservation des espèces en voie de disparition est devenue un défi majeur de nos jours, en raison des impacts anthropiques grandissants sur la biodiversité (changement climatique, pollution, déforestation, ...) et de la prise de conscience que la biodiversité est nécessaire pour une planète en bonne santé. Le singe à ventre rouge (Cercopithecus Erythrogaster), un primate endémique des régions sub-sahariennes, est l'une de ces espèces : le singe a été récemment classé comme "en danger" par la Liste rouge de l'UICN en 2016 (IUCN). Pour assurer sa conservation, de nombreuses approches doivent être combinées et des outils quantitatifs peuvent être d'une grande aide. Les modèles de distribution des espèces (SDM) sont le cadre de modélisation le plus largement utilisé en écologie quantitative, visant à calculer la niche écologique d'une espèce, tant sur la ligne de temps actuelle que sur la ligne de temps future (considérant le changement climatique, la déforestation, ...). Les SDM sont particulièrement pratiques car ils nécessitent de très faibles exigences en matière de données et de stockage, sans utilisation de séquences d'ADN. Les SDM constituent une grande famille et il existe de nombreux modèles et algorithmes qui ont été développés, y compris des algorithmes utilisant des méthodes d'apprentissage automatique. Nous avons réalisé notre étude avec l'un des derniers : l'algorithme des forêts aléatoires (RF), en utilisant le logiciel R. Le projet est disponible dans le repo GitHub suivant : git ou sur le OneDrive d’UCA : Drive.

Mots-clés

Biologie de la conservation, modélisation de la niche écologique, biogéographie, espèce en voie d’extinction, singe à ventre rouge, forêt d’arbres décisionnels, algorithme d’apprentissage

Introduction (biological features)

Biological value of the red-bellied monkey

The red-bellied monkey, Cercopithecus Erythrogaster, is endemic of a small area in the South of Togo, Benin and Nigeria, in an estimated extent of 136 250 km². See ICUN map of geographic range: figure 1. It is classified as « endangered » by the IUCN Red List since 2016 and was even considered extinct back in the 1980s until a small colony was found in 1988. Though we do not have an estimation of the number of monkeys on the regional scale, it has been estimated that the population has declined by 50% since the 1990s in most areas. The population is still in a decreasing trend nowadays and big concerns have been raised by local populations and researchers in primatology. See IUCN.

The monkey has a life expectancy up to 24 years in captivity and is subject to poaching and hunting for its fur, skin and meat in its natural habitat. According to a study conducted in Togo (Agbessi, 2016), monkeys are killed for food, sales, protection of the crops, and for traditional medicine. The species also suffer from human population growth in the area who tend to exploit its habitat by deforestation and the installation of dams in the region (eg the Adjarala dam at the borders of Togo and Benin).

Yet, red-bellied monkey is a keystone species for biodiversity. It contributes to seed dispersal, plant pollination, and insects' regulation. The extinction of the monkey may lead to disruption of the local ecosystem with unknown consequences, as in many case with endangered species (Purvis et al., 2000).

Ecological niche

In ecology, a niche is the match of a species to a specific environmental condition.

In other words, finding the niche of species means exploring all the environmental variables, and find all the values taken by these environmental variables where our species can fit, survive. In a theoretical point of view, the niche can thus be described as an hypervolume of n dimensions, where n is the number of environmental variables. See figure 2 for an example in 2 dimensions.

More specifically, there are two types of niches: the fundamental niche where the species could potentially survive but is not observed on the field; and the realized niche, which is a subset of the fundamental niche, where the species is actually observed on the field. The realized niche is the set of environmental conditions used by the species after interaction with other species (biotic factors, be it competitive, neutral or beneficial) have been added. It corresponds to the area where the species has the best fitness. It is important to note that ecological niches of a species are subject to change due to the change of environmental variables (climate change, deforestation, …) (Mi et al., 2017).

We summarized the main environmental variables and the range of values where the red-bellied monkey has been observed in its natural habitat in the present in tables 1a and 1b. Studies in Benin and Togo (Agbessi, 2016; Kassa et al., 2014) show that more and more monkeys are observed in damaged areas where they were not present before, like on the sides of crops. This suggests that the realized niche of the species has already started to change, and we can expect it to change even more in the future. This is a major concern as the sides of crops render the monkey particularly vulnerable and creates more conflicts with the local population.

Objectives

Our aim is to better understand the spatial dynamics and conservation status of the red-bellied monkey in the sub-Saharan region, in the current and in the future timelines. Spatial distribution modeling (SDM) enables us to evaluate the niche of the monkey for both timelines: current timeline thanks to mapping (interpolation) and future timeline thanks to forecast (extrapolation). This niche then helps us to decide where conservation efforts need to be done as a priority. Note that it is an ongoing debate whether the estimated species-environment relationship given by SDM approximates the fundamental niche or the realized niche (Zurell et al., 2020).

The results of SDM could be used by governments as regards the decision of new policies' enforcement in favor of the protection of the monkey or as regards the creation of new protected areas. Indeed, SDM helps to identify the environmental factors that are most important for the survival of the monkey. This could assist environmental managers in developing more effective conservation plans.

The results could also be used by field observants as regards the evaluation of the number of red-bellied monkeys by pointing out where the monkey could possibly be present. For the moment, the points of observation of the monkeys are often acquired thanks to surveys conducted with local populations and are not always reliable (high chance of confusion between species, vague location, …).

Finally, the results could point out areas where the ecological niche is or will not be interconnected due to the changing environmental variables (climate change, deforestation, …). This is of major relevance since this could lead to isolation of some individuals, which prevent genetic admixture necessary for population dynamics. SDM can help scientists who study the genetic diversity of the monkey by orienting their research towards these isolated populations and see whether they identify genome erosion patterns (cf: GenErode pipeline).

Part 1 (presentation of SDM and our datasets)

A] Principle of SDM

SDM is a statistical method that aims to determine the probability of species presence or absence in a given geographic area. First, both species and environmental data are sampled in geographic space. Second, the data is analyzed using statistical tools to determine the relationships between environmental characteristics and the probability of species presence. A large choice of algorithms and models are available for this step (GLM, RF, etc.). Third, the species-environment relationship is mapped. This 3-step principle is well summarized in figures 3a and 3b.

In practice, we can also establish a 5-step principle of SDM:

(i) conceptualization
(ii) data preparation
(iii) model fitting
(iv) model assessment
(v) prediction

We have seen the first step in the introduction. We are now going to dive into the next steps.

B] Presentation of the datasets

Two datasets are necessary for conducting SDM. One more is necessary for the models we are using.

The first dataset, called occurrence data or species data, contains information on the points of observations (presence) of our species. For our monkey, we have obtained these data from GBIF (Global Biodiversity Information Facility) where the information is stored in tables: there is one line per observation of the monkey and 50 columns indicating the location and date of the observation, the person(s)/institution(s) who has observed it and recorded it into the table, the subspecies, and other relevant observations. The columns which interest us the most are latitude and longitude, and these are the one we store into gbif_coords variable in R, and that we will use as an input of our SDM algorithm. See gbif occurence data or the *'species data'* folder of the project.

The second dataset is one of pseudo-absence data. This is required for the models we will be using (GLM and RF); it is not required for all SDM models (van Proosdij et al., 2016; Barbet-Massin et al., 2012). The dataset contains information on the points of absence of our species, as opposed to the points of presence contained in the occurrence data. Because no such dataset is available for dynamic species, we generate this dataset randomly (see next section 2.C).

The third dataset, called prediction data or environmental data, contains information of the current and future environmental factors of the geographic area of interest. These factors include bioclimatic, topography, vegetation, temperature, precipitation, UV radiation, ... We have obtained two datasets from Worlclim and ESA. It includes 19 bioclimatic variables from Worldclim (see here for the list of the 19 variables) and 2 variables from ESA: trees and land cover. Future predictions are made on different time scales and with different Shared Socio-economic Pathways (SSP) and Global Change Models (GCM). For the dates, we chose the 2041-2060 period. For SSP, we chose SSP2 which is the middle of the road scenario (not too optimistic or too pessimistic). See this paper for a rapid overview on the different SSP: O'Neil, 2015. To find the best GCM models for west Africa, we studied Nigeria (Sanusi Shiru and Chung, 2021) and summarized the results in table 2. Based on this study, we chose the

ACCESS-ESM1-5 model. The reader is invited to take a look at the data on Wordclim and ESA websites (worldclim, esa) or in the *'environmental data'* folder of the project.

C] Data preparation

For data cleaning of the species data, every step is well documented in the following R script: *'cercopitheque_species_data.R'*. Unbiased data is one of the assumptions of the SDM and must be carefully considered for species data. We have followed these steps: removal of duplicates, removal of inconsistent points (eg. points in the ocean), removal of zoo individuals (eg. monkey in the Zoo in Paris), removal of imprecise points. We had 116 points extracted from GBIF, 67 with coordinates and 59 were left after cleaning. Also, while we had many points in Benin thanks to the contribution of scientists in this country, we had fewer points for Nigeria and none for Togo. Thus, we added some missing points in Togo, by hand, thanks to this study (Agbessi, 2016). At the end, we got 65 points. See the points mapped on figure 4a. Though 65 points is not a lot, it is considered to be enough for random forest (van Proosdij et al., 2016).

For pseudoabsence data, we need to generate the data ourselves. Indeed, because primates can move, absence points are never recorded. Generation of absence data is made by a smart balance between randomization, and probabilities: we put probabilities around the absent points using a gradient, and over a certain distance, we use randomization. R packages that use this principle are available, but we were unable to make them work (see: *'cercopitheque_pseudoabsence_data_draft.R'*). Therefore, we produced our own random points, by using a uniform random distribution in well-chosen rectangles of the map, avoiding the ocean and the areas closest to the ocean where the species is mostly present. We produced 100 absence points. This workaround is available in the following script: *'cercopitheque_pseudoabsence_data_kludge.R'*. See the points mapped on figure 4b.

For environmental data, you can refer to the following script: *'cercopitheque_environmental_data.R'*. No data cleaning is required here but there is preparation to do for map resolution and extent. Indeed, any scaling mismatches between species and environmental data or within environmental data, produce errors. In these cases, we need to make decisions about adequate upscaling and downscaling strategies (Zurell et al., 2020). Concerning the red-bellied monkey, we know it is endemic of a single and small area in sub-Saharan land and we do not want to study its invasive potential in other area of the world. Thus, we restricted the geographic extent to its current presence area with larger boarders on the x axis (longitude) and smaller boarders on the y axis (latitude) due to the sub-Saharan geography (ocean in the south and desert in the north constitute natural barriers). We therefore conducted the studies in this extent: 3.5°W-12°E for latitude, 3°N-14°N for longitude. See map on figure 1. As regards map resolution, this is one of the greatest challenges on data preparation for SDM and we have consumed a lot of time on it.

Once these datasets are obtained separately, we need to join them, and especially extract points of the environmental maps to match them with our lat/lon coordinates from the species presence and absence data: see *'cercopitheque_joined_data.R'*. In our case, the joined dataframe contain 165 lines (presence+absence points) and 24 columns: latitude, longitude, presence (1 for presence, 0 for absence), and the 21 environmental variables. We long struggled to extract points of the maps and did not succeed for the maps from ESA (land cover and tree cover). Therefore, the final table of joined_data includes only 22 columns. Because we had problems with coding this part, this process also slowed down our project considerably.

Part 2 (Model fitting and algorithm: GLM and RF)

A] Generalized Linear Model (GLM)

Before using the machine-learning approach of the random forest algorithm, we used the Generalized Linear Model (GLM) to have a first statistical insight of the data.

GLM is a general statistical framework used to model relationships between a response variable, whether continuous or discrete, and one or more predictors (independent variables). See Hastie, Tibshirani and Friedman, 2009. We use it here to find the relationship between species occurrence and the different environmental variables. The goal is to find any multicollinearity between the environmental data, in order to exclude redundance.

In R, packages already exist, like the stats package, and we used it in the R script *'cercopitheque_glm.R'*.

GLM is similar to simple linear regression where the relationship between the response variable and the predictors is considered linear, but it is more flexible. Simple linear regression only works if the error is normally distributed and ranges between $]-\infty,+\infty[$, which is not the case here.

Let us define the Y as the response, ie absence or presence of species, and X as the predictor, ie the matrix of environmental variables. In SDM, we want to know $E[Y|X]$, that is the mean of occurrence probability given all the environmental variables. This expected value is binary: it can only take the value 0 for absence and 1 for presence (as in a Bernouilli scheme). The error will therefore not range over $]-\infty,+\infty[$ and OLS cannot be used. However, mathematicians have shown the GLM with a logit link function can be used to transform binary/Bernouilli response to normality. This is called logistic regression model. In this case, the mean is given as:

$$E[Y|X] = \pi(X) = \frac{e^{\beta X+\varepsilon}}{1+e^{\beta X+\varepsilon}}$$

With β the coefficients for X and ε the error term, assumed to be normally distributed.

And we linearize it with the logit transformation:

$$g(X) = \ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta X + \varepsilon$$

(Zurell et al., 2020)

In our case, the predictor matrix X has a length of 21, since we have 19 bioclimatic variables from WorldClim, plus 2 variables from ESA for trees and land cover. In R, we specified family = "binomial" as a parameter of the *glm* function, so that the algorithm uses the logit link function. We performed GLM on the entire predictor matrix, or a subset of it. Because we obtained no correlation with this method, we tried to refine the model by removing highly correlated variables (correlation matrix). The results we obtained are available in part 3 but note that they are not satisfying. This is probably due to either not enough data points or mistakes in the script *('cercopitheque_GLM.R')*.

B] Random Forest (RF)

The Random Forest (RF) algorithm is a supervised machine learning algorithm used for classification and regression. It is a type of decision tree models, which uses not only one tree but a forest of classification and regression trees (CART), hence the name: random forest.

During the training phase, multiple decision trees are constructed, and they are combined during the testing phase to obtain a final prediction (Liaw and Wiener, 2002). Each decision tree is built from a random subset of the training dataset and a random subset of the prediction variables (subsampling or bootstrapping). The randomness is a great asset and helps reduce overfitting (predictions fit well with the training data, but also with any other datasets), and thus makes the model more robust (Breiman, 2001). The aggregation of trees leads to a more accurate and robust prediction compared to a single decision tree. The output is cross-validated thanks to out-of-bag samples: the prediction for a specific data point is only derived from averaging trees that did not include this data point during tree growing.

In SDM, training data is the species occurrence data, it constitutes the input of the trees. Prediction data is the environmental, bioclimatic data, it constitutes the node of the trees. Thus, one node can represent mean temperature, precipitation, or land cover. The results at the leaf of the trees are binary: presence (1) or absence (0) of the species (Genuer and Poggi, 2020). See figures 6 and 7 for an illustration of random forest for SDM on one single tree.

Key parameters of the algorithm include the number of generated trees, the number of leaves, the size of the prediction dataset, the sampling strategy (subsampling or bootstrapping), and the value of the split. The value of the split is the threshold value for which we decide to create a new group from a node. For mean temperature, the value of the split could be 20° for example.

We were unable to make random forest work in R on our dataset, due to package issues earlier and limited time.

C] Pros and cons

We summarized the pros and cons of RF in table 3 (Scornet et al., 2015; Elith et al., 2006; Luan et al., 2020; Probst et al., 2019; Phillips et al., 2006).

We chose RF over other algorithms like Mahalanobis distance-based model, Maxent or Bioclim for facilitation purposes and because we had very few points. The latter algorithms either require more assumptions, are more obscure in their process or produce hardly interpretable results. For a presentation of the 12 most renowned SDM algorithms and a more detail comparison between them, the reader is largely encouraged to look at figure 5 and read (Pecchi et al., 2019).

Part 3 (Results)

A] First result

As mentioned in data preparation, we struggled to get the table with joined data of location, presence and environmental variables at the points of location. Once we finally succeeded, we only had two weeks left to perform GLM and analyze its results, and we failed to obtain coherent correlation results between presence of the species and environmental variables. However, we still show and analyze them in the Annex below (GLM results). Facing these issues, we decided to make a pairwise correlation matrix of the environmental variables alone (with no presence/absence data); see figure 8 and explanation in the Annex. We then removed the most highly correlated variables in the hope of improving GLM model fitting, but the results were still not satisfying.

Also note that because of limited time we had to stop the project there and we were not able to translate the results back on maps. Therefore, there is no projection on maps available.

B] Other results

No results yet.

The idea here would have been to try other time periods, other environmental variables, or another model. We also had the idea to try a multi-species SDM with another endangered species of the same area (Colobus Vellerosus for example).

C] Model assessment and prediction

Not available yet, though it is a very important part and needs to be carefully conducted.

See table 4 for a list of statistical tests that we intended to do.

Part 4 (Discussion on the expected results)

Due to the lack of results, we are just going to give the results we were expected to have and give discuss cases which deviate from the expected results.

A] Current timeline: comparison with IUCN range maps

On the current timeline, we would expect our results to be similar to IUCN range maps presented in figure 1. Indeed, even though IUCN maps are only maps based on the presence of the monkey in the area and does not consider any ecological niche aspect, the presence of the monkey in this precise area and not in a more extended area suggests that this area is the fittest for the species.

If the results are not similar to IUCN area, the prediction area could be either wider or smaller.

Wider spatial predictions from SDM compared to IUCN maps would either suggest biases in the data (see next section), imprecision of our classification method for environmental variables, too many simplifications, a bad choice of environmental variables for the species, or the importance of other factors such as biotic factors (eg competition with other species), poaching and hunting. The latter prevent the monkey to be distributed over the whole ecological niche provided by SDM. As a reminder, the presence area of the monkey constitutes its realized niche and is a subset of the fundamental niche and thus of SDM prediction areas, thus having a slightly wider area than IUCN would not be surprising.

Smaller spatial predictions from SDM compared to IUCN maps would suggest biases in the data (see next section), a too strict classification method for environmental variables, a bad choice of environmental variables for the species, or that other factors such as biotic factors (eg mutualism or commensalism between two species) enables the monkey to fit well even in a less favorable environment.

Finally, we could also have results where the prediction areas are not connected, and thus suggest the existence of subspecies. We were somewhat expecting these results because we do have two subspecies of the red-bellied monkey: Cercopithecus Erythrogaster Pocoki and Cercopithecus Erythrogaster Gray. The first subspecies consists of the individuals observed in Togo and Benin, the second subspecies consists of the individuals observed in Nigeria. The study in Togo (Agbessi, 2016) states that these 2 subspecies are now completely isolated from each other.

B] Future timeline: impact of climate change and deforestation

We would expect the prediction maps to be smaller in the future timeline (2040-2060) compared to the current timeline due to forecasted climate change and deforestation. The ecological niche could also be moved further West, East, or South. North is unlikely in our case because it reaches the Saharan region and past studies have suggested that the monkey did not accommodate well to a lack of vegetation or hotter temperatures.

Other results would be surprising and would request further investigations in the literature.

C] Simplifications and biases

SDM entails a wide range of simplifications and does not intend to capture all the challenges of species distribution. It only looks at the environmental variables, with no consideration of external factors which play equally important roles in the species distribution (biotic interactions, human presence, poaching, traveling capacity of the species, …). (McSHEA, 2014; Mi et al., 2017; Pecchi et al., 2019; Zurell et al., 2020). On top of these limitations reviewed in literature, our study itself contains a lot of biases and simplified steps.

The biases are present in three parameters of the studies: species data, environmental data, scale.

For species data, there are biases on both presence and absence points. The presence points depend on field observation, yet the region does not have well distributed reliable sources. Most of the observation points had been referenced by Université Abomey-Calavi (Bénin), Direction Generale des Eaux, Forets et Chasse (Bénin) and Université Lagos (Nigeria). This could easily explain why we find clusters of monkeys in some areas and no presence points at all in other areas. The importance of reliable sources is well detailed in the Togo Study (Agbessi, 2016) where they state that field observations can be easily biased depending on the study, points can be imprecise and can even be false because red-bellied monkey can get easily mixed up with other close species if the observant is not an expert. Moreover, the data we got from GBIF included many animals in zoological gardens and though these areas are vast and located in areas where the monkey had been observed in the wild before, it introduces clusters and thus biases in the datasets. There were also measures with a given uncertainty on GBIF and we chose not to include the uncertainty in the dataset, which is an obvious loss of information. Lastly, the way we randomly generated absences with a simple uniform distribution is dumb, and there is large room for improvement there. Many literature reviews cover this topic, and different R packages are available, but we were unable to make them work (van Proosdij et al., 2016).

For environmental data, biases can come from errors in the maps or the lack of key maps, such as fine particle pollution maps in our case. Prediction models can make large biases too, especially in Africa for which the forecast is not always appropriate. As mentioned in data preparation, we tried our best to choose the best predicted model for our region of interest based on literature and we chose the middle of the road scenario, but it is not completely unbiased. Furthermore, predictions are annual mean variables given by chunks of 20 years on Worldclim which imposes a particular time scale, which is not necessarily the most adapted to our case study. In particular, the study in Togo mentions the importance of seasonal change for primates, but this environmental data does not allow us to study this.

For scale, ecological processes are highly scale dependent, but compromises are necessary. Because every environmental map must be on the same scale at the end, different upgrading and downgrading scaling techniques exist. Here, our strategy was to aggregate all GIS layers at the limiting grain and extent, which again conducts to loss of information for most data.

Last but not least, many algorithms and models are provided for SDM. These models and their parameters must be well-chosen in regards of the dataset of interest. In our case, we chose to use the random forest algorithm for its ease of use, but it was not necessarily the best for our datasets.

D] Suggestions of complementary studies

This SDM study only provides predictions for the 2041-2060 period with a middle-of-the-road scenario (SSP2) and one GCM model (ACCESS-ESM1-5). It would be relevant to try other time periods, other SSP scenarios and other GCM models adapted to the sub-Saharan region and compare them with the above results. Similarly, this study only uses two models (GLM, RF). Because these parameters largely affect the species-environment relationship (Elith et al., 2006; Merow et al., 2014), complementary studies with other models and algorithms would be totally relevant.

Furthermore, many complementary studies need to be done for the conservation of the monkey. SDM is a tool among others and cannot be used alone to take decisions for conservation policies. Suggestions of other studies include genomic studies which are necessary to see if patterns of genomic erosion occur, which lead to a loss of genetic diversity and inbreeding, and thus to less adaptation for the species in case of a future change. Socio-economic factors, and especially local populations' opinion, are important and conservation policies cannot be enforced without considering them. SDM alone cannot tell if local initiatives such as ecotourism is a good or bad idea for the conservation of the monkey. Lastly, SDM does not consider the traveling capacity of the species, though it was shown to play a key role in extinction. Indeed, a species may go locally extinct simply in response to stochasticity and the dispersal ability will determine how fast the now empty patch can be recolonized. BAM (biotic-abiotic-movement) diagrams can be used to study the complex interplay between these three factors.

Figures (10 max)

Figure 1: IUCN geographic range of the Cercopithecus Erythrogaster. Extent: 3.5°W-12°E (Lat), 3°N-14°N (Lon). From IUCN

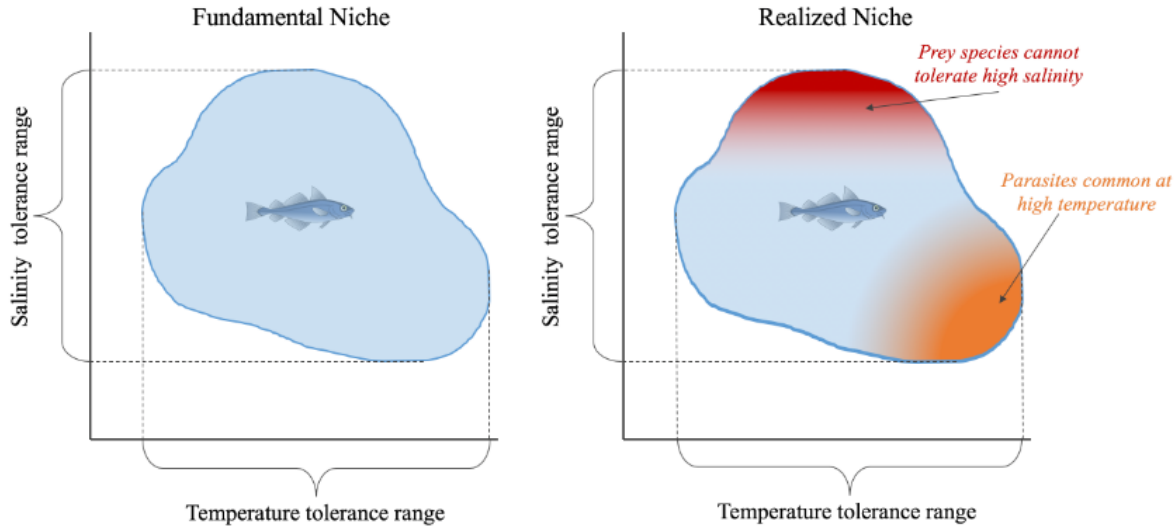


Figure 2: ecological niche: a two-dimensional example for a fictional fish species. Figure by L. Gerhart-Barley, professor at UC Davis

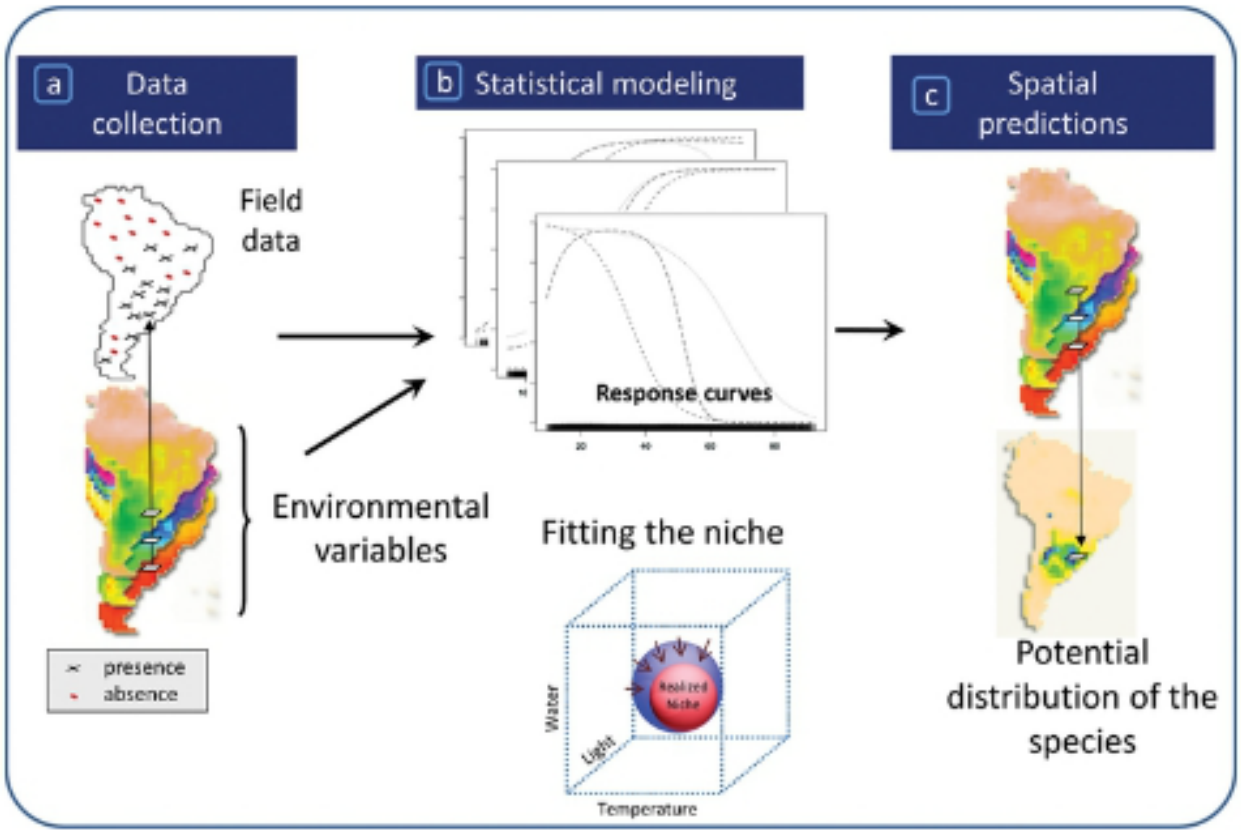


Figure 3a: SDM principle: input – modeling – output. From (Zurell, 2020)

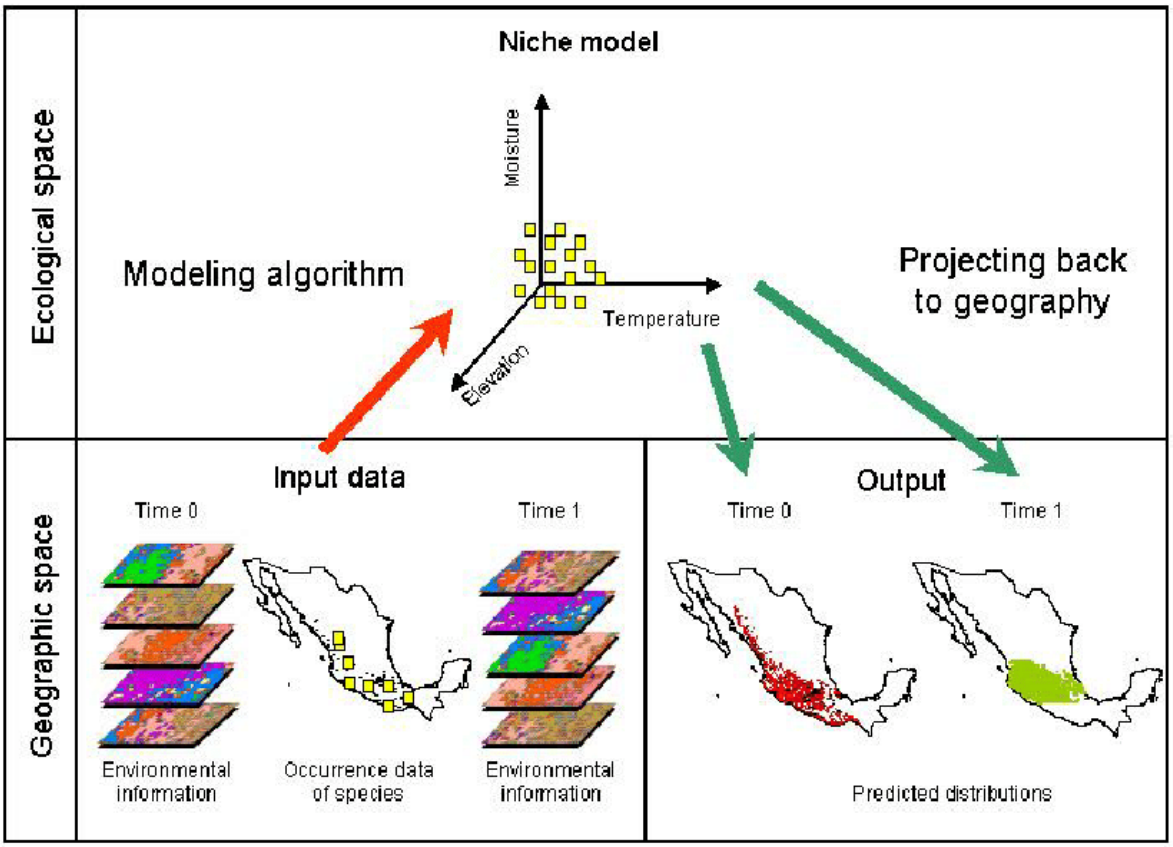


Figure 3b: SDM principle: input – modeling – output. From [Martinez-Meyer, 2005]

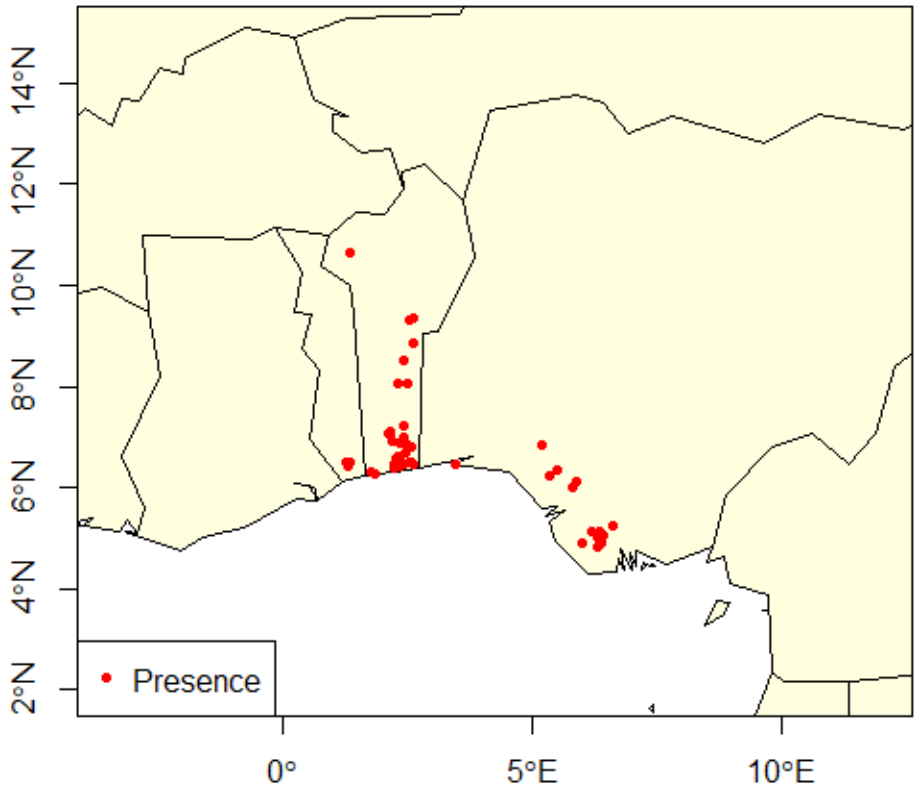


Figure 4a: 65 presence points of the red-bellied monkey. From R script 'cercopitheque_species_data'

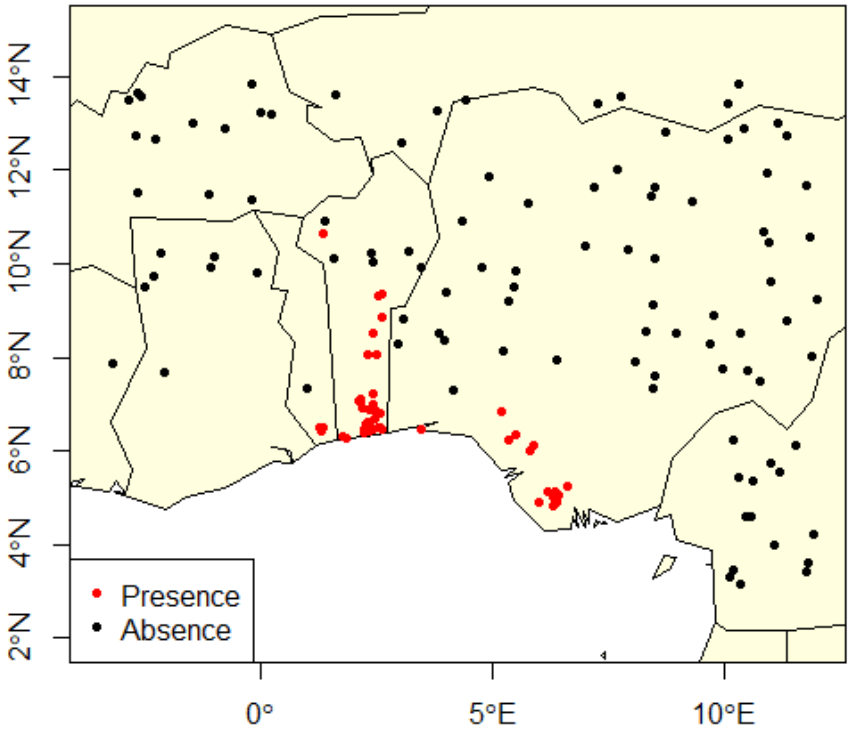


Figure 4b: 65 presence and 100 absence points of the red-bellied monkey. From R script 'cercopitheque_pseudoabsence_data'

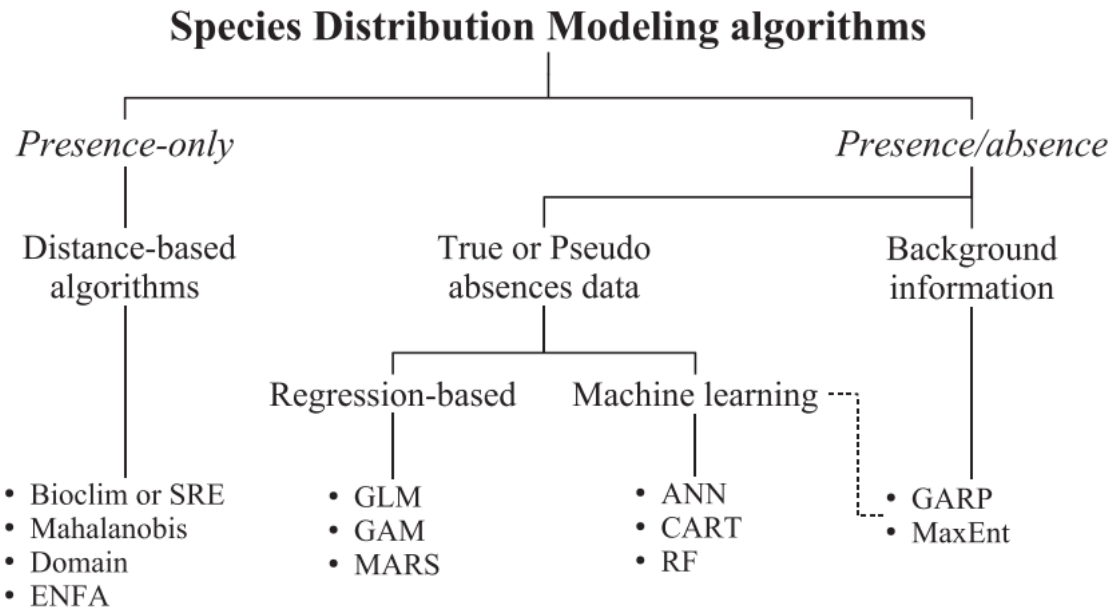


Figure 5: a hierarchical structure of 12 SDM algorithms. From (Pecchi et al., 2019). The reader is largely encouraged to read further the latter article for a detail comparison of these 12 algorithms.

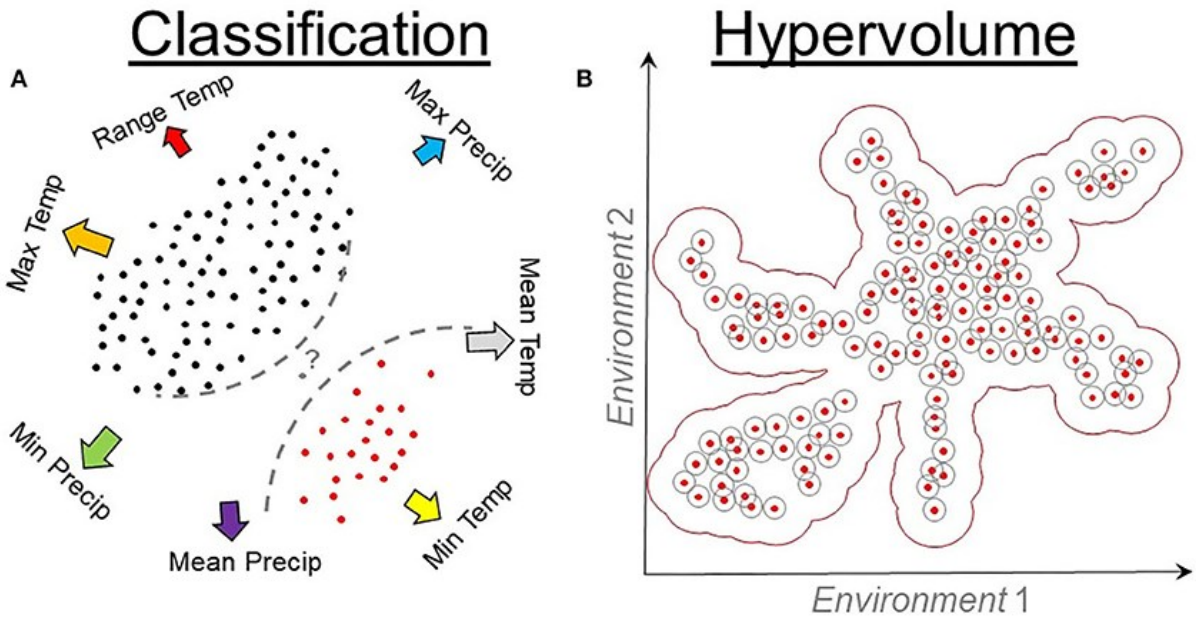


Figure 6: SDM and classification models. Red points = presence, black points = absence. From [(Escobar et al., 2020)]

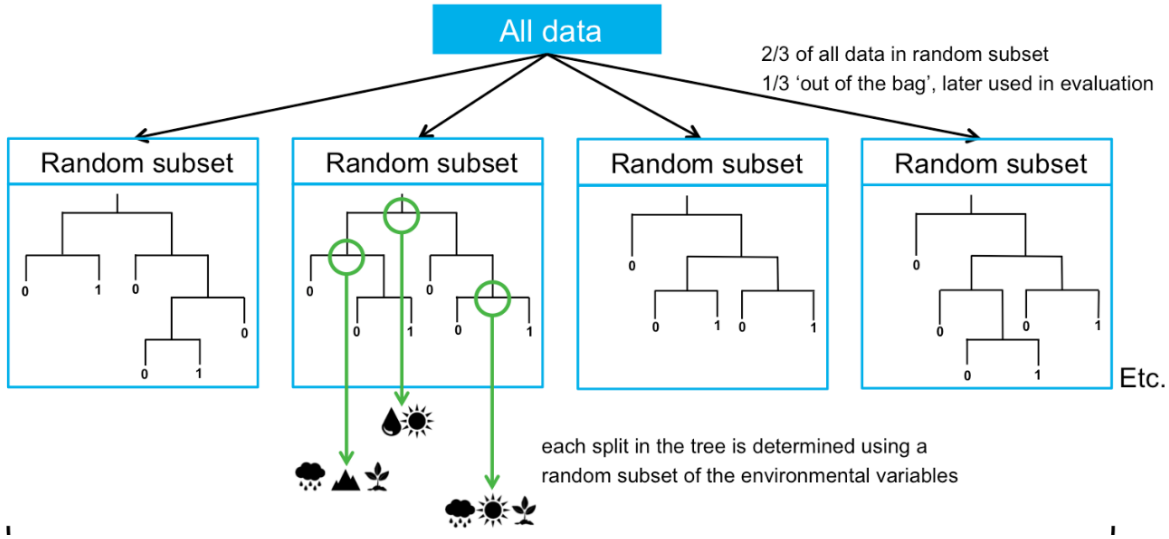


Figure 7: random forest principle for SDM. From [(Cutler et al., 2007)]

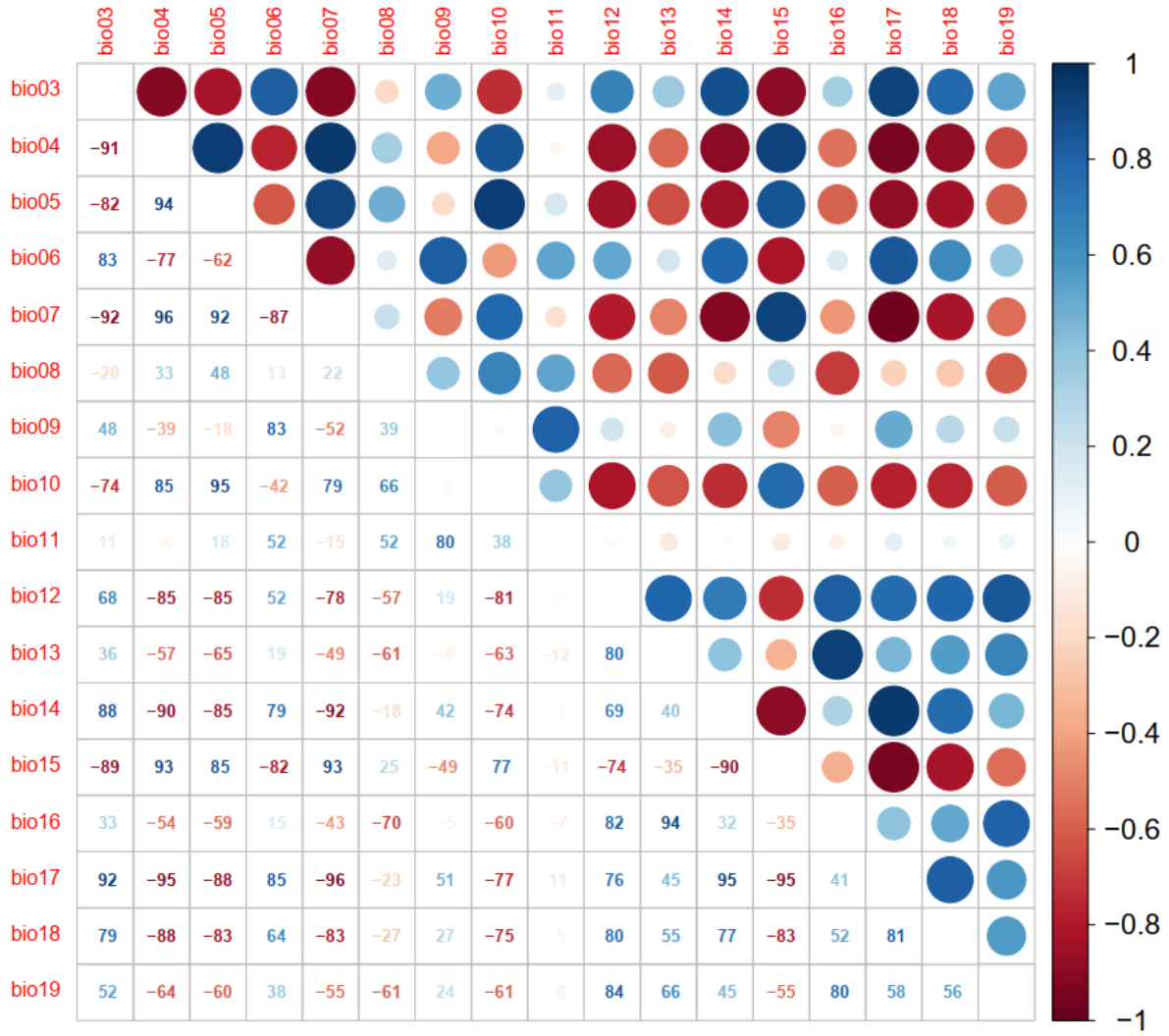


Figure8: pairwise correlation matrix on the 19 bioclimatic variables from worldclim in the region of the red-bellied monkey (this matrix is then use to remove the less correlated variables and make GLM and RF models fit better)

Tables

Environmental variable	Value taken by the environmental variable in the field
Temperature	25°C to 32°C
Habitat type	Swamp forests, westlands, canopy (eg. Lama forest, Lokoli westlands, Ouémé canopy) Dense and diverse vegetation from 2 to 15m
Altitude	0 to 400 m
Humidity	Very high
Precipitation	1 150 mm (Lokoli) 1 112 mm (Lama)
Hour's range of activity (diurnal or nocturnal)	Diurnal
Diet	Mostly fruits, but also leaves and insects

Table 1a: environmental conditions for Cercopithecus erythrogaster (studies in Togo, Benin and NIgeria)

Milieux	Localités	Nombre de strates	Composition floristique	Alimentation disponible	Conditions hygrométriques	Caractéristiques spatiales	Présence du singe
Togbota	Adjossito	2	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Agbodo	2	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Bededji	2	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Doaffa	1	plantation monospécifique	beaucoup de ressources alimentaires	peu humide	proche des autres	absent
	Gbadji	1	plantation monospécifique	peu de ressources alimentaires	peu humide	proche des autres	absent
	Houtan	1	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	absent
	Hounvigué	2	peu diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Isawémé	3	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Sahoua	1	plantation monospécifique	peu de ressources alimentaires	peu humide	proche des autres	absent
	Sotor	1	peu diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Tovouto	2	plantation monospécifique	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Vazoun	3	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	absent
F.C. Lama	Layon 15	2	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	absent
	Layon 14	3	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Lay on 13	3	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Layon 12	3	diversifiée	beaucoup de ressources alimentaires	humidité temporaire	proche des autres	présent
	Lay on 11	2	diversifiée	peu de ressources alimentaires	humidité temporaire	proche des autres	absent
	Lay on 10	1	plantation monospécifique	peu de ressources alimentaires	humidité temporaire	proche des autres	absent
	Layon 9	2	plantation monospécifique	peu de ressources alimentaires	humidité temporaire	isolé	absent
Lokoli	Lokoli	3 strate	diversifiée	beaucoup de ressources alimentaires	humidité permanente	isolé	présent

Table 1b: environmental conditions for Cercopithecus erythrogaster (Kassa et al., 2014)

Environmental variables	Best models
Precipitation	IPSL-CM6A-LR, NESM3, CMCC-CM2-SR5, ACCESS-ESM1-5
Maximum temperature	INM.CM4-8, BCC-CSM2-MR, MRI-ESM2-0, ACCESS-ESM1-5
Minimum temperature	AWI-CM-1-1-MR, IPSL29 CM6A-LR, INM.CM5-0, CanESM5

Table 2: climatic models best fit with different environmental variables in Nigeria (Shiru, 2021)

Pros	Cons
Can handle non-linear relationships	Biases if number of variables is too small or too large
Low biases in general	Issues if number of trees is too large
Can manage different types of data and missing data	Low speed on large datasets
Robustness regarding outliers and measurement errors	
Avoidance of overfitting issues (overfittiong = high variance, noisy performance on non-training data)	
Cross-validated results (out-of-bag samples)	
Always converges (see math demonstration: Breiman, 2001)	
Interpretability: easy identification of the most important variables for the model	
Good speed in general	

Table 3: pros and cons of random forest algorithm (Scornet et al., 2015; Elith et al., 2006; Luan et al., 2020; Probst et al., 2019; Phillips et al., 2006).

Examples of statistical tests for model assessment (GLM and RF)
Threshold-dependant performance measures (convert the continuous predictions in binary, like the absence-presence vector and compare with this vector)
Threshold-independent performance measures (AUC: Area Under the ROC curve)
Assessing novel environment (MESS: Multivariate Environmental Similarity Surface)

Table 4: examples of statistical tests for model assessment of GLM and RF.

Annex

GLM results (analysis)

In the following analysis, the input df is a joined dataframe of presence-absence coordinates of the species and the 19 environmental variables from Bioclim. One line of the dataframe therefore include a location with coordinates (latitude + longitude), presence or absence of the species (1 or 0) and 19 environmental variables at this location (mean temperature, mean precipitation, …).

The variable to be explained ("variable à expliquer") is the third column of the dataframe with the presence/absence of the species (1 for presence, 0 for absence). The explicative variable(s) ("variable(s) explicative(s)") is made of a subset of the 19 columns for environmental variables.

As presented in the principle of GLM for SDM, we always use the parameter family = "binomial" of the *glm* function, so that the algorithm uses the logit link function (logistic regression).

The null hypothesis in GLM is that there is no significative relation between the variable to be explained and the explicative variable(s).

m1 (bio01)

```
Call:
glm(formula = presence_c ~ bio01, family = "binomial", data = df)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
2.409e-06  2.409e-06  2.409e-06  2.409e-06  2.409e-06

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.657e+01  6.081e+05       0        1
bio01       2.131e-09  2.272e+04       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 164  degrees of freedom
Residual deviance: 9.5726e-10  on 163  degrees of freedom
AIC: 4

Number of Fisher Scoring iterations: 25
```

This is a first simplified test of GLM, to have a first insight at the R output of GLM. We adjusted the model with the environmental variable « bio01 » as the only explicative variable. As a reminder, "bio1" is the annual mean temperature.

The R output shows that the slope value between presence probability of the species and bio1 environmental variable is near 0 (2.131e-09), with a p-value of 1. The p-value of 1 shows that we absolutely cannot reject the null hypothesis, that there is not enough proof for rejecting the null hypothesis. The slope of 0 indicates no correlation at all (neither positive nor negative) between

presence of the species and the annual temperature. These results are probably due to either a lack of data (we only have 67 presence points and 100 absence points) or a mistake of ours in the input.

The deviance and AIC are a measurement of model fitting.

The null deviance is the deviance of the model when only considering the intercept/constant, so without the explicative variable “bio1”. The null deviance is 0 (or at least rounded down to 0), which is coherent because we have no correlation here.

The residual variance is the difference between the resulted deviance and the null deviance. It is near 0 for the same reason as above.

The AIC is the Akaike Information Criterion. A lower AIC shows a better model fitting and ability to explain the data. Here, the AIC is 4 and we will be able to compare it with following AIC to select the lowest AIC possible.

The number of Fisher Scoring iterations indicates whether a simpler model (with less variables) could give the same results. It compares the deviance of both models with a chi² distribution. A higher Fisher Scoring shows a better model fitting for the model of interest than a simpler model. Here, the number Fisher Scoring iterations is 25 and we will be able to compare it with following Fisher Scoring to select the highest possible.

m2 (bio01 with squared term)

```
Call:
glm(formula = presence_c ~ bio01 + I(bio01^2), family = "binomial"
    data = df)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
2.409e-06  2.409e-06  2.409e-06  2.409e-06  2.409e-06

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.657e+01  8.233e+06       0        1
bio01       -1.534e-08  6.343e+05       0        1
I(bio01^2)   2.771e-10  1.220e+04       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 164  degrees of freedom
Residual deviance: 9.5726e-10  on 162  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 25
```

This model is a complementary study to the preceding one. It adds a quadratic term on the same explicative variable “bio1” to see if a quadratic relation instead of a linear relation could explain the data better.

The R output shows that effect of bio1 on the presence probability of the species is near 0 (-1.534e-08) and so is the effect of its quadratic term (2.771e-10). They both have a p-value of 1. As before, this suggests no correlation (neither positive nor negative) between presence of the species and the annual temperature.

Again, the deviances are near 0.

The AIC is 6, which is less good than the AIC of 4 obtained in m1.

The number of Fisher Scoring iterations is the same as before.

m3 (19 variables)

```
Call:
glm(formula = presence_c ~ bio01 + bio02 + bio03 + bio04 + bio05 +
    bio06 + bio07 + bio08 + bio09 + bio10 + bio11 + bio12 + bio13 +
    bio14 + bio15 + bio16 + bio17 + bio18 + bio19, family = "binomial",
    data = df)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
2.409e-06  2.409e-06  2.409e-06  2.409e-06  2.409e-06

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.657e+01  3.341e+06       0        1
bio01        8.845e-07  4.395e+05       0        1
bio02       -1.308e-07  3.168e+05       0        1
bio03        4.495e-08  5.058e+04       0        1
bio04        2.963e-09  1.119e+04       0        1
bio05       -1.348e-01  9.732e+10       0        1
bio06        1.348e-01  9.732e+10       0        1
bio07        1.348e-01  9.732e+10       0        1
bio08       -2.276e-07  7.941e+04       0        1
bio09       -1.069e-07  8.115e+04       0        1
bio10        2.066e-07  4.815e+05       0        1
bio11       -4.253e-08  3.446e+05       0        1
bio12        7.113e-10  5.792e+02       0        1
bio13       -1.019e-09  2.623e+03       0        1
bio14       -5.984e-08  2.227e+04       0        1
bio15       -9.601e-09  9.760e+03       0        1
bio16       -2.159e-10  1.292e+03       0        1
bio17        9.755e-09  7.887e+03       0        1
bio18       -1.160e-09  1.010e+03       0        1
bio19       -2.503e-10  3.226e+02       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 164  degrees of freedom
Residual deviance: 9.5726e-10  on 145  degrees of freedom
AIC: 40

Number of Fisher Scoring iterations: 25
```

This model takes all the 19 environmental variables from worldclim as explicative variables.

The R output is unexpected (and disappointing!) with again only p-values of 1 which do not allow us to make any correlation assumptions.

Again, the null deviance is 0, indicating that no information is provided thanks to the 19 explicative variables. No correlation can be made.

The AIC is 40 which indicates a poor model fitting compared to m1 and m2 models.

Again, the number of Fisher scoring iterations is 25.

As a remark, we note the coefficients for "bio05", "bio6" and "bio7" are the most significant (1.348e-01 compared to the other coefficients in range e-08 to e-10) but these coefficients also have a very high standard error (9.732e+10). We can also notice that the values of the coefficients are the same, indicating that they are collinear. This could be explained by the relation bio7 = bio5-bio6 (temperature annual range = max temperature of warmest month – min temperature of coldest month). If the p-value were not 1, the results would suggest that even though the presence of the monkey is not influenced by the annual mean temperature 'bio1' as suggested by models m1 and m2, it is impacted by too low or too high temperature as described by 'bio5' and 'bio6'. It also does not seem to be influenced by any other environmental variables describing precipitation, diurnal range or seasonality.

m4 (3 variables with interactions)

```
Call:
glm(formula = presence_c ~ bio01 * bio02 * bio03, family = "binomial",
    data = df)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
2.409e-06  2.409e-06  2.409e-06  2.409e-06  2.409e-06

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.657e+01  4.055e+07       0        1
bio01             8.848e-06  1.521e+06       0        1
bio02             1.806e-05  4.303e+06       0        1
bio03             3.683e-06  6.420e+05       0        1
bio01:bio02      -6.531e-07  1.606e+05       0        1
bio01:bio03      -1.341e-07  2.409e+04       0        1
bio02:bio03      -2.666e-07  7.159e+04       0        1
bio01:bio02:bio03  9.658e-09  2.672e+03       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 164  degrees of freedom
Residual deviance: 9.5726e-10  on 157  degrees of freedom
AIC: 16
```

This model takes 3 environmental variables 'bio01' (annual mean temperature), 'bio02' (mean diurnal range) and 'bio03' (isothermality) as explicative variables. Moreover, it takes into account the interactions between these 3 variables (two-way interactions between 'bio01'-'bio02', 'bio02'-'bio03' and 'bio01'-'bio03' and a three-way interaction between 'bio01'-'bio02'-'bio03').

The R output again only has p-values of 1 which do not allow us to make any correlation assumptions.

Again, the null deviance is 0, indicating that no information is provided thanks to each of these variables and their interactions. No correlation can be made.

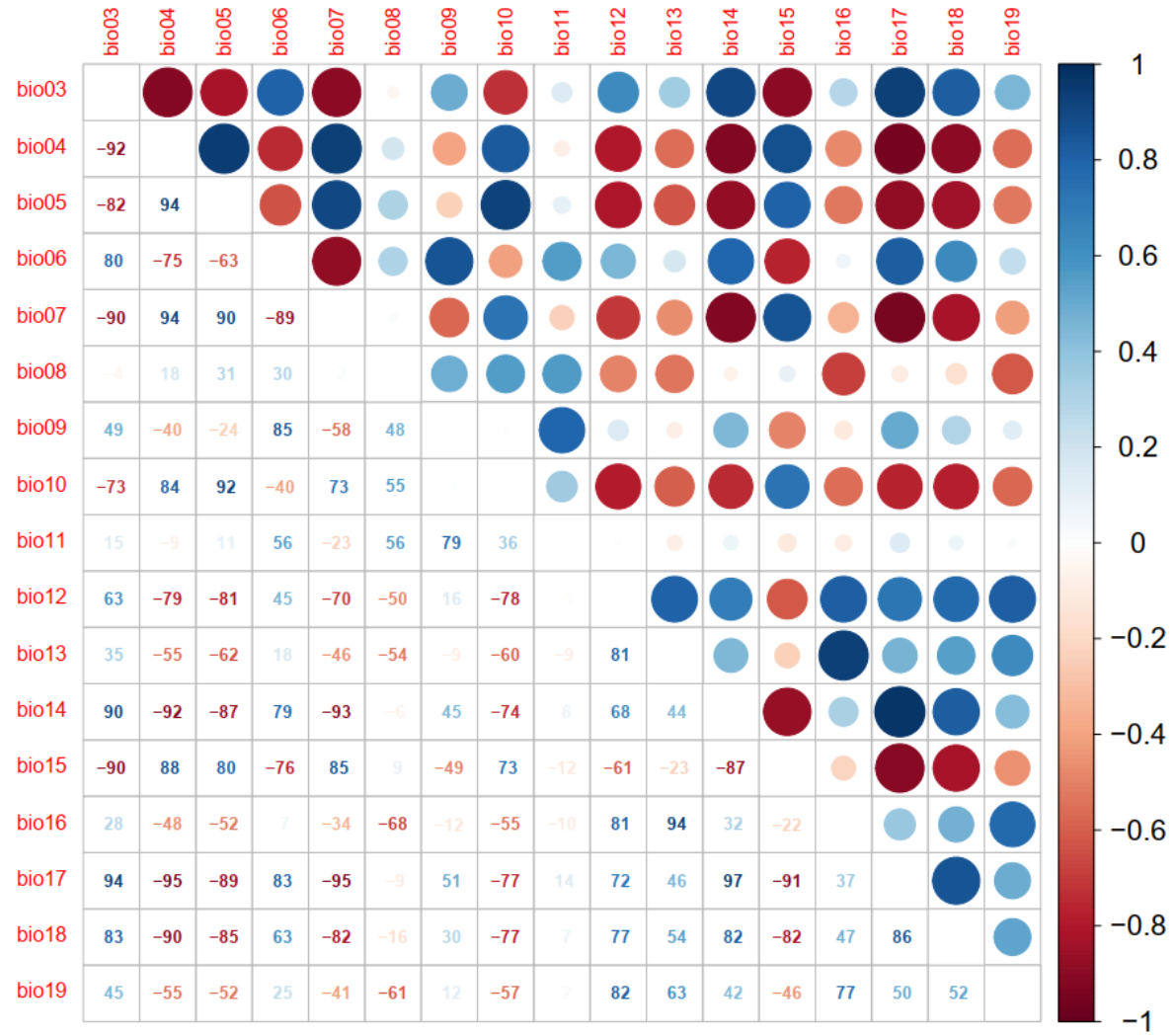
The AIC is 16 which indicates a poor model fitting compared to m1 and m2 models.

Again, the number of Fisher scoring iterations is 25.

Correlation matrix (to help fit GLM model better)

GLM have problems to fit stable parameters if two or more predictor variables are highly correlated, resulting in so-called multicollinearity issues *(Dormann et al. 2013)*. Performing correlation tests on the environmental variables is important (without taking presence/absence data into account, just the region of interest).

Call:
cor(df[, -c(1:3)], method = "spearman")



The aim of the correlation matrix is to obtain pairwise correlation between the environmental variables in the region of interest (at the points of presence and absence of the species). It is useful for identifying

patterns and selecting the variables relevant for modeling. We adjusted the correlation matrix with the Spearman method (Zurell et al. 2020).

The correlation matrix indicates the strength and direction of the relationship between each pair of environmental variables. The values in circles on the upper right range from -1 to 1, where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation. It is very convenient for visualization: the bigger the circle is, the bigger the correlation, and the bluer the circle is, the more positive the correlation. The values in the bottom left represent the same written in numbers; note that they are multiplied by 100 to make it more readable.

For example, the correlation between 'bio03' (isothermality) and 'bio04' (temperature seasonality) is -0.92, which indicates a strong negative correlation between these two variables. Another example is the correlation between 'bio03' (isothermality) and 'bio14' (precipitation of driest month) which is 0.901. This indicates a strong positive correlation between these two variables. Therefore, this suggests that, for example, we could just keep 'bio03' in the GLM and remove 'bio04' and 'bio14' without too many consequences.

Different theoretical frameworks and R packages exist to get the highly correlated variables removed. We used the caret package, and the remaining variables were 'bio06', 'bio08', 'bio11', 'bio14' and 'bio16'.

m5 ('bio06', 'bio08', 'bio11', 'bio14', 'bio16' with their squared term)

Therefore, we performed a new GLM on the remaining variables, which are the least correlated ones, hoping for a better model fitting this time.

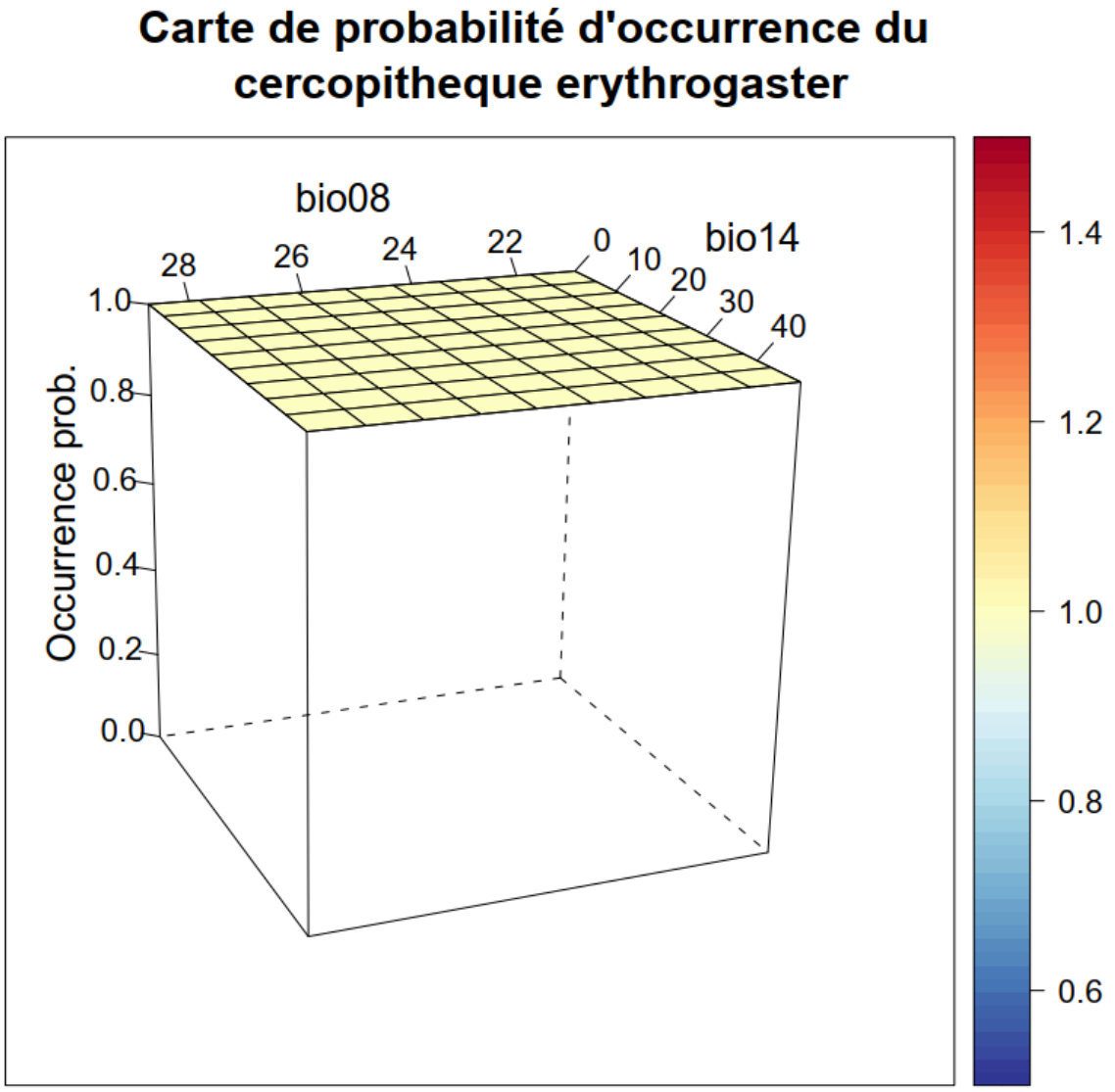
```
Call:
glm(formula = presence_c ~ bio06 + I(bio06^2) + bio08 + I(bio08^2) +
    bio11 + I(bio11^2) + bio14 + I(bio14^2) + bio16 + I(bio16^2),
    family = "binomial", data = df)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
2.409e-06  2.409e-06  2.409e-06  2.409e-06  2.409e-06

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.657e+01  1.148e+07       0        1
bio06       -3.547e-06  2.291e+05       0        1
I(bio06^2)   1.056e-07  6.121e+03       0        1
bio08       -1.353e-05  8.114e+05       0        1
I(bio08^2)   2.733e-07  1.559e+04       0        1
bio11        9.201e-06  1.276e+06       0        1
I(bio11^2)  -1.853e-07  2.588e+04       0        1
bio14       -1.655e-08  1.604e+04       0        1
I(bio14^2)   7.596e-10  3.156e+02       0        1
bio16        2.148e-08  1.223e+03       0        1
I(bio16^2)  -1.423e-11  7.590e-01       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 164  degrees of freedom
Residual deviance: 9.5726e-10  on 154  degrees of freedom
AIC: 22

Number of Fisher Scoring iterations: 25
```

This model takes 4 environmental variables (the least correlated ones) with a linear or quadratic term as explicative variables.

The R output still shows only p-value of 1, which does not allow for any correlation hypothesis.

Occurrence probability with 'bio08' and 'bio014' (a tool for assessment of GLM model)

The occurrence probability predicted thanks to 2 well-chosen environmental variables are also called 'response curves'. It is a look at the ecological niche in a n-dimension hypervolume, before it is reprojected on the maps.



Without surprises, due to the fact that no correlation was found between the different environmental variables, the response curve has no variation. The occurrence probability of the species is 1 no matter the value of bio08 and bio14. This further suggests that there is a mistake in our dataframe of presence/absence data, though we have investigated it and did not find obvious problems.

RF results (analysis)

No results yet.

Our personal experience and work distribution

Despite numerous difficulties encountered on the code with a lot of issues for packages, we both enjoyed the project very much and found it particularly interesting. Conservation biology is a field we were both very keen to work on and we are still convinced that SDM was a great choice of subject.

The biggest regret we have is for math, which we have neglected because of the challenges encountered with coding. We would have wanted to explore more the theoretical aspects of SDM and random forest algorithm, and we are frustrated to have had no results to perform statistics on.

Other than that, we learned a lot. We learned that own-designed research project must have a clear frame and brings its share of surprises: we felt our way along, and discovered fields and questions that we were not expecting at first (dealing with spatial data, new algorithms, new statistical tests, …). Tidjani has been especially amazed by the diversity of models we can use to describe purely biological data. We both became truly aware that creativity, joined efforts and being receptive to each other are important quality in research. Knowing where to seek for help when necessary is also important. Time management is something we found challenging – areas of improvement would be on referring more to the Gantt Chart all along the project and on being able to say 'stop' when needs be, especially when facing coding issues. We both fully realized that research never ends, and each step opens the door to a new starting point.

To summarize, our personal experience can be described by these two African proverbs: *« La connaissance est comme un baobab, personne ne peut l'enlacer complètement »* to refer to the research experience; *« On ne peut traverser la rivière avec une seule main »* to refer to teamwork and the necessity for questions and seeking help from others.

Work distribution was approximately 50/50 between each student.

Tidjani has worked on ecological niche theory and on the studies of the red-bellied monkey in the region of interest, in order to figure out the relevant biological features of the species for SDM. He has then worked in R, finishing the preparation of the datasets, and worked on model fitting by performing GLM on the datasets. He has succeeded to make progress with the code where Anaïs got stuck.

Anaïs has worked on the conceptualization part of SDM, reviewing the different models and algorithms used for SDM and their differences, in order to figure out which model and algorithm we would be using on our case study. She then worked on the collection of the data from GBIF, WorldClim and ESA. She has done some of the preparation and data cleaning, and she struggled with spatial packages in R, without success.

Acknowledgements

We first want to thank Didier FORCIOLI, our teacher in population and evolutionary genetics, for giving us the idea of Spatial Distribution Modeling. We also want to thank the three teachers who supervised us all along the project, Franck DELAUNAY, Christophe BECAVIN, Simon GIREL, with whom we had pleasure to talk over the project and exchange ideas. They were there for us whenever needed. Finally, we want to thank all the scientific community who publishes in open source, and especially the researchers in Togo, Benin and Nigeria who provided us with field data, Damaris ZURELL for her incredible tutorial on SDM as well as all the people who contributed to R documentation on the subject. We also thank Kilian WEINBERGER whose videos on statistics and machine learning helped us understand the mathematical concept of generalized linear models and random forests.

Bibliography (15 max)

[1] Purvis, A., Jones, K. E., & Mace, G. M. (2000). Extinction. *BioEssays*, *22*(12), 1123–1133. https://doi.org/10.1002/1521-1878(200012)22:12<1123::AID-BIES10>3.0.CO;2-C

[2] Kassa, B., Nobimè, G., Hanon, L., Assogbadjo, A. E., & Sinsin, B. (2014). Caractéristiques de l'habitat du singe à ventre rouge (Cercopithecus e. Erythrogaster) dans le Sud-Bénin. In A. Fournier & G. A. Mensah (Eds.), *Quelles aires protégées pour l'Afrique de l'Ouest? : Conservation de la biodiversité et développement* (pp. 262–271). IRD Éditions. https://doi.org/10.4000/books.irdeditions.8037

[3] Agbessi, K. G. E. (n.d.). *Dynamique spatiale des populations de Cercopithecus erythrogaster erythrogaster Gray dans le complexe d'aires protégées Togodo (Togo)*.

[4] Peterson, A. T., & Soberón, J. (2012). Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. *Natureza & Conservação*, *10*(2), 102–107. https://doi.org/10.4322/natcon.2012.019

[5] Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277. https://doi.org/10.1111/ecog.04960

[6] McSHEA, W. J. (2014). What are the roles of species distribution models in conservation planning? *Environmental Conservation*, *41*(2), 93–96. https://doi.org/10.1017/S0376892913000581

[7] Pecchi, M., Marchi, M., Burton, V., Giannetti, F., Moriondo, M., Bernetti, I., Bindi, M., & Chirici, G. (2019). Species distribution modelling to support forest management. A literature review. *Ecological Modelling*, *411*, 108817. https://doi.org/10.1016/j.ecolmodel.2019.108817

[8] Beery, S., Cole, E., Parker, J., Perona, P., & Winner, K. (2021). Species Distribution Modeling for Machine Learning Practitioners: A Review. *ACM SIGCAS Conference on Computing and Sustainable Societies*, 329–348. https://doi.org/10.1145/3460112.3471966

[9] Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ*, *5*, e2849. https://doi.org/10.7717/peerj.2849

[10] van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, *39*(6), 542–552. https://doi.org/10.1111/ecog.01509

[11] Breiman, L. (2001). Random Forests (mathematics). *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[12] Shiru et. Al (2021) *Performance Evaluation of CMIP6 Global Climate Models for Selecting Models for Climate Projection over Nigeria*.

https://doi.org/10.21203/rs.3.rs-642786/v1