DeepData: Machine Learning in the Marine Ecosystems

Leonor Pimenta de Oliveira e Silva
leonor.oliveira.e.silva@tecnico.ulisboa.pt

INESC-ID / Instituto Superior Técnico, Lisboa, Portugal

October 2019

Abstract

This work presents a web-based machine learning tool to facilitate biologists work of building species distribution models. The `DeepData` web-based tool takes into account the way biologists deal with species distribution models nowadays. Biologists mostly use probabilistic algorithms, such as maximum entropy, generalized linear models and generalized addictive models. We propose the use of machine learning algorithms, such as classification and regression trees, random forest and support vector machines. Other steps involved in the species distribution models, such as data preparation and model evaluation, are also discussed. A concrete explanation of the use of the web-based tool is made, as well as the details of implementation and evaluation.

Keywords: machine learning, species distribution models, marine ecosystem

1. Introduction

The world's oceans face increasing pressure from human influences. Marine ecosystems are utilized by several economic sectors, namely commercial and recreational fishing, tourism and passenger transportation. Species are vulnerable to impacts from all these activities due to competition with fisheries, habitat degradation and disturbance [16, 17].

Given its high levels of biodiversity and wealth of resources, spatial planning is recognized as an essential tool for effective management of all human activities occurring in the deep sea and to ensure a sustainable exploitation of its resources [11]. The success of spatial planning and the design of protected areas rely on a good understanding of the spatial distribution patterns of species. Through research and monitoring of species, datasets are created in order to help understanding and managing ecosystems by the characterization of the species habitats. With a reliable dataset consisting of locations where species have been observed, a pattern of the suitable conditions of each species can be inferred. As a result, one can try to infer where each species occurs and does not occur without having to sample the whole ocean. This information can then be used to infer the status of the species. Yet, extensive sampling programs for the deep-sea are costly and technically challenging, in comparison to shallow inshore waters, where spatial planning is a much easier task.

Species distribution models (SDMs) explore these relations between environmental and species, to predict the distribution of species across geographic space. As technology evolves, new methods appear for biologists to model species' distributions.

In order to to help biologists with these new approaches, and as a way to facilitate their use, our aim is to develop a web-based machine learning tool. Nowadays, biologists have to program the species distribution models, which can sometimes be hard as it is not their area of expertise.

2. Background

Species distributions models (SDMs) assume that species distributions depend on the physical environment. The concept that species distribution depends on the environment is known as an ecological niche. Therefore, this area of study is also referred to as ecological niche models. An ecological niche describes how an organism or population responds to the distribution of resources and competitors, and how in turn it alters those same factors. According to the ecological niche theory, species are constrained by their tolerance to environmental factors [15].

SDMs try to understand this ecological niche so that it is possible to explain the environment that each species depends on. By projecting this environment into geographic space, it is possible to estimate species' geographic distribution, predicting where the species could survive. Species distribution models are a very useful mechanism to monitor the variations in habitat suitability of species, impacts of climate change and studies of species delimitation [29]. To do this, SDMs use species oc-

currence data and environmental data. By interpolating both datasets, SDM finds a pattern that describes the ecological niche. Model usefulness and robustness is influenced by the selection of variables and modeling methods and how the relation between environmental and geographic factors is handled [10]. The SDM creation is composed of 3 main steps: (i) data pre-processing, (ii) model selection and training and (iii) model evaluation.
The SDM creation requires that each step is performed multiple times as evaluation is done and knowledge is gained, leading to a better fit of the SDM. There is no known right way to create a SDM, only main steps that serve as guidelines.

2.1. Data pre-processing

SDMs relate occurrence's data with environmental data that is thought to determine the species distribution. Therefore, SDMs assume that the occurrence's data covers the species full ecological range. One of the problems of SDMs is having enough occurrence's records, as well as accurate and relevant environmental variables at a sufficiently high spatial resolution.

Regarding occurrence data, the coordinates of the location data need to be accurate so that the species/environment association is reliable. Even taking into account this precaution, occurrence's data might be biased towards the accessibility of sampling locations. Data may be lacking for remote areas.

There are two types of occurrences data: presence only and presence/absence. Presence only refers to only having the location of where the species are present. Presence/absence refers to when we have the location of both where the species are present and are absent. When dealing with absence, we have to be careful because they can mean that an habitat is unsuitable or it is suitable but unoccupied (maybe because it is inaccessible). This type of data is also tricky to get, because the fact that a species is not detected in a location at a moment in time, does not mean it does not exist there. When absence data is not available, it can be inferred based on the presence data, generating pseudo-absence.

Environmental data need to be in grid type format, where each environmental variable is divided into grid cells representing its value for a location at some resolution. To predict the values of the unknown cells, it is used spatial interpolation, which is possible due to spatial autocorrelation [1]. Spatial autocorrelation states that the closer together two locations are, the more similar are their measures of species occurrences [4, 7]. This similarity is due to biotic processes, such as reproduction, predator-prey interactions, food availability, etc. This similarity phenomenon leads to dependence among samples decaying with distance, which violates the assumption of independence of data. Also leads to underestimation of variance and overestimation of significance of effects.

The success of the prevision depends mostly on the quality of the information used, both of species and environment, because it cannot be biased and it is the base of the learning process. Therefore, preprocessing is the part that takes the longest, and is done various times along the whole process.

2.2. Model selection

Models for prediction need to balance specific fit to the training data against the generality that enables reliable prediction to new cases.

The considered statistical algorithms are:

- Generalized Linear Models (GLMs) [8],
- Generalized Addictive Models (GAMs) [13],
- Maximum Entropy (MaxEnt) [28].

Moreover, the considered machine learning algorithms are:

- Random Forests (RFs) [3],
- Support Vector Machine (SVM) [5],
- Artificial Neural Network (ANN) [25].

While MaxEnt is a presence-only model, the remaining ones are presence-absence models.

2.3. Model evaluation

Assessing the model performance [26, 12, 2] helps determining its suitability and the aspects that need improvement. Without a relevant accuracy assessment, the model has no value. It also allows comparing different models.

The response of the model can be either quantitative (continuous responses) or qualitative (categorical responses), requiring different analysis.

Regarding qualitative response variables, performance can be assessed by constructing a confusion matrix. From the confusion matrix various measures of performance can be derived such as:

- Accuracy,
- Sensitivity,
- Specificity,
- Kappa statistic.

The accuracy measures the proportion of correctly predicted instances. The problem with accuracy is that it is prevalence sensitive. Other measures, such as sensitivity, which measures the proportion of observed presences that are predicted as

[1] Spatial autocorrelation can be assessed through Moran's I measure [22].

such, and specificity, which measures the proportion of observed absences that are predicted as such, are independent of prevalence. The Kappa statistic assesses the extent to which models predict occurrence at a rate higher than expected by chance.

A disadvantage of these metrics is that they are threshold dependent. By using categorical responses, a threshold must be defined to classify an instance as one class or another.

A prevalence and threshold independent measure is the relative operating characteristic (ROC) curve. The ROC curve is a graphical measure that describes the compromise made between sensitivity and false positive as the decision threshold varies. False positive measures the proportion of observed absences that are predicted as present. By knowing the number of observed presences, if we have the sensitivity measure we know the number of correctly predicted presences and we infer the number of absences that were incorrectly predicted as presences. Similarly, by knowing the number of observed absences, if we have the false positive measure we know number of incorrectly predicted presences and we deduct the number of correctly predicted absences. Therefore, the ROC Curve describes the whole model.

The Area Under the ROC Curve (AUC) is used as a global metric predicting the overall discriminatory ability of the model. The AUC is the probability that a randomly chosen presence site will be ranked above a randomly chosen absence site. When no absence is available, such as with MaxEnt, then the AUC is the probability that a randomly chosen presence site will be ranked above a randomly chosen background site.

Regarding quantitative response variables, the error of the model can be evaluated by:

- Mean absolute error (MAE),
- Mean square error (MSE),
- Root mean square error (RMSE).

MAE measures the average magnitude of the error, without considering their distance. MSE and RMSE give more weight to large error, because of the square. The fact that RMSE uses the root makes it easier to interpret its value.

Error assessment can also be done visually by:

- Residual vs fitted plot,
- Partial residuals plot,
- Quantile-quantile plot,
- Scale location plot,
- Residual leverage plot,
- Partial dependence plots.

A residual is the difference between the observed and the predicted. If the residual is not random, then it means that something is missing in the model. This plot tests the assumptions of whether the relationship between your variables is linear and whether there is equal variance along the regression line. Partial residuals represent the residual of a variable in regards to the occurrences after subtracting the contribution of the other variables.

The quantile-quantile plots (QQplots), which lets us assess if the data follows some theoretical distribution. Scale location plots are used to evaluate spatial correlation. These plots show if residuals are spread equally along the ranges of predictors. Residual leverage plots are used to asses influential data points, i.e points whose inclusion or exclusion produce different results on the model. Finally, partial dependence plots allow the visualization of the relationship between the occurrences and each environmental variable, while accounting for the effect of the other variables.

A way to measure how much unexplained variance there is in our model is by using deviance. Since deviance on its own does not say much, we use it when comparing a model that perfectly fits the data and the model that we are testing. Then we evaluate if the reduction on deviance by adding or removing variables is significant.

3. Implementation

`DeepData's` architecture is described in Magda Resende's thesis [30]. This work introduced some changes which are explained in this section.

3.1. Tool architecture

Data from World Ocean Atlas 2013, European Marine Observation and Data Network, Ocean Biogeographic Information System and World Register of Marine Species datasets, concerning the geographic space of the Azores EZZ, is already included in `DeepData` as the environmental variables are commonly used to model species distribution models and as a way to allow the user to experiment `DeepData` without having to insert data. Concerning World Register of Marine Species dataset, it is not used to model the species distribution model but to give information about the taxonomy of the specie that is being modelled.

The existing database is composed of the following tables:

- `Condicoes_ambientais_quarto_grau`, which stores the environmental variables (temperature, salinity, silicate, nitrate, phosphate, density, conductivity, dissolved oxygen, apparent oxygen saturation and apparent oxygen utilization) associated to its `latitude`, `longitude`, depth, source and decade.

- Ocorre_quarto_grau, which stores information about the species' occurrences. It associates the species name from table Taxonomia_Espécie with its latitude, longitude, depth, source, decade, year and month.
- Taxonomia_Reino, which stores the species' kingdom.
- Taxonomia_Filo, which associated the species kingdom with its phylum.
- Taxonomia_Classe, which associated the species phylum with its class.
- Taxonomia_Ordem, which associated the species class with its order.
- Taxonomia_Famímilia, which associated the species order with its family.
- Taxonomia_Género, which associated the species family with its genus.
- Taxonomia_Espécie, which associated the species genus with its species.

To table table Condicoes_ambientais_quarto_grau were added the primary keys resx and resy, meaning the resolution of the latitude and longitude, respectively. This change was made to prevent the situation when data of a variable is uploaded with different resolutions and have the same coordinates. For example, if a variable is uploaded with a resolution of 1° and then with a resolution of 0.5°, both starting at the same coordinates, e.g. $(-34°, 32°)$, then all the coordinates of the first upload belong to the second upload as well.

Changes to the table Ocorre_quarto_grau were also made. This table stores information about the species' occurrences. The primary keys latitude, longitude, profundidade and fonte no longer refer to the primary keys latitude, longitude, profundidade and fonte from the table Condicoes_ambientais_quarto_grau. This change allows for the existence of species data that do not necessarily match the environmental data.

A new table Indice was created, to store the characteristics of each variable, namely name, decade, resx and resy, to make DeepData load faster.

3.2. Tool implementation

For the implementation of the species distribution models, the software used was R. Although python is also commonly used for machine learning, most of the published work on SDMs used the R software. Ecologists tend to use R while computer engineers tend to use python, because R has been established for a long time and includes a broader range of methods employed in ecological analysis as well as numerous routines for data exploration [20]. Python has the advantage that it is better for deployment, and therefore it is used to implement other parts of the application, including fetching the data from the database needed for computing the SDM.

DeepData allows the user to select the inputs of the main categories:

- Species,
- Environmental variables,
- Model parameters,
- Pre-processing parameters,
- Evaluation parameters.

The species can be selected through the taxonomy hierarchy, or by directly selecting the species name. The user can also choose either to generate pseudo-absences or not.

To generate pseudo-absences, we start by classifying the background as suitable or unsuitable according to the environmental conditions of the presence localities. From the unsuitable background we can:

- Define a minimum distance to the presences.
- Select the pseudo-absences at random.
- Select the pseudo-absences with k-means clustering, i.e. taking into account the distance to between each pseudo-absence.

For the environmental variables, the user can select oceanic variables, which can have different resolutions, and terrain variables. DeepData also allows the selection of various oceanic and terrain variables, and the definition of the oceanic zone. The oceanic zone can be:

- Ocean surface, meaning that only 0 to 5 meters of depth is considered.
- Ocean floor, meaning that only 5500 to 5400 meters of depth is considered.
- Average depth of the species occurrence, meaning depth values are prioritized by number of occurrences of the species.

The interval of the ocean floor is much larger than the interval of the ocean surface, since spatial variation in environmental variables decreases with depth [6]. There is also the possibility to calculate the Morans'I correlation coefficient. Given the environmental data, Morans'I evaluates whether the pattern expressed is clustered, dispersed or random. A clustered spatial pattern means most of the values are concentrated to nearby locations or adjacent together. A random spatial pattern means the distribution of the values is homogeneous or independent. A dispersed spatial pattern means that each value from its neighboring values is located far from each other in a uniformed manner.

DeepData also measures the collinearity between variables. Collinearity refers to the existence of correlated environmental variables, which can lead to biased models due to inflated variances. Small changes in the data set can strongly affect results and so the SDM tends to be unstable (high variance) and the relative importance of the variables is difficult to assess [9]. Only checking the collinearity between pairs of variables can be limiting, so the variance inflation factor (VIF) quantifies the extent of correlation between one variable and the other remaining variables. For variance inflation factors larger than 3, which means that the standard error is 1.7 times larger than if the variables were not correlated, the modelling process stops and a pop-up appears asking the user whether he/she wants to remove any variable or not. If the user chooses to maintain all the variables then the collinearity is not verified again, while if the user chooses to remove some, then the collinearity is verified for the remaining.

DeepData allows the specification of the model parameters.

In order to compute generalized addictive models, DeepData allows the specification of the family of the distribution, which can be:

- Binomial, in the case of presence-absence data, which is used as default.
- Poisson, in the case of count data.
- Gaussian, in the case of count data with a normal distribution.

It also allows the specification of the link function, in accordance with the family. Smoothness of fit of each variable can also be controlled, differing on the basis used to represent the smooth function. Possible splines are:

- Thin plate spline, which is used as default,
- Duchon spline,
- Cubic spline,
- Spline on the sphere,
- P-splines,
- Random effects,
- No smoothing.

For computing MaxEnt, DeepData allows uploading a background file, which must be composed of the latitude and longitude. If no file is given, DeepData randomly selects 10000 points of the coordinate space to be used as background.

For computing random forests, DeepData allows tuning the following parameters:

- Number of trees to grow. This should not be set to a number too small, to ensure that every input row gets predicted at least a few times. Default is set to 500.
- Minimum size of the terminal nodes. Setting this parameter to a large number causes smaller trees to be grown (and thus take less time). The default values are 1 for classification and 5 for regression.
- Maximum number of terminal nodes that the trees can have. If not given, trees are grown as much as possible (subject to limits by node size).

For computing neural networks, the structure can be defined by first indicating the number of layers, which corresponds to the sum of the input layer, hidden layers and output layers. Afterwards, the number of perceptrons for each layer is defined. Note that neural networks do not have any default structure.

For computing support vector machines, the kernel can be defined as:

- Linear, which is the default value,
- Polynomial,
- Radial basis,
- Sigmoid.

For the modelling phase, the application allows cross-validation to be performed. DeepData allows the user to select one of the following methods:

- Holdout, which separates the dataset into train set and test set according to the fraction, being the train set larger than the test set. Training is performed as many times as there are test partitions.
- Leave one out, which separates the data in three sections and at each repetition uses two for training and one for testing.
- K-fold, which separates the data in k folds and trains the model over the k number of combinations.

- Years separation, which separates the data according to the years selected for train and test, with the constraint that each year can only be either train or test.

Regarding the model evaluation, the user has to select the metric to compute the binary map threshold and the confusion matrix. DeepData allows this threshold to be defined by:

- SES, which is the threshold value or range in values that maximizes sensitivity equal to specificity.
- Kappa, which is the threshold value or range in values with the maximum Kappa statistic.
- TSS, which is the threshold value or range in values that maximizes sensitivity plus specificity.
- LW, which is the minimum prediction probability for the occurrence (presence) records.
- ROC, which is the threshold value or range in values where the ROC curve is closest to point (0,1).
- CCR, which is the threshold value or range in values with the maximum number of presence and absence records correctly identified.
- No omission, which is the threshold value or range in values with no omission error, meaning no false positives (predicting absences incorrectly).
- Prevalence, which is the threshold value or range in values with the modeled prevalence closest to the observed prevalence.

While the first three metrics (SES, Kappa and TSS) can be applied to all models, the last metrics (No Omission and Prevalence) can only be applied to MaxEnt. The remaining metrics (LW, ROC and CRR) can be applied to all models except MaxEnt, i.e. GAM, GLM, RF, ANN and SVM.

If more than one model is selected, then an ensemble model is computed additionally to the models selected. Ensemble models [31] use multiple learning models to obtain better predictive performance than the performance of a single model. A single model can have biases and inaccuracies that affect the reliability. By combining the decisions of different models, these effects can be reduced, improving the overall performance. This is due to the fact that correct answers are reinforced while incorrect ones then to be blended. This ensemble can be done by:

- Voting,
- Averaging,
- Weighted AUC,
- Weighted Kappa,
- Weighted Sensitivity,
- Weighted Specificity,
- Weighted Proportion correct.

In voting, each model does a prediction to each data point. Each of these predictions is considered a vote. Then the final prediction, meaning the prediction of the ensemble, corresponds to the majority.

Averaging method is similar to maximum voting, but instead of the final prediction being the majority, it is the average of all the single predictions.

In weighted average, instead of being a simple average of the predictions, it gives a weight to each prediction. This weight defines the importance of each model, which can be accessed through some evaluation metric, such as accuracy, kappa statistic, sensitivity, specificity or proportion correct.

For all models DeepData returns the threshold, accuracy, omission rate, sensitivity, specificity, proportion of correctly predicted occurrences and kappa statistic of the best model. To access the overall variation of each model of the cross validation, the mean accuracy and mean threshold and corresponding standard deviations are presented. It also return the calculated VIF of each variable and the number of presences used.

DeepData returns a zip file, with model evaluation plots, specific to each model. In the zip file there is always a png file with the name of the file which plots the predicted occurrence values over the environmental space.

For the GAM model, DeepData returns a zip with also:

- Residuals.png, which is composed of: normal QQplot, residuals vs linear predictors, histogram of residuals and response vs fitted values.
- Patial_dependence_plots.png, which plots the component smooth functions of the model, in the scale of the linear predictor.
- Gam_uncertainty.png, which plots the standard error estimates returned for each prediction over the environmental space.
- Akaike's Information Criterion, which is not on the zip file, but on the evaluation results.txt.

For the RF model, DeepData returns a zip file containing also:

- `Variable_importance.png`, when more than one variable is used to do the modelling. Plots each variable importance according to the mean decrease in accuracy and the mean decrease in node purity.

- `Effect_variable.png`, which plots the marginal effect of a variable on the predicted occurrence.

For the `MaxEnt` model, `DeepData` returns a `zip` with also:

- `Species_omission.png`, which shows how testing and training omission and predicted area vary with the choice of cumulative threshold.

- `Species_roc.png`, which plots the ROC curve.

- `Species_(number of repetition)_(name of the variable).png`, which plots the response curves of a variable for each repetition.

- `Species_(number of repetition)_(name of the variable)_only.png`, which plots the response curve corresponding to a model that only uses the variable, disregarding other variables.

- `Maxent.html`, which opens an html page with all the above plots and explanation. Also, information about the statistical significance of the prediction and analysis of variable contribution is provided.

Regarding the ensemble model, `DeepData` creates the `Rplots.png` which plots the predicted occurrence values of the ensemble method over the environmental space. Finally, a file with a plot with the used presences, called the `Pplots.png`, is also generated.

4. Results

In order to show the usefulness of the `DeepData` tool, we selected papers, whose data we have access to, and tried to obtain the same results. Two papers were examined, the first regarding the use of `MaxEnt`, and the second regarding the use of `Random Forest` and `Generalized addictive model`.

4.1. First case study

4.1.1 Problem description

The first paper selected is entitled ”Habitat modelling of crabeater seals (Lobodon carcinophaga) in the Weddell Sea using the multivariate approach Maxent” [24], which uses `MaxEnt` to identify suitable habitat conditions to the crabeater seal.

Regarding species occurrence data, fifteen crabeater seals of both sexes and different age classes were equipped with satellite-linked dive recorders (SDRs) between 28 January and 6 February 1998. Regarding enviromental variables, a set of 13 environmental variables was used to analyze the habitat preferences of crabeater seals: sea ice concentration, sea ice thickness, sea ice freezing rate, water surface and bottom temperature, surface and bottom salinity, surface and bottom zonal current velocity, surface and bottom meridional current velocity, slope, and distance to shelf break.

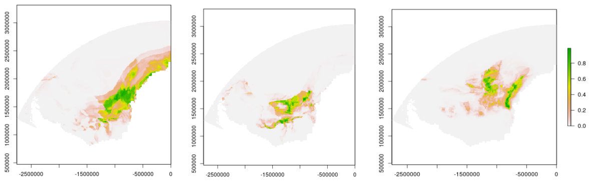


Figure 1: Probability of presence of crabeater seals for each month.

Prior to model building the seal location data were subsampled to diminish potential biases. Therefore, only location data from February, March and April 1998 were used for modelling. All data can be found on pangaea [23].

4.1.2 Tool testing

Both species occurrence and environmental data of each month were loaded into `DeepData`. To recreate the paper, the tool configurations were:

- Not to generate pseudo absences, since `MaxEnt` only uses presences.

- Use average depth of species occurrence.

- Not to calculate moran’s I.

- Use `MaxEnt`, with the background file loaded.

- Use holdout with fraction of 80% and repeat of 20.

This model was used with all variables loaded to verify the influence of each environmental variable contributing to the model by a measure called permutation importance and identify the variables that mattered most concerning the seal distribution. Jackknife test was also used to analyze the relative importance of each variable. from the 13 variables, slope, bottom zonal current velocity and bottom meridional current velocity did not contribute more than 5 % to neither monthly model and therefore they were omitted from the final model.

Regarding the models evaluation, AUC values are high, showing that the predictions are far from random. While standard deviations are low, meaning there is a high degree of uniformity between the repetitions.

4.2. Second case study

4.2.1 Problem description

The second paper selected is entitled "Population Estimates of Trindade Petrel (Pterodroma arminjoniana) by Ensemble Nesting Habitat Modelling" [19], which uses ensemble model to identify suitable habitat conditions to the Trindade Petrel.

Regarding occurrence data, 411 nests were identified between 2000 and 2007 and, afterwards, between September and November of 2014. Regarding environmental variables, a set of 5 environmental variables was used: elevation, slope, flow length, aspect and insulation.

To create the ensemble model, it starts by testing which models best fit the data, so that the best 3 models are used on the ensemble model. The tested models are: texttttGAM, `GLM`, `Multiple adaptive regression splines`, `RF`, `Generalized boosted model`, `ANN`, `MaxEnt Phillips` and `MaxEnt Tsuruoka`. The difference between the two variants of `MaxEnt` is the package that implements each of them. While `MaxEnt Tsuruoka` only uses an `R` package, `MaxEnt Phillips` uses a java software which is called within an `R` package.

4.2.2 Tool testing

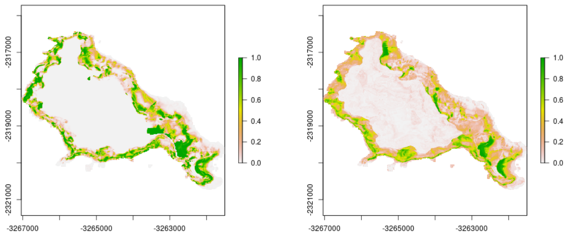
The models the allows for the use of: `GAM`, `GLM`, `RF`, `ANN` and `MaxEnt Phillips`.

Both species occurrence and environmental data were loaded into the tool. To recreate the paper, the tool configurations were:

Do this 20 times:

- Generate pseudo absences, since only maxent uses presences.
- Use average depth of species occurrence.
- Not to calculate moran's I.
- Use `GLM`.
- Use `GAM`, with binomial family.
- Use `MaxEnt`, with default values.
- Use `RF`, with classification and min node size of 5.
- Use `ANN`, with 1 layer with 8 nodes.
- Use holdout with fraction of 80% and repeat of 3.

Since the ensemble model considers one more model that the tool cannot produce, the final distribution cannot be achieved. By examining the partial dependence plots we can see that the overall tendency for the two models are the same. These plots cannot be directly compared to the response curves plots of the paper.



(a) GAM

(b) RF

Figure 2: Spatial distribution of Trindade Petrel for each model.

5. Conclusions and future work

In this work, we present a web-based machine learning tool that allows a simple and efficient way of creating species distribution models, conserving the user domain knowledge and allowing it to experiment different variable combinations and different models, while turning it more efficient as the user does not have to think about programming. It has options for all parts of the modelling process: (i) data pre-processing, (ii) model selection and (iii) model evaluation. Furthermore, it allows to load data of species and environmental data concerning the Azores Exclusive Economic Zone. Furthermore, allows the user to insert its own data of both species and environmental variables.

The developed tool provides a comprehensive interface to perform the entire modelling process using different state-of-the-art approaches. Nowadays, the two most used software tools for SDM modelling are MaxEnt and R [21]. The approaches used by these tools are quite different. While MaxEnt uses a click approach, R uses a syntax driven approach. Our tool is the balance between click and syntax driven approaches. By displaying the available options, the user clicks on the desired options and the tool generates the syntax for the R software. One concern with the click approach is that it works like a 'black-box' software, meaning that the details are hidden from the user. This thesis provides a full specification of all default options and options available for all its processes, so that the user is fully conscious of the model.

Although the interface is mostly composed of click options, flexibility is not compromised. Each modelling phase has its own options allowing great tuning, while making it easier for inexperienced users. It also allows multiple SDMs to be fitted and compared simultaneously. This makes comparison between different models possible because both preprocessing and evaluation methods that are applied are the same.

The major limitation of the `DeepData` tool is that it does not take into account that some users might want to use data that is private. Most of the studies use data that belongs to the government and therefore data that is not for public use.

One way to resolve this problem is to create information access control. Information access control is composed of authentication and authorization. Authentication is concerned with confirming that the user is who it says, while authorization is concerned with the level of access each user is granted.

Another very important and increasing problem is climate change. Successful conservation strategies will require an understanding of climate change and the ability to predict the future. There are two ways of dealing with climate change [1]:

- Mechanistic SDM, which uses physiological information about species to determine the range of environmental conditions that species can tolerate. Then, these tolerances are mapped into geographical space corresponding to the predicted species distribution.

- Climate envelope models, also known as correlative SDM, which rely on statistical correlations between occurrence data and environmental variables to outline a range (envelope) of environmental conditions within which species can exist. Data used for training and testing have a time period different from the data used to project the specie distribution.

Since mechanistic SDMs parameters are not derived from the current distribution of the species, the results are independent of the current climate. Therefore, these models have a more accurate understanding of the relationship between climate and the species life cycle. The problem with mechanistic SDMs is that the type of data it uses is hard to get and that it does not account for non-climatic influences such as biotic interactions.

With climate envelope models, even if biotic interactions are not directly modeled, by considering empirical data of the species distribution, which is constrained by non-climatic variables, these interactions are indirectly considered [27]. So, non-climatic variables are indirectly taken into account. Studies were made to evaluate the accuracy of climate envelope models compared to mechanistic SDMs [14, 18].

To implement climate envelope models, we have to allow the selection of the time period for the desired projection of species distributions.

Acknowledgements

I would like to thank my family for all the support and encouragement through all these years.

I would also like to acknowledge my thesis supervisors Prof. Maria Inês Camarate de Campos Lynce de Faria and Prof. Vasco Miguel Gomes Nunes Manquinho for their insight, support and sharing of knowledge that has made this thesis possible.

Last but not least, to all my friends and colleagues that helped me grow as a person and were always there for me during the good and bad times in my life. Thank you.

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references UID/CEC/50021/2019 and CMU/AIR/0022/2017.

References

[1] S. E. Ahmed, G. McInerny, K. O'Hara, R. Harper, L. Salido, S. Emmott, and L. N. Joppa. Scientists and software – surveying the species distribution modelling community. *Diversity and Distributions*, 21(3):258–267, 2015.

[2] O. Allouche, A. Tsoar, and R. Kadmon. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). *Journal of Applied Ecology*, 43(6):1223–1232, 2006.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[4] M. B. A. Carsten F. Dormann, Jana M. McPherson. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 2007.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[6] M. Costello, Z. Basher, R. Sayre, S. Breyer, and D. Wright. Stratifying ocean sampling globally and with depth to account for environmental variability. *Scientific Reports*, 8, 2018.

[7] B. Crase, A. C. Liedloff, and B. A. Wintle. A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, 35(10):879–888, 2012.

[8] A. J. Dobson. *An Introduction to Generalized Linear Models, Second Edition.* Taylor & Francis, 2010.

[9] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

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

[11] A. Guisan, R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. T. Tulloch, T. J. Regan, L. Brotons, E. McDonald-Madden, C. Mantyka-Pringle, T. G. Martin, J. R. Rhodes, R. Maggini, S. A. Setterfield, J. Elith, M. W. Schwartz, B. A. Wintle, O. Broennimann, M. Austin, S. Ferrier, M. R. Kearney, H. P. Possingham, and Y. M. Buckley. Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12):1424–1435, 2013.

[12] A. Guisan and N. E. Zimmermann. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2):147 – 186, 2000.

[13] T. Hastie and R. Tibshirani. *Generalized Additive Model*. American Cancer Society, 2005.

[14] R. J. Hijamans and C. H. Graham. The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, 12(12):2272–2281, 2006.

[15] A. H. Hirzel and G. Le Lay. Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, 45(5):1372–1381, 2008.

[16] IOC-UNESCO and UNEP. *Large Marine Ecosystems: Status and Trends*. United Nations Environment Programme(UNEP), 2016.

[17] IOC-UNESCO and UNEP. *The Open Ocean: Status and Trends*. United Nations Environment Programme(UNEP), 2016.

[18] M. R. Kearney, B. A. Wintle, and W. P. Porter. Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. *Conservation Letters*, 3(3):203–213, 2010.

[19] L. Krüger. Population estimates of trindade petrel (pterodroma arminjoniana) by ensemble nesting habitat modelling. 2018.

[20] J. Lai, C. J. Lortie, R. A. Muenchen, J. Yang, and K. Ma. Evaluating the popularity of r in ecology. *Ecosphere*, 10(1):e02567, 2019.

[21] E. Meineri, A.-S. Deville, D. Grémillet, M. Gauthier-Clerc, and A. Béchet. Combining correlative and mechanistic habitat suitability models to improve ecological compensation. *Biological Reviews*, 90(1):314–329, 2015.

[22] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[23] D. A. Nachtsheim, K. Jerosch, W. Hagen, J. Plötz, and H. Bornemann. Crabeater seals (Lobodon carcinophaga) in the Weddell Sea during DRE1998 campaign, with link to files of gridded seal locations and environmental parameters for Maxent analyses. PANGAEA, 2015. In supplement to: Nachtsheim, DA et al. (2016): Habitat modelling of crabeater seals (Lobodon carcinophaga) in the Weddell Sea using the multivariate approach Maxent. Polar Biology, 40(5), 961-976, https://doi.org/10.1007/s00300-016-2020-0.

[24] D. A. Nachtsheim, K. Jerosch, W. Hagen, J. Plötz, and H. Bornemann. Habitat modelling of crabeater seals (lobodon carcinophaga) in the weddell sea using the multivariate approach maxent. *Polar Biology*, 40(5):961–976, May 2017.

[25] G. Palm. Warren mcculloch and walter pitts: A logical calculus of the ideas immanent in nervous activity. In G. Palm and A. Aertsen, editors, *Brain Theory*, pages 229–230, Berlin, Heidelberg, 1986. Springer Berlin Heidelberg.

[26] J. Pearce and S. Ferrier. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133(3):225 – 245, 2000.

[27] R. G. Pearson and T. P. Dawson. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, 12(5):361–371, 2003.

[28] S. J. Phillips, M. Dudík, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, New York, NY, USA, 2004. ACM.

[29] L. L. Porfirio, R. M. B. Harris, E. C. Lefroy, S. Hugh, S. F. Gould, G. Lee, N. L. Bindoff, and B. Mackey. Improving the use of species distribution models in conservation planning and management under climate change. *PLOS ONE*, 9(11):1–21, 2014.

[30] M. C. A. Resende. Uma aplicação web para o mar profundo dos açores. Master's thesis, Instituto superior técnico, 2018.

[31] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012.