UFSM

Ciência Florestal

ISSN 1980-5098
OPEN ACCESS

Ci. Fl., Santa Maria, v. 35, e86428, p. 1-21, 2025 • https://doi.org/10.5902/1980509886428
Submitted: 18th/01/2024 • Approved: 7th/04/2025 • Published: 31st/07/2025

Articles

Analysis of algorithms for building a joint model for modeling Amazonian species

Análise de algoritmos para construção de um modelo conjunto para modelagem de espécies amazônicas

Ingrid Lana Lima de Morais[I], Alexandra Amaro de Lima[II], Ivinne Nara Lobato dos Santos[I], Lair Cristina Avelino do Nascimento[III], Santiago Linorio Ferreyra Ramos[I], Carlos Henrique Salvino Gadelha Meneses[IV], Ricardo Lopes[V], Ananda Virgínia de Aguiar[VI], Maria Teresa Lopes[I]

[I]Federal University of Amazonas, Manaus, AM, Brazil
[II]Galileo Institute of Technology and Education of Amazonas, Manaus, AM, Brazil
[III]SOS Amazônia Association, Rio Branco, AC, Brazil
[IV]State University of Paraíba, Campina Grande, PB, Brazil
[V]Embrapa Western Amazon, Manaus, AM, Brazil
[VI]Embrapa Forestry, Colombo, PR, Brazil

ABSTRACT

In order to monitor biodiversity changes in relation to climate change, different ecological niche models (ENMs) are employed. The selection of the most suitable model for a species may be constrained by various factors, such as data availability and resolution. The objective of the study was to analyze 13 algorithms and determine a consensus model to simulate the potential distribution of five deforestation-targeted species in the Amazon: *Aspidosperma desmanthum*, *Cariniana micranta*, *Clarisia racemosa*, *Couratari oblongifolia*, and *Vouchysia guianensis*. To construct the ENMs, bioclimatic and soil variables were used. The information for each species was individually modeled using the 13 algorithms, and subsequently, the average of each algorithm for all species was calculated. The performance was assessed based on metrics such as Area Under the Curve, True Skill Statistics, and Sorensen Index. Based on the results, it was observed that there is no ideal algorithm for all species. Therefore, a consensus model was proposed using the Random Forest, Boosted Regression Trees, Support Vector Machine, Bayesian Gaussian Process, and Maximum Entropy Default algorithms, as they demonstrated better performance on average. It is concluded that it is important to consider the specific characteristics of each species and the individuality of the dataset.

Keywords: Consensus model for modeling; Ecological niche; Potential species distribution; Forest species; Climate change

RESUMO

Para monitorar as mudanças da biodiversidade em relação as mudanças climáticas são utilizadas diferentes modelos de nicho ecológico (ENMs). A seleção do modelo mais adequado para uma espécie pode ser limitada por inúmeros fatores, como disponibilidade e resolução de dados. O objetivo do trabalho foi analisar 13 algoritmos e determinar um modelo consenso para simular a distribuição potencial de cinco espécies alvo do desmatamento na Amazônia: *Aspidosperma desmanthum*, *Cariniana micranta*, *Clarisia racemosa*, *Couratari oblongifolia* e *Vouchysia guianensis*. Para a construção dos ENMs foram utilizadas variáveis bioclimáticas e edáficas. As informações de cada espécie foram modeladas individualmente considerando os 13 algoritmos, posteriormente foi obtida a média de cada algoritmo para todas as espécies onde o desempenho foi analisado a partir das métricas: Area Under the Curve, True Skill Statistics e Índice de Sorensen. Com base nos resultados, observou-se que não existe um algoritmo ideal para todas as espécies, assim, foi proposto um modelo consenso a partir dos algoritmos Random Forest, Boosted Regression Trees, Support Vector Machine, Bayesian Gaussian Process e Maximum Entropy Default, uma vez que estes apresentaram melhor desempenho a partir da média. Concluímos é importante considerar as particularidades de cada espécie e a individualidade do conjunto de dados.

Palavras-chave: Modelo consenso para modelagem; Nicho ecológico; Distribuição potencial de espécies; Espécies florestais; Mudanças climáticas

1 INTRODUCTION

The need to study the effects of climate change on species distribution across different ecosystems has led to the widespread use of Ecological Niche Models (ENMs) (Guo; Li; Zhao; Nawaz, 2019). ENMs are based on the association between biotic and abiotic variables to identify the environmental conditions that allow the persistence of a given species over time (Landa; Castro; Monterrubio-Rico; Lara-Cabrera; Pietro-Torres, 2023). In recent decades, several ENMs have emerged, incorporating different parameters and input criteria (Ndao; Leroux; Hema; Diouf; Bégué; Sambou, 2022; Remya; Ramachandran; Jayakumar, 2015), many of which use presence and absence data of species (Remya; Ramachandran; Jayakumar, 2015). The availability of data from museums and herbaria has made models based on presence data more widely used (Senay; Worner; Ikeda; Novel, 2013).

Thus, the variety of approaches used in these models reflects different levels of complexity and sophistication. Simple models use distance or polygon rules to constrain a species' environmental conditions based on the extent of occurrence points (Senay; Worner; Ikeda, 2013). More refined models relate species occurrence data to environmental predictor variables to represent a species' niche (Zhao; Guo; Wei; Ran; Gu, 2017). Presence-absence models employ techniques to generate pseudo-absence points when true absence data are not available (Senay; Worner; Ikeda, 2013). When mathematically combined, these ENMs can be used to map the potential distribution of species and to extrapolate that distribution across space and time (Guo; Li; Zhao; Nawaz, 2019).

The selection of the most suitable ENMs for specific species may be limited by factors such as data availability, data resolution, and environmental complexity (Ma; You, 2022). Previous studies that have attempted to compare the performance of ENMs have shown that there is no single preferred model to be adopted. These studies emphasize that the predictive capacity of the dataset used should be tested across different algorithms (Konowalik; Nosol, 2021). In this context, adopting a model that integrates multiple algorithms helps minimize errors associated with limitations arising from the use of a single algorithm that may not encompass all the desired characteristics (Guo; Li; Zhao; Nawaz, 2019; Ma; You, 2022).

To survive the edaphoclimatic conditions of the Amazon, with its wide environmental diversification and complexity, forest species share common characteristics, such as adaptations related to genetic diversity and geographic distribution, dense foliage to compete for light in the upper canopy, efficient nutrient cycling due to the rapid decomposition of organic matter, and growth and reproductive cycles synchronized with seasonal rainfall changes, among others (Landa; Castro; Monterrubio-Rico; Lara-Cabrera; Pietro-Torres, 2023). Given the similarity patterns in species occurrence and the shared characteristics of the species used in a study, a higher likelihood of success is expected in developing an efficient consensus model for all species. This is often referred to as ensemble modeling (Estevo; Nagy-Reis; Nichols, 2017).

The aim of this study was to develop a consensus model to simulate the potential distribution of Amazonian forest species — *Aspidosperma desmanthum* Benth. ex Müll. Arg., *Cariniana micrantha* Ducke, *Clarisia racemosa* Ruiz & Pav., *Couratari oblongifolia* Ducke & Knuth, and *Vochysia guianensis* Aubl.—using ENMs available in the ENMTML package (Andrade; Velazco; Marco Júnior, 2020).

2 MATERIALS AND METHODS

2.1 Species occurrence records and data preprocessing

Five target timber forest species, affected by illegal deforestation, from different genera and with distinct distribution areas in the Amazon were identified: *A. desmanthum*, *C. micrantha*, *C. racemosa*, *C. oblongifolia*, and *V. guianensis*. The occurrence data used were obtained from the Global Biodiversity Information Facility (GBIF) database (GBIF, 2023), the Center for Environmental Information Reference (CRIA) (CRIA, 2023), via the SpeciesLink network, and the Botanical Information and Ecology Network (BIEN) database (Maitner; Boyle; Casler; Condit; Donoghue; Durán; Guaderrama; Hinchliff; Jorgensen; Kraft; McGill; Merow; Morueta-Holme; Peet; Sandel; Schildhauer; Smith; Svenning; Thiers; Violle; Wiser; Enquist, 2018). These data were restricted to the South American continent and underwent a rigorous verification process, during which points lacking geographic coordinates, duplicate points, and outlier data were removed.

To reduce autocorrelation between occurrence data and sampling bias, the occurrence locations were spatially reduced to 5 km using the “thin occ” argument. To avoid sampling bias, the data were partitioned into 4 folds using the K-fold method, where validation was performed according to the total number of folds (Andrade; Velazco; Marco Júnior, 2020).

2.2 Environmental variables used in model construction

For the construction of the ENMs, both bioclimatic and edaphic variables were used. As bioclimatic variables, 19 variables provided by the Global Climate Data (WorldClim), version 2.1, with a resolution of 2.5 arc minutes or ~0.041º (~4 km² per pixel), were used. These variables include minimum, mean, and maximum temperatures, and precipitation. The baseline period was simulated using data with 30-year intervals (1970–2000) from a set of 9,000 to 60,000 meteorological stations (Fick; Hijmans, 2017). The prediction of climatic variables for future scenarios was based on climate change projections provided in the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), generated using the atmospheric circulation models HadGEM-GC31-LL, IPSL-CM6A-LR (Firpo; Guimarães; Dantas; Silva; Alvez; Chadwick; Llopart; Oliveira, 2022), and MIROC6 (Monteverde; De Sales; Jones, 2022) for the periods 2021-2040 and 2041-2060, under two different scenarios for Greenhouse Gas (GHG) emissions: SSP2-4.5 and SSP5-8.5.

The effect of edaphic variables on the species was demonstrated through the use of 9 variables for 2 soil depths (0 to 20 cm, 20 to 40 cm), with a dataset of 18 pieces of information related to the physical and chemical properties of the soil, which have the same resolution as the bioclimatic variables (FAO and IIASA, 2023). These data are available in the Harmonized World Soil Database with a spatial resolution of 1 km² (30 seconds) (version 2.0; FAO and IIASA, 2023).

The environmental variables exhibit high collinearity, which is undesirable in the modeling process. To reduce the high collinearity among these variables, Principal Component Analysis (PCA) was applied, a statistical technique used to transform correlated variables into uncorrelated principal components. Thus, the first fourteen components were selected, which explained more than 95% of the variance in the original data, according to the criterion of maximizing explained variance. These components were used as representative environmental layers in the modeling, ensuring the inclusion of the main environmental gradients with reduced redundancy and greater statistical independence (Andrade; Velazco; Marco Júnior, 2020).

2.3 Analysis of algorithms for constructing the ensemble model

Data processing was performed using RStudio, integrated with R software (version 4.2), through the package Create Ecological Niche Models with TheMetaLand EcologyLab (ENMTML) (Andrade; Velazco; Marco Júnior, 2020). The ENMTML package provided access to thirteen algorithms for constructing individual and combined ENMs: Bioclim (BIO), Mahalanobis (MAH), Domain (DOM), Generalized Linear Model (GLM), Generalized Additive Models (GAM), Support Vector Machine (SVM), Boosted Regression Trees (BRT), Random Forest (RDF), Bayesian Gaussian Process (GAU), Maximum Likelihood (MLK), Maximum Entropy Simple (MXS), Ecological Niche Factor Analysis (ENF), and Maximum Entropy Default (MXD) (Andrade; Velazco; Marco Júnior, 2020).

For each species, the fundamental niche was estimated using the 13 algorithms, which can be classified based on the type of input data required by the model: presence, presence and pseudo-absence, and presence and background. Due to the lack of absence data for the studied species, the methodology of combining geographic and environmental data was used to allocate pseudo-absences and backgrounds (Lobo; Jiménez-Valverde; Hortal, 2010). To do this, a 50 km circular buffer was defined around the presence points, and all locations that did not share similarity with the presence points were extracted as a potential background for pseudo-absence selection. These dissimilar locations were grouped by K-means and used to select a representative sample (Senay; Worner; Ikeda, 2013). Additionally, it was determined that the number of pseudo-absences and backgrounds would be equal to the number of presence points.

Subsequently, the analysis of the metrics Area Under the Curve (AUC), True Skill Statistics (TSS), and Sorensen Index was performed. AUC was a metric obtained from the integration of the Receiver Operating Characteristic (ROC) curve (Allouche; Tsoar; Kadmon, 2007). AUC values range from 0 to 1 (Fielding; Bell, 1997), while TSS values can range from -1 to +1 (Allouche; Tsoar; Kadmon, 2007). Both metrics exhibit prevalence dependence, which led to the use of the Sorensen Index as a third option due to its independence from prevalence. The Sorensen Index ranges from 0 to 1, with values equal to or below 0.7 indicating poor performance (Leroy; Hugueny; Meynar; Barhoumi; Massi; Bellard, 2018).

The AUC (Area Under the Curve), TSS (True Skill Statistics), and Sorensen Index metrics were chosen because they offer complementary perspectives on the performance of the models. AUC is derived from the ROC curve and assesses the model's ability to distinguish between presence and absence, ranging from 0 to 1, with values closer to 1 indicating high accuracy (Fielding; Bell, 1997; Allouche; Tsoar; Kadmon, 2007). TSS measures model performance by simultaneously considering sensitivity and specificity, with values ranging from -1 to +1, where +1 indicates perfect prediction (Allouche; Tsoar; Kadmon, 2007). Although widely used, both metrics exhibit prevalence dependence, which may affect the interpretation of results in imbalanced datasets. For this reason, the Sorensen Index was included, as it is independent of prevalence and measures the similarity between predicted and observed distributions, making it especially useful in contexts with sparse occurrence data. Values below 0.7 indicate unsatisfactory model performance (Leroy *et al*., 2018). Thus, the combination of these three metrics provided a more robust and reliable assessment of species distribution models.

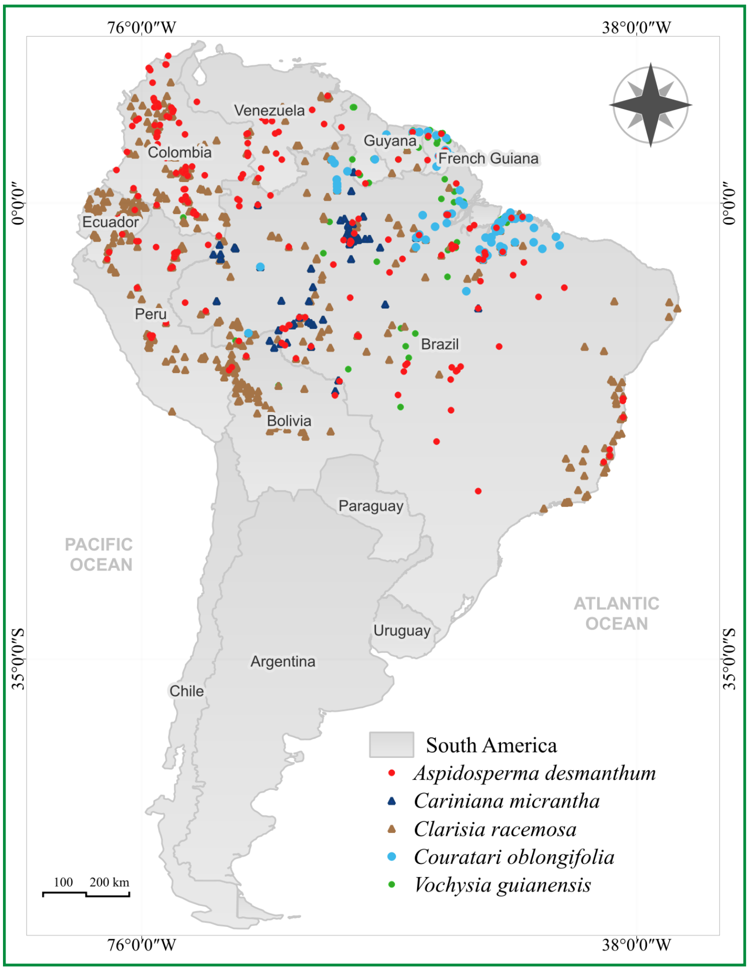
Following the aforementioned procedures, a consensus model was created using the five algorithms that performed best for ecological niche modeling of the five species mentioned earlier. The selection of the five algorithms to be used in the consensus model was based on the analysis of the mean performance of each algorithm, considering the metric values obtained for each species. Thus, the means of each algorithm for AUC, TSS, and the Sorensen Index were analyzed. Models with metric values above 0.7 were considered satisfactory (Allouche; Tsoar; Kadmon, 2007; Thuiller; Guéguen; Renaud; Karge; Zimmermann, 2019).

3 RESULTS AND DISCUSSIONS

3.1 Natural occurrence of the species

The occurrence distribution of the species can be visualized in Figure 1. The highest number of occurrence points (455) was observed for the species *C. racemosa*, distributed across Bolivia (11%), Brazil (36%), Colombia (20%), Ecuador (14%), Peru (17%), Suriname (1%), and Venezuela (2%).

Figure 1 – Distribution of Species Occurrence Points *Aspidosperma desmathum*, *Cariniana micrantha*, *Clarisia racemosa*, *Couratari oblongifolia* e *Vouchysia guianenses*



Source: Authors (2023)

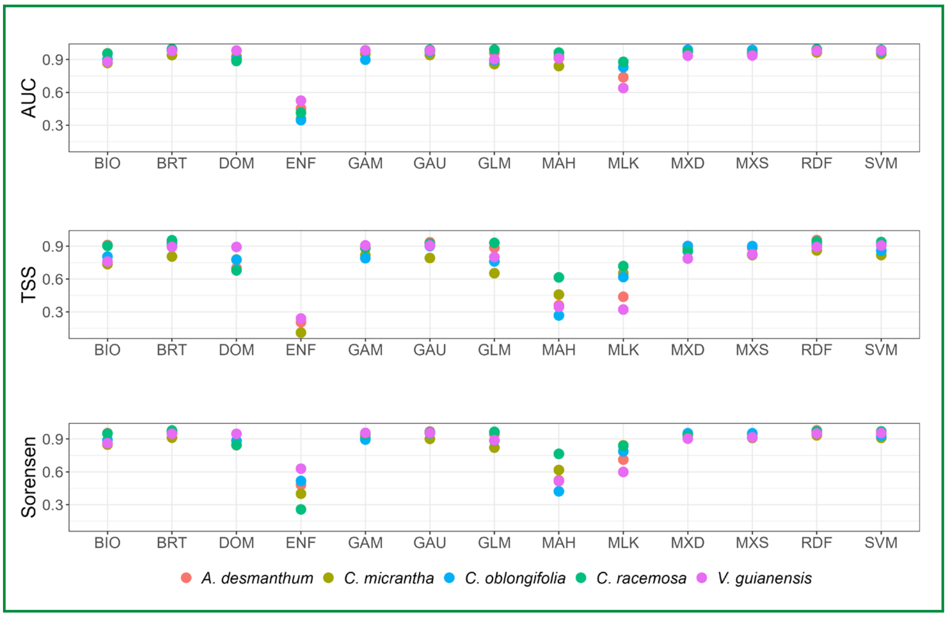
For *A. desmanthum*, 203 occurrence points were used, located in Bolivia (1%), Brazil (81%), Colombia (33%), Ecuador (4%), Guyana (1%), French Guiana (2%), Suriname (1%), Peru (10%), and Venezuela (7%). The species *C. micrantha* had 72 occurrence points, distributed across Bolivia (6%), Brazil (89%), Colombia (3%), Peru (1%), and Guyana (1%). For *V. guianensis*, 75 occurrence points were considered, with 1% of these points in Bolivia, 55% in Brazil, 1% in Colombia, 4% in Ecuador, 8% in Guyana, 17% in French Guiana, and 13% in Suriname. The lowest number of occurrence points was obtained for *C. oblongifolia*, with 80% of these points distributed in Brazil, 1% in Guyana, 10% in French Guiana, and 8% in Suriname. Spatial distribution differences were observed among the occurrence points of the species, taking into account the geographical, geological, and climatic characteristics of South America.

3.1 Comparison of ecological niche modeling algorithms

Based on the analysis and comparison of the thirteen Ecological Niche Modeling algorithms for the species *A. desmanthum, C. micrantha, C. racemosa, C. oblongifolia,* and *V. guianensis,* five algorithms were identified as most suitable for composing the consensus model. Analyzing the AUC, TSS, and Sorensen metrics, it was found that there was no consensus among the ideal algorithms for all species presented in this study (Figure 2), thus confirming the theory that there is no single ideal algorithm. On the contrary, it depends on the location and species being modeled (Konowalik; Nosol, 2021; Ndao; Leroux; Hema; Diouf; Bégué; Sambou, 2022; Qiao; Soberón; Peterson, 2015).

The algorithms used in species modeling are widely applied; however, authors often do not cite the criteria used for their selection or fail to conduct tests to identify and use the most appropriate algorithm. The first step in the modeling process is the evaluation of a set of algorithms (Qiao; Soberón; Peterson, 2015; Konowalik; Nosol, 2021). This step is important in the process due to the need to understand each algorithm, as there are various algorithms for adjusting ENMs (Andrade; Velazco; Marco Júnior, 2020).

Figure 2 – Evaluation of the thirteen algorithms available in the ENMTML package according to the metrics and species, *Aspidosperma desmanthum, Cariniana micrantha, Clarisia racemosa, Couratari oblongifolia e Vouchysia guianensis*



Source: Authors (2023)

In where: Bioclim = BIO; Mahalanobis = MAH; Domain = DOM; Generalized Linear Mode = GLM; Generalized Additive Models = GAM; Support Vector Machine = SVM; Boosted Regression Trees = BRT; Random Forest = RDF; Bayesian Gaussian Process = GAU; Maximum Likelihood = MLK;, Maximum Entropy simple = MXS; Ecological Niche Factor Analysis = ENF; Maximum Entropy default = MXD; Under the Curve = AUC; True Skill Statistics = TSS.

The results revealed that ENM algorithms are used in studies both with and without standardized criteria, which is not ideal. Predictions can differ depending on the applied model, the type of data used (presence, presence-background, and presence-pseudo-absence), the spatial area, and the availability of information (Qiao; Soberón; Peterson, 2015; Konowalik; Nosol, 2021). These factors can lead to potential underestimation or overestimation in the modeling outcomes.

From the analysis of the efficiency and performance of the ENF algorithm, it was possible to identify that the metric results for the individual species (Figure 2) were below 0.7, which indicates that this algorithm is unsatisfactory for the modeling in this study. This performance may have been influenced by the scope of the data, considering that the algorithm compares the species' distribution spatially with environmental conditions. Additionally, it is sensitive to the extrapolation of results due to an error in the formulation of the covariance matrix (Mugo; Saitoh; Igarashi; Toyoda; Masuda; Awaji; Ishikawa, 2020).

Furthermore, the BIO and DOM algorithms showed metric values above 0.7 for all species. Contrastingly, when *A. desmanthum* and *C. racemosa* were modeled with BIO, the performance values were higher than those of the other species, which can be explained by the density of their occurrence points. However, when the DOM algorithm modeled *C. micrantha*, *C. oblongifolia*, and *V. guianensis*, it was observed that despite the reduced number of occurrence points for these species, it performed better compared to those with a higher number of occurrence points (Figure 2). Considering the distribution of the species presented in Figure 2, it was possible to identify the difference in performance between DOM and BIO. The superiority of BIO may have occurred due to the use of occurrence points to create a hyper-space in the calculation of the similarity of environmental conditions in areas where a given species is present (Motta; Braga; Braga, Da Silva, and Christofaro, 2017). In contrast, for the DOM model, the similarity calculation is based on Gower's distance (Allouche, Tsoar, Steinitz, Rotem, and Kadmon, 2007). Despite species such as *C. micrantha*, *C. oblongifolia*, and *V. guianensis* exhibits a reduced number of occurrence points, are in close proximity to one another.

When analyzing the MAH algorithm, it was observed that the highest metrics for the species occurred when they had a larger number of occurrence points. However, considering the TSS and Sorensen Index metrics, the values reveal that the model is not suitable, thus highlighting the importance of using more than one evaluation metric. The MAH algorithm uses the multivariate sample mean and the covariance matrix, which are sensitive to outliers, potentially influencing the results observed for species with fewer presence points (Leys; Klein; Dominicy; Ley, 2018).

The MXD algorithm showed the best performance for the species *C. micrantha* and *C. oblongifolia*. This algorithm uses a machine learning technique that identifies the most uniform probability distribution for the species, relating it to the constraints of the observed data, making it a model with good performance in ENMs (Elith; Graham; Anderson; Durík; Ferrier; Guisan; Hijmans; Huettmann; Leathwick; Lehmann; Li; Lohmann; Loiselle; Manion; Moriz; Nakamura; Nakazawa; Overton; Townsend; Phillips; Richardson; Scachetti-Pereira; Schapire; Soberón; William; Wisz; Niklaus, 2006), and is widely used in the potential prediction of species (Qiao; Soberón; Peterson, 2015). MXD is the predominant algorithm employed in models characterized by limited presence data (Fois; Fenu; Lombraña; Cogoni; Bacchetta, 2015), corroborating the findings illustrated in Figure 2, where the algorithm demonstrated superior performance for species with fewer occurrence points.

However, the best performance for the species *V. guianensis* was observed when it was modeled using the GAM algorithm, as it is a generalized additive model that allows capturing non-linear relationships by using smooth functions for each predictor variable. It also has a parametric structure, making it more flexible and capable of capturing more complex patterns in the data (Ingram; Vukcevic; Golding, 2020).

The BRT algorithm showed the best performance for the species *A. desmanthum* and *C. racemosa*, and although it did not present the best performance for the other species, the metrics were still adequate, showing values above 0.7. This algorithm uses a boosting technique aimed at improving the model's prediction. It is capable of selecting important variables, adjusting functions, and identifying and modeling interactions, providing predictive advantages over other models (Elith; Leathwick; Hastie, 2008).

In the evaluation of the SVM, GAU, and RDF algorithms, it was observed that although their performance was not superior for any of the species, as seen with the MXD, GAM, and BRT algorithms, their performance was still adequate as their metric values were above 0.7. The SVM algorithm maps input data into a high-dimensional space to find a hyperplane that best separates the data into different classes, and its effectiveness depends on maximizing the margin of separation between classes and the ability to handle nonlinear data using a Kernel function (Amiri; Pourghasemi; Ghanbarian; Afzali, 2019). The GAU algorithm uses Gaussian processes with Bayesian interference to provide probabilistic estimates for predictions in the study region, being highly flexible and capable of modeling nonlinear relationships (Golding; Purse, 2016). The RDF is a machine learning algorithm developed from decision trees, where each tree is built from a bootstrap sample. In RDF, each tree is constructed using a data subsample, and a random selection of features is made at each node, preventing overfitting (Mi; Huettmann; Guo; Han; Wen, 2017).

In the analysis of the average metric values for the five species, the RDF algorithm showed superiority compared to the other algorithms in the ENMTML package, presenting values closest to 1 for all evaluation metrics. This result aligns with those presented in Aguiar; Alencar; Santana; and Teles (2023), Guo; Li; Zhao; and Nawaz (2019), and Mi; Huettmann; Guo; Han; and Wen (2017), which indicated good performance when using RDF, compared to other algorithms in the distribution modeling of *Scirtothrips dorsalis*, *Polyporus umbellatus*, and three species of Asian cranes, respectively.

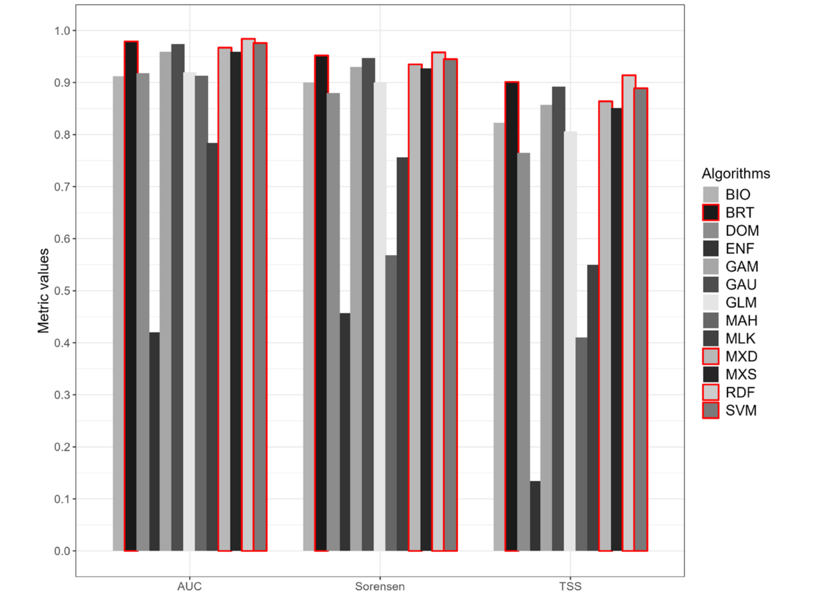
It was possible to identify that the performance of the algorithms differed from the particularities of each dataset, namely, the number of occurrence points, the spatial extent of the presence data, and the environmental layer data. This result was also observed by Qiao, Soberón, and Peterson (2015), where the authors identified variations in model predictions, and the lack of similarity could be related to the sensitivity of the algorithm. This confirms the results obtained, as there is an inconsistency in the number of presence points and spatial extent for each species analyzed, with each algorithm showing different performance for each dataset used.

The averages for the species presented in this study are comparable to those obtained individually for the species. It is possible to identify that the five algorithms (RDF, BRT, SVM, GAU, and MXD) that performed best when the average of the algorithms was analyzed are also the ones that showed superiority in the analysis of the metrics of the individual species, except for the GAM algorithm. Although GAM showed superiority for the species *V. guianensis*, it did not stand out among the other algorithms for the other species.

Thus, the algorithms that showed the best performance for all evaluation metrics were RDF (AUC = 0.984 ± 0.012; Sorensen = 0.958 ± 0.018; TSS = 0.914 ± 0.038), BRT (AUC = 0.979 ± 0.023; Sorensen = 0.952 ± 0.025; TSS = 0.901 ± 0.058), SVM (AUC = 0.976 ± 0.015; Sorensen = 0.945 ± 0.025; TSS = 0.889 ± 0.049), GAU (AUC = 0.974 ± 0.020; Sorensen = 0.947 ± 0.026; TSS = 0.892 ± 0.058), and MXD (AUC = 0.967 ± 0.020; Sorensen = 0.935 ± 0.020; TSS = 0.864 ± 0.047) (Figure 3). In other words, these algorithms performed well for the characteristics presented by the dataset of the individual species, since their determination was based on the average of the metrics obtained for each species.

After constructing the consensus model using the RDF, BRT, SVM, GAU, and MDX algorithms, based on the AUC, TSS, and Sorensen index metrics, it was observed that the consensus model was satisfactory for the species, presenting values greater than 0.7 and low standard deviation (Table 1). Therefore, it can be concluded that using the average of the metrics and standard deviations to determine a consensus model for different species may be an effective alternative to reduce uncertainties generated by the models.

Figure 3 – Evaluation of the average of the thirteen algorithms available in the ENMTML package, according to the average of the metrics and species *Aspidosperma desmanthum, Cariniana micrantha, Clarisia racemosa, Couratari oblongifolia* e *Vouchysia guianensis*



Source: Authors (2023)

In where: Bioclim = BIO; Mahalanobis = MAH; Domain = DOM; Generalized Linear Mode = GLM; Generalized Additive Models = GAM; Support Vector Machine = SVM; Boosted Regression Trees = BRT; Random Forest = RDF; Bayesian Gaussian Process = GAU; Maximum Likelihood = MLK;, Maximum Entropy simple = MXS; Ecological Niche Factor Analysis = ENF; Maximum Entropy default = MXD; Under the Curve = AUC; True Skill Statistics = TSS.

Table 1 – Results of the AUC, TSS, and Sorensen Index metrics generated from the use of the consensus model for the species *Aspidosperma desmanthum, Cariniana micrantha, Clarisia racemosa, Couratari oblongifolia* e *Vouchysia guianensis*

Species	AUC	TSS	Sorensen
Aspidosperma desmanthum	0.995 ±0.00	0.961 ±0.01	0.980 ±0.00
Cariniana micrantha	0.965 ±0.02	0.847 ±0.05	0.927 ±0.02
Clarisia racemosa	0.989 ±0.00	0.932 ±0.02	0.966 ±0.01
Couratari oblongifolia	0.986 ±0.00	0.916 ±0.06	0.956 ±0.03
Vouchysia guianensis	0.981 ±0.02	0.906 ±0.09	0.951±0.05

Source: Authors (2023)

In where: Area Under the Curve = AUC; True Skill Statistics = TSS.

4 CONCLUSIONS

Based on the analysis of five algorithms from the ENMTML package—Random Forest (RDF), Boosted Regression Trees (BRT), Support Vector Machine (SVM), Bayesian Gaussian Process (GAU), and Maximum Entropy default (MXD)—and the metrics Area Under the Curve (AUC), True Skill Statistics (TSS), and Sorensen Index, it is possible to propose a consensus model for the species *Aspidosperma desmanthum*, *Cariniana micrantha*, *Clarisia racemosa*, *Couratari oblongifolia*, and *Vouchysia guianensis*. A method to lower the uncertainties produced by each model is to use the average of the metrics above 0.7 to establish a consensus model for several species.

The performance of the algorithms varies with the particularities of each dataset, such as the number and spatial extent of occurrence points and the environmental layer data.

ACKNOWLEDGEMENTS

We wish to thank the Foundation for Research Support of the State of Amazonas (FAPEAM) for their support and for granting the scholarship. The authors' also wish to thank CNPq for providing the research productivity grant that enabled the following to complete this research: Carlos Henrique Salvino Gadelha Meneses (Process nº 313075/2021-2); Santiago Linorio Ferreyra Ramos (Process nº 305280/2022-8); Ricardo Lopes (Process nº 308815/2023-8); and Maria Teresa Gomes Lopes (Process nº 310307/2018-0) and (Process nº 442914/2020-2).

REFERENCES

ALLOUCHE, O.; TSOAR, A.; STEINITZ, O.; ROTEM, D.; KADMON, R. A comparative evaluation of presence-only methods for modelling species distribution. **Diversity and Distributions**, v. 13, pp. 397-405, 2007. DOI: 10.1111/j.1472-4642.2007.00346.x.

AMIRI, M.; POURGHASEMI, H. R.; GHANBARIAN, G. A.; AFZALI, S. F. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. **Geoderma,** v. 340, pp. 55-69, 2019. DOI: 10.1016/j.geoderma.2018.12.042.

ANDRADE, A. F. A.; VELAZCO, S. J. E.; MARCO JÚNIOR, P. D. ENMTLM: An R package for a straightforward construction of complex ecological niche models. **Environmental Modelling & Software**, v. 125, p. 104615, 2020. DOI: 10.1016/j.envsoft.2019.104615.

CRIA - Centro de Referência e Informação Ambiental**. Specieslink - simple search**, 2023. Available in: https://specieslink.net/search/. Accessed in: 2 April 2023.

ELITH, J.; GRAHAM, C. H.; ANDERSON, R. P.; DURÍK, M.; FERRIER, S.; GUISAN, A.; HIJMANS, R.; HUETTMANN, F.; LEATHWICK, J. R.; LEHMANN, A.; LI, J.; LOHMANN, L. G.; LOISELLE, B. A.; MANION, G.; MORIZ, C.; NAKAMURA, M.; NAKAZAWA, Y.; OVERTON, J. M.C.,; TOWNSEND, P.; PHILLIPS, S.; RICHARDSON, K.; SCACHETTI-PEREIRA, R.; SCHAPIRE, R. E.; SOBERÓN, J.; WILLIAM, S.; WISZ, M. S.; NIKLAUS, E. Z. Novel methods improve prediction of species distributions from occurrence data. **Ecography,** v. 29, pp. 129-151, 2006. DOI: 10.1111/j.2006.0906- 7590.04596.x.

ELITH, J.; LEATHWICK, J. R.; HASTIE, T. A working guide to boosted regression trees. **Journal of Animal Ecology**, v. 77, pp. 802-813, 2008. DOI: 10.1111/j.1365-2656.2008.01390.x.

ESTEVO, C. A.; NAGY-REIS, M. B.; NICHOLS, J. D. When habitat matters: Habitat preferences can modulate co-occurrence patterns of similar sympatric species. **Plos One**, v. 12, 2017. DOI: 10.1371/journal.pone.0179489.

FAO; IIASA. **Harmonized World Soil Database Version 2.0**. Rome and Laxenburg. 2023. 69 p. Available in: https://www.fao.org/documents/card/en/c/cc3823en. Accessed: 30 Mar. 2023.

FICK, S. E.; HIJMANS, R. J. WordClim 2: new 1-km spatial resolution climate surfaces for global land areas. **International Journal of Climatology,** v. 37, n. 12, pp. 4302-4315, 2017. DOI: 10.1002/joc.5086.

FIELDING, A. H.; BELL, J. F. A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. **Environmental Conservation**, v. 24, pp. 38-49, 1997. DOI: 10.1017/S0376892997000088.

FIRPO, M. A. F.; GUIMARÃES, B. S.; DANTAS, L. G.; DA SILVA, M. G. B.; ALVEZ, L. M.; CHADWICK, R.; LLOPART, M. P.; DE OLIVEIRA, G. S. Assessment of CPIP6 models performance in simulating present day climate in Brazil. **Frontiers in Climate,** v. 4, 2022. DOI: 10.3389/fclim.2022.948499.

FOIS, M.; FENU, G.; LOMBRAÑA, A. C.; COGONI, D.; BACCHETTA, G. A practical method to speed up the discovery of unknown populations using Species Distribution Models. **Journal for Nature Conservation**, v. 24, pp. 42-48, 2015. DOI: 10.1016/j.jnc.2015.02.001.

GBIF. **The Global Biodiversity Information Facility**, 2023. Available in: https://www.gbif.org/pt/occurrence/search. Accessed in: April 05, 2023.

GOLDING, N.; PURSE, B. V. Fast and flexible Baysian species distribution modelling usind Gaussian processes. **Methods in Ecology and Evolution**, v. 7, pp. 598-608, 2016. DOI: doi.org/10.1111/2041-210X.12523.

GUO, Y.; LI, X.; ZHAO, Z.; NAWAZ, Z. Predicting the impacts of climate change, soils and vegetation types on the geographic distribution of Polyporus umbellatus in China. **Science of the total environment**, v. 648, p. 1-11, 2019. DOI: 10.1016/j.scitotenv.2018.07.465.

INGRAM, M.; VUKCEVIC, F.; GOLDING, N. Multi-output Gaussian processes for species distribution modelling. **Methods in Ecology and Evolution**, v. 11, pp. 1587-1598, 2020. DOI: 10.1111/2041-210X.13496.

KONOWALIK, K., NOSOL, A. Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage**. Sci Rep**, v. 11, 2021. DOI: 10.1038/s41598-020-80062-1.

LANDA, M. L. N.; CASTRO, J. C. M.; MONTERRUBIO-RICO, T. C.; LARA-CABRERA, S. I.; PIETRO-TORRES, D. A. Predicting co-distribution patterns of parrots and woody plants under global changes: The case of the Lilac-crowned Amazon and Neotropical dry forests. **Journal for Nature Conservation**, v. 71, 2023.

LEROY, B. DELSOL, R.; HUGUENY, B.; MEYNARD, C. N.; BARHOUMI, C.; MASSIN, M. B.; BELLARD, C. Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. **Journal of Biogeography**, v. 45, n. 9, p. 1994-2002, 2018. DOI: 10.1111/jbi.13402.

LEYS, C.; KLEIN, O.; DOMINICY, Y.; LEY, C. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. **Journal of Experimental Social Psychology,** v. 74, p. 150-156, 2018. DOI: 10.1016/j.jesp.2017.09.011.

LOBO, J.; JIMÉNEZ-VALVERDE.; HORTAL, J. The uncertain nature of absences and their importance in species distribution modelling, **Ecography**, v. 33, pp. 103-114, 2010. DOI: 10.1111/j.1600-0587.2009.06039.x.

MA, Y.; YOU, X. A sustainable conservation strategy of wildlife in urban ecosystems: Case of Gallinua chloropus in Beijing-Tianjin-Hebei region. **Ecological Informatics**, v. 68, 2022. DOI: 10.1016/j.ecoinf.2022.101571.

MAITNER, B. S.; BOYLE, B.; CASLER, N.; CONDIT, R.; DONOGHUE, J.; DURÁN, S. M.; GUADERRAMA, D.; HINCHLIFF, C. E.; JORGENSEN, P. M.; KRAFT, N. J. B.; MCGILL, B.; MEROW, C.; MORUETA-HOLME, N.; PEET, R. K.; SANDEL, B.; SCHILDHAUER, M.; SMITH, S. A.; SVENNING, J. C.; THIERS, B.; VIOLLE, C.; WISER, S.; ENQUIST, B. The BIEN R package: A tool to access the Botanical Information and Ecology Network (BIEN) database. **Methods in ecology and Evolution**, v. 9, pp. 373-379, 2018. DOI: 10.1111/2041-210X.12861.

MI, C.; HUETTMANN, F.; GUO, Y.; HAN, X.; WEN, L. 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,** v. 5, 2017. DOI: 10.7717/peerj.2849.

MONTEVERDE, C.; DE SALES, F.; JONES, C. Evaluation of the CMIP6 Performance in Simulating Precipitation in the Amazon River Basin. **Climate**, v. 10, 2022. DOI: 10.3390/cli10080122.

MOTTA, A. Z.; BRAGA, S. R.; DA SILVA, N. D. M.; CHRISTOFARO, C. Avaliação do desempenho de modelos de distribuição potencial da espécie *Wunderlichia azulenzis*. In: Anais do XVIII Simpósio Brasileiro de Sensoriamento Remoto, 2017, São Paulo. **Anais** [...]. São Paulo: INPE Santos, 2017.

MUGO, R.; SAITOH, S. I.; IGARASHI, H.; TOYODA, T.; MASUDA, S.; AWAJI, T.; ISHIKAWA, Y. Identification of skipjack tuna (Katsuwonus pelamis) pelagic hotspots applying a satellite remote sensing-driven analysis of ecological niche factors: A short-term run. **PLoS One**, v.15, 2020. DOI 10.1371/journal.pone.0237742.

NDAO, B.; LEROUX, L.; HEMA, A.; DIOUF, A. A.; BÉGUÉ, A.; SAMBOU, B. Tree species diversity analysis using species distribution models: A Faidherbia albida parkland case study in Senegal. **Ecological Indicators**, v. 144, 2022.

QIAO, H.; SOBERÓN, J.; PETERSON, A. T. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. **Methods in Ecology and Evolution**, v. 6, n. 10, pp. 11126-1136, 2015. DOI: 10.1111/2041-210X.12397.

REMYA, K.; RAMACHANDRAN, A.; JAYAKUMAR, S. Predicting the current and future suitable habitat distribution of Myristica dactyloides Gaertn. Using Maxent model in the Eastern Ghats, **India. Ecological Engineering**, v. 82, pp. 184-188, 2015. DOI: 10.1016/j.ecoleng.2015.04.053.

SENAY, S. D.; WORNER, S. P.; IKEDA, T. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. **Plos ONE**, v. 8, 2013. DOI 10.1371/journal.pone.0071218.

THUILLER, W.; GUÉGUEN, M.; RENAUD, J.; KARGE, D. N.; ZIMMERMANN, N. E. Uncertainty in ensembles of global biodiversity scenarios. **Nature Communication**, v. 10, 2019. DOI: 10.1038/s41467-019-09519-w.

ZHAO, Z.; GUO, Y.; WEI, H.; RAN, Q.; GU, W. Predictions of the Potential Geographical Distribution and Quality of Gynostemma pentaphyllum Base on the Fuzzy Matter Element Model in China. **Sustainability,** v. 9, 2017. DOI: 10.3390/su9071114.

Authorship Contribution

1 Ingrid Lana Lima de Morais

Forest Engineer, Master in Environmental and Forest Sciences

https://orcid.org/0000-0001-8005-273X • ingridmorais24@gmail.com

Contribution: Conceptualization; Data curation; Data analysis; Investigation; Methodology; Writing – original draft

2 Alexandra Amaro de Lima

Meteorologist, Ph.D. in Climate and Environment

https://orcid.org/0000-0003-3918-0013 • xanduca@gmail.com

Contribution: Conceptualization; Methodology; Data validation; Review of the original manuscript; Supervision; Writing – review and editing

3 Ivinne Nara Lobato dos Santos

Forest Engineer, Ph.D. in Environmental Sciences and Sustainability in the Amazon
https://orcid.org/0000-0002-7993-4345 • ivinne.lobato@gmail.com
Contribution: Conceptualization; Data curation; Data analysis; Investigation; Methodology; Writing – review and editing

4 Lair Cristina Avelino do Nascimento

Forest Engineer, Master in Forest Sciences
https://orcid.org/0000-0002-8920-4661 • laircristina1@gmail.com
Contribution: Conceptualization; Data curation; Software implementation and testing

5 Santiago Linorio Ferreyra Ramos

Agronomist, Ph.D. in Genetics and Plant Breeding
https://orcid.org/0000-0003-0364-316X • slfr@ufam.edu.br
Contribution: Investigation; Data validation; Writing – review and editing

6 Carlos Henrique Salvino Gadelha Meneses

Biologist, Ph.D. in Plant Biotechnology
https://orcid.org/0000-0001-8394-1305 • carlos.meneses@servidor.uepb.edu.br
Contribution: Investigation; Data validation; Writing – review and editing

7 Ricardo Lopes

Agronomist, Ph.D. in Agronomy (Genetics and Plant Breeding)
https://orcid.org/0000-0002-5559-9685 • ricardo.lopes@embrapa.br
Contribution: Investigation; Data validation; Writing – review and editing

8 Ananda Virgínia De Aguiar Lopes

Agronomist, Ph.D. in Genetics and Plant Breeding
https://orcid.org/0000-0003-1225-7623 • ananda.aguiar@embrapa.br
Contribution: Investigation; Data validation; Writing – review and editing

9 Maria Teresa Lopes

Agronomist, Ph.D. in Genetics and Plant Breeding
https://orcid.org/0000-0003-1988-7126 • mtglopes@ufam.edu.br
Contribution: Conceptualization; Methodology; Data validation; Review of the original manuscript; Supervision; Writing – review and editing

How to quote this article

MORAIS, I. L. L.; LIMA, A. A.; SANTOS, I. N. L.; NASCIMENTO, L. C. A.; RAMOS, S. L. F.; MENESES, C. H. S. G.; LOPES, R.; AGUIAR, A. V.; LOPES, M. T. Analysis of algorithms for building a joint model for modeling Amazonian species. **Ciência Florestal**, Santa Maria, v. 35, e86428, p. 1-21, 2025. DOI 10.5902/1980509886428. Available from: https://doi.org/10.5902/1980509886428. Accessed in: day month abbr. year.

Data Availability Statement:

Datasets related to this article will be available upon request to the corresponding author.

Evaluators in this article:

Editorial Board: