Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

María Martha Torres Martínez

Applying "Ensemble of Small Models" to Species Distribution Models with limited data

Curitiba
2024

María Martha Torres Martínez

Applying "Ensemble of Small Models" to Species Distribution Models with limited data

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Paulo Justiniano Ribeiro Junior

Curitiba
2024

Applying "Ensemble of Small Models" to Species Distribution Models with limited data

María Martha Torres Martínez[1], Paulo Justiniano Ribeiro Junior[2]

[1]Data Science & Big Data Specialization Course, Statistics Department, Universidade Federal do Paraná, Postal box 19031, CEP 81531-990, Curitiba, PR, Brazil, mariam94@hotmail.com.
[2]Statistics Department, Universidade Federal do Paraná, Postal box 19031, CEP 81531-990, Curitiba, PR, Brazil, paulojus@ufpr.br

Os modelos de distribuição de espécies (SDMs) são ferramentas amplamente utilizadas para prever a distribuição potencial de espécies, com diversas aplicações em ecologia. Modelar espécies raras é desafiador devido às suas áreas de distribuição restritas e pequenos conjuntos de dados, que podem levar ao sobreajuste do modelo. A abordagem Ensemble of Small Models (ESMs) surgiu para mitigar essas limitações, utilizando conjuntos de modelos pequenos e demonstrando resultados satisfatórios. Este estudo aplica ESMs para construir SDMs para espécies comuns e raras, usando como modelo de estudo o gênero de roedores neotropicais *Coendou*, que inclui 16 espécies variando de comuns (>100 localidades) a raras (<30 localidades). Utilizamos variáveis climáticas, topográficas e de cobertura vegetal como preditores e o algoritmo Maxent. Numerosos pequenos modelos (bivariados) foram calibrados e avaliados, formando um conjunto ponderado com base nos índices AUC e Boyce's, e suas previsões de conjunto avaliadas com o índice Somers'D. Para reduzir a sobrepredição, aplicamos a metodologia MSDM "*a posteriori*". Os modelos apresentaram bom desempenho (0,76 AUC - 0,63 Boyce's), especialmente para espécies raras. A metodologia MSDM ajustou a sobrepredição, embora com algumas restrições artificiais à realidade biológica das espécies. A abordagem ESM, com algoritmos como o Maxent e ajustes de sobrepredição com MSDM, permitiu o desenvolvimento de modelos de alto desempenho, melhorando a confiabilidade dos SDMs, especialmente para espécies raras.
Palavras-chave: Conjunto de dados pequeno, Desempenho do modelo, Sobreajuste, Sobrepredição.

Species Distribution Models (SDMs) are widely used tools for predicting the potential distribution of species, with various applications in ecology. Modeling rare species is challenging due to their restricted distribution areas and small datasets, which can lead to model overfitting. The Ensemble of Small Models (ESMs) approach emerged to overcome these limitations by using small model ensembles and demonstrating satisfactory results. This study applies ESMs to construct SDMs for both common and rare species, using the neotropical rodent genus *Coendou* as a case study, which includes 16 species ranging from common (>100 locations) to rare (<30 locations). We used climatic, topographic, and vegetation cover variables as predictors and the Maxent algorithm. Numerous small (bivariate) models were calibrated and evaluated, forming a weighted ensemble based on AUC and Boyce's indices, and their ensemble predictions were assessed with the Somers'D index. To reduce overprediction, we applied the MSDM "*a posteriori*" methodology. The models showed good performance (0.76 AUC - 0.63 Boyce's), particularly for rare species. The MSDM methodology adjusted for overprediction, though with some artificial constraints on the biological reality of the species. We found that the ESM approach, with algorithms like Maxent and overprediction adjustments with MSDM, enables the development of high-performance models, thereby improving the reliability of SDMs, especially for rare species.
Keywords: Model performance, Overfitting, Overprediction, Small dataset.

1. Introduction

Species distribution models (SDMs) are crucial tools in ecology for analyzing the environmental factors that determine the distribution of biological species and for predicting the areas where a species might find suitable environmental conditions for its distribution [13], [62]. These models use a wide range of algorithms to statistically describe the relationship between species and their environment, supported by various software

packages designed to connect ecological theory with data analysis [11], [62]. The algorithms quantify the correlation between species and environmental factors. The choice of data types and analysis methods is crucial, as it determines the resulting output. Therefore, these choices should align with the model's objectives [35], [51].

Building on this, SDM methods include techniques ranging from simple models based solely on presence data [25], [40] to advanced approaches that differentiate between probability predictions (e.g., logistic regression, [61]) and fitness predictions (e.g., classification trees, [49]). Some approaches, such as artificial neural networks, are iterative and can be improved by external iterative methods applied to machine learning algorithms, such as decision trees, to increase accuracy [54], [60], [61].

Recent development includes generalized regression and machine learning methods have been developed or extended to handle presence-only data, often requiring the use of "background"or pseudo-absence data [40], [54]. Maximum entropy (Maxent) [43] has been shown to be very effective in SDM studies with presence-only data [14], [43], especially with small sample sizes [26], [43]. Additionally, it can model complex and nonlinear relationships between response variables and predictors [14]. However, its ease of use and simplicity have driven Maxent to become one of the most prominent and widely used SDM techniques in scientific research [22], [26]. Despite criticisms regarding its incorrect application and oversimplification [36], [59] and concerns about overfitting [19], [34], Maxent continues to be frequently used for modeling across various taxa, geographic areas, time periods, and environmental scenarios [26].

In practical applications, SDMs are widely used in several areas, being key tools in conservation planning [4], testing biogeographic hypotheses [29], investigating evolutionary processes [27], identifying suitable areas for species reintroduction [15], mapping fuels and fire regimes [46], and increasing the probability of detection of rare species [10], [18].

Modeling rare species, which have limited distribution range (small data set) compared to common and widely distributed species (larger data set), presents challenges. Modeling rare species generally presents higher accuracy compared to common species [17], [6]. However, considering that model accuracy increases asymptotically with sample size [50], [21], [6], several studies have shown that with occurrences below approximately 30, model accuracy is often low and the variability of model accuracy increases between species [50], [21], [58], [6].

This can be explained because many times, with species with a small data set, the number of explanatory environmental variables is greater than the number of occurrences, which can lead to overfitting of the model, leading to a lower generalization of the model restricting its applicability to new data [56], [6].

Considering that the rule of thumb indicates that the sample size, number of occurrences, be 10 times larger than the number of predictors used to make the models [20], [6], the problem could be solved by reducing the number of predictors by means of selection methods (Akaike Information Criterion - AIC [2]; Bayesian Information Criterion - BIC, [47]; Lasso, [52], etc.). However, this is not applicable for rare species since, in order to maintain that relationship, it would imply keeping only one or two predictors per 20 occurrences, being that many times that number corresponds to the total sample size [6].

For that Lomba and collaborators [31] proposed an Ensemble of Small Models (ESMs) approach, which is based on fitting a larger number of small bivariate models (models with only two predictors at a time), averaging them into an ensemble prediction using weights based on model performances [6]. This makes the predictor-occurrence relationship much improved with this method because the number of predictors used in each small model remains low [6], [7]. Therefore, the estimation of model coefficients was expected to be much more robust and the risk of model overfitting was minimized without losing explanatory power because the number of predictors remains low for a single model, but high within the ensemble framework [7]. Therefore, this work aims to apply the ESMs approach to construct SDMs for both common and rare species, evaluating its effectiveness in handling datasets with varying occurrences.

2. Materials and Methods

2.1. Study model

We use the Neotropical rodent genus *Coendou* as a biological study model. This genus encompasses 16 recognized species [57], [44], [33], of which we have identified localities for 13. Among these, some species have been recorded in over 100 recorded localities (e.g., *C. longicaudatus*, n= 137; *C. rufescens*, n= 142; *C. spinosus*, n=160), while others have fewer than 60 confirmed

localities (e.g., *C. bicolor*, n= 56 and *C. quichua*, n=56), and others have less than 30 confirmed localities (e.g., *C. baturitensis*, n= 21; *C. ichillus*, n= 10; *C. insidiosus*, n= 19; *C. melanurus*, n=23; *C. nycthemera*, n=29; *C. pruinosus, n= 25*; *C. prehensilis*, n= 13 and *C. vestitus*, n=10). The latter are considered rare species due to their restricted distribution and limited available information [45].

2.2. Pre-analysis

2.2.1. Data Acquisition

Occurrence data were obtained from global online database such as the Mammal Networked Information System (MaNIS) and the Global Biodiversity Information Facility (GBIF). We performed distribution modelling using the confirmed localities as occurrence data and environmental variables such climate, topographical (slope and elevation) and land cover (NDVI, Normalized difference vegetation index). We obtained the 19 climatic variables from WorldClim (worldclim.org), the 2 topographic variables from EarthEnv (earthenv.org), and NDVI from NEO Nasa Earth Observations (nasa.gov), (Table S1 in Supplementary section).

As spatial limits for the cuts of the environmental variables and spatial definition of the calibration area of the models, shapefiles of the American continent, of the countries of Central and South America obtained from DIVA-GIS (diva-gis.org), as well as of the ecoregions obtained from Morrone and collaborators (2022) [37] were used.

2.2.2. Data Preparation

The occurrence data were cleaned and duplicates were removed using data manipulation packages such as "*dplyr*" and "*sf*" in R Software. Subsequently, with the cleaned dataset, we corrected for spatial autocorrelation. For this, we applied spatial filtering analysis, retaining occurrences that were greater than 2km^2 in proximity to one another [38]. This analysis was performed using "*spThin*" package [1].

The variables were standardized 1km^2 resolution and comprised 22 environmental layers. To reduce errors related to the correlation between the environmental predictors, we used a Principal Component Analysis (PCA) to synthesize the variation of environmental predictors into principal components [9]. We used the first seven principal components, as predictor variables, which explained 96% of the original variation.

2.3. Species Distribution Models (SDMs)

To avoid model overfitting that occurs because of the low number of occurrences and considering that most of the species evaluated have few occurrences (<30) or are considered rare, we adopted the Ensembles of Small Models (ESM) [6],[7] as the modelling framework. This approach uses small calibrated models (usually bivariate) with different combinations of predictors pairs making a final ensemble model [6] (Figure 1). We used Maximum Entropy method - Maxent [43] because it has been good performance compared with other algorithms and indicating for small datasets [14], [43], [39]. Additionally, Maxent is a presence background algorithm (works only with presence data that we have available), since it considers that compares the environmental variables of the true presence with the background pixel to create area predictions of environmental suitability [43],[39], [42]. For the background points, we created five sets of 10,000 random points in the area included in the shapefile of the ecoregions.

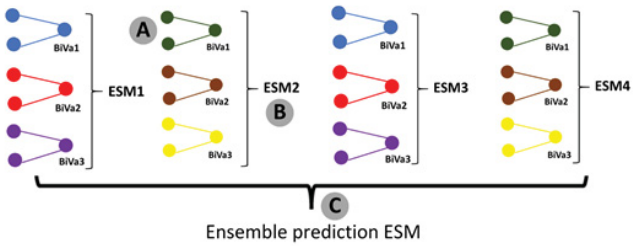

Figura 1: A: Calibration and evaluation of all small (here bivariate) models. **B**: Weighted average of all predictions per each ensemble and evaluation of ESM predictions. **C**: Weighted average of ESM predictions and evaluation (Adapted from Breiner et al. 2015)

Each model was calibrated and projected to the ecoregions corresponding to the intersection of the occurrence points of each species [5]. For model validation, the occurrence data were randomly divided in a proportion of 70% for model training and 30% for model testing, and this procedure was repeated 5 times. For each bivariate model we considered the AUC and Boyce's index as performance measures. The AUC discriminates between presences and absences considered >0.5 acceptable model [16], [12]. The Boyce's index is a complement to the usual evaluation of presence/absence models and a reliable measure of presence-only based predictions [23]. Boyce's index is based on Sperman's correlation coefficient, and its values range from -1 to 1 with positive values indicating that the model is able to predict correctly according to the ac-

tual presence of the evaluation data [23]. For the final ensemble of each model, we used the Somers'D that averages simple bivariate models by weighted means to Ensemble Small Models [31], [6], [7]). Somers'D index gives more weight to models that perform well and less to those that perform poorly. Bivariate models with Somers'D less than 0 (i.e., AUC < 05) were set to zero and were not used to construct ensembles [6].

To reduce the overprediction of the model output (continuous suitability map), we used the MSDM methodology proposed by Mendes and collaborators (2020) [32]. We applied "*a posteriori*"method Buffered Minimum Convex Polygon (MCPB), compiled and adapted from Kremen and collaborators (2008) [28] to the final models of six species (*C. ichillus*, *C. longicaudatus*, *C. pruinosus*, *C. rufescens*, *C. quichua* and *C. vestitus*). The method uses presences to build a minimum convex polygon, with interior angles less than 180^o, and excludes pixels outside the polygon. This method includes a buffer zone around the polygon, with size based on the maximum distance of minimal pairwise distances between occurrences [32]. We kept the sensitivity at 1 (to avoid losing any locality) for species with less than 60 localities (*C. ichillus*, *C. pruinosus*, *C. quichua* and *C. vestitus*) and sensitivity of 0.8 (chance of losing 20% localities) for *C. longicaudatus* and *C. rufescens* (133 and 129 localities, respectively). For all analyses we used R software, the "*Biomod2*" and "*Ecospat*" packages for the distribution models [6], [7], and "*MSDM*"package to reduce overprediction [32].

3. Results and discussion

3.1. Modelling analysis

The models presented a good performance, with average AUC and Boyce's index of 0.76 and 0.63, respectively (Figure 2 and 3, Table 1). The use of the ESM modeling approach was adequate for the species data set, considering that we obtained species with data sets <30 localities, and some of them with <15 confirmed localities. The use of ESM is widely recommended in the literature for modeling small datasets or subsampled species because they perform well in both internal (the entire dataset) and external (the transferability assessment evaluation dataset) cross-validation [6]. Consequently, species with fewer locations showed the greatest improvement in model performance when using ESM.

However, despite the recommendations for the ESM approach [6], [7], it is important to note some additional shortcomings or considerations [6]. The first, the ESM approach has a higher computational effort compared to standard SDM approaches. Another limitation is that we randomly selected 10,000 background points located in the ecoregions defined for each species to calibrate the models. Of the species modeled, three of them were considered rare being restricted to small geographic regions (*C. ichillus*, *C. prehenislis* and *C. vestitus*). As a consequence, for these species many background points could be located far from presence points, leading to high AUC values by easily distinguishing background from presence points [6], [55]. Therefore, other strategies for selecting background data and the use of actual absence or abundance data could be tested in future evaluations of the ESM approach [6].

The AUC score and Boyce's index are both widely used methods for evaluating distribution models, as they provide a single numerical value from which the best model can be selected based on the highest value [24]. In our results, the AUC score is greater than 0.5, indicating that Maxent models perform better than random models. The highest AUC values (approx. 0.90), corresponded to some of the species with <30 localities (*C. baturitensis*, *C. insidiosus*, *C. nycthemera*, *C. prehensilis* and *C. vestitus*; Table 1), compared to the more common species (*C. longicaudatus* and *C. quichua*). In contrast to the AUC, rare species underperform compared to less rare species when evaluated with Boyce's index (e.g., *C. vestitus*; Table 1). This may be mainly due as common assessment indices based on presence-absence information, such as the AUC, generally give higher scores than presence-only indices, such as Boyce's, when the number of presences is low [6], [23]. It has been documented that the AUC index being one of the most widely used, applied to measure performance of calibrated models for rare species tends to overestimate the performance of these models [30], [41], [48], [6].

The low Boyce's values in the rare species should be considered, given that both the model calibration and its projection were performed in a spatial region where there is confirmation of the presence of the species, which is considered a highly informative space in the construction of the models [24]. This could be a problem derived from how the Maxent algorithm assigns suitability values, based on occurrences, to the rest of the cells in the study area, rather than a problem with the evaluation methods [24]. However, considering the performance of AUC, the good performance of the mo-

Tabela 1: Average and Standard deviation of AUC, Somers'D, and Boyce's by each *Coendou* distribution model.

Species	AUC		Somers'D		Boyce's	
	average	*sd*	average	*sd*	average	*sd*
C. ichillus	0.75	0.08	0.51	0.17	0.67	0.10
C. longicaudatus	0.69	0.03	0.38	0.06	0.78	0.08
C. pruinosus	0.81	0.08	0.63	0.16	0.62	0.18
C. quichua	0.63	0.06	0.26	0.13	0.44	0.17
C. rufescens	0.81	0.03	0.62	0.05	0.85	0.05
C.vestitus	0.89	0.04	0.78	0.07	0.42	0.42
C. bicolor	0.76	0.04	0.51	0.09	0.77	0.10
C. baturitensis	0.91	0.05	0.81	0.11	0.60	0.27
C. insidiosus	0.88	0.06	0.76	0.13	0.45	0.35
C. melanurus	0.66	0.08	0.33	0.15	0.35	0.31
C. nycthemera	0.90	0.05	0.81	0.10	0.79	0.11
C. prehensilis	0.91	0.06	0.82	0.12	0.80	0.12
C. spinosus	0.85	0.02	0.71	0.04	0.75	0.13

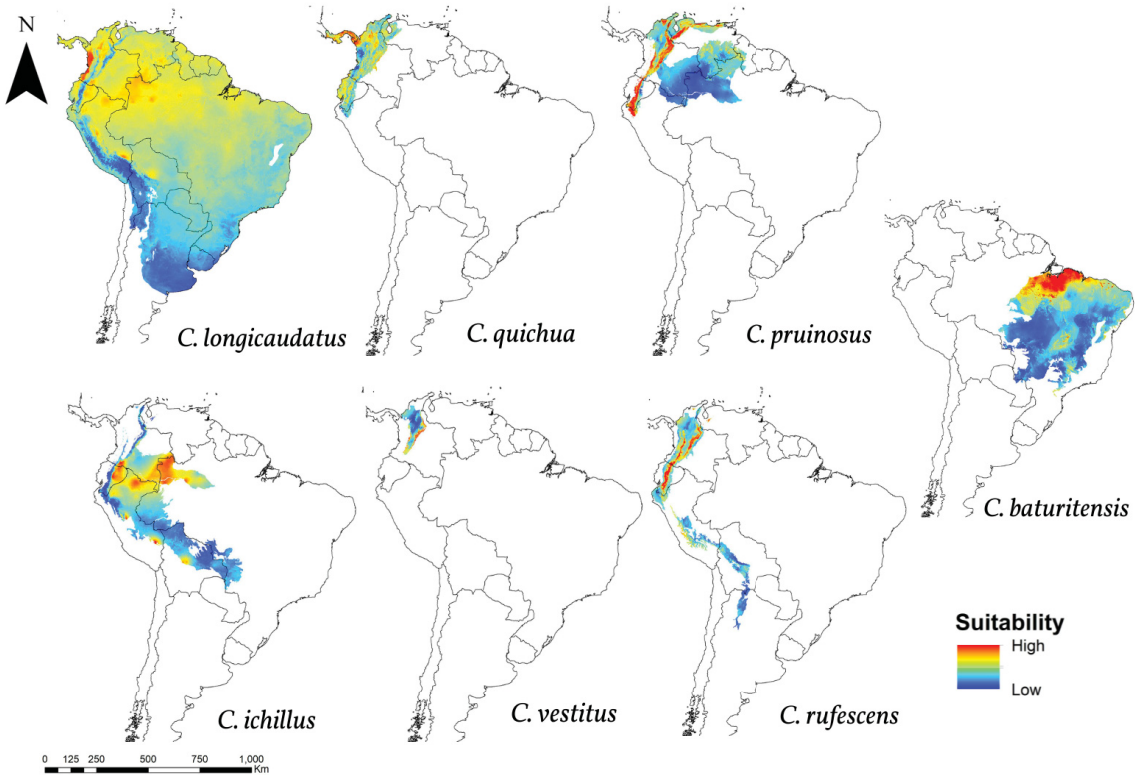

Figura 2: Environmental suitability maps of seven *Coendou* species resulting from the SDMs. Warm colors (red) indicate high environmental suitability while cool colors (blue) indicate sites with less suitable environmental conditions for the presence of the species.

dels could be explained by the algorithm used in the ESM approach (Maxent), since it is an algorithm that performs well when dealing with small data sets [14], [43], [39], [24].

This illustrates the challenges associated with modeling rare species as they are often not fully recognized when only indices such as AUC are considered to measure model performance. Therefore, there are studies that recommend not relying solely on the AUC if true absences are not available [21]. It is also recommended to use Boyce's index more systematically and choosing ESM when modeling rare species [6].

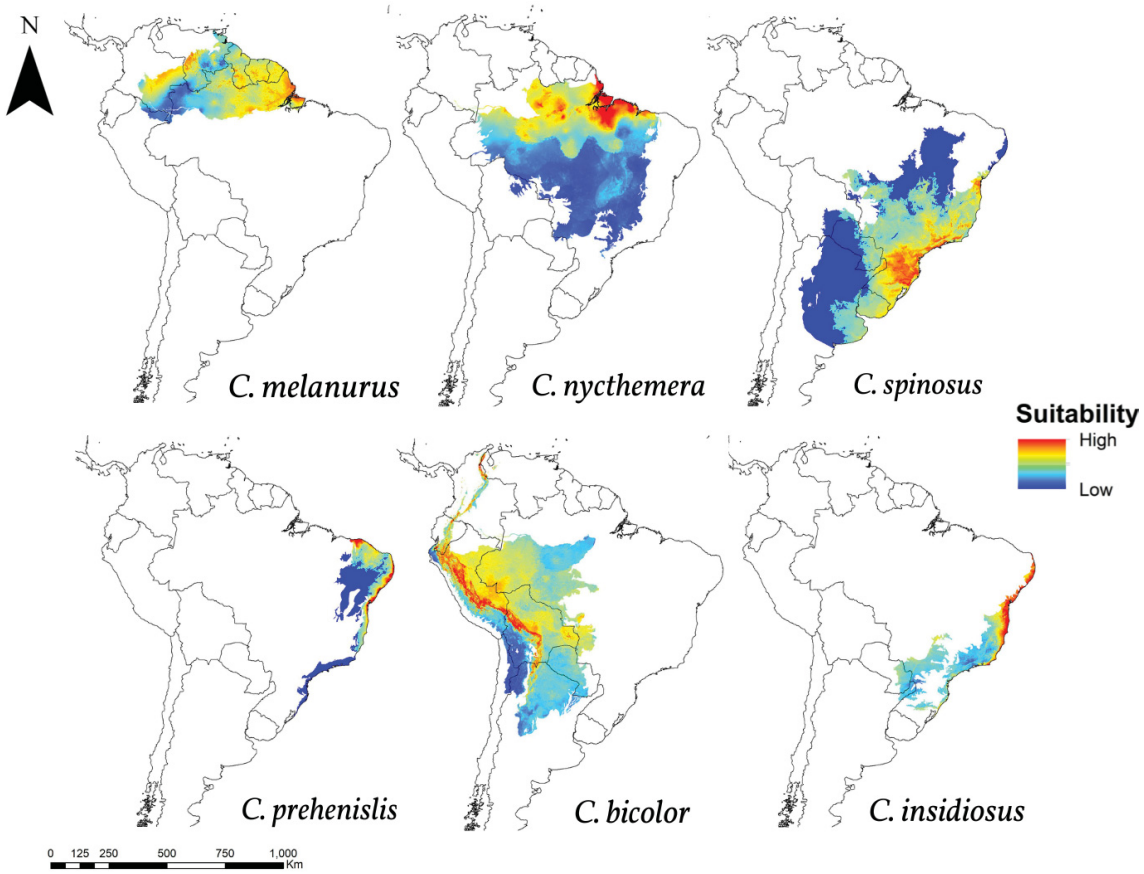

Figura 3: Environmental suitability maps of six *Coendou* species resulting from the SDMs. Warm colors (red) indicate high environmental suitability while cool colors (blue) indicate sites with less suitable environmental conditions for the presence of the species.

For the models to which the overprediction adjustment was applied by the MSDM "*a posteriori*"method, the result was maps with straight cuts corresponding to the polygons constructed based on the most extreme occurrence points of the known localities for the species, this being a method that only requires data on the presence of species (Figure 4 and 5, [5]. This "*a posteriori*"method as a spatial information strategy superimposed on the suitability maps has purely visual purposes.

These cuts were left with artificial visual, which does not correspond to the biological and ecological reality of the distribution of a species [5], however as they are constructed with the occurrence data themselves as a proxy for accessibility, because the occurrences of native species are found only in places they could access [3], [8], this spatial filter, serves as a species dispersal distance constraint, which could help increase the predictive power of the model [5].

4. Conclusions and future work

By applying the ESM approach, we obtained models with good performance for small data sets. The introduction of modern machine learning techniques, such as Maxent, together with the application of the ESM approach and its subsequent fitting using the "*a posteriori*" MSDM method, are important steps to improve rare species distribution models. Small model ensembles represent a powerful strategy applicable with several SDM techniques being ESM one of them. The strength of ESMs is their high performance with small sample sizes, a characteristic of rare and endangered species data, or species that are difficult to detect. The potential of ESMs was further demonstrated by their ability to predict independent areas.

For future work, small model ensembles should not be limited to bivariate models, but could consider univariate or trivariate (or higher) models to build ESM models. Regarding model performance evaluation metrics, the use of several indices is recommended to have a clearer view of one's own performance. It is also re-

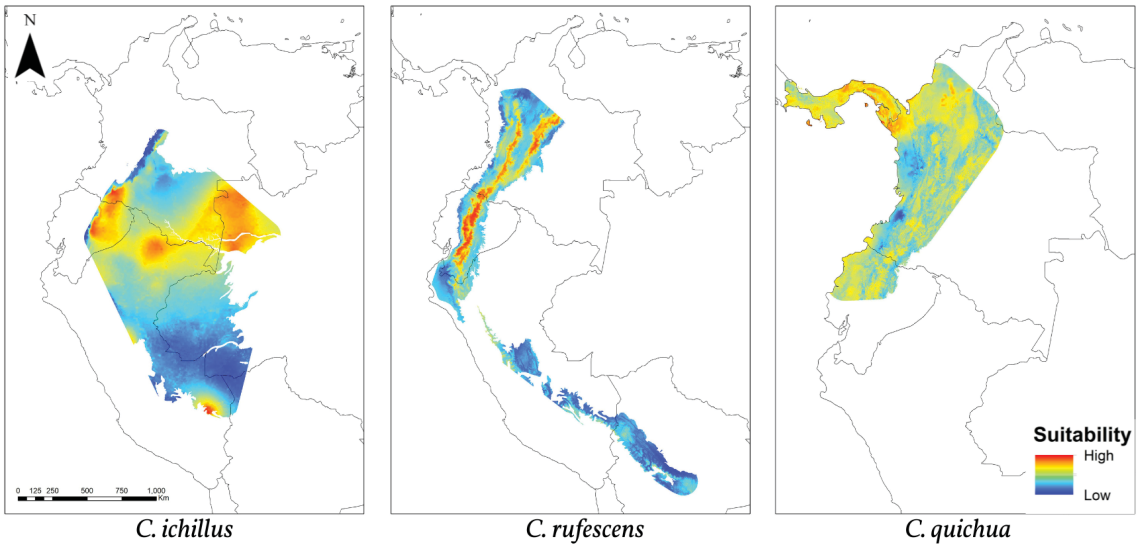

Figura 4: Environmental suitability maps of three *Coendou* species (*C. ichillus*, *C. rufescens* and *C. quichua*) resulting from the overprediction adjustment with the MSDM "*a posteriori*" method. Warm colors (red) indicate high environmental suitability, while cool colors (blue) indicate locations with less suitable environmental conditions for the presence of the species.

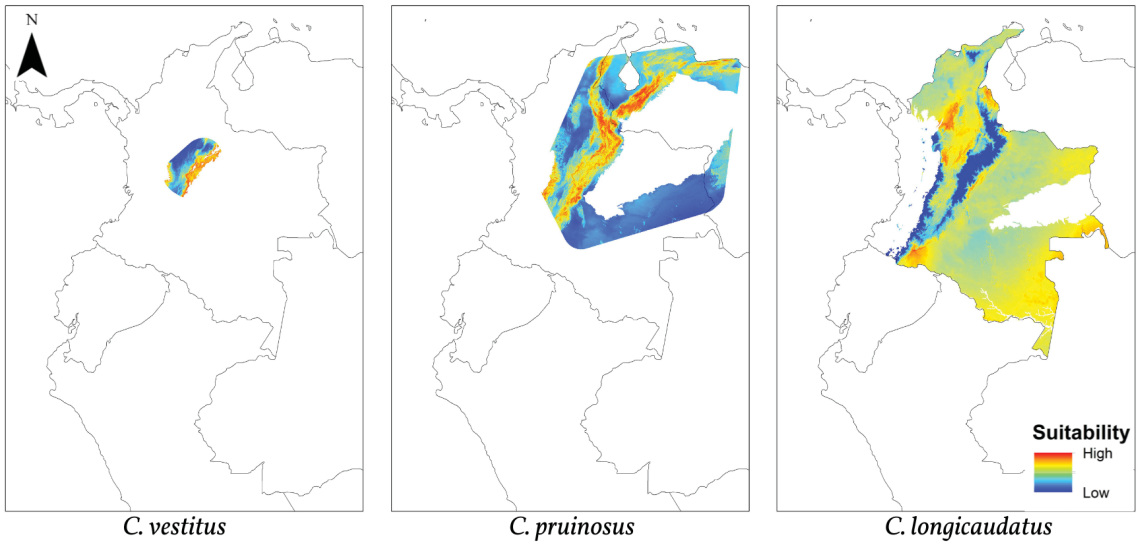

Figura 5: Environmental suitability maps of three *Coendou* species (*C. vestitus*, *C. pruinosus* and *C. longicaudatus*) resulting from the overprediction adjustment with the MSDM "*a posteriori*" method. Warm colors (red) indicate high environmental suitability, while cool colors (blue) indicate locations with less suitable environmental conditions for the presence of the species.

commended to explore other methods of the MSDM ("*a priori*"and "*a posteriori*") for model overprediction adjustments that are more in line with the biological and ecological reality of the modeled species.

5. Acknowledgments

We thank MSc. Licet Calambás and Dr. Fernanda Brum for their assistance with model development. MMTM thank to Professor Dr. Paulo Justiniano Ribeiro Junior for his orientation, support and advice for the realization of this work, and to Professor Dr. Wagner Hugo Bonat for his support in the completion of the specialization course. MMTM also thank to their husband, family and colleagues for their unwavering support and encouragement.

Referências

[1] Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A. Vilela, B. & Anderson, R. P. (2015). spThin: an R

package for spatial thinning of species occurrence records for use in ecological niche models. Ecography, 38, 541–545.

[2] Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716-723.

[3] Allouche, O., Steinitz, O., Rotem, D., Rosenfeld, A., & Kadmon, R. (2008). Incorporating distance constraints into species distribution models. Journal of Applied Ecology, 45(2), 599-609.

[4] Araújo, M. B., Williams, P. H., & Fuller, R. J. (2002). Dynamics of extinction and the selection of nature reserves. Proceedings of the Royal Society of London. Series B: Biological Sciences, 269(1504), 1971-1980.

[5] Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberon, J. & Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecological Modelling, 222, 1810-1819.

[6] Breiner, F. T., Guisan, A., Bergamini, A., & Nobis, M. P. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. Methods in Ecology and Evolution, 6(10), 1210-1218.

[7] Breiner, F. T., Nobis, M. P., Bergamini, A. & Guisan, A. (2018). Optimizing ensembles of small models for predicting the distribution of species with few occurrences. Methods in Ecology and Evolution, 9, 802-808.

[8] Cardador, L., Sardà-Palomera, F., Carrete, M., & Mañosa, S. (2014). Incorporating spatial constraints in different periods of the annual cycle improves species distribution model performance for a highly mobile bird species. Diversity and distributions, 20(5), 515-528.

[9] De Marco, P. J. & Nóbrega, C. C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. PloS One, 13, e0202403.

[10] Edwards Jr, T. C., Cutler, D. R., Zimmermann, N. E., Geiser, L., & Alegria, J. (2005). Model-based stratifications for enhancing the detection of rare ecological events. Ecology, 86(5), 1081-1090.

[11] Elith, J., & Franklin, J. (2013). Species distribution modeling. In Encyclopedia of Biodiversity: Second Edition (pp. 692-705). Elsevier Inc.

[12] Elith, J., & Leathwick, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions, 13, 265-275.

[13] Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. Annual review of ecology, evolution, and systematics, 40(1), 677-697.

[14] Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., ... & E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29(2), 129-151.

[15] Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of applied ecology, 41(2), 263-274.

[16] Fielding, A. H. & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24, 38-49.

[17] Franklin, J., & Miller, J. A. (2009). Mapping species distributions: spatial inference and prediction. Cambridge University Press.

[18] Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T. et al. (2013) Predicting species distributions for conservation decisions (H. Arita, Ed.). Ecology Letters, 16, 1424–1435

[19] Halvorsen, R. (2013). A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. Sommerfeltia, 36(1), 1-132.

[20] Harrell. Jr., F. E., Lee, K. L., & Mark, D. B. (1996). Tutorial in biostatistics: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med, 15, 361-387.

[21] Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography, 29(5), 773-785.

[22] Hijmans, R. J., & Elith, J. (2013). Species distribution modeling with R. R Cran Project.

[23] Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. Ecological Modeling, 199, 142-152.

[24] Jiménez, L., & Soberón, J. (2020). Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence only data. Methods in Ecology and Evolution, 11(12), 1571-1586.

[25] Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and distributions, 14(6), 885-890.

[26] Kaky, E., Nolan, V., Alatawi, A., & Gilbert, F. (2020). A comparison between Ensemble and MaxEnt species distribution modelling approaches for conservation: A case study with Egyptian medicinal plants. Ecological Informatics, 60, 101150.

[27] Kozak, K. H., Graham, C. H., & Wiens, J. J. (2008). Integrating GIS-based environmental data into evolutionary biology. Trends in ecology & evolution, 23(3), 141-148.

[28] Kremen, C., Cameron, A., Moilanen, A., Phillips, S. J., Thomas, C. D., Beentje, H., Dransfield, J. et al. (2008). Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools. Science, 320, 222-226.

[29] Leathwick, J. R. (1998). Are New Zealand's Nothofagus species in equilibrium with their environment?. Journal of Vegetation Science, 9(5), 719-732.

[30] Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography, 17(2), 145-151.

[31] Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J., & Guisan, A. (2010). Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant. Biological conservation, 143(11), 2647-2657.

[32] Mendes, P., Velazco, S. J. E., Andrade, A. F. A. & De Marco, P. (2020). Dealing with overprediction in species distribution models: How adding distance constraints can improve model accuracy. Ecological Modelling, 431, 109180.

[33] Menezes, F. H., Feijó, A., Fernandes-Ferreira, H., da Costa, I. R., & Cordeiro-Estrela, P. (2021). Integrative systematics of Neotropical porcupines of Coendou prehensilis complex (Rodentia: Erethizontidae). Journal of Zoological Systematics and Evolutionary Research, 59(8), 2410-2439.

[34] Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., & Vanaverbeke, J. (2011). Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. Ecological Modelling, 222(3), 588-597.

[35] Miller, J. (2010). Species distribution modeling. Geography Compass, 4(6), 490-509.

[36] Morales, N. S., Fernández, I. C., & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. PeerJ, 5, e3093.

[37] Morrone, J. J., Escalante, T., Rodríguez-Tapia, G., Carmona, A., Arana, M. & Mercado-Gómez, J. D. (2022). Biogeographic regionalization of the Neotropical region: New map and shapefile. Anais da Academia Brasileira de Ciências, 94, e20211167.

[38] Narváez-Romero, C., Puig, C. R., Valle, D. & Brito, J. (2018). New records and estimation of the potential distribution of the Stump-Tailed Porcupine Coendou rufescens. Therya, 9, 137-146.

[39] Pearson, R. G., Raxworthy, C. J., Nakamura, M. & Peterson, A. T. (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography, 34, 102-117.

[40] Pecchi, M., Marchi, M., Burton, V., Giannetti, F., Moriondo, M., Bernetti, I., ... & Chirici, G. (2019). Species distribution modelling to support forest management. A literature review. Ecological Modelling, 411, 108817.

[41] Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecological modelling, 213(1), 63-72.

[42] Phillips, S. J. & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31, 161-175.

[43] Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological modelling, 190(3-4), 231-259.

[44] Ramírez-Chaves, H. E., Suárez-Castro, A. F., Morales-Martínez, D. M., & Vallejo-Pareja, M. C. (2016). Richness and distribution of porcupines (Erethizontidae: Coendou) from Colombia. Mammalia, 80(2), 181-191.

[45] Raphael, M. G. & Molina, R. (2007). Conservation of rare or little-known species: biological, social, and economic considerations. Washington, DC: Island Press.

[46] Rollins, M. G., Keane, R. E., & Parsons, R. A. (2004). Mapping fuels and fire regimes using remote sensing, ecosystem simulation, and gradient modeling. Ecological Applications, 14(1), 75-95.

[47] Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.

[48] Smith, A. B. (2013). On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. Diversity and Distributions, 19(7), 867-872.

[49] Srivastava, V., Lafond, V., & Griess, V. C. (2019). Species distribution models (SDM): applications, benefits and challenges in invasive species management. CABI Reviews, (2019), 1-13.

[50] Stockwell, D. R., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. Ecological modelling, 148(1), 1-13.

[51] Syphard, A. D., & Franklin, J. (2009). Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. Ecography, 32(6), 907 -918.

[52] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

[53] Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence only data with random forests. Ecography, 44(12), 1731-1742.

[54] Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological monographs, 92(1), e01486.

[55] VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know?. Ecological modelling, 220(4), 589-594.

[56] Vaughan, I. P., & Ormerod, S. J. (2005). The continuing challenges of testing species distribution models. Journal of applied ecology, 42(4), 720-730.

[57] Voss, R. S. (2015). Superfamily Erethizontoidea Bonaparte, 1845. In Patton, J. L., Pardiñas, U. F. J. & D'Elía, G.

(Ed.). Mammals of South America Volume 2. Rodents. USA: The University of Chicago Press.

[58] Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. Diversity and Distributions, 14(5), 763-773.

[59] Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., & Veran, S. (2013). Presence-only modelling using MAXENT: when can we trust the inferences?. Methods in Ecology and Evolution, 4(3), 236-243.

[60] Yu, H., Cooper, A. R., & Infante, D. M. (2020). Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. Ecological Modelling, 432, 109202.

[61] Zhang, L., Huettmann, F., Zhang, X., Liu, S., Sun, P., Yu, Z., & Mi, C. (2019). The use of classification and regression algorithms using the random forests method with presence-only data to model species' distribution. MethodsX, 6, 2281-2292.

[62] Zimmermann, N. E., Edwards Jr, T. C., Graham, C. H., Pearman, P. B., & Svenning, J. C. (2010). New trends in species distribution modelling. Ecography, 33(6), 985-989.

Tabela 2: Variables used to construct distribution models of *Coendou* species.

Source	Class	Abbreviation	Variable	Units
EarthEnv	Topographic	Elevation	Elevation	1 Km
		Slope	Slope	1 km
NEO Nasa Earth Observations	Vegetation index	NDVI	Normalized Difference Vegetation Index	1.0 degrees
WorldClim	Climatic	BIO1	Annual Mean Temperature	30 seconds (0.86 km² at the equator)
		BIO2	Mean Diurnal Range	
		BIO3	Isothermality	
		BIO4	Temperature Seasonality	
		BIO5	Max Temperature of Warmest Month	
		BIO6	Min Temperature of Coldest Month	
		BIO7	Temperature Annual Range	
		BIO8	Mean Temperature of Wettest Quarter	
		BIO9	Mean Temperature of Driest Quarter	
		BIO10	Mean Temperature of Warmest Quarter	
		BIO11	Mean Temperature of Coldest Quarter	
		BIO12	Annual Precipitation	
		BIO13	Precipitation of Wettest Month	
		BIO14	Precipitation of Driest Month	
		BIO15	Precipitation Seasonality	
		BIO16	Precipitation of Wettest Quarter	
		BIO17	Precipitation of Driest Quarter	
		BIO18	Precipitation of Warmest Quarter	
		BIO19	Precipitation of Coldest Quarter	