An Acad Bras Cienc (2023) 95(2): e20201421 DOI 10.1590/0001-3765202320201421
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

ECOSYSTEMS

Accurate species distribution models: minimum required number of specimen records in the Caatinga biome

AUGUSTO CÉSAR P. SAMPAIO & ARNÓBIO DE M.B. CAVALCANTE

Abstract: Species distribution models (SDMs) are one of the most widely used tools to predict areas with potential for occurrence of native, invasive and endangered species, based on current and future environmenal and climate conditions. Despite their global use, evaluating the accuracy of SDMs based only on presence records is still a challenge. The performance of models depends on the sample size and species prevalence. Recently, studies to model the distribution of species in the Caatinga biome in Northeast Brazil have gained force, raising the question about the minimum number of presence records adjusted to different prevalences that are necessary to generate accurate SDMs. In this context, the objective of this study was to indicate minimum numbers of presence records for species with different prevalences in the Caatinga biome to obtain accurate SDMs. For that purpose, we used a method involving simulated species and performed repeated evaluations of the models' performance in function of the sample size and prevalence. The results indicated that for this approach in the Caatinga biome, the minimum required numbers of specimen records were 17 and 30 for species with narrow and widespread distributions, respectively.

Key words: MaxEnt, minimum number, prevalence, sample size, simulated species.

INTRODUCTION

Species distribution modeling with presence-only algorithms relies on environmental preferences of the target species obtained from places of known occurrence to estimate the similarity of environmental conditions of any area (Phillips et al. 2006, Hijmans & Elith 2017). The result is expressed as the habitat suitability index of the species at each point, and can be used to predict probability of presence only by using specific modeling settings (Merow et al. 2013, Phillips et al. 2017).

In this respect, it is necessary for the presence records to closely represent the environmental scope of the species being modeled (Raes 2012), which is not always the case of species for which data have been collected from museums and herbaria, where the number of specimen records is generally deficient (Elith et al. 2006, Oliveira et al. 2016). Therefore, the bias from the sampling effort or collecting bias can be transposed to environmental bias, where only a fraction of the environmental conditions suitable to the species' occurrence is really represented in the sample, thus impairing the accuracy of the final model (Kadmon et al. 2003, 2004, Phillips et al. 2009).

Various studies have confirmed that the accuracy of species distribution models (SDMs), with the area held constant, increases with the number of presence records (McPherson et al. 2004) Hernandez et al. 2006, Pearson et al. 2007, Wisz et al. 2008, Bean et al. 2012). On the other hand, smaller sample sizes cause an increase in

environmental bias, which negatively affects the accuracy of the models' predictions (Hernandez et al. 2006, Pearson et al. 2007, Wisz et al. 2008).

However, the increase in environmental bias depends on the species' range sizes. Widespread species tend to require a larger number of presence records for good representation (Pearson et al. 2007, Wisz et al. 2008), while specialist species tend to generate more accurate models even with samples considered small, since these species occupy a much narrower niche (Hernandez et al. 2006, van Proosdij et al. 2016).

Hernandez et al. 2006 and van Proosdij et al. 2016, using subsampling and species simulation approaches respectively, demonstrated that the minimum number of presence records to generate accurate models also depends on the species prevalence, defined as the fraction of the modeled area occupied by the species (Phillips et al. 2009, Elith et al. 2011).

The relationship between the species prevalence and sample size reflects a more accurate ecological property (Allouche et al. 2006). However, when using models generated only with presence records, there is concern regarding the effects caused by the species prevalence on the accuracy indicators of SDMs (Vanderwal et al. 2009, van Proosdij et al. 2016). Therefore, assessment of the accuracy of SDMs that only use presence records is challenging (Elith et al. 2006, Hernandez et al. 2006). Objectively, that pattern results from the lack of data for independent tests, as well as the dependence on the species prevalence, intrinsic to the conventional evaluation methods, which use data from sampling tests (Lobo et al. 2008, Vanderwal et al. 2009, Jiménez-Valverde 2012, van Proosdij et al. 2016). This is the case of the Area Under the Receiver Operating Characteristic Curve (AUC), a method that is widely used to assess the predictive performance of SDMs (McPherson et al. 2004) Allouche et al. 2006), but that has been criticized, particularly when applied to presence-only algorithms, since in these cases we do not have absences. These are replaced by background points or pseudo-absences (Lobo et al. 2008, Vanderwal et al. 2009, Jiménez-Valverde 2012).

In light of this context, the use of simulated species can be useful to test adequate parameters for modeling. The simulation of species in a computational environment has been a recurring tool in studies of ecology and biogeography in recent years (Duan et al. 2015, Leroy et al. 2016, van Proosdij et al. 2016). The use of simulated species assumes control over the environmental parameters that define the presence of species, and their true distribution is known, permitting systematic evaluation of the models without the interference of the errors commonly found in data on real species (Hirzel et al. 2001, Jiménez-Valverde et al. 2009, Miller 2014).

Some studies have used the simulated species approach to evaluate the performance of SDMs with different sample sizes and prevalence classes (Jiménez-Valverde & Lobo 2007, Jiménez-Valverde et al. 2009, van Proosdij et al. 2016), as well as to reduce the effects of sampling bias on the modeling (Varela et al. 2014).

The use of SDMs has been growing worldwide, including in the Caatinga biome, a unique ecological region located in Northeast Brazil that contains the largest tropical seasonally dry forest in South America (Silva et al. 2018, Nascimento et al. 2020, Cavalcante et al. 2020). The minimum number of presence records considered in these studies has generally been a number recommended by the algorithm applied in the model. Therefore, it is opportune to indicate numbers that offer more statistical quality of biotic data, and hence greater precision of the models generated.

The objective of this study was to indicate the minimum numbers of presence records for species of different prevalence classes in the Caatinga biome, for the purpose of improving the accuracy and reliability of the SDMs produced.

MATERIALS AND METHODS

Modeling and Simulation

We used a species simulation approach implemented in the R computational environment (R Core Team 2019) as suggested by van Proosdij et al. (2016), as well as their proposal for assessment of the SDMs generated based on simulated species. The scripts in R adapted for this study are identified separately (Supplementary Material - Appendices S1-S4).

As predictor layers we used 19 climatic variables obtained from WorldClim 1.4 (2018) (Hijmans et al. 2005), with spatial resolution of 30 arc seconds (~ 1 km), and three topographic variables derived from the digital elevation model of the Shuttle Radar Topographic Mission, with spatial resolution of 3 arc seconds (~ 90 m) (CGIAR-CSI 2018). The spatial resolutions of the layers were standardized to 1 arc minute (~ 2 km), resulting in a total of 245,086 cells (pixels) in the area studied (Caatinga biome). Furthermore, we tested two other spatial resolutions (2.5 and 5 arc minutes), whose results are presented separately in tables (Appendices S5 and S6), because they did not present significant differences with the results reported here.

To reduce the correlation between the predictor layers, we used a 0.7 thresshold to Spearman's rank correlation coefficient to define sets of correlated variables (Spearman's |rho| > 0.7) (Dormann et al. 2013). Thus, we performed a principal component analysis (PCA) to select the variable with the strongest explanatory power within each set of correlated variables. Thus, from the initial set of 22 variables, we selected 7: altitude (Alt); standard deviation of altitude (Alt-sd); mean daily thermal amplitude (Bio2); temperature seasonality (Bio4); maximum temperature in the hottest month (Bio5); annual precipitation (Bio12); and seasonality of precipitation (Bio15). These variables were used to construct two orthogonal variables (van Proosdij et al. 2016), corresponding to PCA axes 1 and 2, which together represented 64.7% of the environmental variability in the study area.

The habitat suitability of each simulated species was defined based on a bivariate normal function, for which the values of the two orthogonal variables (Figures 1a, b) under each cell chosen randomly within the boundaries of the biome were used as the midpoints of the bivariate normal response curve. For this purpose, we used the "*dmvnorm*" function of the R library "*mvtnorm*" (Genz et al. 2019). The result was the defined habitat suitability (Figure 1c).

The defined presence (Figure 1d), in turn, was limited to the central interval of the bivariate normal density curve, which had probability of 68%. Therefore, we defined species' prevalence as the fraction of raster cells where the species is present and adjusted five prevalence classes (0.1, 0.2, 0.3, 0.4 and 0.5) utilizing increments in the standard deviation of the bivariate normal response, until approximation of the estimated prevalence of less than 1%.

Sampling

For each of the five prevalence classes, we simulated 100 species, for a total of 500 species, and for each of these species we get the presence records from the defined presence cells (Figure 1d) with 30 different sizes (3 to 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75 and 100), for a total of 15,000 samples sets with different sample sizes. The habitat suitability scores (Figure 1c) were used as the probability of selection of the

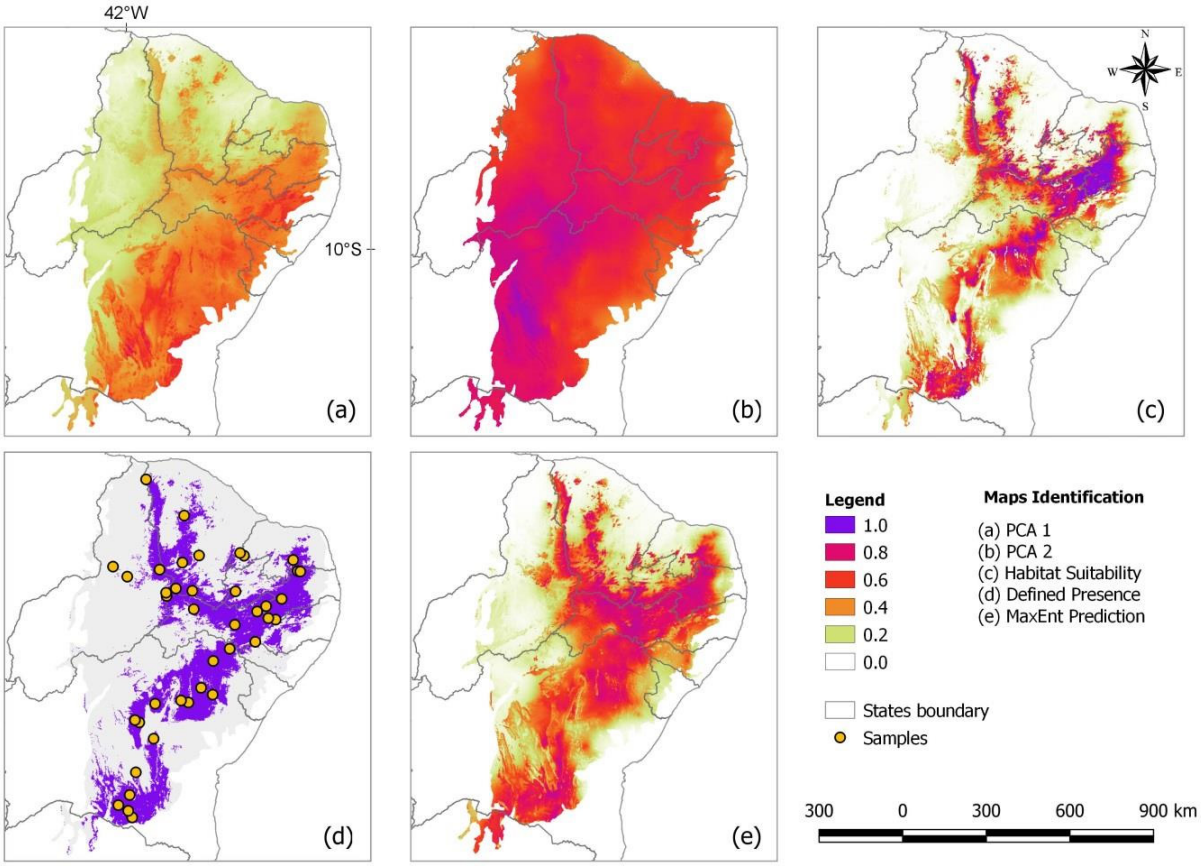

Figure 1. Methodological steps for species simulation with prevalence of 0.3 in the Caatinga biome. 1a, 1b - orthogonal predictor variables (PCA 1 and 2), 1c - habitat suitability defined by the simulation, 1d – presence defined by the simulation and selected presence records, 1e - habitat suitability predicted by MaxEnt (SDM).

presence records, based on the assumption that specimens should be more probably to find in places with higher suitability scores (Lomolino et al. 2010).

All the models were generated with the Maximum Entropy algorithm - MaxEnt (Phillips et al. 2006), indicated as one of the best algorithm, even with few presence records, compared to others algorithms (Hernandez et al. 2006, Pearson et al. 2007, Wisz et al. 2008). The MaxEnt algorithm only requires presence records and background points (Elith et al. 2006, Phillips et al. 2006). For each sample set, we generated 10,000 random background points for use in the modeling, resulting in 15,000 SDMs generated with MaxEnt (Figure 1e). Models replicate were not calculated for the same sample set, since random test points were not extracted from the sample sets. All presence records were used in models training.

Assessment of the models' accuracy

We used the AUC as the main method to evaluate the accuracy of the SDMs generated. The AUC is based on the estimate of the correct presences (sensitivity) and false absences (1 - specificity) forecast (Allouche et al. 2006, Phillips et al. 2006). Hence, it represents the chance that a randomly chosen presence receives a higher suitability score than a randomly chosen absence (Merow et al. 2013). Accordingly, AUC values of around 0.5 indicate the model has a nearly random

distribution, while values near 1.0 indicate high ability to distinguish between presences and absences (Elith et al. 2006).

In the step of evaluating presence-only models, as is the case of MaxEnt, the background points are utilized as pseudo-absences (Phillips et al. 2006, Elith et al. 2006, Merow et al. 2013). This procedure causes the number of background samples classified as having false absences to increase with species prevalence (Lobo et al. 2008, Vanderwal et al. 2009, Jiménez-Valverde 2012), so that the maximum AUC value in these cases is not 1, but 1 minus half of the prevalence ($1 - a/2$), where a is the species prevalence, unknown *a priori* (Phillips et al. 2006, Raes & ter Steege 2007). Because of this dependence on prevalence, the AUC has been questioned as a single evaluation criterion in models with only presence (Lobo et al. 2008, Vanderwal et al. 2009).

On the other hand, the Real AUC was calculated based on the cross-validation between the suitability scores predicted by MaxEnt and a random subsample of 10% of the presences and absences defined by the simulation. For that purpose, we used the "*evaluate*" function of the R library "*dismo*" (Hijmans et al. 2017).

An alternative to using the MaxEnt AUC as the only evaluation criterion is to compare it with the performance of null models and check whether the AUC of the model being evaluated is better than the random expectation (Raes & ter Steege 2007). SDM null models have been used to improve testing for statistical significance (Raes & ter Steege 2007, Merckx et al. 2011, Bohl et al. 2019). For each sample size, we generated 99 null models with randomly selected presence records in the entire area studied, in a number equal to the sample sizes considered. Random presence records were modelled similarly as the species, resulting in 99 null models AUCs for each sample size. The Real AUC and MaxEnt AUC values are regarded significantly better than random expectation if those values exceed rank number 95, when ranked with the 99 null model AUC values, corresponding a 95% statistical significance threshold (p-value 0.05).

In complement, we directly compared the predictive result of MaxEnt with the true distribution of the simulated species through the Spearman correlation coefficient (*rho*), calculated by cell-by-cell comparison between the defined and predicted habitat suitability layers.

Criteria to define the minimum number of presence records

We disregarded 5% of the values of Real AUC, MaxEnt AUC, Real AUC Rank, MaxEnt AUC Rank and Spearman correlation for the SDMs with worst performance of the 100 repetitions for each combination between sample size and prevalence. We then smoothed the lower and upper limits of the remaining 95% of the models to attenuate the stochastic effects, employing the "*loess*" function of the R library 'stats' (R Core Team 2019).

When considering only the lower limit of the range of values corresponding to 95% of the SDMs evaluated as best for each sample size (p-value 0.05), the minimum number of presence records to generate accurate SDMs was defined as the sample size where: a) the Real AUC of the SDM exceeded 0.9, characterizing high accuracy of the model (Lobo et al. 2008); b) the Real AUC value of the SDM surpassed 95% of the null models (Real AUC Rank), corresponding to a significantly better performance than expected randomly (Raes & ter Steege 2007); and c) the value of the Spearman correlation between the SDM and the defined habitat suitability exceeded 0.9 (van Proosdij et al. 2016). As discussed above and demonstrated shortly, the values of MaxEnt AUC and MaxEnt AUC Rank were not ideal to define

the minimum number of presence records, so they were disregarded for this purpose.

RESULTS AND DISCUSSION

The accuracy of the models was evaluated considering only the lower limit of the range of values corresponding to the upper 95% of the Real AUC, MaxEnt AUC, Real AUC Rank, MaxEnt AUC Rank and Spearman correlation results. According to the evaluation criteria used in this study, the minimum number of presence records to generate accurate SDMs in the Caatinga biome increased with higher observed species prevalence (Table I).

We found many minimum numbers of presence records as low as 17 for species with low prevalence (0.1) and of 30 for species with high prevalence (0.5), considering an Real AUC > 0.9. The minimum numbers indicated by the Real AUC Rank and MaxEnt AUC Rank criteria were significantly lower (Table I). This can partly be attributed to the nature of these criteria, which classify SDMs as better than the random expectation instead of just being good on their own.

According to the patterns of Real AUC, MaxEnt AUC, Real AUC Rank, MaxEnt AUC Rank and Spearman correlation obtained for the SDMs generated with various sample sizes and number of presence records, it was possible to identify the number of presence records where the curve approached the asymptote in the graphical representation (Figures 2, 3, 4). This number rises as the species prevalence increases. Sample sizes smaller than the minimum threshold

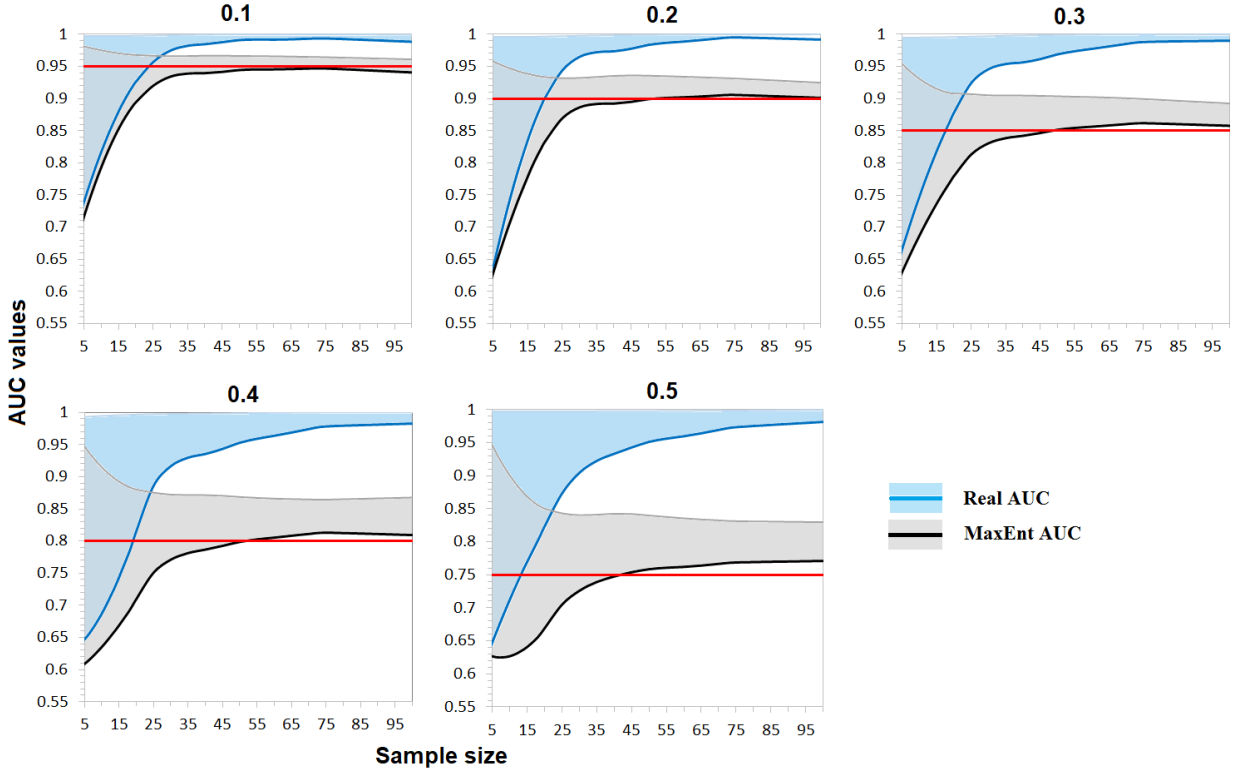

Figure 2. Predictive performance of the SDMs based on the AUC values (Real and MaxEnt) for each combination of sample size and prevalence (separate squares) of species in the Caatinga biome. The solid lines represent the lower limits of the ranges of values corresponding to the 95% of SDMs with best performance, represented by the shaded area in the graph. The horizontal red line identifies the maximum AUC attained by MaxEnt, which corresponds to 1 - *a* / 2 (where *a* is the species prevalence).

produce sampling bias, which can be translated into environmental bias, impairing the predictive accuracy of the models. According to Luoto et al. (2005) and Segurado et al. (2006), widely distributed species (high prevalence) are more vulnerable to environmental bias caused by deficient samples, an observation corroborated by our results.

Furthermore, there was large variability in the results of the accuracy indicators of the SDMs for very low presence record numbers. This variability is denoted graphically by the shaded regions between the lower and upper limits of the range of values for the upper 95% of the results (Figures 2, 3, 4). These regions are larger for small samples and high prevalences, indicating the imprecision of the models generated with insufficient presence records for the respective prevalence class, as also reported by Pearson et al. (2007) and van Proosdij et al. (2016).

The performance of the SDMs increased progressively with rising sample size until reaching stability, represented by the asymptote in the graph, as of which an increase in sample size no longer produced significant improvement of the models (Figures 2, 3, 4). This is in line with the finding of Oliveira et al. (2016), who reported lower environmental deficit for species with higher number of presence records.

The accuracy indicated by the MaxEnt AUC value was considerably worse for species with high prevalence, although this did not correspond to an equally inferior performance when considering the accuracy indicated by the Real AUC (Figure 2), causing a need to evaluate

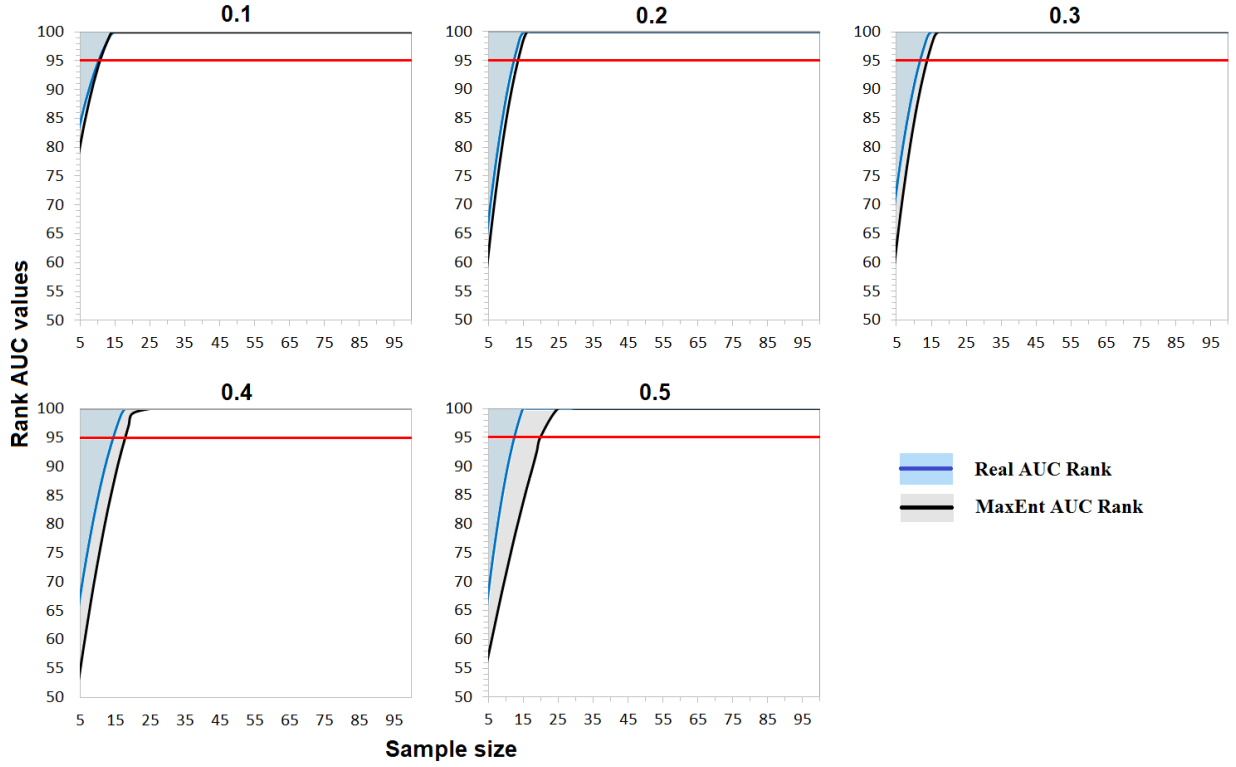

Figure 3. Predictive performance of the SDMs based on the ranking of the AUC values (Real and MaxEnt) in relation to the null models for each combination of sample size and prevalence (separate squares) of species in the Caatinga biome. The solid lines represent the lower limits of the ranges of values corresponding to the 95% of SDMs with best performance, represented by the shaded area in the graph. The horizontal red line identifies the 95% critical ranking value, considering significance of 0.05.

it in relation to the theoretical level for each prevalence defined by Phillips et al. (2006).

The MaxEnt AUC values slightly exceeded the predicted maximum AUC based on prevalence (a) advocated by Phillips et al. (2006), where: $AUC_{MaxEnt} = 1 - a/2$. This can be partly attributed to the optimistic method used in selecting the presence records, with probability defined by the habitat suitability (Figure 2). Due to its dependence on prevalence, the MaxEnt AUC value can incorrectly reject adequate SDMs, principally for highly prevalent species. For this reason, we avoided using it as a criterion to define the minimum number of presence records in this study.

Furthermore, the minimum number of presence records based on the values of the Real AUC Rank and MaxEnt AUC Rank (Figure 3) also increased with higher species prevalence. However, based on the nature of this evaluation method, which classifies models as better than the random expectation, the minimum values were substantially lower and some stochastic effects could be noted.

The Spearman correlation coefficient (Figure 4) was the most demanding criterion in terms of minimum numbers of presence records. This was expected, since it is the only one that permits an absolute cell-by-cell comparison between the defined habitat suitability and the value predicted by MaxEnt. According to this criterion, we defined minimum numbers of presence records as low as 19 for species with low prevalence (0.1) and up to 50 for species with high prevalence (0.5). The Spearman coefficient is a very conservative criterion in relation to the

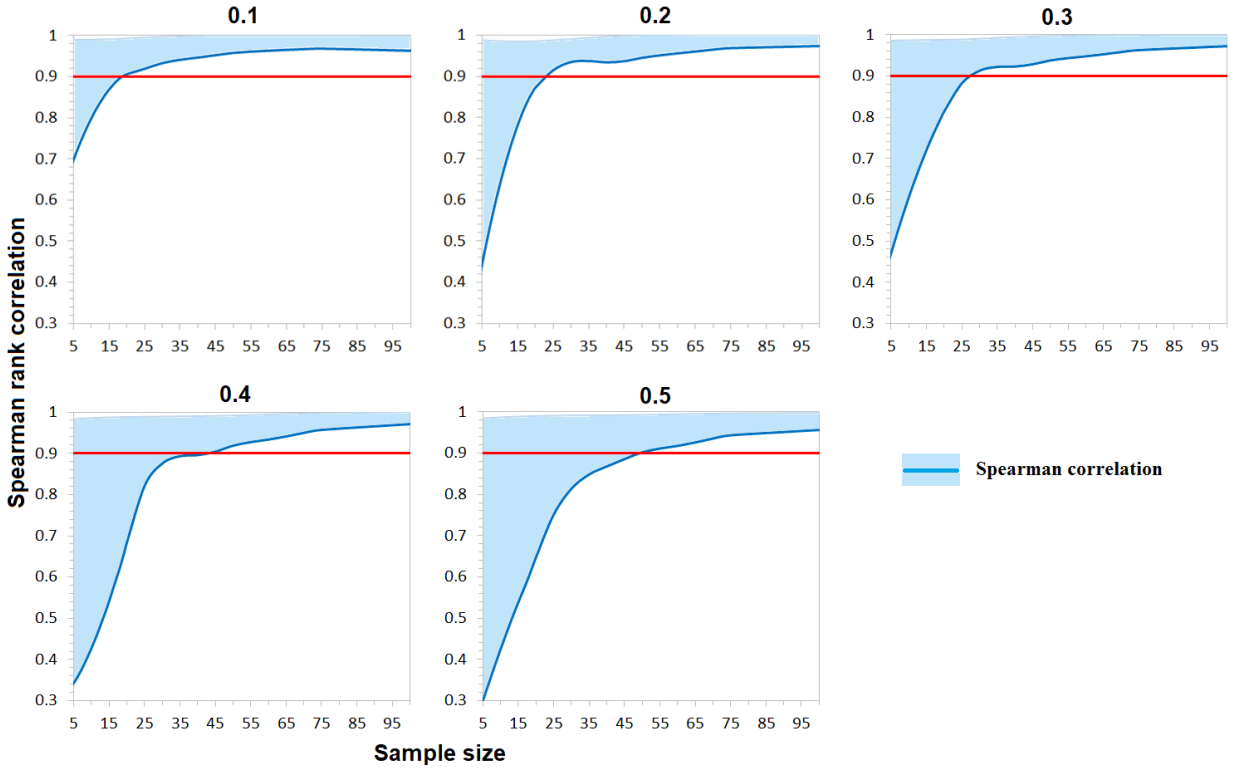

Figure 4. Predictive performance of the SDMs based on the Spearman correlation coefficient for each combination of sample size and prevalence (separate squares) of species in the Caatinga biome. The solid lines represent the lower limits of the ranges of values corresponding to the 95% of SDMs with best performance, represented by the shaded area in the graph. The horizontal red line identifies the value of the Spearman correlation coefficient equal to 0.9.

representation of the real distribution by the SDM, with the numbers 19/50 being applied to increase the accuracy of the models even more, but without invalidating the numbers 17/30 indicated by the AUC Real values, which were adequate to construct accurate SDMs in the Caatinga biome.

The results of this study corroborate the occurrence of a relationship between rising sample size and more precise models, in line with the findings of other researchers (McPherson et al. 2004) Hernandez et al. 2006, Pearson et al. 2007, Wisz et al. 2008, Bean et al. 2012, Van Proosdij et al. 2016). The results also showed that the species prevalence in the area studied significantly influenced the performance of the models, corroborating the theoretical expectation that species with wider distribution require more presence records due to the greater environmental complexity of their niches (Wisz et al. 2008, Hernandez et al. 2006). According to Allouche et al. (2006), this reflects a true ecological characteristic.

CONCLUSION

The species simulation method was applied successfully to the Caatinga biome, helping to verify the necessary levels of quantitative and qualitative data, as well as improving the accuracy and reliability of the SDMs. As was expected, the sample size required to generate accurate SDMs depended on the species prevalence considered. Therefore, it is fundamental to estimate the prevalence in advance using exploratory SDMs.

The species prevalence was decisive as an evaluation criterion based only on presence records, represented here by the MaxEnt AUC value, which had lower reliability for species with high prevalence. This dependence of the MaxEnt AUC value in relation to the prevalence level supports the importance of using simulated species to define in advance the minimum numbers of presence records in models of species distribition that only use presences.

In turn, the minimum required numbers of specimen records for species in the Caatinga biome was 17 for species with low prevalence (narrow distribution) and 30 for species with high prevalence (widespread distribution). Therefore, this study provides a secure estimate of the minimum number of presence records for modeling the distribution of species in the Caatinga biome. However, it should be noted that: (1) alternative modeling methods to MaxEnt still need to be tested; and (2) these numbers should be taken as a base for coarser resolutions, common in online climate databases.

Acknowledgments

We are grateful for the PCI-DC research grant provided by the Instituto Nacional de Pesquisas Espaciais (INPE).

REFERENCES

ALLOUCHE O, TSOAR A & KADMON R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43 (6): 1223-1232.

BEAN WT, STAFFORD R & BRASHARES JS. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography 35(3): 250-258.

BOHL CL, KASS JM & ANDERSON RP. 2019. A new null model approach to quantify performance and significance for ecological niche models of species distributions. J Biogeogr 46(6): 1101-1111.

CAVALCANTE AMB, DUARTE AS & OMETTO JPHB. 2020. Modeling the potential distribution of Epiphyllum phyllanthus (L.) Haw. under future climate scenarios in the Caatinga biome. An Acad Bras Cienc 92: 1-12.

CGIAR-CSI. 2018. CGIAR - Consortium for Spatial Information (Version 4): SRTM Data. Available at: <www.srtm.csi.cgiar.org>. Last access: August 2019.

DORMANN CF ET AL. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36(1): 27-46.

DUAN RY, KONG XQ, HUANG MY, WU GL & WANG ZG. 2015. SDMvspecies: a software for creating virtual species for species distribution modelling. Ecography 38(1): 108-110.

ELITH J ET AL. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2): 129-151.

ELITH J, PHILLIPS SJ, HASTIE T, DUDIK M, CHEE YE & YATES CJ. 2011. A statistical explanation of MaxEnt for ecologists. Divers Distrib 17(1): 43-57.

GENZ A, BRETZ F, MIWA T, MI X, LEISCH F, SCHEIPL F, BORNKAMP B, MAECHLER M & HOTHORN T. 2019. mvtnorm: multivariate normal and t distributions. R package ver. 1.0-11. Available at: <https://CRAN.R-project.org/package=mvtnorm>. Last access: May 2019.

HERNANDEZ PA, GRAHAM CH, MASTER LL & ALBERT DL. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29(5): 773-785.

HIJMANS RJ, CAMERON SE, PARRA JL, JONES PG & JARVIS A. 2005. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25(15): 1965-1978.

HIJMANS RJ & ELITH J. 2017. Species distribution modeling with R. R CRAN Project, 78 p. Available at: <https://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf>. Last access: July 2020.

HIJMANS RJ, PHILLIPS S, LEATHWICK J & ELITH J. 2017. dismo: species distribution modeling. – R package ver. 1.1-4. Available at: <https://CRAN.R-project.org/package=dismo>. Last access: May 2019.

HIRZEL AH, HELFER V & METRAL F. 2001. Assessing habitat-suitability models with a virtual species. Ecol Model 145(2-3): 111-121.

JIMÉNEZ-VALVERDE A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecol Biogeogr 21(4): 498-507.

JIMÉNEZ-VALVERDE A & LOBO JM. 2007. Threshold criteria for conversion of probability of species presence to either-or presence-absence. Acta Oecol 31(3): 361-369.

JIMÉNEZ-VALVERDE A, LOBO JM & HORTAL J. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecol 10(2): 196-205.

KADMON R, FARBER O & DANIN A. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecol Appl 13(3): 853-867.

KADMON R, FARBER O & DANIN A. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol Appl 14(2): 401-413.

LEROY B, MEYNARD CN, BELLARD C & COURCHAMP F. 2016. virtualspecies, an R package to generate virtual species distributions. Ecography 39(6): 599-607.

LOBO JM, JIMÉNEZ-VALVERDEL A & REAL R. 2008. AUC: A misleading measure of the performance of predictive distribution models. Global Ecol Biogeogr 17(2): 145-151.

LOMOLINO MV, RIDDLE BR, WHITTAKER RJ & BROWN JH. 2010. Biogeography, 4th ed., Sunderland: Sinauer Associates, 878 p.

LUOTO M, POYRY J, HEIKKINEN RK & SAARINEN K. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. Global Ecol Biogeogr 14(6): 575-584.

MCPHERSON JM, JETZ W & ROGERS DJ. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J Appl Ecol 41(5): 811-823.

MERCKX B, STEYAERT M, VANREUSEL A, VINCX M & VANAVERBEKE J. 2011. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. Ecol Model 222(3): 588-597.

MEROW C, SMITH M J & SILANDER JUNIOR JA. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36(10): 1058-1069.

MILLER JA. 2014. Virtual species distribution models: using simulated data to evaluate aspects of model performance. Prog Phys Geogr 38(1): 117-128.

NASCIMENTO FAO, MOURA-JÚNIOR EG, NASCIMENTO ES & RODRIGUES RG. 2020. Modeling the potential distribution of Anamaria heterophylla (Giul. & V.C. Souza) V.C. Souza (Plantaginaceae) in the Caatinga. Oecol Australis 24(1): 76-87.

OLIVEIRA U ET AL. 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. Divers Distrib 22(12): 1232-1244.

PEARSON RG, RAXWORTHY CJ, NAKAMURA M & PETERSON AT. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J Biogeogr 34(1): 102-117.

PHILLIPS SJ, ANDERSON RP, DUDÍK M, SCHAPIRE RE & BLAIR ME. 2017. Opening the black box: An open-source release of Maxent. Ecography 40(7): 887-893.

PHILLIPS SJ, ANDERSON RP & SCHAPIRE R. 2006. Maximum entropy modeling of species geographic distributions. Ecol Model 190(3-4): 231-259.

PHILLIPS SJ, DUDIK M, ELITH J, GRAHAM CH, LEHMANN A, LEATHWICK J & FERRIER S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19(1): 181-197.

R CORE TEAM. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/>. Last access: May 2019.

RAES N. 2012. Partial versus full species distribution models. Nat Conserv 10(2): 127-138.

RAES N & TER STEEGE H. 2007. A null model for significance testing of presence only species distribution models. Ecography 30(5): 727-736.

SEGURADO PAGE, ARAUJO MB & KUNIN WE. 2006. Consequences of spatial autocorrelation for niche-based models. J Appl Ecol 43(3): 433-444.

SILVA UBT, DELGADO-JARAMILLO M, AGUIAR LMS & BERNARD E. 2018. Species richness, geographic distribution, pressures, and threats to bats in the Caatinga drylands of Brazil. Biol Conserv 221: 312-322.

VANDERWAL J, SHOO LP, GRAHAM C & WILLIAMS SE. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? Ecol Model 220(4): 589-594.

VAN PROOSDIJ AS, SOSEF M, WIERINGA JJ & RAES N. 2016. Minimum required number of specimen records to develop accurate species distribution models. Ecography 39(6): 542-552.

VARELA S, ANDERSON RP, GARCÍA-VALDÉS R & FERNÁNDEZ-GONZÁLEZ F. 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography 37(11): 1084-1091.

WISZ M, HIJMANS RJ, LI J, PETERSON AT, GRAHAM CH, GUISAN A & NCEAS. 2008. Effects of sample size on the performance of species distribution models. Divers Distrib 14(5): 763-773.

WORLDCLIM. 2018. Worldclim Version 1.4. Available at: <http://worldclim.org/version1 >. Last access: November 2018.

SUPPLEMENTARY MATERIAL

Appendices S1-S6.

How to cite

SAMPAIO ACP & CAVALCANTE. AMB 2023. Accurate species distribution models: minimum required number of specimen records in the Caatinga biome. An Acad Bras Cienc 95: e20201421. DOI 10.1590/0001-3765202320201421.

Manuscript received on November 2, 2020;
accepted for publication on January 11, 2021

AUGUSTO CÉSAR P. SAMPAIO
https://orcid.org/0000-0003-2309-4292

ARNÓBIO DE M.B. CAVALCANTE
https://orcid.org/0000-0001-5541-6677

Instituto Nacional de Pesquisas Espaciais, Coordenação Geral de Ciências da Terra, Divisão de Impactos, Adaptação e Vulnerabilidade, Av. dos Astronautas, 1758, 12227-010 São José dos Campos, SP, Brazil

Correspondence to: **Arnóbio de Mendonça Barreto Cavalcante**
E-mail: arnobio.cavalcante@inpe.br

Author contributions

Cavalcante AMB contributed to the design, choice of methodological strategy, interpretatoion of the results and writing of the manuscript. Sampaio ACP contributed to the methodological adaptation and construction of the models, along with interpretatoion of the results and writing of the manuscript.