iForest
Biogeosciences and Forestry

Research Article
doi: 10.3832/ifor4084-015
vol. 15, pp. 363-371

Predictive capacity of nine algorithms and an ensemble model to determine the geographic distribution of tree species

Juan Carlos Montoya-Jiménez [(1)], **José René Valdez-Lazalde** [(1)], **Gregorio Ángeles-Perez** [(1)], **Héctor Manuel de los Santos-Posadas** [(1)], **Gustavo Cruz-Cárdenas** [(2)]

The different models that predict the distribution of species are a useful tool for the evaluation and monitoring of forest resources as they facilitate the planning of their management in a changing climate environment. Recently, a significant number of algorithms have been proposed for this purpose, making it difficult to select the most appropriate to use. The evaluation of performance and predictive stability of these models can elucidate this problem. Distribution data of 17 pine species with high economic importance for Mexico were collected and distribution models were carried out. We carried out a pre-modeling design to select the prediction variables (climatic, edaphic and topographic), after which nine algorithms and an ensemble model were contrasted against one another. The true skill statistic (TSS) and the area under the curve (AUC) were used to evaluate the predictive performance of the models, and the coefficient of variation of the predictions was used to evaluate their stability. The number of predictive variables in the final models fluctuated from 6 to 12; the mean diurnal range and the maximum temperature of warmest month were included in the models for most species. Random forests, the ensemble model, generalized additive models and MaxEnt were the ones that best described the distribution of the species (AUC> 0.92 and TSS> 0.72); the opposite was found in Bioclim and Domain (AUC<0.75 and <0.82; and TSS<0.5 and <0.55). Support vector machine, Mahalanobis distance, generalized linear models and boosted regression trees obtained intermediate settings. The coefficient of variation indicated that Bioclim, Domain and Support vector machine have low predictive stability (CV>0.055); on the contrary, Maxent and the ensemble model attained high predictive stability (CV<0.015). The ensemble model obtained greater performance and predictive stability in the predictions of the distribution of the 17 species of pines. The differences found in performance and predictive stability of the algorithms suggest that the ensemble model has the potential to model the distribution of tree species.

Keywords: TSS, AUC, BRT, SVM, MaxEnt, Random Forests, GAM, Ensemble Model

Introduction

Species distribution models (SDM) are statistical methods or machine learning algorithms used to model and map past, present or future species distributions (Elith et al. 2006, Pecchi et al. 2019). These tasks are relevant for forest resources management (Pecchi et al. 2020) and for forest conservation planning (Ramos-Dorantes et al. 2017).

The SDM are performed by three different approaches: correlative, mechanistic, and process-oriented or hybrid (Peterson et al. 2012). The correlative approach is the most widely used due to the availability of databases of the presence records of a species, climatic and meteorological data, and computerized algorithms that facilitate the modeling process (Hijmans & Elith 2017). Additionally, the models with a correlative approach are the simplest to apply, since they rely on the statistical relationship that exists between the presence records of the species and the data that describes the environment they inhabit (Peterson et al. 2012). Recent studies showed that the choice of the algorithm used for SDM is a source of variation in the process of the spatial prediction of the species, which can significantly affect the performance of the model (Jarnevich & Young 2019, Pecchi et al. 2020). This is the reason why the current trend is to examine multiple algorithms and select the most appropriate for their application, in accordance with the objective of each research (Ren-Yan et al. 2014, Shabani et al. 2016).

The algorithms available to implement SDM under the correlative approach are diverse (Tab. 1), all of which present advantages and disadvantages in its application,

(1) Postgrado en Ciencias Forestales. Colegio de Postgraduados. Campus Montecillo (México); (2) Instituto Politécnico Nacional, CIIDIR-IPN-Michoacán, COFAA, Justo Sierra 28, 59510 Jiquilpan, Michoacán (México)

@ José René Valdez-Lazalde (valdez@colpos.mx)

Received: Feb 22, 2022 - Accepted: Jul 12, 2022

Citation: Montoya-Jiménez JC, Valdez-Lazalde JR, Ángeles-Perez G, De Los Santos-Posadas HM, Cruz-Cárdenas G (2022). Predictive capacity of nine algorithms and an ensemble model to determine the geographic distribution of tree species. iForest 15: 363-371. - doi: 10.3832/ifor4084-015 [online 2022-09-20]

Communicated by: Maurizio Marchi

Tab. 1 - Evaluated algorithms to build the spatial distribution models of 17 species of pines. *For these analyzes, for each evaluated species we selected 20.000 pseudo-absences at random.

Statistical approach – General description	Data	Algorithm	Reference
Distance – they are considered climate envelope algorithms and calculate the similarity that exists between candidate pixels with respect to the selected presence records.	Presence	Bioclim (BIO)	Nix (1986)
		Domain (DOM)	Carpenter et al. (1993)
		Mahalanobis distance (MD)	Etherington (2019)
Regression – they are algorithms that model the median of a response variable regarding prediction variables; they use the Logit Link Function to relate the expected value of the response variable with included predictors.	Presence/ absence*	Generalized linear model (GLM)	Guisan et al. (2002)
		General additive models (GAM)	
Machine Learning – within SDM they are algorithms that focus on classification. Their main objective is to automatically improve the classification of training data until finally obtaining a better model.	Presence/ absence*	Sector-vector machine (SVM)	Betancourt (2005)
		Boosted regression trees (BRT)	Hijmans & Elith (2017)
		Random forests (RF)	Mi et al. (2017)
	Presence/ Background*	MaxEnt (MAX)	Phillips et al. (2006)

performance and predictive stability (Peterson et al. 2012). In SDM, an algorithm is rarely identified as the best one in a consistent manner (Ren-Yan et al. 2014), since they are sensitive to the data and mathematical functions used to describe the distribution of the species based on environmental variables. In order to reduce this variation, it has been suggested to combine the predictions of different algorithms in a composite called ensemble model (Araújo & New 2007). The way to build the ensemble model can be through the median, the mean, and the weighted mean based on the predictive performance of the individual algorithms, measuring the performance based on statistics such as the area under the curve (AUC) and the true skill statistic (TSS – Araújo & New 2007, Hao et al. 2019). Recent studies report that evaluating ensemble models from two or more algorithms is a viable alternative to improve prediction and consequently reduce uncertainty (Pecchi et al. 2020).

The genus *Pinus* has a notable importance in the Mexican economy by contributing around 6.3 million cubic meters of wood per year (CONAFOR 2019). According to the Biometric System for Planning Sustainable Forest Management (SIBIFOR), 17 pine species contribute with about 70% of the timber production in the country (Vargas-Larreta et al. 2017); these 17 species were incorporated to our SDM modelling analysis.

Recently, efforts have been made in Mexico to model the distribution areas of some pine species by means of the MaxEnt algorithm (Aceves-Rangel et al. 2018, García-Aranda et al. 2018, Reynoso Santos et al. 2018, Manzanilla-Quiñones et al. 2019). However, it has not been considered to evaluate the performance and predictive stability of other algorithms, which could mean that the most robust and reliable models designed to carry out this task are not being used. Consequently, forest resource managers would be dispensing with the best information in their process of decision-making. Therefore, the objective of this study was to evaluate the performance and predictive stability of nine algorithms and an ensemble model in defining the geographic distribution area of 17 coniferous species with high economic importance in Mexico.

Materials and methods

Species and presence records

The seventeen pine species that most contribute to timber production in Mexico (up to 70 %) were considered in the study: *Pinus arizonica* (P.ar), *P. ayacahuite* (P.ay), *P. cembroides* (P.ce), *P. devoniana* (P.de), *P. douglasiana* (P.do), *P. durangensis* (P.du), *P. hartwegii* (P.ha), *P. herrerae* (P.he), *P. leiophylla* (P.le), *P. maximinoi* (P.ma), *P. montezumae* (P.mo), *P. oocarpa* (P.oo), *P. patula* (P.pa), *P. pseudostrobus* (P.ps), *P. strobiformis* (P.st), *P. strobus* var. *chiapensis* (P.sc), and *P. teocote (P.te)*. A database containing 17,908 presence records for the 17 species (Tab. 2) was assembled with information from the National Forest and Soil Inventory of Mexico (INFyS 2009-2014), the Global Biodiversity Information Facility (GBIF 2020), the National Commission for the Knowledge and Use of the Biodiversity (CONABIO 2020). We used two land use and vegetation type maps scale 1:250,000 corresponding to years 1985 and 2014 (INEGI's series 1 and 6 respectively – INE/INEGI 1997, INEGI 2016) to eliminate repeated, incomplete and poorly georeferenced data. For species distributed beyond the limits of Mexico, we use of a global land cover map (Hansen et al. 2000). For coarse scale presence record validation, we used the Atlas of the World's Conifers (Farjon & Filer 2013), but the available scientific literature (Ramos-Dorantes et al. 2017, Aceves-Rangel et al. 2018, García-Aranda et al. 2018, Reynoso Santos et al. 2018, Manzanilla-Quiñones et al. 2019) was used for finer scale validation of the records collected for each species so that they coincided with their reported natural distribution.

The databases for each species were entered into the Diva-GIS program ver. 7.5 (Hijmans et al. 2012). Through jacknife resampling (Chapman 2005) atypical climatic values were excluded, all of which might be related to poor georeferencing or to problems in the taxonomic identification of the species. For all species, records that contained three or more atypical climate conditions were excluded from the analysis (Chapman 2005). Finally, the density of the

Tab. 2 - Number of occurrences by species used to estimate the distribution area of 17 species of pines. (NP): Number of presences.

Species	NP	Species	NP
Pinus arizonica (P.ar)	1165	*Pinus maximinoi* (P.ma)	370
Pinus ayacahuite (P.ay)	234	*Pinus montezumae* (P.mo)	560
Pinus cembroides (P.ce)	1846	*Pinus oocarpa* (P.ooc)	2182
Pinus devoniana (P.de)	540	*Pinus patula* (P.pa)	479
Pinus douglasiana (P.do)	394	*Pinus pseudostrobus* (P.ps)	1667
Pinus durangensis (P.du)	1426	*Pinus strobiformis* (P.st)	1430
Pinus hartwegii (P.ha)	448	*Pinus strobus* var. *chiapensis* (P.sc)	117
Pinus herrerae (P.he)	803	*Pinus teocote* (P.te)	1786
Pinus leiophylla (P.le)	2461	-	-

presence records was reduced to one record per km² to reduce the effects of sampling bias and thus avoid over-fitting the models due to redundant environmental information (Hijmans & Elith 2017).

Selection of environmental variables and calibration area

The environmental variables that characterize the areas where the species presence were recorded were obtained from the WorldClim version 2 repository, with an approximate resolution of 1 km² (Fick & Hijmans 2017). From 19 available variables, isothermality (Bio3), temperature seasonality (Bio4), temperature annual range (Bio7) and precipitation seasonality (Bio15) were eliminated, since biologically they are difficult to interpret. Escobar et al. (2014) reported that the mean temperature of wettest quarter (Bio8), mean temperature of driest quarter (Bio9), precipitation of warmest quarter (Bio 18) and precipitation of coldest quarter (Bio19) show discontinuities between neighboring pixels, and therefore these variables were also excluded (Tab. 3). We incorporated to the analysis edaphic variables available in the SoilGrids database (Hengl et al. 2017), since it has been shown that the combined use of edaphic and climatic variables improve the precision of predictions, compared to those constructed only with climatic variables (Velazco et al. 2017). We also considered topographic variables in the modeling process (altitude, slope, diurnal anisotropic heat, Convergence index, Terrain ruggedness index, Topographic wetness index) derived from the digital elevation model available at the WorldClim website in SAGA-GIS v. 7.5.0 (Conrad et al. 2015), as they were formerly identified important for modeling the spatial distribution of some pine species analyzed in our study (Ramos-Dorantes et al. 2017).

The selection of the environmental variables to be used in SDM is an important aspect in the modeling process, since they have a significant effect on the predictive performance of the models (Cobos et al. 2019). In this study, for each species a pre-modeling exercise was initially performed to identify the five most important predictor variables according to each tested algorithm. We fitted models iteratively (including or excluding each variable) and monitored the area under the curve (AUC) statistic value obtained to identify the five variables with the greatest contribution to predict the potential distribution area of the species. Variables that were identified in the top five by more than one algorithm were added only once to conform a set of likely predictor variables for a given pine species. We ended up with preliminary sets of 12-19 predictor variables depending on the pine species analyzed. We also implemented a variance inflation factor analysis (VIF <5 – Cobos et al. 2019) to identify and minimize the presence of multicollinearity in the subset of relevant predictor variables. The final number of uncorrelated predictive variables was different for each species (Tab. S2 in Supplementary material).

In the process of building species distribution models, the definition of the accessible area for the species is a critical factor for the result of the calibration, evaluation and comparison of the model (Peterson et al. 2012). In this study, the accessible area for each species was made from the terrestrial ecoregions of the world that delimit the distribution of each of the species, since they are zones with common physiographic, biological and historical characteristics. In this sense, climatological, geological and edaphological conditions are similar, and they are of great importance in the distribution of species and communities (Farjon & Filer 2013).

Spatial modeling process: algorithms and predictions

The modeling process for each species was carried out with the following configuration: 20,000 randomly pseudo-absences were generated and the evaluation of the models was executed with the bootstrap resampling method with 10 repetitions (Naimi & Araújo 2016), since using 70% and 30% of the presence data for the calibration and evaluation of the model respectively can overestimate the evaluation parameters of the models (Radosavljevic & Anderson 2014). The SDM was performed for each of the species of interest using nine algorithms (Tab. 1) and an ensemble model, which was built from the weighted arithmetic mean of the true skill statistic (TSS) values obtained for the three algorithms with the highest predictive performance (Marmion et al. 2009).

The performance of the algorithms was evaluated through the TSS (Hao et al. 2020) and the AUC (Pecchi et al. 2020). In order to calculate the TSS, binary predictions (presence/absence) are required, which is the reason we applied the cut-off threshold criterion to generate continuous predictions; this maximizes the sum of sensitivity and specificity, since it has been shown to be useful in modeling methods that employ only presence data (Manzanilla-Quiñones et al. 2019). To denote differences between algorithms we performed the non-parametric Kruskall-Wallis test on the AUC values. Then we grouped the algorithms using the Fisher's least significant

Tab. 3 - Description of climatic, edaphic and topographic variables employed to build the spatial distribution models of 17 species of pines.

Category	Variable	Key
Climatic	Annual mean temperature (°C)	bio1
	Mean diurnal range (°C)	bio 2
	Maximum temperature of warmest month (°C)	bio 5
	Minimum temperature of coldest month (°C)	bio 6
	Mean temperature of warmest quarter (°C)	bio10
	Mean temperature of coldest quarter (°C)	bio11
	Annual precipitation (mm)	bio12
	Precipitation of wettest month (mm)	bio13
	Precipitation of driest month (mm)	bio14
	Precipitation of wettest quarter (mm)	bio16
	Precipitation of driest quarter (mm)	bio17
Soil proprieties	Sand (g kg^{-1})	sa
	Cation exchange capacity (mmol(c) kg^{-1})	cec
	Clay content (g kg^{-1})	clc
	Organic carbon soil (dg kg^{-1})	ocs
	Bulk density (cg cm^{-3})	bd
	Organic carbon density (g dm^{-3})	ocd
	Silt (g kg^{-1})	sil
	Nitrogen (cg kg^{-1})	nit
	pH water (pH·100)	pH
	Soil organic carbon stock (t ha^{-1})	socs
Topographic	Altitude (m a.s.l.)	alt
	Slope (°)	slo
	Diurnal anisotropic heat	dah
	Convergence index	ci
	Terrain ruggedness index (m)	tri
	Topographic wetness index	twi

Tab. 4 - Important and uncorrelated variables used in the spatial distribution models of 17 pine species in Mexico. (bio1): annual mean temperature (°C); (bio2): mean diurnal range (°C); (bio5): maximun temperature of warmest month (°C); (bio6): minimun temperature of coldest month (°C); (bio10): mean temperature of warmest quarter (°C); (bio11): mean temperature of coldest quarter (°C); (bio12): annual precipitation (mm); (bio13): precipitation of wettest month (mm); (bio14): precipitation of driest month (mm); (bio16): precipitation of wettest quarter (mm); (bio17): precipitation of driest quarter (mm); (sa): sand (g kg^{-1}); (cec): cation exchange capacity (mmol (c) kg^{-1}); (clc): clay content (g kg^{-1}); (ocs): organic carbon soil (dg kg^{-1}); (bd): bulk density (cg cm^{-3}); (ocd): organic carbon density (g dm^{-3}); (sil): silt (g kg^{-1}); (nit): nitrogen (cg kg^{-1}); (pH): pH water (pH·100); (socs): soil organic carbon stock (t ha^{-1}); (alt): altitude (m); (slo): slope (°); (dah): diurnal anisotropic heat; (ci): convergence index; (tri): terrain ruggedness index (m); (twi): topographic wetness index.

Variable	*P. arizonica*	*P. ayacahuite*	*P. cembroides*	*P. devoniana*	*P. douglasiana*	*P. durangensis*	*P. hartwegii*	*P. herrerae*	*P. leiophylla*	*P. maximinio*	*P. montezumae*	*P. oocarpa*	*P. patula*	*P. pseudo strobus*	*P. strobi formis*	*P. strobus chiapensis*	*P. teocate*
bio1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
bio2	-	x	x	x	x	x	x	x	-	x	x	-	x	x	x	x	x
bio5	-	x	x	x	-	x	x	x	-	x	x	x	x	x	-	x	x
bio6	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-
bio10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
bio11	-	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	x
bio12	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-
bio13	x	x	-	x	x	x	-	x	-	x	-	x	x	-	x	-	x
bio14	x	x	x	-	-	x	x	-	-	x	-	x	x	x	-	-	-
bio16	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	x	-
bio17	-	-	-	x	x	-	-	x	-	-	-	-	-	-	x	-	-
sa	-	x	-	x	-	-	-	-	-	-	-	x	-	-	-	x	-
cec	-	-	x	x	-	-	x	-	x	-	x	x	-	x	-	x	x
clc	-	x	x	-	-	-	-	-	-	-	-	x	-	-	-	x	-
ocs	-	-	-	-	-	x	-	x	-	-	-	-	-	x	-	-	x
bd	x	x	-	x	-	-	x	-	-	x	x	-	x	x	x	x	-
ocd	-	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
sil	-	-	-	x	-	-	x	-	-	x	x	x	x	-	-	x	-
nit	-	-	-	x	x	-	-	x	-	x	-	-	-	-	-	x	x
pH	-	-	x	x	x	x	x	x	x	-	-	-	-	-	-	-	x
socs	x	x	-	-	-	-	x	-	-	-	x	-	x	-	x	x	-
alt	x	-	-	-	x	-	-	-	x	-	-	-	-	-	x	-	-
slo	x	x	-	x	-	-	x	-	-	-	-	-	-	-	x	-	-
dah	-	-	-	x	x	-	-	-	-	x	x	x	-	-	-	-	-
ci	-	-	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-
tri	-	-	x	-	-	-	-	-	-	-	-	x	x	x	-	-	x
twi	-	-	-	-	x	-	x	x	x	-	-	-	-	x	-	-	-

difference criterion after correcting the p-value through the Bonferroni method. To compare the predictive stability of the algorithms including the ensemble model, we followed the methodology reported by Ren-Yan et al. (2014). For each species and each algorithm, we calculated the coefficient of variation (CV = $\sigma_x/\bar{x}$) for the TSS statistic values obtained from the modeling process through bootstrap resampling with 10 repetitions. A scatter plot was generated to show the results, including the mean and the standard error of the CV (*y*-axis) and TSS (*x*-axis) of the 17 species. The predictions of the best algorithm found for each species were applied to the cutoff threshold that maximizes the sum of sensitivity and specificity, and were projected on binary maps (suitable-not suitable).

Software

The pre-modeling and modeling processes were carried out using the "sdm" package (Naimi & Araújo 2016), whereas for secondary information processes we used "raster" packages (Hijmans et al. 2020), "ntbox" (Osorio-Olvera et al. 2020), "rgdal" (Bivand et al. 2020), all of which were implemented in the R software (R Core Team 2020).

Results

Presence records and relevant environmental variables

The models were fitted with different number of presence records. In this regard, *P.sc* and *P.le* were the species with the lowest (117) and highest (2461) number of records; the models for the rest of the species were adjusted with a number of records that ranged between 234 and 2182 (Tab. 2). The premodeling and VIF analyses identified the important variables (Tab. S1 in Supplementary material) and with low collinearity (Tab. S2) to fit the models of the 17 species of pines. The number of variables used for the final models fluctuated between 6 variables for *P.ar* and up to 12 variables for *P.de*. Variables bio2 and bio5 were included as predictors in a large number of the SDM (Tab. 4). In contrast, variables bio6, bio12, bio16 and socs were included in very few of the generated models.

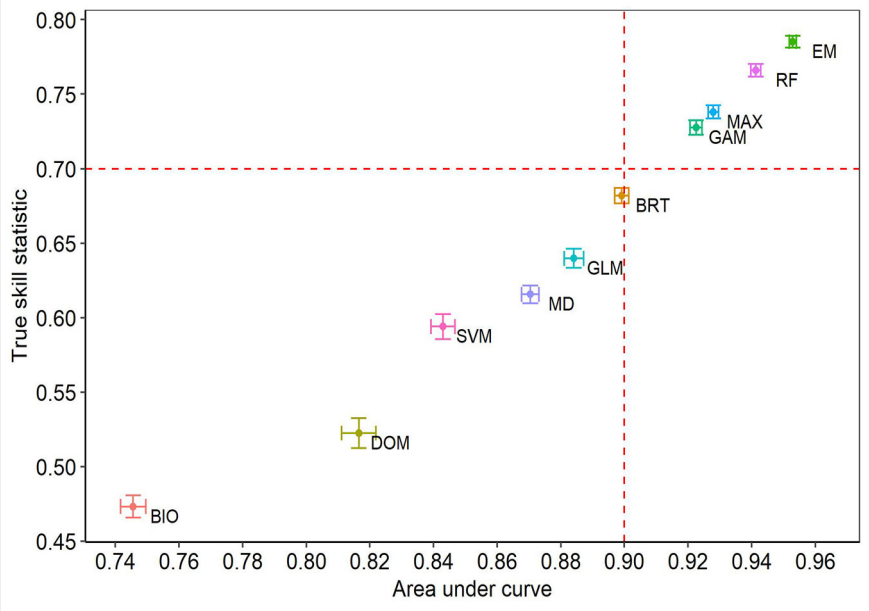

Fig. 1 - True skill statistic (TSS) and area under the curve (AUC) of nine algorithms and ensemble model for predicting the geographic range of 17 pine species. (BIO): Bioclim; (DOM): Domain; (SVM): Support Vector Machine; (MD): Mahalanobis distance; (GLM): Generalized linear models; (BRT): Boosted regression trees; (GAM): Generalized additive models; (MAX): MaxEnt; (RF): Random forests; (EM): Ensemble model.

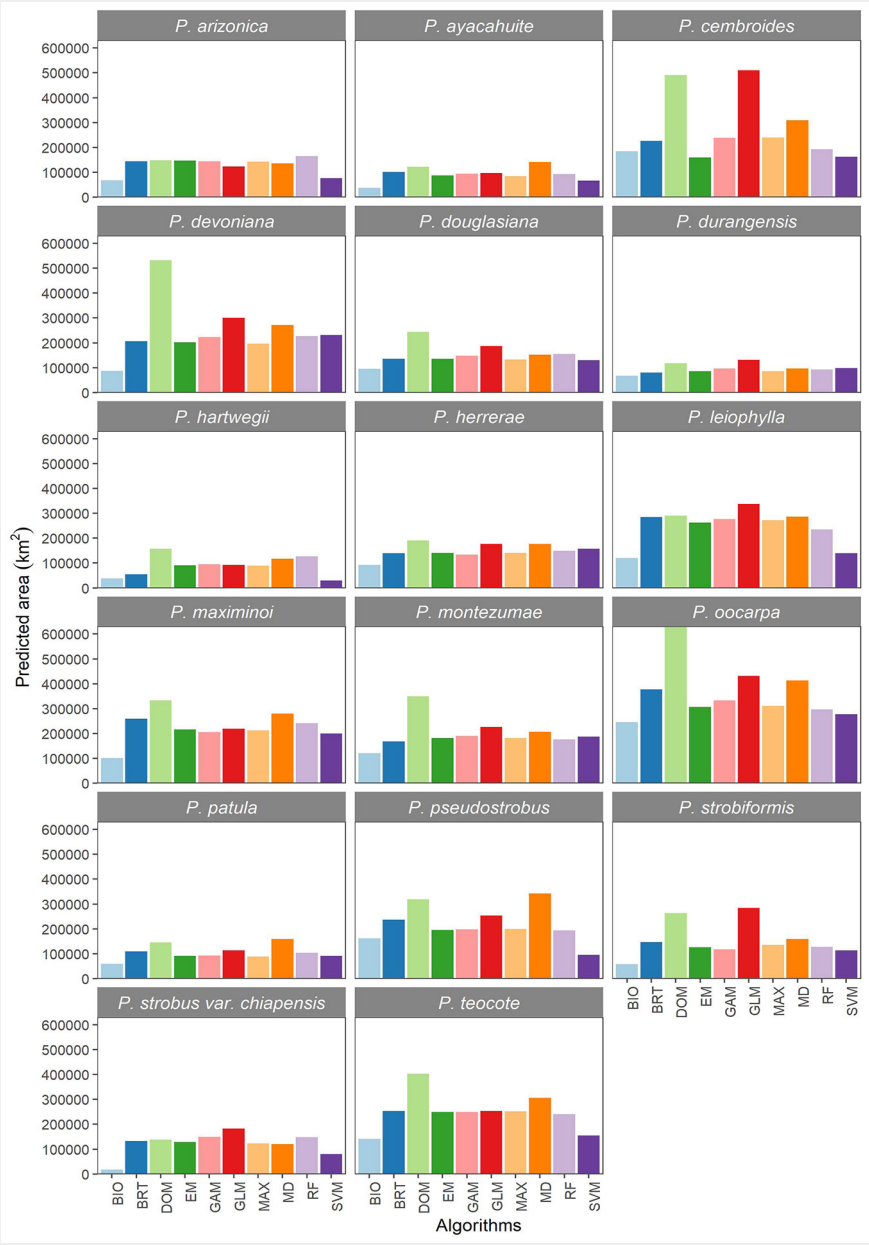

Fig. 2 - Distribution area predicted by nine algorithms and ensemble model for the modeling of 17 pine species in Mexico. (Bio): Bioclim; (BRT): Boosted regression trees; (EM): Ensemble model; (MD): Mahalanobis distance; (DOM): Domain; (GAM): Generalized additive models; (GLM): Generalized linear models; (MAX): MaxEnt; (RF): Random forests; (SVM): support vector machine.

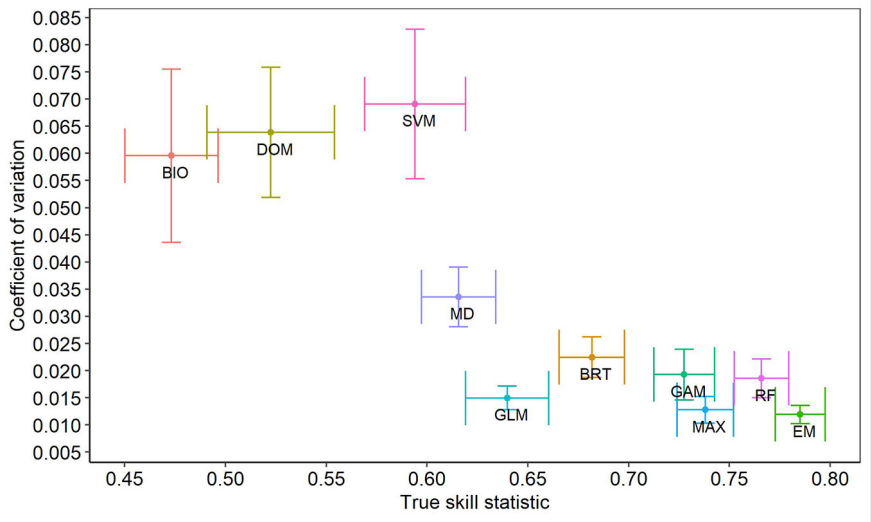

Fig. 3 - Coefficient of variation and TSS of nine algorithms and ensemble model for the modeling of the geographic distribution of 17 pine species in Mexico. (Bio): Bioclim; (DOM): Domain; (SVM): support vector machine; (MD): Mahalanobis distance; (GLM): Generalized linear models; (BRT): Boosted regression trees; (GAM): Generalized additive models; (MAX): MaxEnt; (RF): Random forests; (EM): Ensemble model.

Predictive performance of the algorithms

The evaluating statistics of the predictive performance of the algorithms as well as the ensemble model obtained different values. Six of the analyzed algorithms generated low values in the AUC (<0.90) and TSS (<0.70) statistics with respect to the rest of the algorithms (Fig. 1). GAM, MAX, RF and the ensemble model attained the highest values in AUC and TSS (>0.90 and >0.70, respectively). The algorithms with the lowest predictive performance were BIO and DOM, which obtained low values in AUC (<0.75 and <0.82) and TSS (<0.5 and <0.55). The SVM, DMAH, GLM and BRT algorithms obtained intermediate values for both statistics. The non-parametric Kruskall-Wallis test denoted that the performance between the algorithms was statistically different; the *post-hoc* test allowed us to rank the performance of the algorithms in descending order, and the assembled model obtained the best performance (Fig. S3a in Supplementary material). In contrast, BIO obtained the lowest performance (Fig. S3h), while MAX and GAM obtained a similar performance (Fig. S3c).

Predictions and stability of the algorithms

The nine algorithms and the ensemble model differed in predicting spatial ranges for all species. For most species DOM predicted a larger distribution area (Fig. 2). Conversely, BIO and SVM predicted lesser areas of distribution for all species. The rest of the algorithms projected surfaces of different magnitude with a relatively lower difference than the previously mentioned algorithms. The algorithms with the highest performance (GAM, MAX, RF and the ensemble model) predicted similar areas (Fig. 2). According to the performance, the ensemble model better described the distribution of the pine species (Fig. S1 in Supplementary material).

The coefficient of variation indicated that BIO, DOM and SVM have the lower stability in predictions (CV >0.055); also, their error bars are large, indicating that CV values differ widely between species (Fig. 3). EM and MAX obtained smaller variation (CV <0.015) and the standard error was low. MD, GLM, BRT, GAM, and RF were the algorithms that obtained a mean coefficient of variation in the TSS statistic (CV ≥0.015 and CV <0.040 – Fig. 3).

Discussion

Environmental variables and final models

The SDMs with their correlative algorithms are the result of the projection of an ecological niche model created in the environmental space with different variables (Soberón et al. 2017). It is clear that the selection of variables is an important factor in SDM and a set of important and uncorrelated variables can increase the predictive power and reduce the complexity of the model (Cobos et al. 2019). Watling et al. (2012) found that selecting uncorrelated variables with biological significance did not affect the predictive performance between the models; however, they did not find the same in spatial predictions, as using the RF algorithm with uncorrelated predictors, more stable predictions were obtained. In the same study, GLM provided more unstable predictions, which indicates that the algorithms are sensitive to the selection of environmental variables that affect the stability.

In our study, final models for all species used different combinations of variables, although it was observed that bio2 and bio5 were important variables for many species. This suggests that the distribution of the pine species considered in the present study is explained by variables related to temperature, which coincides with Ramos-Dorantes et al. (2017), who reported temperature as one of the most important variables for the distribution of seven pine species in Mexico. In another study by Aceves-Rangel et al. (2018), altitude was the most important variable in the models of 11 species of the same genus in Mexico. Due to the fact that altitude has a high degree of correlation (r=0.9) with temperature in an indirect manner, it can be inferred that it is a key variable in the distribution of pine species. The two formerly cited studies and our research obtained similar results due to the fact that the pine species of Mexico are distributed in temperate habitats where temperature is an important ecological factor (Farjon & Filer 2013). In this context, Perry (1991) indicated that the pine species distributed in Mexico and part of Central America were strongly influenced by climatic fluctuations during the Paleogene period. This historical fact indicates that temperature and precipitation are key factors in the current distribution of pine trees and that the variables selected in our models were adequate. It is also worthwhile noting that for all the species analyzed, variables bio1, bio10 and bio15 were not considered in the model because they have a high collinearity with other variables.

Predictive performance of algorithms

Testing several algorithms to perform SDM helps to have a broader view of the advantages and disadvantages of the use of each single algorithm (Jarnevich & Young 2019). Numerous studies on SDM of Mexican pines (Aceves-Rangel et al. 2018, García-Aranda et al. 2018, Reynoso Santos et al. 2018, Manzanilla-Quiñones et al. 2019) only used the MaxEnt algorithm, though the prediction of the distribution areas can be improved by employing an ensemble or an RF model. In this study, we found that all the analyzed algorithms obtained better predictions than those obtained by chance (Fig. 1). Nevertheless, it was observed that RF and the ensemble model were superior than the rest of the algorithms. In other studies (Marmion et al. 2009, Ren-Yan et al. 2014, Pecchi et al. 2020), RF presented a superior predictive performance with respect to other algorithms, which is consistent with our results. However, it is important to mention that RF does not always attains the best prediction performance (Shabani et al. 2016, Marchi & Ducci 2018). Discrepancies found between these studies can be explained by the different configurations of the algorithm (Mi et al. 2017). On the other hand, the ensemble model achieved the highest predictive performance (TSS >0.78 and AUC >0.95), although for *P. ooc*

and *P. ps* it obtained a lower predictive performance than RF (Fig. S2 in Supplementary material). These findings are consistent with Hao et al. (2020), who found that the ensemble model did not always obtain a higher predictive performance than the individual models.

Regarding the remaining evaluated algorithms, it was observed that SVM, MD, GLM and BRT obtained a medium predictive performance. BIO and DOM were the algorithms with the lowest predictive performance. These results are similar to those of Ren-Yan et al. (2014) and Pecchi et al. (2020), who classified algorithms according to their predictive performance (high and low). Nonetheless, we consider that this may be ambiguous as those findings could not represent the potential of the algorithms, since under certain conditions (predictors, size and sample quality) a method may or may not be effective (Peterson et al. 2012, Hijmans & Elith 2017). As explained by Jarnevich & Young (2019), the evaluation of algorithms is a practice that should be done in each case study and, depending on the obtained results and the objectives of the research, the most appropriate one should be selected in order to make the necessary inferences, since under certain conditions one of them can be better or worse than the others.

Prediction and variability of the algorithms

The predicted distribution areas for each of the 17 analyzed pine species (Fig. S1 in Supplementary material) coincided with the distribution reported by Farjon & Filer (2013). Nonetheless, they differed for *P.ce, P.do, P.du, P.ha* and *P.st* with respect to the distribution described by Perry (1991), since he identified more restricted distributions than those mentioned by Farjon & Filer (2013). The distribution of the remaining species (*P.ar, P.ay, P.de, P.he, P.le, P.ma, P.mo, P.ooc, P.pa, P.ps, P.sc* and *P.te*) was similar.

The different algorithms showed significant differences in the predicted spatial distribution areas (Fig. 2). These results are due to the fact that the correlative models depend to a great extent on the chosen modeling algorithms (Araújo & New 2007), since each one starts from different assumptions for its construction. Therefore, it can be said that the prediction differences are inherent to the models (Peterson et al. 2012, Hijmans & Elith 2017). Because of the above, it will always be difficult to choose *a priori* the best algorithm to model the distribution of species. However, the results of this study can be helpful when deciding which is the best algorithm to use (Jarnevich & Young 2019).

The results indicated that three of the tested algorithms have a high prediction variability, except for EM and MAX (Fig. 3). The ensemble model had a low variation and high performance compared to the other algorithms, which provides an advantage in the prediction of the distribution of species. If species conservation problems or predictions are to be addressed under climate change scenarios, it is important to have a smaller variation caused by the modeling algorithms, since more robust inferences can be made with less variation (Hao et al. 2020). In several studies, ensemble models have been used as an alternative to reduce variability between the different algorithms, and they have even been proposed as promising techniques for species distribution modeling (Araújo & New 2007, Marmion et al. 2009, Hao et al. 2019). Yet another way to minimize variability problems between algorithms is to repeatedly evaluate and compare multiple algorithms and, based on the obtained results, select the most adequate for modeling (Jarnevich & Young 2019).

Conclusions

The ensemble model showed the highest predictive performance in modeling the spatial distribution of 17 pine species in Mexico, although RF, MAX and GAM also provided good predictions. The rest of the applied algorithms (BRT, GLM, MD, SVM, DOM and BIO) presented a lower accuracy; BIO, DOM and SVM were the algorithms with the greatest variability in predictions and, on the other hand, MAX and EM obtained the smallest variation. The rest of the algorithms attained a medium variability in the prediction of the distribution areas. For most of the species (16), the ensemble model showed the best predictive performance, although for *P. ooc* the most accurate predictions were obtained using RF.

The assessment of algorithms' performance for predicting the spatial species distribution is an important step in the modeling process and must be carried out carefully, since the selection of one algorithm over another could lead to different results and therefore to different conclusions. Because the predictive differences between the algorithms are relatively large, the choice of one over the other should be based on the study objectives. The results derived from this research suggest that it is not convenient to choose an algorithm *a priori*, but rather to carry out tests among the available algorithms in order to increase the confidence in the prediction performance and stability of SDM.

Acknowledgements

The first author thanks CONACYT (*Consejo Nacional de Ciencia y Tecnología*, México, México) for the scholarship granted to carry out a doctorate in forest sciences. We also thank CONAFOR (*Comisión Nacional Forestal*) for providing data from the National Forest and Soil Inventory.

References

Aceves-Rangel LD, Méndez-González J, García-Aranda MA, Nájera-Luna JA (2018). Distribución potencial de 20 especies de pinos en México [Potential distribution of 20 pine species in Mexico]. Agrociencia 52: 1043-1057. [in Spanish]

Araújo MB, New M (2007). Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22: 42-47. - doi: 10.1016/j.tree.2006.09.010

Betancourt GA (2005). Las máquinas de soporte vectorial (SVMs) [Vector support machines (SVMs)]. Scientia et Technica 11 (27): 67-72. [in Spanish] [online] URL: http://www.researchgate.net/publication/49588125

Bivand R, Keitt T, Rowlingson B, Pebesma E, Sumner M, Hijmans R, Rouault E, Warmerdam F, Ooms J, Rundel C (2020). Package "rgdal". R package version 1:5-18. [online] URL: http://cran.r-project.org/package=rgdal

Carpenter G, Gillison AN, Winter J (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation 2: 667-680. - doi: 10.1007/BF00051966

Chapman AD (2005). Principles and methods of data cleaning - Primary species and species-occurrence data (version 1.0). Report for the Global Biodiversity Information Facility - GBIF, Copenhagen, Denmark, pp. 72.

Cobos ME, Peterson AT, Osorio-Olvera L, Jiménez-García D (2019). An exhaustive analysis of heuristic methods for variable selection in ecological niche modeling and species distribution modeling. Ecological Informatics 53: 100983. - doi: 10.1016/j.ecoinf.2019.100983

CONABIO (2020). Sistema Nacional de Información sobre Biodiversidad. Registros de ejemplares [National Information System on Biodiversity. Specimen Records]. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), Ciudad de México, México, web site. [in Spanish] [online] URL: http://www.snib.mx/ejemplares/descarga/

CONAFOR (2019). Estado que guarda el sector forestal en México [Status of the Forest Sector in Mexico]. Comisión Nacional Forestal (CONAFOR), Zapopan Jalisco, México, pp. 406. [in Spanish]. [online] URL: http://www.conafor.gob.mx:8080/documentos/docs/1/7743Estado

Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Böhner J (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific Model Development 8: 1991-2007. - doi: 10.5194/gmd-8-1991-2015

Elith J, H. Graham C, P. Anderson R, Dudík M, Ferrier S, Guisan A, J. Hijmans R, Huettmann F, R. Leathwick J, Lehmann A, Li J, G. Lohmann L, A. Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, Mcc. M. Overton J, Townsend Peterson A, J. Phillips S, Richardson K, Scachetti-Pereira R, E. Schapire R, Soberón J, Williams S, S. Wisz M, E. Zimmermann N (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography 29 (2): 129-151. - doi: 10.1111/j.2006.0906-7590.04596.x

Escobar LE, Lira-Noriega A, Medina-Vogel G, Peterson AT (2014). Potential for spread of the white-nose fungus (*Pseudogymnoascus destructans*) in the Americas: use of Maxent and NicheA to assure strict model transference. Geospatial Health 9: 221-229. - doi: 10.4081/gh.2014.19

Etherington TR (2019). Mahalanobis distances and ecological niche modelling: correcting a chi-squared probability error. PeerJ 7: 1-8. - doi: 10.7717/peerj.6678
Farjon A, Filer D (2013). An atlas of the world's conifers. Brill, Leiden, The Netherlands, pp. 512. - doi: 10.1163/9789004211810
Fick SE, Hijmans RJ (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37: 4302-4315. - doi: 10.1002/joc.5086
García-Aranda MA, Méndez-González J, Hernández-Arizmendi JY (2018). Distribución potencial de *Pinus cembroides*, *Pinus nelsonii* y *Pinus culminicola* en el Noreste de México [Potential distribution of *Pinus cembroides*, *Pinus nelsonii* and *Pinus culminicola* in northeastern Mexico]. Ecosistemas y Recursos Agropecuarios 5: 3-13. [in Spanish] - doi: 10.19136/era.a5n13.1396
GBIF (2020). Occurrence download. Global Biodiversity Information Facility, Copenhagen, Denmark, web site. [online] URL: http://data.gbif.org
Guisan A, Edwards TC, Hastie T (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157: 89-100. - doi: 10.1016/S0304-3800(02)00204-1
Hansen MC, Defries RS, Townshend JRG, Sohlberg R (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. International Journal of Remote Sensing 21: 1331-1364. - doi: 10.1080/014311600210209
Hao T, Elith J, Guillera-Arroita G, Lahoz-Monfort JJ (2019). A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Diversity and Distributions 25: 839-852. - doi: 10.1111/ddi.12892
Hao T, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. Ecography 43: 549-558. - doi: 10.1111/ecog.04890
Hengl T, Mendes De Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, Guevara MA, Vargas R, MacMillan RA, Batjes NH, Leenaars JGB, Ribeiro E, Wheeler I, Mantel S, Kempen B (2017). SoilGrids250m: global gridded soil information based on machine learning. PLoS One 12: 1-40. - doi: 10.1371/journal.pone.0169748
Hijmans RJ, Guarino L, Bussink C, Mathur P, Cruz M, Barrantes I, Rojas E (2012). DIVA-GIS, ver. 7.5. A geographic information system for the analysis of species distribution data. Versão 7: 476-486.
Hijmans RJ, Elith J (2017). Species distribution modeling with R. Web site. [online] URL: http://rspatial.org/raster/sdm/
Hijmans RJ, Etten Van J, Sumner M, Cheng J, Bevan A, Bevan R, Busetto L, Canty M, Forrest D, Ghosh A, Golicher D, Gray J, Greenberg JA (2020). Package "raster". R Package version 3:3-13. [online] URL: http://cran.r-project.org/web/packages/raster/raster.pdf
INE/INEGI (1997). Conjunto nacional de uso del suelo y vegetación a escala 1:250.000, Serie I [National collection of land use and vegetation scale 1:250.000, Series I]. DOEG-INEGI, México City, Mexico. [in Spanish]
INEGI (2016). Conjunto nacional de uso del suelo y vegetación a escala 1:250.000, Serie VI [National collection of land use and vegetation scale 1:250.000, Series VI]. Instituto Nacional de Estadística y Geografía (INEGI), México City, Mexico. [in Spanish]
Jarnevich CS, Young NE (2019). Not so normal normals: species distribution model results are sensitive to choice of climate normals and model type. Climate 7: 1-15. - doi: 10.3390/cli7030003
Manzanilla-Quiñones U, Aguirre-Calderón OA, Jiménez-Pérez J, Treviño-Garza EJ, Yerena-Yamallel JI (2019). Distribución actual y futura del bosque subalpino de *Pinus hartwegii* Lindl en el Eje Neovolcánico Transversal [Current and future distribution of the *Pinus hartwegii* Lindl subalpine forest in the Tranverse Neovolcanic Belt]. Madera y Bosques 25: 1-16. [in Spanish] - doi: 10.21829/myb.2019.2521804
Marchi M, Ducci F (2018). Some refinements on species distribution models using tree-level national forest inventories for supporting forest management and marginal forest population detection. iForest 11: 291-299. - doi: 10.3832/ifor2441-011
Marmion M, Parviainen M, Luoto M, Heikkinen RK, Thuiller W (2009). Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions 15: 59-69. - doi: 10.1111/j.1472-4642.2008.00491.x
Mi C, Huettmann F, Guo Y, Han X, Wen L (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. PeerJ 5: 1-22. - doi: 10.7717/peerj.2849
Naimi B, Araújo MB (2016). SDM: a reproducible and extensible R platform for species distribution modelling. Ecography 39: 368-375. - doi: 10.1111/ecog.01881
Nix HA (1986). A biogeographic analysis of Australian elapid snakes. In: "Atlas of elapid snakes of Australia" (Longmore R ed). Australian Flora and Fauna Series 7, Bureau of Flora and Fauna, Canberra, Australia, pp. 4-15.
Osorio-Olvera L, Lira-Noriega A, Soberón J, Peterson AT, Falconi M, Contreras-Díaz RG, Martínez-Meyer E, Barve V, Barve N (2020). ntbox: an R package with graphical user interface for modelling and evaluating multidimensional ecological niches. Methods in Ecology and Evolution 11: 1199-1206. - doi: 10.1111/2041-210X.13452
Pecchi M, Marchi M, Moriondo M, Forzieri G, Ammoniaci M, Bernetti I, Bindi M, Chirici G (2020). Potential impact of climate change on the spatial distribution of key forest tree species in Italy under RCP4.5 for 2050s. Forest 11: 1-19. - doi: 10.3390/f11090934
Pecchi M, Marchi M, Burton V, Giannetti F, Moriondo M, Bernetti I, Bindi M, Chirici G (2019). Species distribution modelling to support forest management. A literature review. Ecological Modelling 411: 108817. - doi: 10.1016/j.ecolmodel.2019.108817
Perry JP (1991). The pines of Mexico and Central America. Timber Press, Portland, OR, USA, pp. 231. [online] URL: http://www.cabdirect.org/cabdirect/abstract/19910653262
Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, Araújo MB (2012). Ecological niches and geographic distributions (MPB-49). Series "Monographs in Population Biology", vol. 49, Princeton University Press, Princeton, NJ, USA, pp. 316. - doi: 10.1515/9781400840670
Phillips SJ, Anderson RP, Schapire RE (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling 190: 231-259. - doi: 10.1016/j.ecolmodel.2005.03.026
R Core Team (2020). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [online] URL: http://www.R-project.org/
Radosavljevic A, Anderson RP (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. Journal of Biogeography 41: 629-643. - doi: 10.1111/jbi.12227
Ramos-Dorantes DB, Villaseñor JL, Ortiz E, Gernandt DS (2017). Biodiversity, distribution, and conservation status of Pinaceae in Puebla, Mexico. Revista Mexicana de Biodiversidad 88: 215-223. - doi: 10.1016/j.rmb.2017.01.028
Ren-Yan D, Xiao-Quan K, Min-Yi H, Wei-Yi F, Zhi-Gao W (2014). The predictive performance and stability of six species distribution models. PLoS One 9. - doi: 10.1371/journal.pone.0112764
Reynoso Santos R, Pérez Hernández MJ, López Baez W, Hernández Ramos J, Muñoz Flores HJ, Cob Uicab JV, Reynoso Santos MD (2018). El nicho ecológico como herramienta para predecir áreas potenciales de dos especies de pino [The ecological niche as a tool for predicting potential areas of two pine species]. Revista Mexicana de Ciencias Forestales 9: 47-68. [in Spanish] - doi: 10.29298/rmcf.v8i48.114
Shabani F, Kumar L, Ahmadi M (2016). A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. Ecology and Evolution 6: 5973-5986. - doi: 10.1002/ece3.2332
Soberón J, Osorio-Olvera L, Peterson T (2017). Diferencias conceptuales entre modelación de nichos y modelación de áreas de distribución [Conceptual differences between ecological niche modeling and species distribution modeling]. Revista Mexicana de Biodiversidad 88: 437-441. [in Spanish] - doi: 10.1016/j.rmb.2017.03.011
Vargas-Larreta B, Corral-Rivas JJ, Aguirre-Calderón OA, López-Martínez JO, De los Santos-Posadas HM, Zamudio-Sánchez FJ, Treviño-Garza EJ, Martínez-Salvador M, Aguirre-Calderón CG (2017). SiBiFor: forest biometric system for forest management in Mexico. Revista Chapingo Serie Ciencias Forestales y Del Ambiente 23: 437-455. - doi: 10.5154/r.rchscfa.2017.06.040
Velazco SJE, Galvão F, Villalobos F, De Marco Júnior P (2017). Using worldwide edaphic data to model plant species niches: An assessment at a continental extent. PLoS One 12: 1-24. - doi: 10.1371/journal.pone.0186025
Watling JI, Romañach SS, Bucklin DN, Speroterra C, Brandt LA, Pearlstine LG, Mazzotti FJ (2012). Do bioclimate variables improve performance of climate envelope models? Ecological Modelling 246: 79-85. - doi: 10.1016/j.ecolmodel.2012.07.018

Supplementary Material

Fig. S1 - Binary prediction of the ensemble model for 17 species of pines.

Fig. S2 - True statistical skill and AUC of nine algorithms and ensemble model for modeling 17 species of pines.

Fig. S3 - Kruskall-Wallis non-parametric test and classification of algorithms.

Tab. S1 - Relative importance of the variables (auc_test) obtained in the pre-modelling process of 17 pine species.

Tab. S2 - Variance inflation factor of the variables included in final models of spatial distribution of 17 species of pines.

Link: Montoya_4084@supplo01.pdf