Evaluación de la configuración de tres algoritmos para realizar modelos de distribución potencial de especies forestales

Evaluation of the configuration of three algorithms to model the potential distribution of forest species

Juan Carlos Montoya-Jimenez[1], José René Valdez-Lazalde[2*], Gregorio Ángeles-Pérez[2], Héctor Manuel De los Santos-Posadas[2], Gustavo Cruz-Cárdenas[3]

[1]Tecnológico Nacional de México-TES Valle de Bravo, División de Ingeniería Forestal. Carretera federal Monumento-Valle de Bravo, km 30, Ejido San Antonio de la Laguna, CP. 51200. Valle de Bravo, Estado de México. México.
[2]Colegio de Postgraduados. Postgrado en Ciencias Forestales. Campus Montecillo. km 36.5 Carretera México-Texcoco, Montecillo. CP. 56264. Texcoco, Estado de México. México.
[3]Instituto Politécnico Nacional, CIIDIR-IPN-Michoacán, COFAA, Justo Sierra 28, CP. 59510. Jiquilpan, Michoacán, México.

*Autor de correspondencia: valdez@colpos.mx

Artículo científico

Recibido: 01 de abril 2024
Aceptado: 17 de mayo 2025

RESUMEN. Los modelos de distribución potencial son una herramienta útil para identificar condiciones ambientales óptimas para que un organismo prevalezca. El objetivo del estudio fue evaluar la configuración de parámetros de los algoritmos Maxent, Random Forest y Modelos Aditivos Generalizados (GAM) a través de la generación de modelos de distribución de seis especies forestales de México. Con registros de presencia de seis especies forestales se ajustaron modelos de distribución potencial con tres algoritmos, se usaron dos configuraciones en sus parámetros (afinados y predeterminados), los modelos se evaluaron a través del área bajo la curva, para comparar la configuración afinada y predeterminada se realizaron gráficos de violín, gráficos de valores de idoneidad predicha con ambas configuraciones y un análisis de coincidencia global difusa. Los conjuntos de cuatro, cinco y seis variables mejoraron las predicciones. Los mejores valores en el multiplicador de regularización oscilaron entre 0.1 y 0.4, las clases de características que describen mejor la distribución potencial de las seis especies fueron lineal, cuadrática y producto. Random Forest mostró que con 750 y 1 000 árboles y dos variables en cada división no mejora el ajuste de los modelos para las seis especies. Los mejores valores de suavización del algoritmo GAM oscilaron entre 0.0001 para *P. pseudostrobus* hasta 1.5 para *P. durangensis*, pero no se encontraron diferencias entre el ajuste de modelos con configuración afinada y predeterminada. Los algoritmos obtuvieron buen rendimiento, sin embargo, el efecto del ajuste de los parámetros en la capacidad predictiva fue marginal y varió según el algoritmo.
Palabras clave: Random Forest, Maxent, Modelos aditivos generalizados, afinación, *Pinus*.

ABSTRACT. Potential distribution models are a useful tool to identify optimal environmental conditions for an organism to prevail. The objective of the study was to evaluate the parameter configuration of the Maxent, Random Forest and Generalized Additive Models (GAM) algorithms through the generation of distribution models of six forest species from Mexico. With presence records of six forest species, potential distribution models were adjusted with three algorithms, two configurations were used in their parameters (tuned and default), the models were evaluated through the area under the curve, to compare the tuned and default configuration. violin plots, plots of predicted fitness values with both configurations, and a fuzzy global matching analysis were performed. Sets of four, five, and six variables improved predictions. The best values in the regularization multiplier ranged between 0.1 and 0.4, the feature classes that best describe the potential distribution of the six species were linear, quadratic and product. Random Forest showed that with 750 and 1 000 trees and two variables in each division the fit of the models for the six species does not improve. The best smoothing values for the GAM algorithm ranged from 0.0001 for *P. pseudostrobus* to 1.5 for *P. durangensis*, however, no differences were found between fitting models with fine-tuned and default settings. The algorithms obtained good performance, however, the effect of parameter tuned on the predictive capacity was marginal and varied depending on the algorithm.
Keywords: Random Forest, Maxent, generalized additive model, tuning, *Pinus*.

Como citar: Montoya-Jiménez JC, Valdez-Lazalde JR, Ángeles-Pérez G, De los Santos-Posadas HM, Cruz-Cárdenas G (2025) Evaluación de la configuración de tres algoritmos para realizar modelos de distribución potencial de especies forestales. Ecosistemas y Recursos Agropecuarios 12(2): e4127. DOI: 10.19136/era.a12n2.4127.

INTRODUCCIÓN

Los modelos de distribución potencial de especies (SDM) son una herramienta útil para identificar las condiciones ambientales óptimas para que un organismo prevalezca. A partir de este conocimiento es posible predecir cambios en el espacio y en el tiempo de la distribución de las especies debido a factores como el cambio climático (Pecchi *et al*. 2020). Lo anterior ha hecho que este tipo de modelos se apliquen con diversos objetivos en distintas áreas del conocimiento, resultando en un incremento exponencial en el número de artículos publicados a nivel global en los últimos 10 años (Urbina-Cardona *et al*. 2019).

Existe una variedad amplia de algoritmos para ajustar SDM, lo cual representa en primera instancia una ventaja para los desarrolladores de modelos. Sin embargo, las múltiples opciones algorítmicas traen consigo la necesidad de evaluar el rendimiento de estos (Montoya-Jiménez *et al*. 2022), además de mejorar su capacidad predictiva a través de prácticas novedosas. Los algoritmos más usados en el campo de los SDM tienen parámetros de configuración predeterminados y poco se conoce del efecto de dicha configuración. Sabemos que esos valores desempeñan una función importante en el ajuste de los modelos y deberían seleccionarse cuidadosamente (Hallgren *et al*. 2019), sin embargo, a pesar de los avances tecnológicos actuales, esta práctica puede representar un desafío en su implementación dadas las numerosas opciones posibles y los altos requerimientos computacionales (Vignali *et al*. 2020).

La búsqueda de los mejores valores en los parámetros (afinación) de los algoritmos para ajustar SDM es una práctica novedosa que se realiza con el objetivo de disminuir la complejidad de los modelos y mejorar su capacidad predictiva (Cobos *et al*. 2019a, Hallgren *et al*. 2019, Vignali *et al*. 2020) y transferibilidad a otros espacios y tiempos (Low *et al*. 2020). Estudios recientes han documentado que Random Forest (RF), Maxent (MAX) y los modelos aditivos generalizados (GAM) son algoritmos prometedores para ajustar SDM (Probst *et al*. 2019, Schratz *et al*. 2019, Low *et al*. 2020). RF ha demostrado que sus estadísticos de ajuste son superiores a otros algoritmos (Pecchi *et al*. 2020). Mientras que MAX, es un algoritmo ampliamente utilizado para ajustar SDM debido a su utilidad para explicar las relaciones ambientales-ecológicas con los patrones de distribución de las especies (Urbina-Cardona *et al*. 2019). En contraparte a los algoritmos que se consideran de aprendizaje automático, los GAM son considerados como semi-paramétricos y han resultado ser útiles en los SDM cuando se espera que la relación entre las variables sea de una forma más compleja y no se ajuste fácilmente a los SDM estándar (Hallgren *et al*. 2019).

La utilidad demostrada de los algoritmos RF, MAX y GAM para construir SDM justifica la búsqueda de una configuración óptima en sus parámetros que ayude a dilucidar el efecto que se obtiene realizando dicha práctica. La estrategia a seguir para afinar los algoritmos implica la construcción de varios modelos con múltiples combinaciones en los valores de parámetros, y su evaluación mediante una medida de desempeño que permita identificar la combinación de parámetros que produce el mejor ajuste (Vignali *et al*. 2020). La importancia de evaluar la configuración de los algoritmos también radica en que ésta afecta el área de idoneidad predicha y las transferencias del modelo a otro tiempo sin importar la especie que se modela (Derville *et al*. 2018).

El ajuste de modelos parsimoniosos y optimizados en sus parámetros es de importancia en los SDM. Por lo anterior, el objetivo de este trabajo fue evaluar la configuración de parámetros (afinados y predeterminados) de los algoritmos Maxent, Random Forest y Modelos aditivos generalizados a través de la generación de modelos de distribución de seis especies forestales que tienen una contribución importante en el sector forestal de México.

MATERIALES Y MÉTODOS

Área de estudio y datos

Las especies incluidas en este trabajo se distribuyen en las principales regiones montañosas de México y son endémicas de esos lugares (Figura 1). Estas cadenas montañosas se caracterizan por una gran heterogeneidad ambiental y variabilidad altitudinal (SEMARNAT 2010). Los parámetros de los algoritmos de interés se afinaron implementando modelos de distribución potencial (SDM) para seis especies forestales: *Pinus durangensis* Martínez (P.du), *P. leiophylla* Schltdl. and Cham (P.le), *P. oocarpa* Schiede ex Schltdl. (P.oo), *P. patula* Schltdl. and Cham (P.pa), *P. pseudostrobus* Gordon (P.ps), y *P. teocote* Cham and Schltdl. (P.te). Para ajustar los SDM se utilizaron los registros de presencia de las especies colectados y pre-procesados por Montoya-Jiménez *et al.* (2022) (Figura 1).

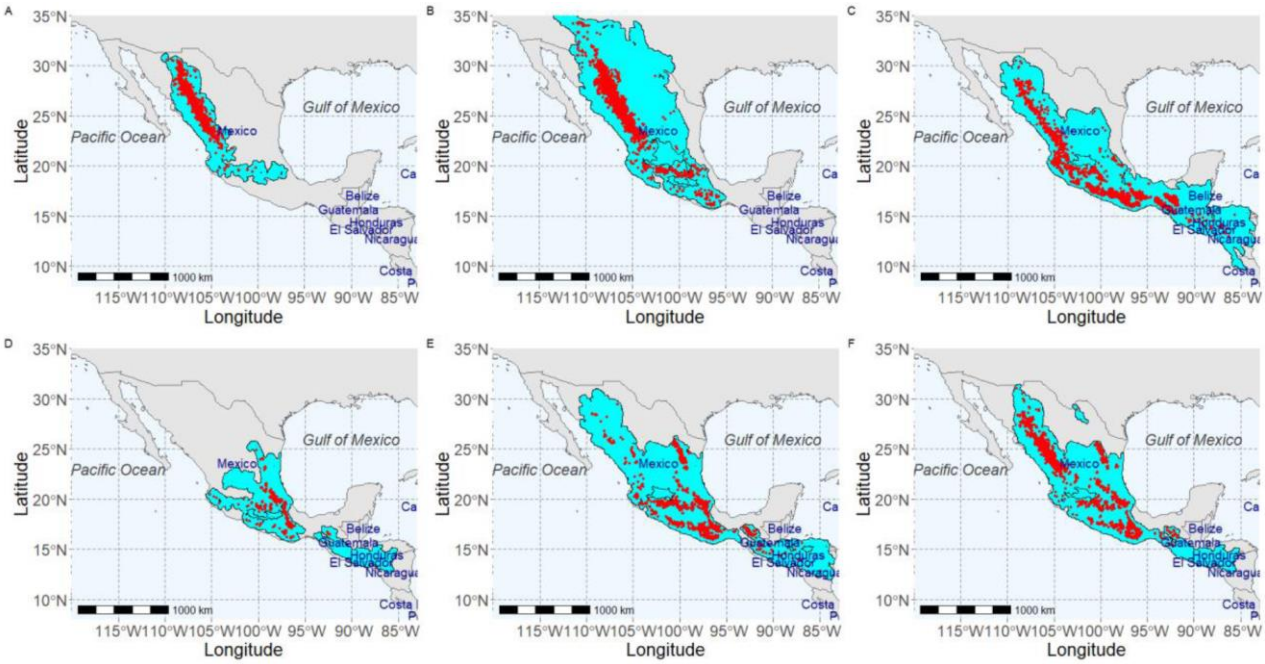

Figura 1. Registros de presencia (rojo) y área de calibración (cyan) para ajustar los modelos de distribución potencial para *P. durangensis* (A), *P. leiophylla* (B), *P. oocarpa* (C), *P. patula* (D), *P. Pseudostrobus* (E) y *P. teocote* (F).

Variables ambientales y análisis preliminar de datos

Los SDM se ajustaron incorporando 11 variables climáticas de WorldClim (Fick y Hijmans 2017), 10 variables de edafología de SoilsGrids (Hengl *et al.* 2017) y seis variables topográficas derivadas de los datos de elevación de WorldClim (Tabla 1). Las áreas accesibles para las especies se construyeron a partir de las ecorregiones del mundo.

Tabla 1. Variables consideradas inicialmente para ajustar los modelos de distribución potencial para *P. durangensis*, *P. leiophylla*, *P. oocarpa*, *P. patula*, *P. pseudostrobus* y *P. teocote*.

Categorías	Variables (unidades)	Clave
Climáticas	Temperatura media anual (°C)	bio1
	Intérvalo de temperaturas diurnas (°C)	bio 2
	Temperatura máxima del mes más cálido (°C)	bio 5
	Temperatura mínima del mes más frío (°C)	bio 6
	Temperatura media del trimestre más cálido (°C)	bio10
	Temperatura media del trimestre más frío (°C)	bio11
	Precipitación anual (mm)	bio12
	Precipitación del mes más lluvioso (mm)	bio13
	Precipitación del mes más seco (mm)	bio14
	Precipitación del trimestre más lluvioso (mm)	bio16
	Precipitación del trimestre más seco (mm)	bio17
Edáficas	Arena (g/kg)	are
	Capacidad de intercambio catiónico (cmmol/kg)	cic
	Contenido de arcilla (g/kg)	coar
	Carbono orgánico del suelo (dg/kg)	cos
	Densidad aparente (cg/cm^3)	deap
	Densidad de carbono orgánico (g/dm^3)	deco
	Llimo (g/kg)	limo
	Nitrógeno (cg/kg)	nit
	pH del agua (pH*100)	ph
	Reserva de carbono orgánico del suelo (t/ha)	rcos
Topográficas	Altitud (m)	alt
	Pendiente (°)	pen
	Calentamiento anisotrópico diurno	cad
	índice de convergencia	ic
	Índice de rugosidad del terreno (m)	irt
	índice topográfico de humedad	ith

Las variables finales que conformaron el SDM para cada una de las especies se eligieron considerando tres criterios: 1) mínima correlación entre los predictores; para esto se realizó un análisis de correlación en el software R (R Core Team 2024) eliminando una de cada par de variables ambientales que mostraron correlación de Pearson, ($r \geq 0.7$. 2) porcentaje preliminar de contribución explicativa; a través de un análisis de pre-modelacón realizado en MAX se seleccionaron las seis variables con mayor capacidad explicativa. 3) combinación de variables con el mejor ajuste en cada modelo (Cobos *et al*. 2019b). Se realizó un análisis exhaustivo para identificar, para cada especie, la mejor combinación de dos o más variables obtenidas de los dos criterios anteriores y con base en los estadísticos Receiver Operating Characteristic (ROC parcial), tasa de omisión y el criterio de información de Akaike corregido para tamaños de muestra pequeños. El último criterio se estimó en el paquete kuenm (Cobos *et al.* 2019a).

Algoritmos

Se utilizaron los algoritmos MAX, RF y GAM para ajustar modelos de distribución potencial (SDM) para cada una de las seis especies forestales de interés. La afinación de los parámetros asociados a los algoritmos se realizó en el software R (R Core Team 2024) a través de los paquetes kuenm (Cobos *et al.* 2019a), SDMtune (Vignali *et al.* 2020) y mgcv (Wood 2021), respectivamente (Tabla 2).

Tabla 2. Valores predeterminados y parámetros explorados para la afinación de los algoritmos Random Forest (RF), Maxent (MAX) y Modelos Aditivos Generalizados (GAM).

Algoritmo (paquete)	Parámetros	Valores predeterminados	Valores explorados
RF (SDMtune)	MTRY	$\sqrt{p}$	2, 3, 4, 5, 6.
	ntree	500	50, 75, 100, 150, 200, 300, 500, 750, 1000.
	node.size	1	1, 2, 3, 4, 5, 10, 15
Maxent (kuenm)	Multiplicador de regularización (RM)	1	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 8, 10
	Combinaciones de características (FC)	l, q, p, t, h, lq, lp, lt, lh, qp, qt, qh, pt, ph, th, lqp, lqt, lqh, lpt, lph, lth, qpt, qph, qth, pth, lqpt, lqph, lqth, lpth, qpth, lqpth.	l, q, p, lq, lp, qp, lqp
GAM (mgcv)	Suavización (Sp)	*	0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 1, 1.5, 2, 10, 100, 1000, 10000

MTRY = número de variables para dividir en cada nodo del árbol; ntree = número de árboles; node.size = número mínimo de observaciones en un nodo terminal; l = lineal, q = cuadrática, p = producto, t = umbral, h = bisagra; * el valor es calculado durante el ajuste del modelo a través de validación cruzada; $\sqrt{p}$ = raíz cuadrada del número de predictores.

Los valores explorados para realizar la afinación de parámetros en MAX fueron todas las combinaciones de 17 valores del parámetro de regularización con siete tipos de respuesta en los modelos (todas las combinaciones de tres clases de características: lineal, cuadrática y producto) (Tabla 2). La mejor configuración de parámetros fue la que obtuvo los valores más bajos en la significancia (p) de la ROC parcial, en la tasa de omisión y el criterio de información de Akaike corregido (AICc) (Cobos *et al.* 2019a).

Los valores explorados para realizar la afinación de RF fueron todas las combinaciones de cinco valores en el parámetro MTRY, con nueve valores en el parámetro ntree y siete valores en el parámetro node.size (Tabla 2). La mejor combinación de parámetros fue la que obtuvo el valor más alto en el estadístico área bajo la curva (AUC) (Vignali *et al.* 2020). Para afinar el algoritmo GAM se exploraron 13 valores en el parámetro sp (suavización), el mejor valor del parámetro fue aquel que obtuvo el valor más alto en el estadístico AUC.

Comparación de algoritmos con configuración afinada y predeterminada.

Las predicciones del mejor SDM afinado (ajustado con la mejor combinación identificada de parámetros) para cada una de las especies de interés, se compararon con las predicciones obtenidas con modelos de configuración predeterminada. Para ambas configuraciones, los modelos se ajustaron utilizando 10 000 pseudo-ausencias de fondo y se evaluaron a través del estadístico AUC calculado a través de validación cruzada con 10 pliegues (división de datos de forma aleatoria en diez conjuntos de aproximadamente el mismo tamaño, nueve se emplean para entrenar el modelo y uno se emplea para la validación). Para facilitar el análisis de resultados se construyeron gráficos de violín a partir de los valores de AUC de los modelos ajustados.

La comparación estadística de los modelos ajustados se complementó con gráficos de dispersión de los valores predichos de idoneidad climática. Los mapas de idoneidad climática (mediana de 10 repeticiones) de ambas configuraciones (predeterminada/afinada) en las seis especies se analizaron a través del sistema de inferencia difusa (SID) (Bodbyl-Roels *et al*. 2011) implementado en el software Conjunto de Comparación de Mapas (MCK) (Visser y De-Nijs 2006). La configuración del SID fue estándar, i.e., el mapa resultado de la configuración predeterminada se consideró el de referencia y el mapa resultado de la configuración afinada se utilizó como el de comparación.

El cálculo de la similitud de dos mapas considera características como: área de intersección, área de desacuerdo y tamaño del polígono, la similitud se resume en un valor denominado coincidencia global difusa el cual obtiene valores que van de 0 (totalmente distinto) hasta 1 (totalmente idéntico). Para facilitar su análisis los valores obtenidos se graficaron por especie y algoritmo.

RESULTADOS

Selección de variables

El análisis de las variables explicativas por el método de correlación por pares permitió identificar, para las seis especies, a las variables alt, bio5, bio6, bio10, bio11, bio13, bio14, bio16, bio17, deco, scos e irt como altamente correlacionadas ($r > 0.7$) (Figura 2). Esta información sirvió de base para acotar el número de variables explicativas para cada especie: 16 variables para *P. durangensis* y *P.*

patula, 17 variables para *P. leiophylla*, *P. oocarpa* y *P. teocote*, y 18 variables para *P. pseudostrobus* las cuales presentaron baja correlación (< 0.7). Para cada una de las especies en estudio la prueba Jackniffe permitió identificar las seis variables con mayor (> 1.5) y menor porcentaje de contribución explicativa (< 1.4) en los modelos (Figura 3). Por otro lado, el análisis exhaustivo de combinaciones posibles de variables explicativas reveló que el conjunto de seis variables tuvo el mejor ajuste en los modelos para *P. durangensis*, *P. occarpa*, *P. patula* y *P. teocote* (Tabla 3). En contraparte para *P. leiophylla* y *P. pseudostrobus* un conjunto de cuatro y cinco variables se ajustaron los mejores modelos para estas especies (Tabla 3).

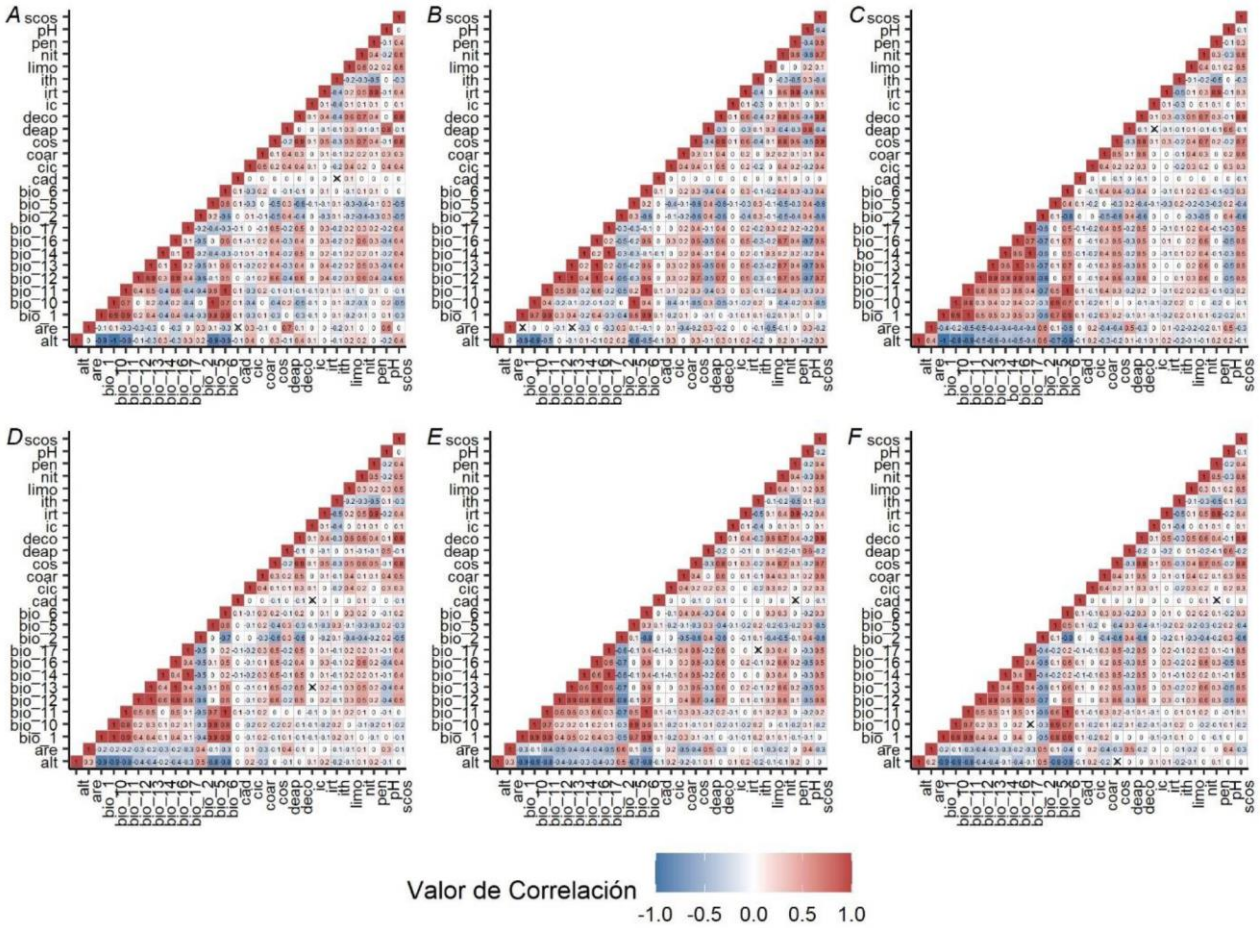

Figura 2. Correlación entre variables ambientales utilizadas para ajustar los modelos de distribución potencial para *P. durangensis* (A), *P. leiophylla* (B), *P. oocarpa* (C), *P. patula* (D), *P. pseudostrobus* (E) y *P. teocote* (F).

Comparación de algoritmos afinados y predeterminados

Maxent

A partir de los estadísticos de evaluación con menor valor (P_pROC, TO y AICc) se encontró que los mejores valores del multiplicador de regularización para las seis especies oscilaron entre 0.1 y 0.4 (Tabla 3), mismos que difirieron del valor predeterminado (1) del algoritmo. Las clases de características que mejor describen la distribución potencial de las seis especies fueron lineal,

cuadrática y producto (Tabla 3), en ningún modelo se incluyeron las características de umbral y bisagra las cuales están predeterminadas.

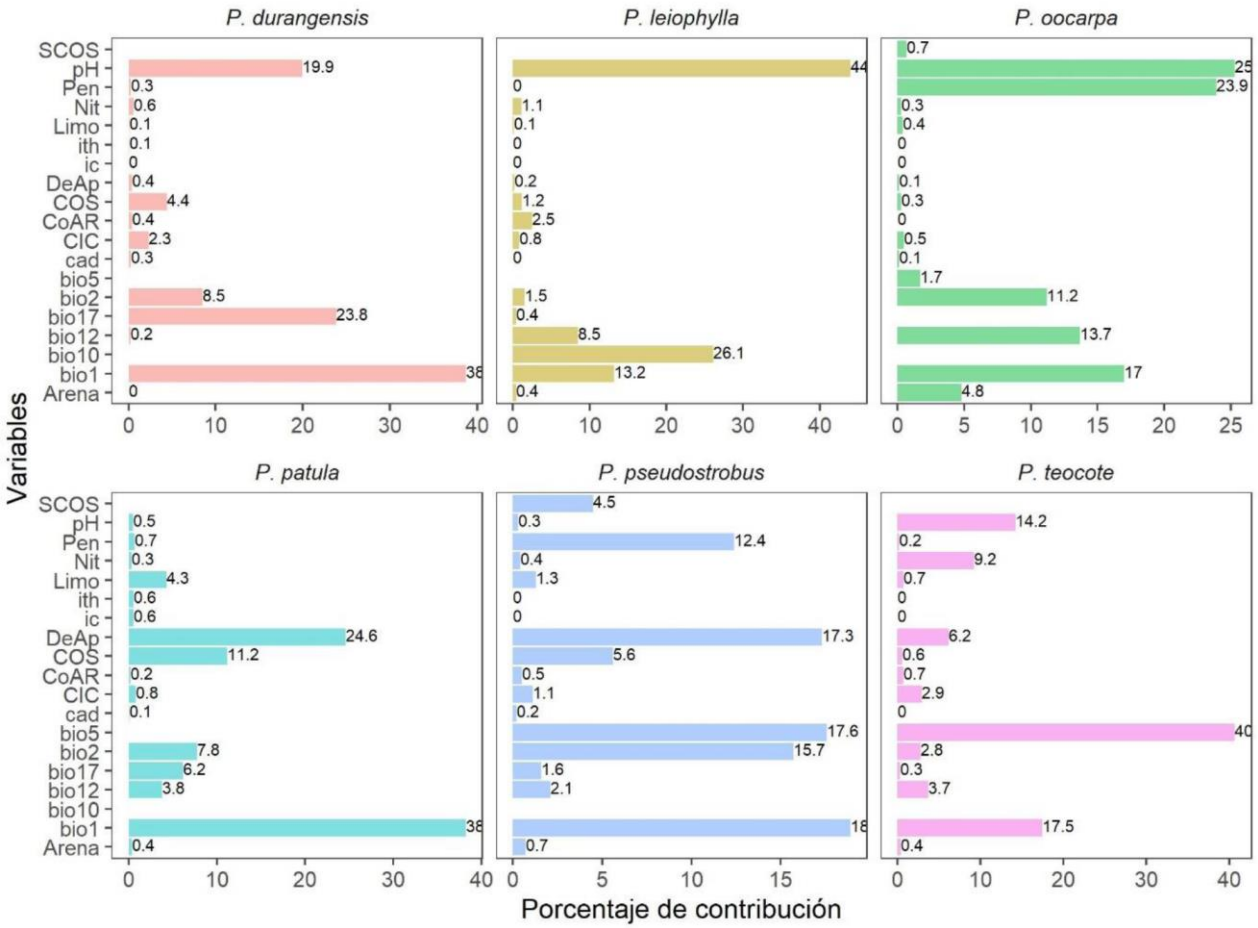

Figura 3. Porcentaje de contribución explicativa de las variables ambientales en los modelos de distribución potencial obtenido a través de la prueba Jaccknife.

Los SDM desarrollados para las seis especies a partir de MAX afinado y predeterminado demostraron tener muy buen ajuste (AUC > 0.85), lo que sugiere que ambos tipos de modelos tienen alta precisión de predicción. El estadístico AUC demostró que el modelo ajustado a partir del algoritmo afinado no difirió del modelo ajustado con la configuración predeterminada ya que ambas configuraciones obtuvieron valores similares en las seis especies forestales (Figura 4).

Aunque en el presente estudio no se encontraron diferencias estadísticas en el ajuste de los modelos MAX con configuración afinada y predeterminada, es importante destacar que los mapas de idoneidad climática de las seis especies difieren entre las configuraciones predeterminada y afinada. Por otro lado, los valores de coincidencia global difusa oscilaron entre 0.69 para *P. durangensis* y 0.53 para *P. pseudostrobus.* Para *P. leiophylla*, *P. oocarpa*, *P. patula* y *P. teocote* los valores fueron de 0.63, 0.57, 0.68 y 0.54, respectivamente.

Tabla 3. Selección de conjunto de variables y estadísticos de ajuste de la mejor combinación de parámetros para el algoritmo Maxent (MAX).

sp	MR	CC	MCV	MAR	P_pROC	TO	AICc	NP	NV
P.du	0.1	Lq	42	3.149	0	0.04907	23793.46	12	6
P.le	0.1	lqp	32	2.482	0	0.04878	42023.00	10	4
P.oo	0.1	lqp	42	1.972	0	0.04885	37567.29	24	6
P.pa	0.2	lqp	42	3.247	0	0.04861	7344.25	21	6
P.ps	0.4	qp	36	1.924	0	0.05000	27883.25	10	5
P.te	0.1	lqp	42	1.963	0	0.02425	30475.97	23	6

MR = Multiplicador de regularización, CC = Clase de característica, MCV = Mejor combinación de variables, MAR = Media del AUC ratio, P_pROC = Valor de p de la ROC parcial, TO = Tasa de omisión en 5 %, AICc = Criterio de información de akaike, NP = Número de parámetros, NV = Número de variables.

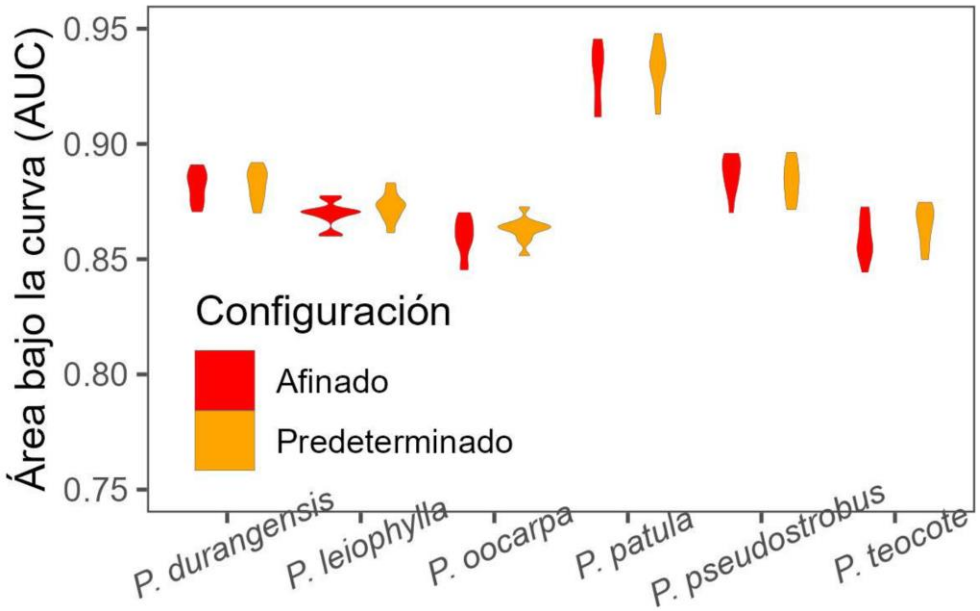

Figura 4. Comparación de modelos ajustados a través del algoritmo Maxent (MAX) con configuración afinada y predeterminada para predecir la distribución potencial de seis especies forestales.

Random Forest (RF)

La afinación de los parámetros del algoritmo RF mostró que respecto al parámetro por defecto (500), 750 y 1 000 árboles (ntree) no mejoró significativamente el ajuste de SDM de las seis especies (Figura 5). Respecto al número de variables muestreadas aleatoriamente como candidatas en cada división (MTRY), se encontró que dos variables mejoran ligeramente el ajuste en las seis especies

(Figura 5). El tamaño del nodo no influyó significativamente en el rendimiento de los modelos ya que los tamaños de nodo dos y 15 proporcionan un ajuste similar (AUC > 0.81) (Figura 6). En conjunto los parámetros conformaron modelos en los que no se encontraron diferencias en el ajuste de ambas configuraciones (afinados y predeterminados) (Figura 7).

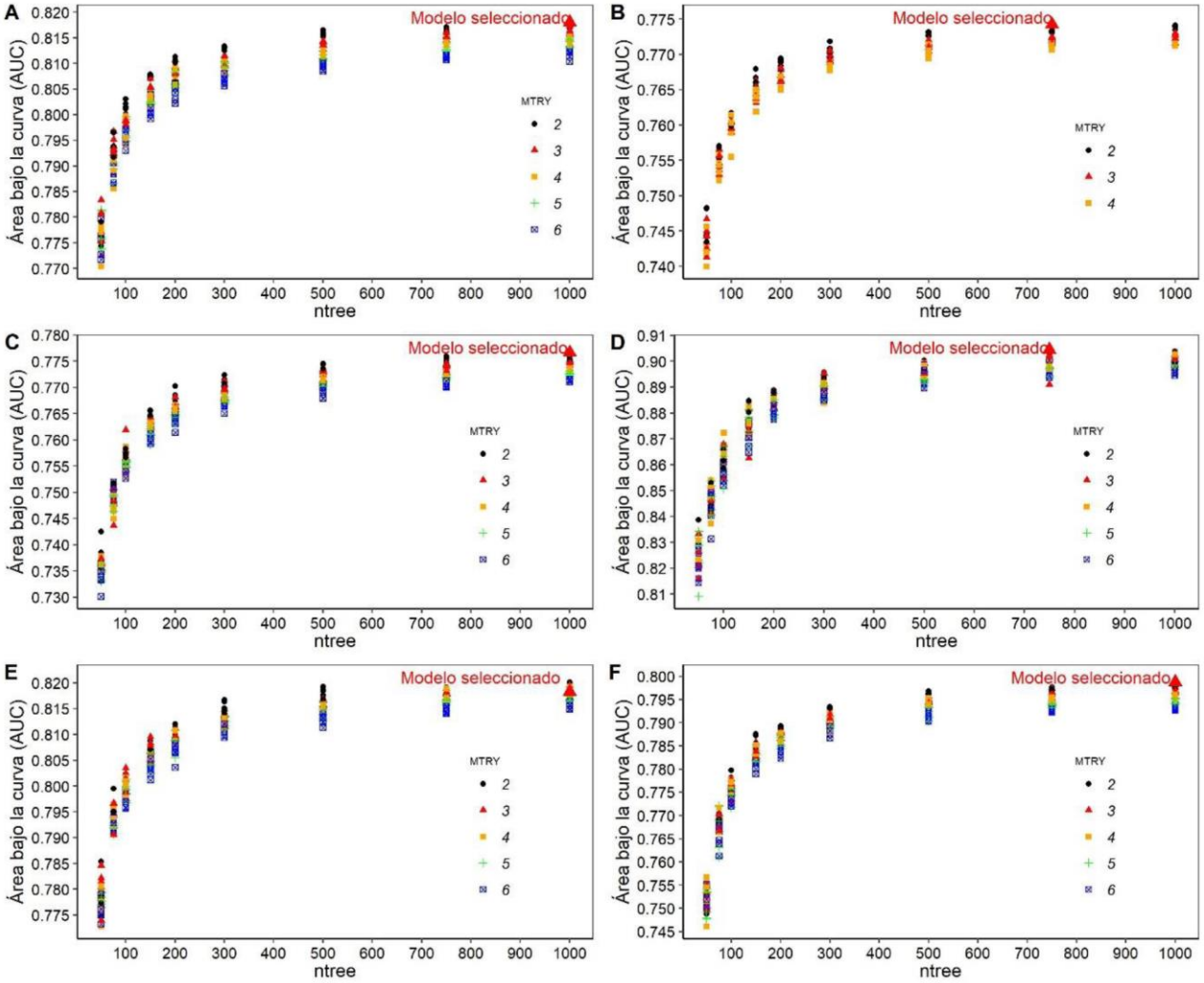

Figura 5. Selección del mejor número de árboles (ntree) y de variables para dividir en cada nodo del árbol (MTRY) para modelar la distribución potencial de *P. durangensis* (A), *P. leiophylla* (B), *P. oocarpa* (C), *P. patula* (D), *P. pseudostrobus* (E) y *P. teocote* (F) a través del algoritmo Random Forest.

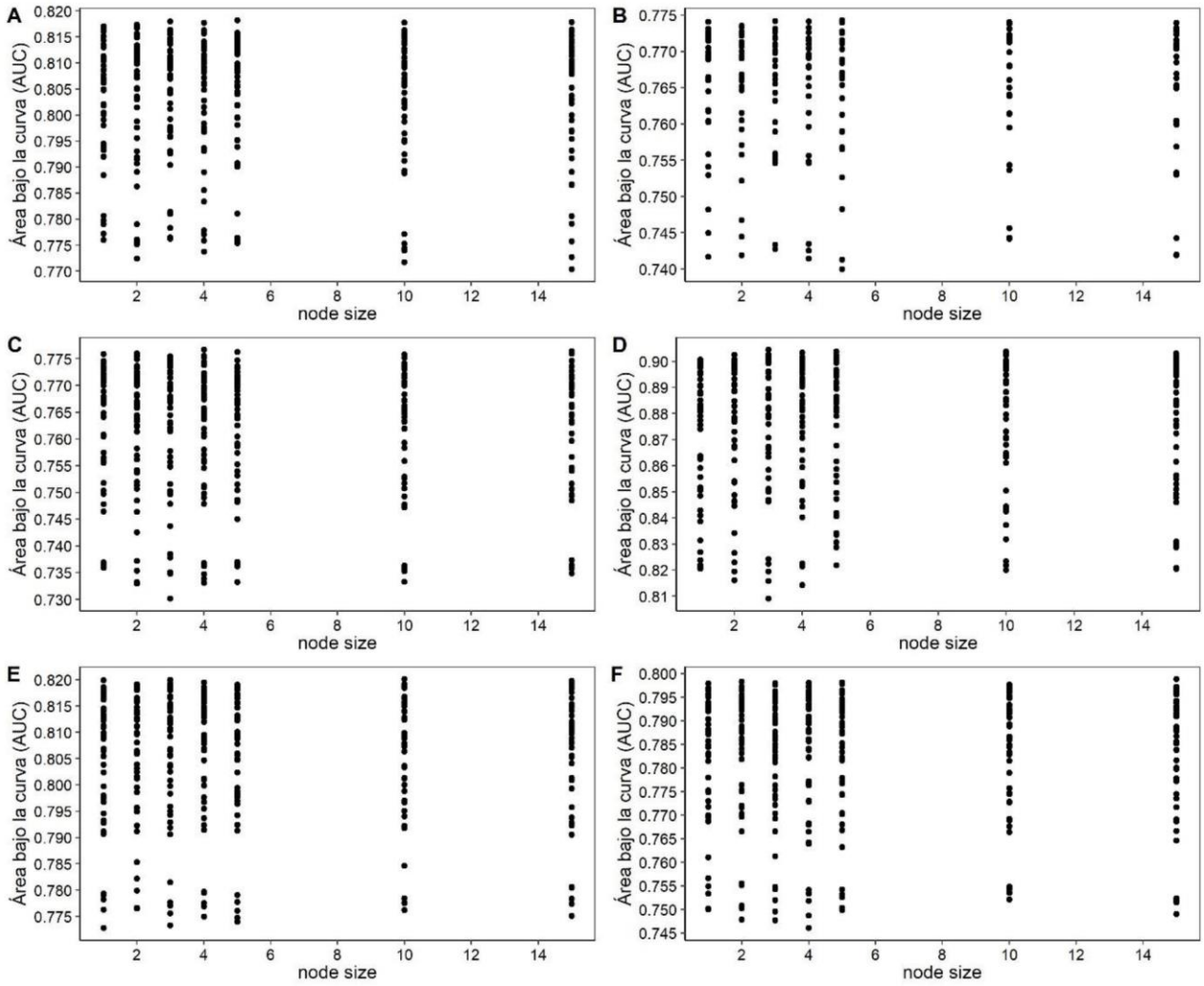

Figura 6. Selección del número mínimo de observaciones en un nodo terminal (node size) para modelar la distribución potencial de *P. durangensis* (A), *P. leiophylla* (B), *P. oocarpa* (C), *P. patula* (D), *P. pseudostrobus* (E) y *P. teocote* (F) a través del algoritmo Random Forest.

Modelos aditivos generalizados (GAM)

A través del estudio se encontró que la suavización de los predictores no produjo un efecto en el ajuste de los modelos de distribución de las seis especies forestales. Los mejores valores en este parámetro oscilaron entre 0.0001 para *P. pseudostrobus* hasta 1.5 para *P. durangensis;* con mayor frecuencia se obtuvieron valores bajos (0.1, 0.01, 0.001, 0.0001) (Figura 8). La comparación de modelos indicó que el ajuste es similar para todas las especies, obteniéndose valores de AUC muy parecidos entre modelos con configuración afinada y predeterminada (Figura 9).

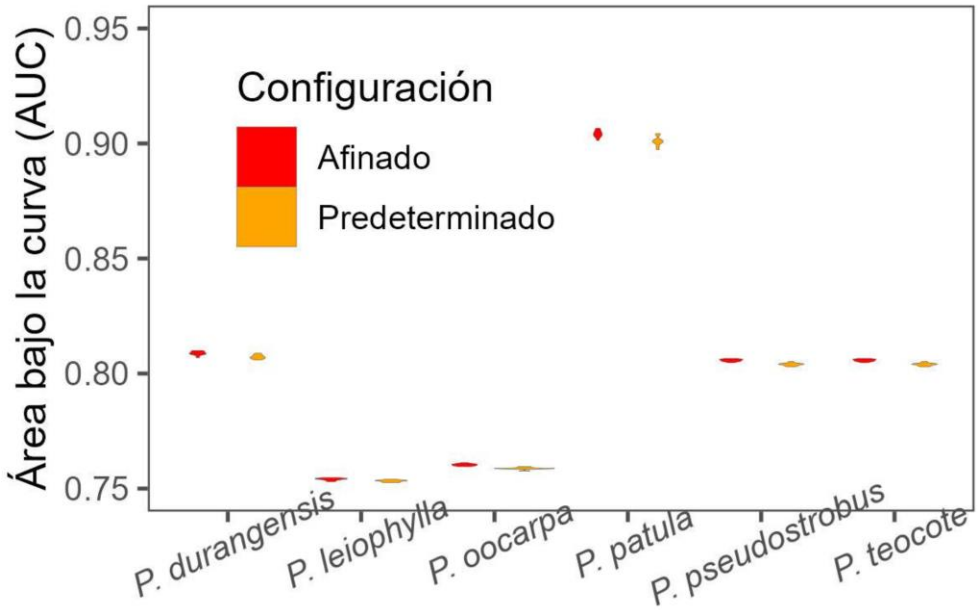

Figura 7. Comparación de modelos ajustados a través del algoritmo Random Forest con configuración afinada y predeterminada para predecir la distribución potencial de seis especies forestales.

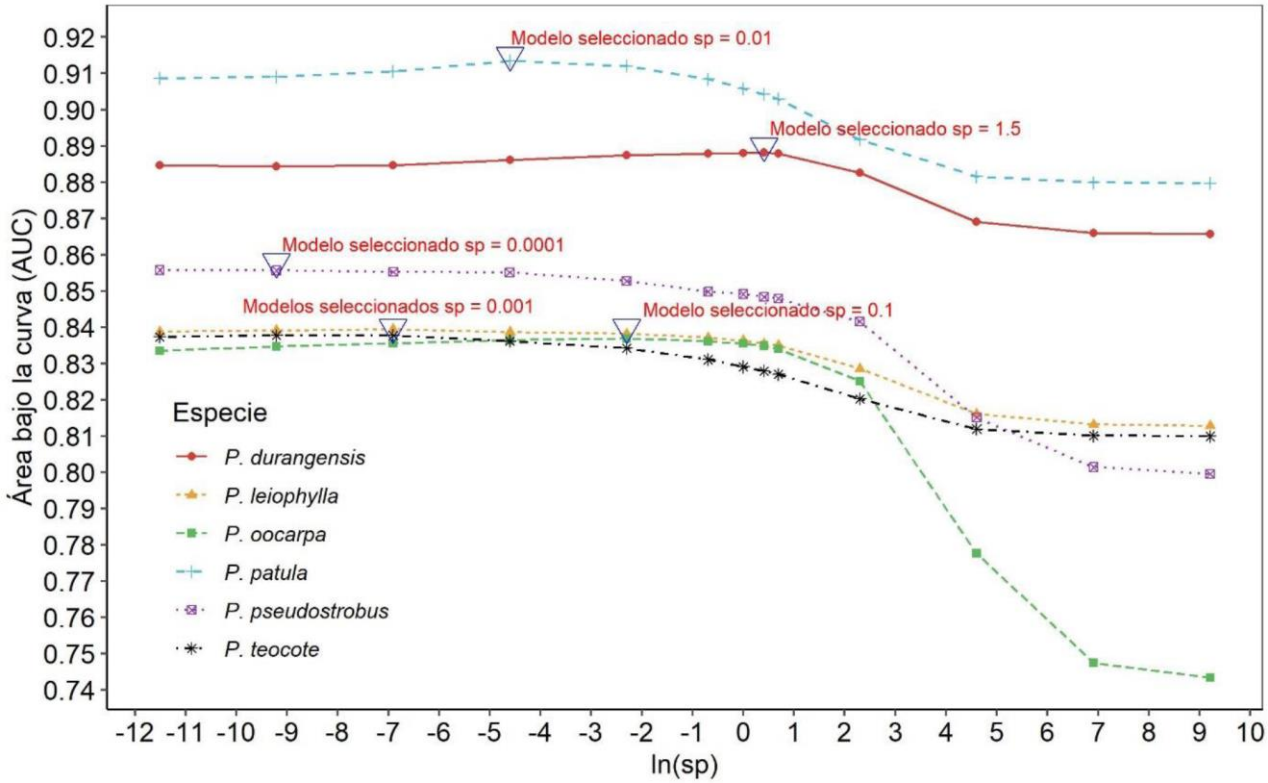

Figura 8. Selección del mejor valor en el parámetro de suavización (sp) para modelar la distribución potencial de seis especies forestales a través del algoritmo GAM.

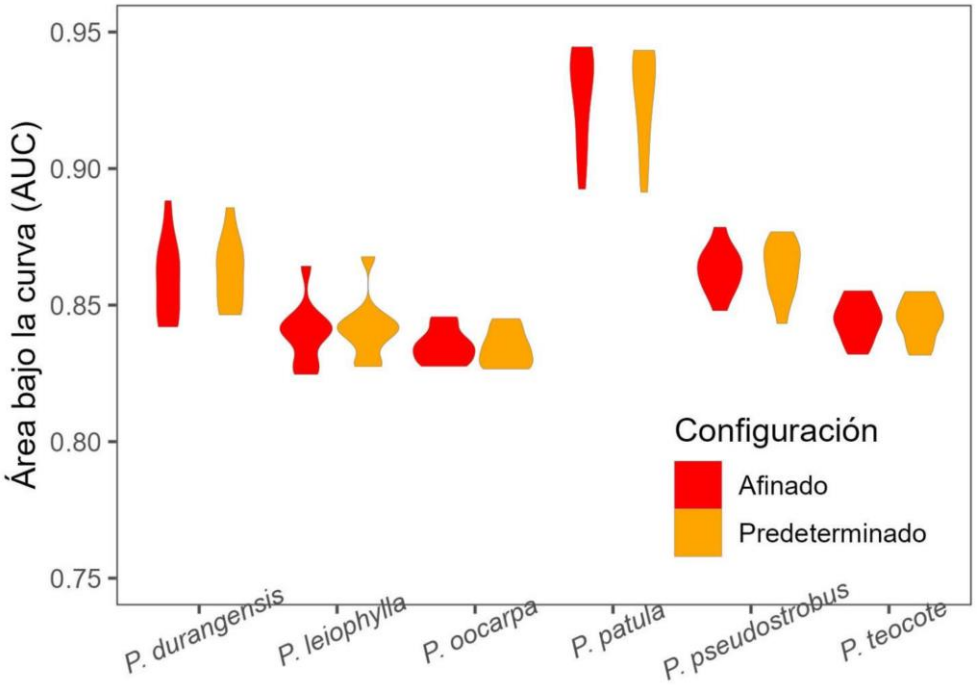

Figura 9. Comparación de modelos ajustados a través del algoritmo GAM con configuración afinada y predeterminada para predecir la distribución potencial de seis especies forestales.

Los valores predichos de idoneidad climática con configuración predeterminada y afinada mostraron congruencia con los valores de coincidencia global difusa debido a que los valores de idoneidad predichos presentaron una relación cercana a 1:1 (Figura 10). La comparación de los mapas de idoneidad climática de las seis especies dilucidó que las configuraciones predeterminada y afinada del algoritmo RF produjeron mapas similares. Los valores de coincidencia global difusa oscilaron de entre 0.815 para *P. oocarpa* y 0.831 para *P. patula*. Las especies *P. durangensis*, *P. leiophylla*, *P. pseudostrobus* y *P. teocote* obtuvieron valores de: 0.821, 0.82, 0.821 y 0.816, respectivamente (Figura 11).

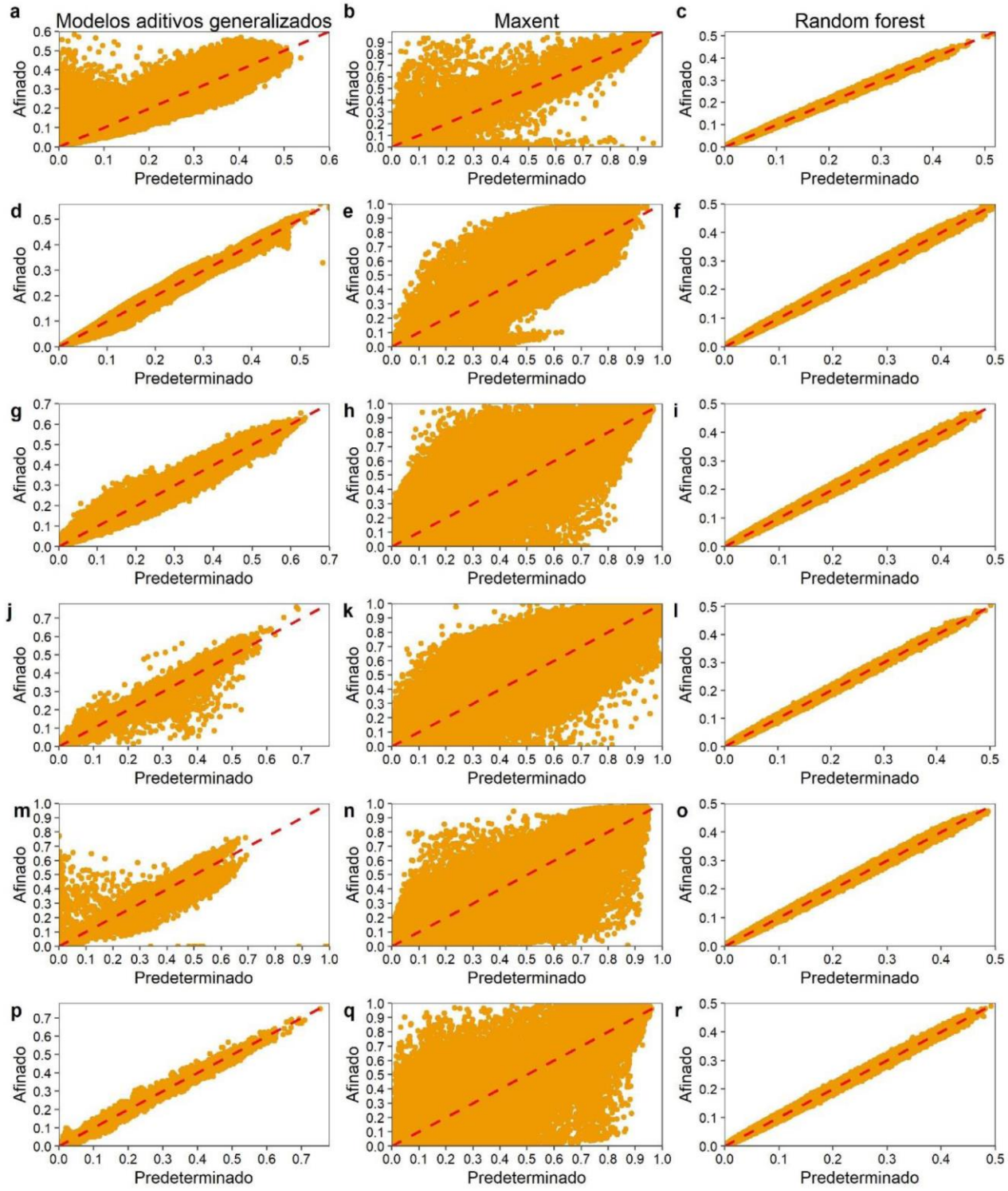

Figura 10. Idoneidad climática predicha con los algoritmos con configuración predeterminada y afinada para *P. durangensis* (a, b, c); *P. leiophylla* (d, e, f); *P. oocarpa* (g, h, i); *P. patula* (j, k, l); *P. pseudostrobus* (m, n, o); *P. teocote* (p, q, r).

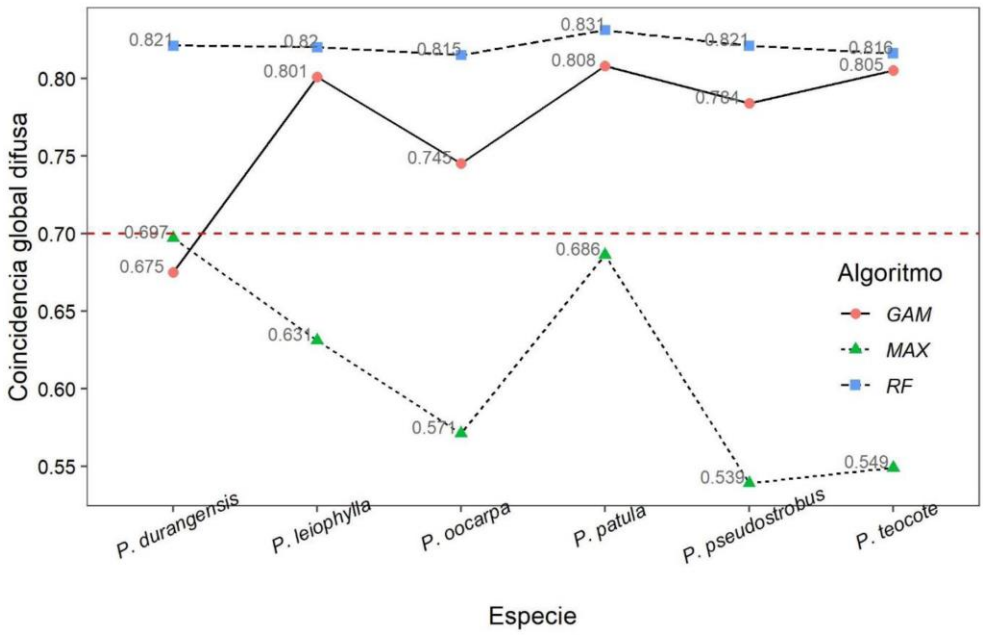

Figura 11. Coincidencia global difusa para analizar la similitud geográfica de la idoneidad climática pronosticada por algoritmo (predeterminado y afinado) y especie.

DISCUSIÓN

Selección de variables ambientales

La selección de las variables explicativas con los métodos utilizados en el presente estudio es recomendada en la literatura (Cobos *et al.* 2019b), ya que incluir muchas covariables podría conducir a problemas de sobreajuste (Merow *et al*. 2013) y colinealidad entre las variables. De manera contraria, la inclusión de variables no importantes para explicar la distribución de las especies disminuye la transferibilidad del modelo (Low *et al*. 2020). Al respecto, Cobos *et al*. (2019b) compararon varios procedimientos para la selección de variables predictoras y encontraron que los modelos creados a partir de variables seleccionadas con el método exhaustivo obtuvieron mejor ajuste y menor sub y sobreestimación geográfica respecto a los otros enfoques. Lo anterior denota que la selección de las variables explicativas a través de múltiples filtros permite disminuir el número de variables, identificando solo las más relevantes en los SDM, aspecto fundamental para incrementar el ajuste, la precisión, y la transferibilidad de los modelos (Fournier *et al*. 2017).

Comparación de algoritmos afinados y predeterminados

Maxent

Se sugiere prestar atención a las clases de características y el factor de regularización en el algoritmo MAX, ya que las primeras tienen efecto sobre las transformaciones disponibles para las covariables y el segundo controla la complejidad de las clases de entidades elegidas que se utilizan para construir el modelo (Vignali *et al*. 2020). De acuerdo con Merow *et al*. (2013) no es recomendable utilizar MAX con la configuración por defecto ya que su configuración debe ser específica para cada especie que se modela, así como para el objetivo del estudio.

A pesar de la importancia que tienen los parámetros anteriormente mencionados, en muy pocos estudios sobre distribución potencial del género *Pinus* se han ajustado modelos afinados (Jiménez y Méndez 2021, Méndez-Encina *et al*. 2021). La mayoría de los trabajos utilizan la configuración predeterminada del algoritmo (Ávila *et al*. 2018, García Aranda *et al*. 2018, Manzanilla-Quiñones *et al*. 2019, Martínez-Sifuentes *et al*. 2020, Pecchi *et al*. 2020), mismos que pueden estar sobre-ajustados debido a la cantidad de clases que se utilizaron para ajustar los modelos. Así mismo, pueden estar careciendo de los valores óptimos en los parámetros reflejándose en los valores predichos de idoneidad.

Con base en los siguientes criterios de evaluación del AUC definidos por Swets (1988): excelente (0.90 – 1.00), muy bueno (0.8 – 0.9), bueno (0.7 – 0.8), regular (0.6 – 0.7) y deficiente (0.5 – 0.6), los modelos ajustados a través del algoritmo MAX mostraron rendimientos muy buenos. Es importante denotar que la nula diferencia entre ambas configuraciones del algoritmo (afinada y predeterminada) concuerdan con lo reportado por Li *et al*. (2020) quienes obtuvieron diferencias pequeñas entre modelos afinados (AUC = 0.885) y modelos predeterminados (0.888). Sin embargo, difieren de lo encontrado por Fan *et al*. (2018) ya que los modelos basados en la configuración ajustada de MAX mostraron mejor ajuste respecto a los basados en la configuración predeterminada. Breiner *et al*. (2018) mencionan que el ajuste de parámetros puede mejorar el rendimiento y la transferibilidad de los modelos, pero se necesita buenos recursos de cómputo y alto tiempo de cálculo, en consecuencia, el ajuste de parámetros podría usarse cuando los recursos informáticos no son un factor limitante.

De acuerdo con Bodbyl-Roels *et al*. (2011) valores de coincidencia global difusa < 0.7 indican diferencia entre mapas, mientras que valores > 0.7 sugieren similitud entre mapas. Estos resultados revelan que un valor de AUC similar para ambas configuraciones de MAX no está asociado con una similitud en la predicción de los valores y mapas de idoneidad ya que la configuración predeterminada y afinada produjeron resultados diferentes.

El algoritmo MAX con parámetros afinados permitió obtener modelos más parsimoniosos. Esto favorece la proyección de la distribución potencial de especies a escenarios futuros (generalmente cambio climático), ya que se ha demostrado que los modelos complejos afectan de forma negativa dicho proceso (Merow *et al*. 2013, Fournier *et al*. 2017, Breiner *et al*. 2018, Fan *et al*. 2018, Li *et al*. 2020, Brun *et al*. 2020).

Random Forest

Los resultados indicaron que el rendimiento de los modelos fue bueno (Swets 1988), sin embargo no se encontraron diferencias entre las configuraciones afinada y predeterminada lo cual coincide con lo reportado en estudios como los de Schratz *et al*. (2019) y Brun *et al*. (2020). Probst *et al*. (2019) también mencionan que RF es insensible a la afinación de sus parámetros. En contraparte, Valavi *et al*. (2021) encontraron que el ajuste de este algoritmo mejoró optimizando los valores de algunos de sus parámetros. A pesar de las discrepancias anteriormente mencionadas, no se debe subestimar la búsqueda de los mejores valores en los parámetros del algoritmo ya que esta práctica metodológica permitirá que se mejore la trasferencia de los modelos a otros espacios o tiempos (Brun *et al*. 2020). Al respecto, Heikkinen *et al*. (2012) mencionan que RF es capaz de proporcionar predicciones más precisas que otros métodos dentro del área que se ajustó el modelo, pero se

requiere mucha precaución si el método se va a utilizar para realizar transferencias en tiempo y espacio.

Por otra parte, Probst *et al*. (2019) mencionan que los parámetros MTRY y tamaño de nodo controlan la aleatoriedad de RF y deben configurarse para lograr una fuerza razonable de los árboles individuales sin demasiada correlación entre ellos. El tamaño del nodo rara vez tiene una influencia en el ajuste de los modelos, pero en muchos casos vale la pena ajustarlo. La RF se ajusta a través de árboles descorrelacionados, en consecuencia, según Breiman (2001) el algoritmo tiende a no presentar sobreajuste, incluso cuando se usan muchos árboles para la clasificación. Contrariamente, Heikkinen *et al*. (2012) han argumentado que RF puede ser propenso a sobreajustarse para ciertos conjuntos de datos, por lo tanto, RF puede hacer que el modelo se comporte de forma errática cuando se extrapola a nuevas áreas. Mientras que Valavi *et al*. (2021) mencionan que los parámetros de RF pueden configurarse de acuerdo a las diferentes implementaciones del algoritmo en determinado lenguaje o programa estadístico. En el software R los parámetros más importantes para ajustar modelos de distribución son ntree, MTRY y node size, mismos que controlan la complejidad del algoritmo (Vignali *et al*. 2020). De acuerdo Bodbyl-Roels *et al*. (2011) valores altos de coincidencia global difusa indican que los mapas comparados presentan alto grado de similitud, en este sentido a través del presente estudio, se dilucidó que la comparación de los mapas con las dos configuraciones (afinada y predeterminada) para RF son muy similares y se puede explicar debido a la insensibilidad del algoritmo al ajuste de los parámetros (Brun *et al*. 2020).

Modelos aditivos Generalizados

Los GAM son utilizados para predecir la distribución de especies debido a su flexibilidad para modelar la variable de respuesta a través de múltiples predictores, los cuales son suavizados de tal forma que las curvas de respuesta se ajusten mejor a los datos (Wood 2017). De acuerdo con Larsen (2015) y Wood (2017) el parámetro de suavización en las funciones de predicción puede evitar el sobreajuste, ya que cuanto mayor sea el valor del parámetro, más suave es la curva y viceversa.

El rendimiento de los modelos afinados y predeterminados encontrados en este estudio se debe a que la implementación de GAM utilizada en el paquete mgcv, realiza una validación cruzada generalizada (GCV) para encontrar el mejor parámetro de suavizado (sp) para cada predictor (Wood 2017). Sobre lo mismo, Schratz *et al*. (2019) ajustaron GAM y encontraron que estos modelos obtuvieron mejor ajuste cuando se afinó el parámetro de suavización. En contraparte Brun *et al*. (2020) mencionan que GAM produjo predicciones menos confiables conforme la parametrización del algoritmo se hacía más compleja. De acuerdo con Derville *et al*. (2018) el algoritmo GAM puede ofrecer una valiosa compensación de complejidad y predicciones precisas.

La comparación de los mapas mediante análisis de coincidencia global difusa indica alta similitud en los mapas generados a través del algoritmo GAM para la mayoría de las especies, con excepción de *P. durangensis*. Esta discrepancia podría deberse a que los mapas generados con el algoritmo afinado predicen valores de idoneidad más altos en áreas donde la configuración predeterminada produce valores muy bajos, estas discrepancias las denota el valor de coincidencia global difusa ya que compara los pixeles de ambos mapas (Bodbyl-Roels *et al.* 2011).

CONCLUSIONES

El desempeño predictivo de los modelos ajustados con configuración predeterminada para definir la distribución potencial de especies arbóreas superó a la configuración afinada; el efecto de la afinación de los parámetros en el ajuste predictivo fue marginal con variaciones según el algoritmo analizado. La comparación de los mapas y valores de idoneidad climática obtenidos a partir de RF reveló que las configuraciones analizadas producen mapas y valores similares para todas las especies. GAM produjo mapas similares en la mayoría de las especies excepto para *P. durangensis;* respecto a los valores predichos de idoneidad climática, se encontraron diferencias entre ambas configuraciones para *P. durangensis*, *patula* y *pseudostrobus*, lo opuesto ocurrió para *leoiphylla*, *oocarpa* y *teocote*. Para MAX se observó que un valor similar de AUC en ambas configuraciones no está asociado con una similitud en la predicción de los valores y mapas de idoneidad ya que produjeron resultados diferentes.

AGRADECIMIENTOS

El primer autor agradece al Consejo Nacional de Humanidades, Ciencias y Tecnologías por la beca otorgada para realizar los estudios de doctorado.

CONFLICTO DE INTERÉS

Los autores declaran que no tienen intereses en competencia.

LITERATURA CITADA

Ávila Coriaz R, Villavicencio-García R, Ruiz-Corral JA (2018) Distribución potencial de Pinus herrerae Martínez en el occidente del estado de Jalisco. Revista Mexicana de Ciencias Forestales 5(24): 92-109. https://doi.org/10.29298/rmcf.v5i24.322.

Bodbyl-Roels S, Peterson AT, Xiao X (2011) Comparative analysis of remotely-sensed data products via ecological niche modeling of avian influenza case occurrences in Middle Eastern poultry. International Journal of Health Geographics 10(21): 1-12. https://doi.org/10.1186/1476-072X-10-21.

Breiman L (2001) Random Forests. Machine Learning 45: 5-32. https://doi.org/https://doi.org/10.1023/A:1010933404324.

Breiner FT, Nobis MP, 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(4): 802-808. https://doi.org/10.1111/2041-210X.12957.

Brun P, Thuiller W, Chauvier Y, Pellissier L, Wüest RO, Wang Z, Zimmermann NE (2020) Model complexity affects species distribution projections under climate change. Journal of Biogeography 47(1): 130-142. https://doi.org/10.1111/jbi.13734.

Cobos ME, Peterson AT, Barve N, and Osorio-Olvera L (2019a) Kuenm: An R package for detailed development of ecological niche models using Maxent. PeerJ 7: e6281. https://doi.org/10.7717/peerj.6281.

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

Derville S, Torres LG, Iovan C, Garrigue C (2018) Finding the right fit: Comparative cetacean distribution models using multiple data sources and statistical approaches. Diversity and Distributions 24(11): 1657-1673. https://doi.org/10.1111/ddi.12782.

Fan JY, Zhao NX, Li M, Gao WF, Wang ML, Zhu GP (2018) What are the best predictors for invasive potential of weeds? Transferability evaluations of model predictions based on diverse environmental data sets for Flaveria bidentis. Weed Research 58(2): 141-149. https://doi.org/10.1111/wre.12292.

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37(12): 4302-4315. https://doi.org/10.1002/joc.5086.

Fournier A, Barbet-Massin M, Rome Q, Courchamp F (2017) Predicting species distribution combining multi-scale drivers. Global Ecology and Conservation 12: 215-226. https://doi.org/10.1016/j.gecco.2017.11.002.

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. Ecosistemas y Recursos Agropecuarios 5(13): 3-13. https://doi.org/10.19136/era.a5n13.1396.

Hallgren W, Santana F, Low-Choy S, Zhao Y, Mackey B (2019) Species distribution models can be highly sensitive to algorithm configuration. Ecological Modelling 408: 108719. https://doi.org/10.1016/j.ecolmodel.2019.108719.

Heikkinen RK, Marmion M, Luoto M (2012) Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35(3): 276-288. https://doi.org/10.1111/j.1600-0587.2011.06999.x.

Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez-Gonzalez M, Kilibarda M, Blagotić A, Kempen B (2017) SoilGrids250m: Global gridded soil information based on machine learning. In PLoS ONE 12(2): e0169748. https://doi.org/10.1371/journal.pone.0169748.

Jiménez SMÁ, Méndez GJ (2021) Distribución actual y potencial de Pinus engelmannii Carriére bajo escenarios de cambio climático. Madera y Bosques 27(3): e2732117. https://doi.org/10.21829/myb.2021.2732117.

Larsen K (2015) GAM: the Predictive Modeling Silver Bullet. MultiThreaded 1-27. Retrieved from https://multithreaded.stitchfix.com/blog/2015/07/30/gam/

Li Y, Li M, Li C, Liu Z (2020) Optimized maxent model predictions of climate change impacts on the suitable distribution of cunninghamia lanceolata in China. Forests 11(3): 1-25. https://doi.org/10.3390/f11030302.

Low BW, Zeng Y, Tan HH, Yeo DCJ (2020) Predictor complexity and feature selection affect Maxent model transferability: Evidence from global freshwater invasive species. Diversity and Distributions 27(3): 497-511. https://doi.org/10.1111/ddi.13211.

Manzanilla-Quiñones U, Aguirre-Calderón ÓA, 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. Madera y Bosques 25(2): 1-16. https://doi.org/10.21829/myb.2019.2521804.

Martínez-Sifuentes AR, Villanueva-Díaz J, Manzanilla-Quiñones U, Becerra-López JL, Hernández-Herrera JA, Estrada-ávalos J, Velázquez-Pérez AH (2020) Spatial modeling of the ecological niche of pinus greggii engelm. (pinaceae): A species conservation proposal in mexico under climatic change scenarios. IForest 13(5): 426-434. https://doi.org/10.3832/ifor3491-013.

Méndez-Encina FM, Méndez-González J, Mendieta-Oviedo R, López-Díaz JÓM, Nájera-Luna JA (2021) Ecological niches and suitability areas of three host pine species of bark beetle dendroctonus mexicanus hopkins. Forests 12(4): 1-18. https://doi.org/10.3390/f12040385.

Merow C, Smit MJ, Silander 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. https://doi.org/10.1111/j.1600-0587.2013.07872.x.

Montoya-Jiménez JC, Valdez-Lazalde JR, Ángeles-Pérez 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(5): 363-371. https://doi.org/https://doi.org/10.3832/ifor4084-015.

Pecchi M, Marchi M, Moriondo M, Forzieri G, Ammoniaci M, Bernetti I, Chirici G (2020) Potential impact of climate change on the forest coverage and the spatial distribution of 19 key forest tree species in italy under RCP4.5 IPCC trajectory for 2050s. Forests 11(9): 1-19. https://doi.org/10.3390/F11090934.

Probst P, Wright MN, and Boulesteix AL (2019) Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(3): 1-15. https://doi.org/10.1002/widm.1301.

R Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Fecha de consulta: 5 de mayo de 2023.

Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A (2019) Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling 406(24): 109-120. https://doi.org/10.1016/j.ecolmodel.2019.06.002.

SEMARNAT (2010) Atlas Geográfico del Medio Ambiente y Recursos Naturales. Secretaría de Medio Ambiente y Recursos Naturales. https://biblioteca.semarnat.gob.mx/janium/Documentos/Ciga/Libros2011/CG006431.pdf. Fecha de consulta: 7 mayo de 2023.

Swets JA (1988) Measuring the accuracy of diagnostic information. Science 240(4857): 1285-1293. https://doi.org/10.1002/9781118341544.ch5.

Urbina-Cardona N, Blair ME, Londoño MC, Loyola R, Velásquez-Tibatá J, Morales-Devia H (2019) Species Distribution Modeling in Latin America: A 25-Year Retrospective Review. Tropical Conservation Science 12(1): 1-19. https://doi.org/10.1177/1940082919854058.

Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2021) Modelling species presence-only data with random forests. Ecography 44(12): 1731-1742. https://doi.org/10.1111/ecog.05615.

Vignali S, Barras AG, Arlettaz R, Braunisch V (2020) SDMtune: An R package to tune and evaluate species distribution models. Ecology and Evolution 10(20): 11488-11506. https://doi.org/10.1002/ece3.6786.

Visser H, De-Nijs T (2006) The map comparison kit. Environmental Modelling and Software 21(3): 346-358. https://doi.org/10.1016/j.envsoft.2004.11.013.

Wood SN (2017) Generalized Additive Models An Introduction with R. Chapman and Hall/CRC. New York, USA. 496p. https://doi.org/10.1201/9781315370279.

Wood SN (2021) Package mgcv. R package version 1.8-33. https://doi.org/10.1201/9781315370279.