TERRA
LATINOAMERICANA

Covariables Ambientales que Definen los Principales Grupos de Suelo en México
Environmental Covariates that Define the Main Soil Groups in Mexico

Cristina Bonilla-Gaviño[1], Demetrio S. Fernández-Reynoso[1‡], Lenom Cajuste-Bontemp[2] y Carlos Ramírez-Ayala[1]

[1] Colegio de Postgraduados, Posgrado de Hidrociencias, [2] Programa de Edafología. Carretera México-Texcoco km 36.5, Montecillo. 56230 Texcoco, Estado de México, México; (C.B.G), (D.S.F.R), (L.C.B.), (C.R.A.).
[‡] Autor para correspondencia: demetrio@colpos.mx

RESUMEN

La variabilidad de los suelos depende de la interacción de covariables ambientales que intervienen en su formación. En México, se registran 25 grupos de suelo de los 32 que incluye la Base Referencial Mundial (BRM). Este estudio identifica el orden de importancia de 11 covariables ambientales, utilizando el modelo no paramétrico basado en un algoritmo de aprendizaje automático supervisado denominado random forest, el cual caracteriza 19 grupos de suelo que incluyen el 99.2% del territorio nacional. Las covariables que se incluyeron fueron, curvatura, densidad de drenaje, distancia al cauce más cercano, geología, índice de aridez, índice de humedad topográfica, índice de posición topográfica, índice de vegetación de diferencia normalizada (NDVI), radiación, rugosidad y temperatura. Los resultados mostraron un total de 100 árboles de clasificación, con una precisión global de 81.83% del modelo a través de random forest y un valor de Kappa de 0.80, expresado como muy bueno. La precisión de disminución promedio mostró que, las cinco covariables analizadas más importantes que clasifican los 19 grupos de suelo son, índice de posición topográfica, índice de aridez, curvatura, radiación y densidad de drenaje.

Palabras clave: *bosque aleatorio, kappa, precisión, WRB, variabilidad.*

SUMMARY

The variability of soils depends on the interaction of environmental covariates involved in their formation. In Mexico, 25 out of the 32 soil groups included in the World Reference Base (WRB) are recorded. This study identified the importance order of eleven environmental covariates that using the non-parametric model based on a supervised machine learning algorithm called random forest, which characterize 19 soil groups covering 99.2% of the national territory. The covariates included were, curvature, drainage density, distance to the nearest stream, geology, aridity index, topographic humidity index, topographic position index, NDVI, radiation, roughness, and temperature. The results showed 100 classification trees, with accuracy of 81.83% of the random forest model and 0.80 of Kappa, which is considered as very good. The mean decrease accuracy, showed that, the five most important covariates for classifying the 19 analyzed soil groups are topographic position index, aridity index, curvature, radiation, and drainage density.

Index words: *random forest, kappa, precision, WRB, variability.*

Cita recomendada:

Bonilla-Gaviño, C., Fernández-Reynoso, D. S., Cajuste-Bontemp, L., & Ramírez-Ayala, C. (2023). Covariables Ambientales que Definen los Principales Grupos de Suelo en México. *Terra Latinoamericana, 41*, 1-13. e974. https://doi.org/10.28940/terra.v41i0.974

Recibido: 30 de abril de 2021.
Aceptado: 3 de junio de 2023.
Artículo. Volumen 41.
Septiembre de 2023.

Editor de Sección:
Dr. Juan Pedro Flores Margez
Editor Técnico:
Dr. Bernardo Murillo Amador
Dr. Gerardo Cruz Flores

INTRODUCCIÓN

La formación del suelo es producto de la continua interacción de cinco factores formadores: clima, organismos, relieve, material parental y tiempo (Jenny, 1961; Buol, Hole y McCraken, 1989; McBratney, Mendonca y Minasny, 2003; Porta, López y Poch, 2014). El Instituto Nacional de Estadística Geografía e Informática (INEGI, 2007) indica que, en México se encuentran 25 de los 32 grupos de suelos que aparecen en la Base Referencial Mundial del Recurso Suelos (WRB, por sus siglas en inglés World Reference Base for Soil Resources), de la FAO-UNESCO-ISRIC. Esta diversidad edáfica se explica por las combinaciones múltiples de los factores que forman el suelo (Cruz, Balboltin, Paz, Etchevers y Krasilnikov, 2007[1]).

La distribución de suelos en México, basada en el relieve, guarda una relación significativa entre las topoformas y las unidades de clasificación (Cajuste-Bontemps y Gutiérrez, 2011); sin embargo, factores como la altitud, clima, patrón de drenaje y vegetación entre otros, influyen en el desarrollo y clasificación final de los suelos (Cajuste-Bontemps y Gutiérrez, 2011) y está directamente vinculada con los factores formadores de suelo en cada región y ecosistema correspondiente (García-Calderón, 2011). El relieve, entendido como el conjunto de deformaciones, desniveles e irregularidades de la superficie del terreno, es determinante en la formación de suelos porque influye en la posición y disposición de los materiales originales; de tal manera que, dependiendo del tipo de relieve, puede presentarse erosión en planos inclinados o modificar el clima mediante control de los escurrimientos, el nivel freático y la vegetación (IMT, 1998).

Los suelos en México son diversos y variados. En el Estado de Oaxaca, la diversidad de suelos se explica en parte por la complejidad del área en cuanto a procesos geológicos diferentes como la formación de la Sierra Madre del Sur y Sierra Norte de Oaxaca (Vásquez-Rasgado y Rodríguez, 2018). En la finca cafetalera el Nueve, ubicada en la Sierra Madre del Sur, Krasilnikov, García y Galicia (2007) identificaron que, la formación de los suelos de la zona está regulada por el origen del material parental. Por su parte, Colín-García *et al*. (2017) y Figueroa-Jáuregui, Martínez, Ortiz y Fernández (2018), mediante un análisis de componentes principales, identificaron que las covariables ambientales que influyen en la variabilidad de los suelos son aquellas de origen geológico, topográfico y climático.

En la porción este de la Península de Yucatán con ambiente tectokárstico, el relieve tiene un rol importante en la distribución de los suelos. En las planicies del norte dominan los Leptosols y Cambisols, en la base de los lomeríos del sur Phaeozems con Vertisols y en las zonas intermedias (planicies acolinadas), se encuentra la diversidad mayor de suelos (Fragoso-Servón, Bautista, Pereira y Frausto, 2016). En el estado de Baja California, los factores formadores del suelo con más relevancia son, material parental, condiciones climáticas y topográficas (INEGI, 2001).

En los métodos de análisis estadísticos de minería de datos se incluyen la regresión lineal, análisis de componentes principales, análisis discriminantes, análisis de correlación espacial y random forest. Entre estos métodos sobresale el random forest el cual destaca por su capacidad para incluir conjuntos de datos complejos, incluir variables correlacionadas y proporcionar medidas de importancia de las variables predictoras sobre las demás. El método random forest se basa en la generación de árboles de clasificación o regresión (Breiman, 2001). Un árbol de decisión es una estructura jerárquica compuesta por nodos internos que representan evaluaciones sobre ciertos atributos, características del problema que se busca resolver, así como nodos hoja que representan las decisiones finales en base a las evaluaciones de los nodos. El algoritmo para construir un random forest, consiste básicamente en seleccionar aleatoriamente *m* grupos distintos de variables aleatorias independientes, sobre cada uno de los cuales se creará un árbol. Posteriormente, la capacidad de predicción de todos los árboles formados se promedia, dando como resultado un modelo que incluyes desde una variable que posiblemente sea relevante para la predicción del objetivo; sin embargo, si se construye un solo árbol no se tomaría en cuenta debido a su frecuencia baja con relación a la variable de salida (Andrade-Saltos y Flores, 2018).

En este estudio se utilizó el modelo de minería de datos random forest (bosque aleatorio), el cual se aplica en áreas diversas de la ciencia. Entre estos estudios relacionados con la clasificación de suelo sobresalen los de Stum, Boettinger, White y Ramsey (2010) y Heung, Bulmer y Schmidt (2014). El término covariable generalmente se utiliza para referirse a una variable independiente o predictora que se incluye en un modelo para controlar

[1] Cruz, C., Balboltin, C., Paz, F., Etchevers, J., & Krasilnikov, P. (2007). Variabilidad Morfogenética de los Suelos de México y su relación con el modelo fisiográfico nacional. En *XVII Congreso Latinoamericano de la Ciencia del Suelo.* León, Guanajuato, México: SLCS.

o ajustar su efecto sobre la variable dependiente de interés. En este estudio, las covariables se consideraron como los factores de formación que se analizan para entender su importancia en la clasificación de suelos. Para este estudio, se seleccionaron variables que se considera tienen una influencia importante en la formación de los suelos. Por ejemplo, la geología afecta la composición mineralógica de los suelos, mientras que, el índice de aridez y el índice de humedad topográfica se relacionan con la disponibilidad de agua y la evapotranspiración, lo cual influye en la formación y características de los suelos. Asimismo, se asume que, las covariables analizadas (curvatura, densidad de drenaje, distancia al cauce más cercano, geología, índice de aridez, índice de humedad topográfica, índice de posición topográfica, índice normalizado de vegetación -NDVI-, radiación, rugosidad y temperatura), representan los procesos formadores de suelo y que el método de random forest permite identificar el orden de importancia, de estas covariables en la formación de los principales grupos de suelo en México. El objetivo de este estudio fue identificar las covariables de importancia mayor en la formación de los grupos principales de suelos de México, en el esquema de clasificación de la Base Referencial Mundial del Recurso Suelos WRB, por sus siglas en inglés World Reference Base for Soil Resources.

MATERIALES Y MÉTODOS

Grupos de Suelos

En México, debido a la complejidad de los factores fisiográficos, climáticos, biológicos, y geológicos se presenta una diversidad amplia de grupos de suelo. En este estudio se incluyeron 19 grupos de suelos, Acrisol (AC), Andosol (AN), Arenosol (AR), Cambisol (CM), Chernozem (CH), Durisol (DU), Fluvisol (FL), Gipsisol (GY), Gleysol (GL), Kastañozem (KS), Leptosol (LP), Luvisol (LV), Phaeozem (PH), Planosol (PL), Regosol (RG), Solonchak (SC), Umbrisol (UM), Vertisol (VR), Calcisol (CL), que se encuentran distribuidos en todo el país (Figura 1) y comprenden el 99.2% del territorio nacional (INEGI, 2007).

La base de datos de los grupos de suelos se obtuvo del conjunto de datos de perfiles de suelo, escala 1:250 000 Serie II (Continuo Nacional) del Instituto Nacional de Estadística y Geografía (INEGI, 2013), clasificado con base a la WRB (FAO, 1999).

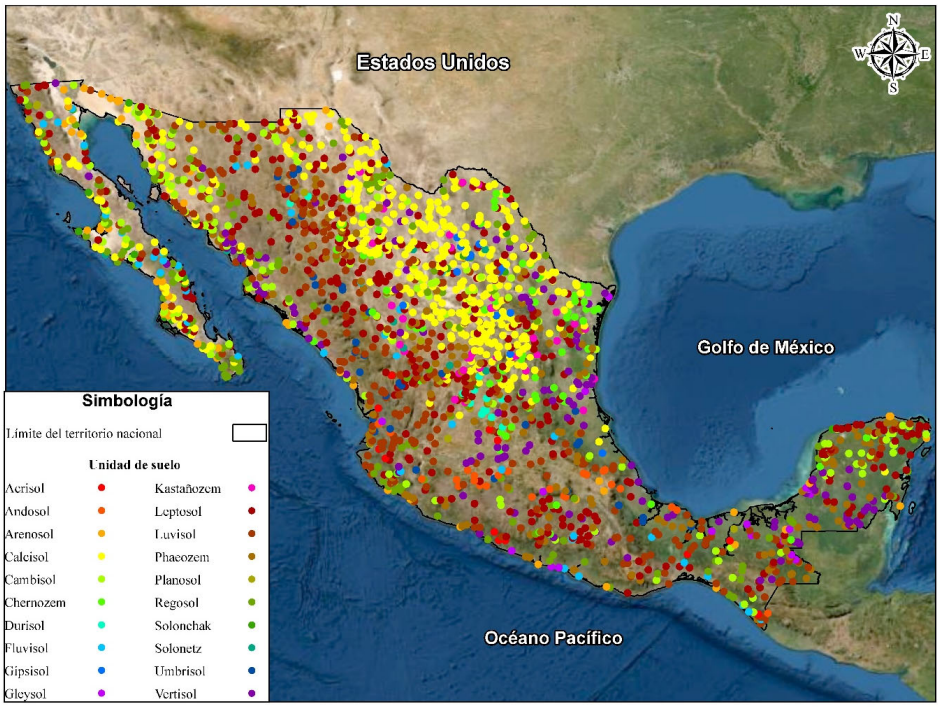

Figura 1. Distribución de los perfiles de suelo para los grupos estudiados (INEGI, 2013).
Figure 1. Distribution of soil profiles for the studied groups (INEGI, 2013).

Obtención de las Covariables por Factor Formador

En esta investigación se consideraron los factores de formación, clima, relieve, material parental y organismo.

Clima

El clima es uno de los factores que influyen directamente en la formación del suelo a escala planetaria (Blanco-Sepúlveda y Sensiales, 2001), porque condiciona la velocidad de meteorización de la roca madre. Los parámetros que se consideraron en este factor son, temperatura, precipitación, índice de aridez y radiación solar. Los datos de temperatura media y precipitación anual se reportan en la base de datos del conjunto de perfiles de suelo (INEGI, 2013) y la radiación solar se obtuvo utilizando el software ArcGIS versión 10.5 (Esri, 2016) así como la herramienta Spatial Analysts Tools, utilizando el comando Solar Radiation- Area Solar Radiation para realizar este análisis.

La temperatura y la radiación solar son factores que influyen en los procesos de meteorización y descomposición de la roca madre. La temperatura afecta la velocidad de las reacciones químicas y los procesos biológicos, mientras que, la radiación solar proporciona energía para la actividad biológica e influye vía la evapotranspiración (ET) en la disponibilidad de agua en el suelo. Las variables consideradas capturan aspectos relevantes de la influencia del clima en la formación de los suelos; además, varían significativamente en diferentes regiones y a lo largo de gradientes espaciales. Al incluir estas variables, es posible capturar patrones espaciales y diferencias regionales en la clasificación de suelos, que tome en cuenta la variabilidad climática y su influencia en la formación de los suelos en diferentes áreas del país.

El índice de aridez se obtuvo con la fórmula:

$$Ia = Pma / Evt \quad (1)$$

La precipitación media anual (Pma) y la evaporación (Ev) provienen de 1941 estaciones climatológicas del periodo 1981-2010 del Servicio Meteorológico Nacional (SMN) de la Comisión Nacional del Agua (CONAGUA, 2023). La evapotranspiración potencial (Evt) se calculó mediante la expresión:

$$Evt = Ev \times 0.8 \quad (2)$$

(Díaz-Padilla *et al*., 2011). Los datos del índice de aridez resultantes se interpolaron mediante el método de Kriging para obtener la cobertura nacional (Figura 2) del cual se obtuvo el índice de aridez correspondiente a cada perfil de suelo.

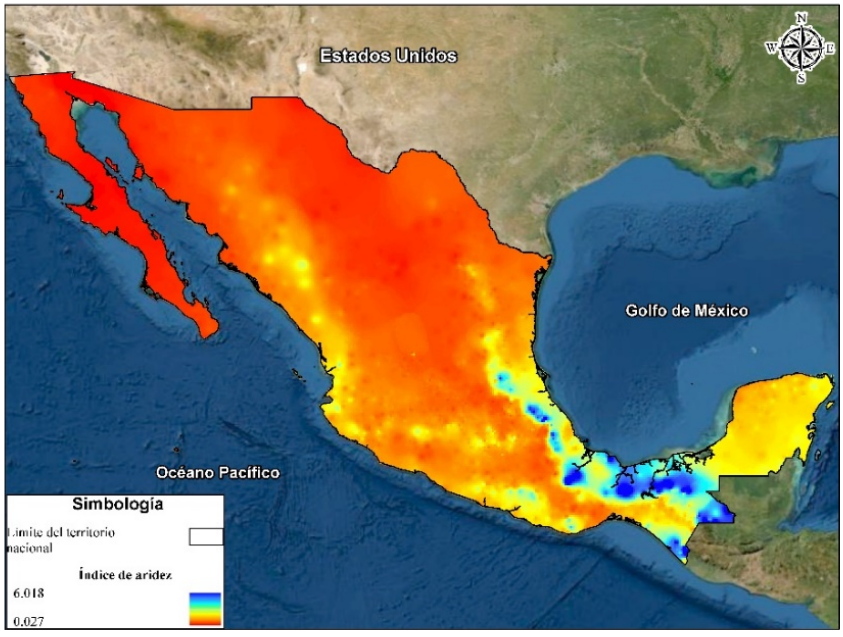

Figura 2. Índice de aridez de los Estados Unidos Mexicanos.
Figure 2. Aridity index of the Mexican United States.

El índice de aridez se utilizó para evaluar la relación entre la precipitación y la evaporación, lo que proporciona información sobre la disponibilidad de agua en un área determinada. El uso de este índica radica en su capacidad para caracterizar la humedad relativa de un área, lo cual es relevante para comprender la formación y distribución de los suelos de diversos sitios, especialmente aquellos suelos donde domina la ET sobre la precipitación y viceversa.

Organismos

Los aspectos más importantes que tienen los organismos en la formación del suelo radican en las relaciones de intercambio de sustancias y energía entre los seres vivos y los minerales presentes en el material parental (IMT, 1998). En este estudio se consideró como covariable, para este factor, el índice normalizado de vegetación (NDVI), obtenido del sensor MODIS (Copernicus, 2020) cuya resolución es de 333 m y el periodo de datos fue del 1 al 10 de enero de 2014. Esta fecha se considera cercana a la publicación del vectorial de perfiles de suelos, permitiendo una coincidencia temporal mayor entre los datos de vegetación y los datos del suelo (INEGI, 2013).

Relieve

La forma de la superficie de la tierra tiene un rol importante en la formación del suelo (Montgomery, 2007). Las covariables para este factor se calcularon a partir del modelo de elevación digital (INEGI, 2011) con resolución del píxel de 30 × 30 m. La función Fill, se utilizó con el objetivo de rellenar vacíos en la superficie del ráster y eliminar imperfecciones del modelo de elevación digital. Las covariables que se obtuvieron con la extensión ArcSIE 10®, utilizando un tamaño de vecindad de 100 m y el método de Shi para el cálculo (Shi *et al*., 2007) fueron las siguientes:

1. Gradiente. Inclinación del terreno, midiendo la inclinación en un lugar. Los valores se expresan en porcentaje (45 grados = 100%).
2. Curvatura general. Mide la forma general de un sitio, los valores positivos indican formas convexas y los valores negativos indican formas cóncavas.
3. Rugosidad. Esta se definió a través del Índice de Rugosidad Topográfica (TRI) desarrollado por Riley, DeGloria y Elliot (1999) utilizando la diferencia en valores de elevación de una celda central respecto a las ocho celdas que lo rodean.
4. Índice de humedad. Este índice también se denomina índice topográfico compuesto (CTI) y se calcula de la siguiente forma:

$$w = \ln(\text{Acumulación de flujo}/\text{gradiente de pendiente}) \quad (3)$$

5. Índice de posición topográfica (IPT). Esta covariable identifica la porción geográfica de una región que maximiza la continuidad y la diversidad de las unidades de paisaje definidas por rasgos topográficos (Jenness, Brost y Beier, 2013). Para la obtención del IPT se utilizó la extensión Land Facet Corridor Designer, de ArcGIS, para radios de acción de 100 metros.
6. Densidad de drenaje. Es la longitud de los cauces presentes en un área específica a una escala determinada, sus unidades son km km^{-2}. La densidad se calculó utilizando la red hidrográfica escala 1:50 000 (INEGI, 2010).
7. Distancia al cauce más cercano. Es la longitud que existe entre un perfil del suelo al cauce más cercano. Para obtenerlo, se utilizó la red hidrográfica escala 1:50 000 (INEGI, 2010) y la distancia en kilómetros.

Estas covariables se seleccionaron porque capturan características topográficas clave que influyen en la formación del suelo, incluyendo la pendiente, la curvatura, la rugosidad, la acumulación de agua y los patrones espaciales del relieve. Al considerar estas covariables, se obtiene una comprensión mejor de cómo el relieve afecta la formación y las características de los suelos en el área de estudio.

Material Parental

El material parental es el punto de partida para la formación de un suelo. La composición química, mineralógica, textural y estructural de la roca, tiene correlación con propiedades físicas y química del suelo como permeabilidad, textura, fertilidad (Jenny,1961). La información del tipo de roca se encontró adjunta en la base de datos del conjunto de perfiles de suelo (INEGI, 2013).

Balanceo de Datos

Una de las complicaciones más comunes al trabajar con bases de datos, es que no se cuenta con el mismo número de observaciones por clases y tratamiento. Para evitar el desbalance de datos, se utilizó la técnica sintética de sobre muestreo con reemplazo Smote, el cual es un algoritmo que sobre muestra la clase minoritaria generando instancias sintéticas con el objetivo de equilibrarla con la mayoritaria. Las instancias sintéticas nuevas se generan a través de la interpolación entre varias instancias de clases minoritarias basándose en la regla del vecino más cercano (Chawla, Bowyer, Hall y Kegelmeyer, 2002).

División de Datos

La base de datos de los grupos de suelos se dividió en la porción 70-30; el 70% de los datos para entrenamiento y el 30% restante para realizar la calibración, con el propósito de evaluar el desempeño del modelo.

Modelado con Random Forest

El análisis de la información de las covariables de los perfiles de suelo para entrenamiento se realizó con el procedimiento de bosque aleatorio a través del algoritmo random forest en el programa "R" (Liaw y Wiener, 2002) con el propósito de entrenar los modelos correspondientes a cada grupo de suelos y posteriormente aplicar estos modelos al conjunto de datos para validación. Los parámetros usados en random forest fueron 100 árboles y 4 covariables predictoras elegidas al azar por cada árbol generado, el cual se calculó considerando que los valores predeterminados para la clasificación es la raíz cuadrada de las covariables. El proceso del algoritmo es el siguiente:

a) Se seleccionaron individuos al azar (usando muestreo con reemplazo) para crear diferentes muestras de datos; b) Se crearon los árboles, eligiendo las variables al azar en cada nodo del árbol, dejando crecer el árbol hasta la máxima profundidad, con cada muestra de datos, obteniendo diferentes árboles, porque cada muestra contiene diferentes individuos y diferentes variables y, c) Se predicen los nuevos datos usando el voto mayoritario, donde clasificará como positivo si la mayoría de los árboles predicen la observación como positiva.

Los dos indicadores de importancia de las covariables que vienen con el algoritmo random forest son, la reducción media de exactitud (Mean Decrease Accuracy), la cual muestra el impacto o pérdida que tiene sobre el rendimiento de la predicción la modificación de una covariable en la base de datos de entrenamiento, que permite encontrar las covariables que tienen mayor poder predictivo en el modelo y la reducción media del índice Gini, o reducción de impureza (Mean Decrease Gini) mide la contribución de cada covariable a la homogeneidad de los nodos y las hojas en el bosque aleatorio resultante.

RESULTADOS Y DISCUSIÓN

El conjunto total de datos balanceados fue de 6685 perfiles de suelos; el 70% corresponde a 4682 perfiles para entrenamiento y el 30% a 2003 perfiles para la base de prueba. El modelo random forest se suministró con once covariables. La ejecución del algoritmo Random Forest en R, implica que se obtenga un resumen que incluye el valor OOB (Out of Bag) el cual se calcula a partir de un muestreo aleatorio con reemplazo. En este caso, el valor OOB alcanzó un 18.3%, lo que representa la proporción de perfiles que fueron predichos correctamente mediante el procedimiento de bootstrapping. El valor OOB no considera los datos de entrenamiento que se descartaron o excluyeron durante el proceso (Out of Bag) los cuales ayudan a minimizar la varianza del modelo.

En la matriz de confusión resultante, se reportan los resultados obtenidos en la muestra de entrenamiento. El error reportado indica que, cuando el modelo se aplique a observaciones nuevas, se espera que la precisión del modelo sea del 81.7%, el cual se considera un modelo bueno. El valor OOB proporciona una estimación del rendimiento del modelo utilizando datos que no se incluyeron en el proceso de entrenamiento. En este caso, el modelo logró una precisión del 81.7% en la muestra de entrenamiento, lo cual sugiere un desempeño satisfactorio. En la Figura 3 se presenta el error de predicción por cada grupo de suelo en función del número de árboles generados. Se observa que, la clase con el error mayor es el Luvisol, con un valor de 0.8 mientras que, las clases Acrisol, Durisol y Gleysol muestran un error menor, con un valor de 0.0.

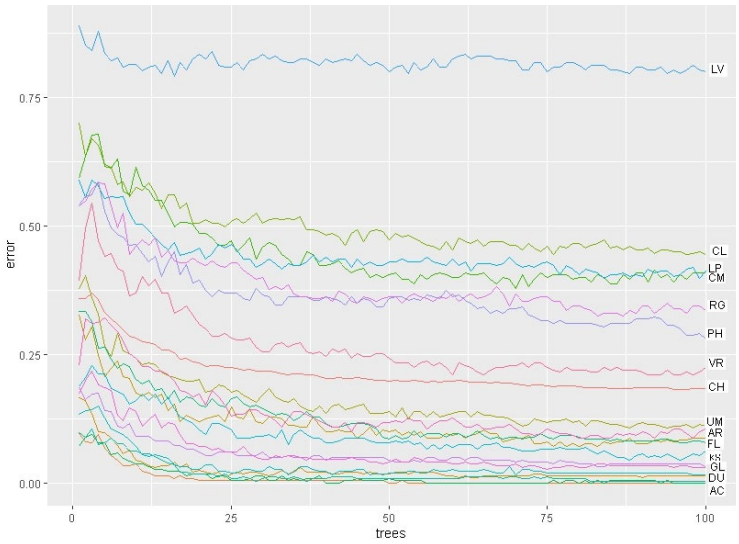

Figura 3: Tasa de error estimada contra el número de árboles.
Figure 3: Estimated error rate versus number of trees.

Importancia de las Covariables

Las covariables de mayor importancia según la reducción media de exactitud son, índice de posición topográfica, índice de aridez, curvatura, radiación y densidad de drenaje. Estas covariables tienen un valor predictivo alto para caracterizar los 19 grupos de suelo (Figura 4).

En la Figura 5 se muestran las cinco covariables más importantes según la reducción media del índice Gini, las cuales son, índice de posición topográfica, índice de aridez, radiación, temperatura y NDVI. Estas covariables son fundamentales para crear categorías diferenciadas, porque miden la homogeneidad en los nodos y hojas del árbol de decisión; además, estas covariables están directamente relacionadas con la forma en que se construyen los árboles de decisión.

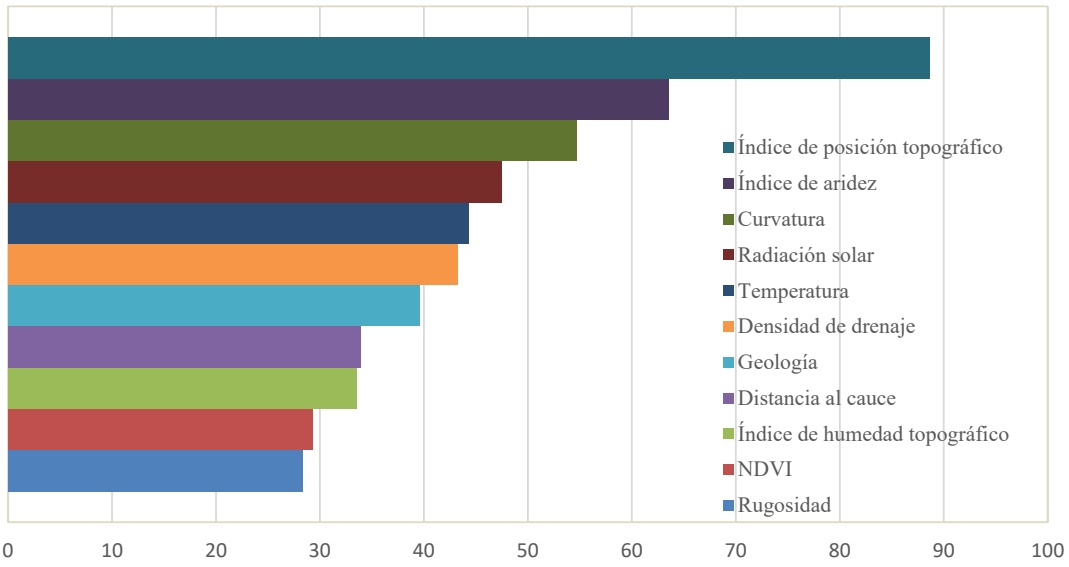

Figura 4. Reducción media de la exactitud de las once covariables del modelo.
Figure 4. Mean reduction in the accuracy of the eleven covariates of the model.

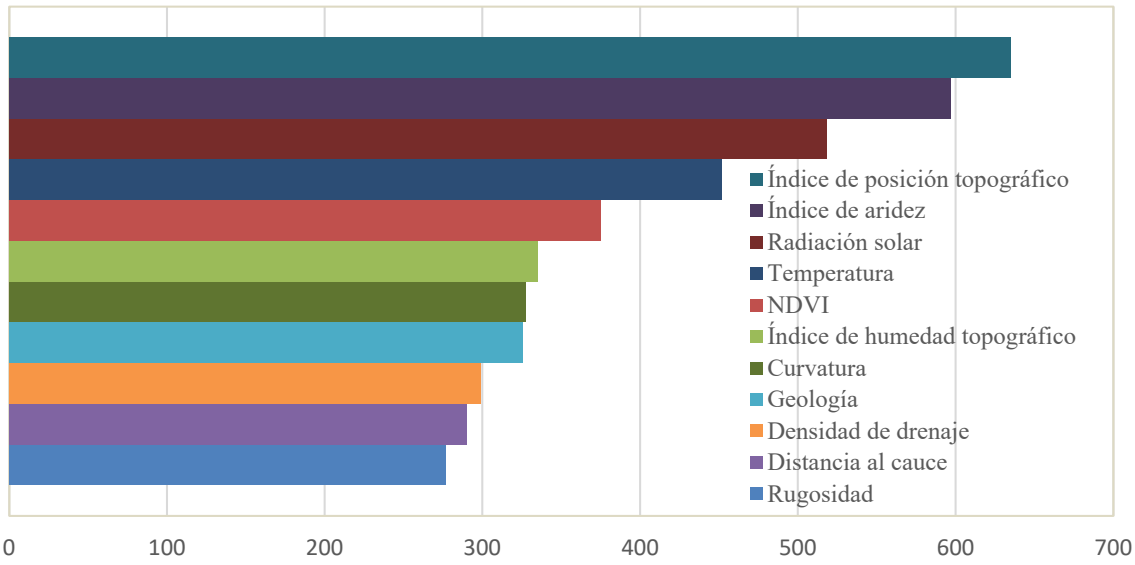

Figura 5. Reducción media de Gini de las once covariables del modelo.
Figure 5. Mean Gini reduction of the eleven covariates of the model.

Evaluación del Modelo

La evaluación del modelo se realizó mediante la generación de una matriz de confusión utilizando la base de datos de prueba. En el Cuadro 1 se presenta la matriz, la cual analiza el rendimiento del modelo en la clasificación de los grupos de suelo. En la matriz de confusión, se observa que, algunos grupos de suelo presentan una precisión menor en la clasificación, específicamente los grupos Leptosol y Phaeozem con una precisión de 62.1 y 62.5%, respectivamente. Esto indica que el modelo puede presentar incidencias para distinguir adecuadamente entre estas clases de suelo. El grupo de suelo Durisol destaca por su precisión alta, con un valor de 94.8%, indicando que, el modelo es capaz de clasificar de manera muy precisa las muestras relacionadas a este grupo.

Los resultados son importantes para comprender las fortalezas y limitaciones del modelo de clasificación utilizado. La precisión menor en ciertos grupos de suelo se atribuye a factores diversos como la variabilidad intrínseca de las muestras o la presencia de características similares entre los grupos. Por su parte, la precisión alta en el grupo Durisol indica que, este grupo presenta características distintivas y bien definidas en relación con los demás grupos.

En general, estos resultados permiten tener una visión más completa del desempeño del modelo y proporcionan información para futuras investigaciones y mejoras en la clasificación de suelos. Las limitaciones del modelo se deben tomar en cuenta y buscar formas de optimizar el modelo para mejorar la precisión en la clasificación de todos los grupos de suelo.

La precisión global del modelo alcanzó un valor de 81.83% (Cuadro 2), indicando que el modelo es capaz de clasificar correctamente aproximadamente el 81.83% de las muestras de suelo en la base de datos de prueba. Este resultado es satisfactorio y sugiere que el modelo tiene un desempeño correcto en la clasificación de suelos.

El índice de Kappa también se calculó y es un parámetro utilizado para medir la concordancia entre la clasificación realizada por el modelo y la clasificación real de las muestras. El valor del índice de Kappa obtenido en este estudio es superior a 0.80 (Cuadro 2) y, de acuerdo con la interpretación propuesta por Landis y Koch (1977), indica una concordancia muy buena entre el modelo y la clasificación real de las muestras.

Los resultados respaldan la eficacia y la robustez del modelo de clasificación utilizado en este estudio. La precisión global alta y el valor de Kappa superior a 0.80 sugieren que, el modelo es capaz de capturar y utilizar de manera efectiva las relaciones y patrones que se encuentran en los datos para realizar una clasificación precisa de los grupos de suelo. Sin embargo, se debe considerar que, la precisión del modelo difiere en función de las características específicas de los conjuntos de datos utilizados y de las condiciones particulares del estudio. Por lo anterior, se recomienda realizar análisis adicionales y validar el modelo en diferentes contextos antes de generalizar los resultados. Los resultados indican que, el modelo muestra un nivel alto de precisión y concordancia en la clasificación de los grupos de suelo y respaldan la utilidad y aplicabilidad del modelo en la caracterización de los suelos.

Cuadro 1. Matriz de confusión de la base de datos de prueba.
Table 1. Confusion matrix of the test database.

Referencia Predicción	AC	AN	AR	CH	CL	CM	DU	FL	GL	GY	KS	LP	LV	PH	PL	RG	SC	UM	VR	%
AC	100	1	0	0	0	0	0	0	0	0	0	3	6	0	0	0	0	2	0	89.3
AN	0	92	0	0	0	0	0	0	0	0	0	2	3	0	0	0	0	3	0	92.0
AR	0	0	104	0	4	1	0	0	0	0	0	0	0	0	0	6	0	0	0	90.4
CH	0	0	0	98	7	1	0	0	0	0	0	0	1	1	1	1	0	0	12	80.3
CL	0	0	0	4	71	10	0	0	0	0	0	1	3	1	0	8	0	0	7	67.6
CM	0	0	0	0	10	60	0	0	0	0	1	2	7	0	0	4	0	0	2	69.8
DU	0	0	0	1	0	0	92	0	0	0	0	0	0	0	1	1	0	0	2	94.8
FL	0	0	0	0	7	1	0	99	0	0	0	0	1	0	0	2	0	0	4	86.8
GL	0	0	1	0	0	0	0	0	111	0	0	0	1	0	0	0	0	0	10	90.2
GY	0	0	0	0	7	1	0	0	0	91	0	0	0	0	0	0	2	0	0	90.1
KS	0	0	0	0	1	0	0	0	0	0	96	3	0	11	0	3	0	0	0	84.2
LP	0	0	0	0	2	1	0	0	0	0	1	64	13	14	0	8	0	0	0	62.1
LV	0	0	0	0	1	1	0	0	0	0	0	3	20	5	0	1	0	0	0	64.5
PH	0	0	0	5	3	2	0	0	0	0	1	17	13	85	0	9	0	1	0	62.5
PL	0	0	0	1	4	3	1	0	0	0	1	0	0	0	89	1	0	0	2	87.3
RG	0	0	0	0	3	2	0	0	0	0	0	7	10	2	0	75	0	0	0	75.8
SC	0	0	2	0	4	4	0	3	0	0	0	0	0	0	0	0	115	0	0	89.8
UM	0	0	0	0	0	0	0	0	0	0	0	4	12	1	0	1	0	106	0	85.5
VR	0	0	4	3	2	2	0	1	0	0	0	0	4	0	1	3	0	0	71	78.0

AC = acrisol; AN = andosol; AR = arenosol; CH = chernozem; CL = calcisol; CM = cambisol; DU = durisol; FL = fluvisol; GL = gleysol; GY= gipsisol; KS = kastañozem; LP = leptosol; LV = luvisol; PH = phaeozem; PL = planosol; RG = regosol; SC = solonchak; UM = umbrisol; VR = vertisol.
AC = acrisols; AN = andosols; AR = arenosols; CH = chernozems; CL = calcisols; CM = cambisols; DU = durisols; FL = fluvisols; GL = gleysols; GY = gypsisols; KS = kastanozems; LP = leptosols; LV = luvisols; PH = phaeozems; PL = planosols; RG = regosols; SC = solonchaks; UM = Umbrisols; VR = Vertisols.

En el Cuadro 3 se muestran las estadísticas por grupo de suelo, donde se destaca la sensibilidad que varía entre 0.21 y 1.0. La sensibilidad representa la tasa de valores que se clasifican correctamente dentro de cada grupo de suelo. Un valor más alto de sensibilidad indica una clasificación mejor de los grupos de suelo. Los grupos de suelo Acrisol, Gleysol y Gipsisol presentan una sensibilidad de 1.0, que significa que, todos los valores de estos grupos se clasifican correctamente por el modelo. Lo anterior indica que, el modelo tiene una capacidad alta para identificar y clasificar adecuadamente las muestras correspondientes a estos grupos de suelo.

Cuadro 2. Estadísticas generales de la base de datos de prueba.
Table 2. General statistics of the test database.

Parámetro	Valor
Precisión:	81.83
95% CI :	(0.8007, 0.8349)
Tasa de no información:	0.0629
Valor P [Acc > NIR]:	$< 2.2e^{-16}$
Kappa:	0.8081
Prueba de Mcnemar Valor P:	NA

Cuadro 3. Estadísticas por grupo de suelos de la base de datos de prueba.
Table 3. Statistics by soil unit from the test database.

	AC	AN	AR	CH	CL	CM	DU	FL	GL	GY
Sensibilidad	1.00	0.99	0.94	0.88	0.56	0.67	0.99	0.96	1.00	1.00
Especificidad	0.99	1.00	0.99	0.99	0.98	0.99	1.00	0.99	0.99	0.99
Valor Predictor Positivo	0.89	0.92	0.90	0.80	0.68	0.70	0.95	0.87	0.90	0.90
Valor Predictor Negativo	1.00	1.00	1.00	0.99	0.97	0.98	1.00	1.00	1.00	1.00
Prevalencia	0.05	0.05	0.06	0.06	0.06	0.04	0.05	0.05	0.06	0.05
Tasa de detección	0.05	0.05	0.05	0.05	0.04	0.03	0.05	0.05	0.06	0.05
Prevalencia de detección	0.06	0.05	0.06	0.06	0.05	0.04	0.05	0.06	0.06	0.05
Exactitud equilibrada	1.00	0.99	0.97	0.93	0.77	0.83	0.99	0.98	1.00	1.00
	KS	LP	LV	PH	PL	RG	SC	UM	VR	
Sensibilidad	0.96	0.60	0.21	0.71	0.97	0.61	0.98	0.95	0.65	
Especificidad	0.99	0.98	0.99	0.97	0.99	0.99	0.99	0.99	0.99	
Valor Predictor Positivo	0.84	0.62	0.65	0.63	0.87	0.76	0.90	0.85	0.78	
Valor Predictor Negativo	1.00	0.98	0.96	0.98	1.00	0.97	1.00	1.00	0.98	
Prevalencia	0.05	0.05	0.05	0.06	0.05	0.06	0.06	0.06	0.05	
Tasa de detección	0.05	0.03	0.01	0.04	0.04	0.04	0.06	0.05	0.04	
Prevalencia de detección	0.06	0.05	0.02	0.07	0.05	0.05	0.06	0.06	0.05	
Exactitud equilibrada	0.98	0.79	0.60	0.84	0.98	0.80	0.99	0.97	0.82	

El valor de 1.0 de sensibilidad no necesariamente implica una clasificación perfecta para todos los grupos de suelo, pues cada grupo de suelo tiene características específicas y distintivas, y algunos grupos pueden ser más fácilmente distinguibles y clasificables que otros. El modelo tiene un desempeño satisfactorio en la clasificación de los grupos de suelo, especialmente para los grupos Acrisol, Gleysol y Gipsisol, donde la sensibilidad es máxima (Cuadro 3).

Los datos clasificados por grupo de suelo muestran que, los grupos de suelo que están más cohesionados en la clasificación presentan una precisión mayor, indicando que, las muestras de estos grupos se clasificaron de manera acertada y consistente. Por otro lado, los grupos de suelo con precisión menor muestran una dispersión mayor, sugiriendo que, algunas muestras de estos grupos se clasificaron en otros grupos edáficos (Figura 6). Los resultados destacan la importancia de la coherencia y la homogeneidad dentro de cada grupo de suelo en la clasificación. Las muestras de un grupo de suelo que comparten características similares y forman un conjunto más compacto, es más probable que sean clasificadas correctamente; en contraste, cuando las muestras de un grupo de suelo presentan variabilidad mayor y se dispersan en el espacio de clasificación, es más difícil asignarlas de manera precisa a un único grupo edáfico, lo que conlleva a una precisión menor en la clasificación.

En un estudio realizado en Haití a través del modelo random forest con una precisión global de 52%, Jeune, Francelino, Souza, Fernandes y Rocha (2018) concluyen que, las covariables nivel de base de la red de canales (CNW), elevación, textura de la superficie del terreno, precipitación, NDVI, distancia vertical a la red de canales (VDCN), pendiente, litología y curvatura fueron las más importantes en la clasificación de los ocho grupos de suelos, Cambisoles, Chernozems, Fluvisoles, Gleysoles, Leptosoles, Luvisoles, Nitisoles y Vertisoles clasificados mediante la World Reference Base.

En una cuenca hidrográfica Guapi-Macacu en Brasil, Pinheiro, Owens, Anjos, Carvalho y Chagas (2017) utilizaron los métodos de árbol de decisión y random forest para predecir los grupos de suelo en la cuenca. La clasificación de los suelos se basó en la clasificación de la World Reference Base y los órdenes taxonómicos de suelos predominantes identificados fueron Ferrasoles, Acrisoles, Gleysoles, Cambisoles, Fluvisoles y Regosoles. El algoritmo random forest demostró ser un modelo de rendimiento mayor, pues los índices estadísticos resultantes

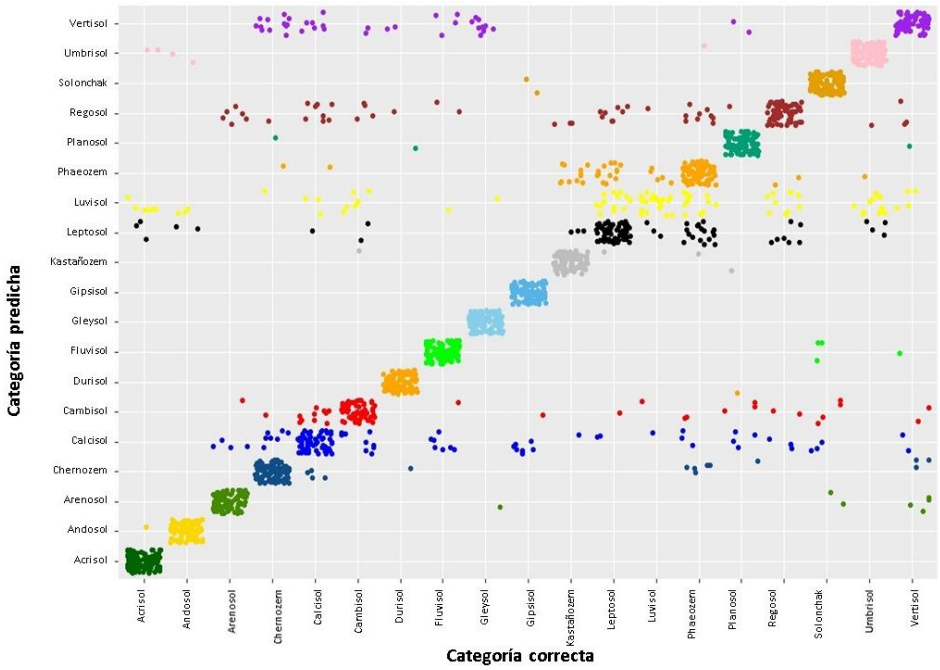

Figura 6. Clasificación de los grupos de suelo.
Figure 6. Classification of soil units.

se consideran excelentes, con una precisión global de 0.966 y un índice Kappa de 0.962. Estos valores indican una concordancia alta entre las clasificaciones realizadas por el modelo y las clasificaciones de referencia. En cuanto a las covariables utilizadas en el modelo, se incluyeron varios atributos relacionados con la topografía, la geomorfología y la composición del suelo. Estos atributos incluyeron la elevación, la pendiente, la curvatura, el índice topográfico compuesto, la distancia euclidiana de las redes de arroyos, el mapa de formas del terreno, el índice de minerales de arcilla, el índice de óxido de hierro y el índice de vegetación de diferencia normalizada, así como información sobre la geología de la zona de estudio. La inclusión de estas covariables permitió capturar la variabilidad espacial de los suelos en la cuenca hidrográfica Guapi-Macacu y contribuyó a la precisión y robustez del modelo de clasificación. Los resultados reportados en la cuenta Guapi-Macacu respaldan la importancia de considerar atributos múltiples relacionados con la topografía, la composición del suelo y las características del paisaje para mejorar la predicción de los grupos de suelo en México.

En una cuenca árida al oste de Utha, USA, Stum *et al*. (2010) identificaron 24 clases de suelo utilizando el método de random forest y de las covariables NDVI, aspecto, pendiente y curvatura, obteniendo una precisión general de 44.8% y definiendo 19 grupos de suelo con las covariables, índice de posición topográfica, índice de aridez, curvatura, radiación solar, temperatura, densidad de drenaje, geología, distancia al cauce más cercano, índice de humedad topográfico, NDVI, y rugosidad, obteniendo una precisión de 0.96 por ciento.

CONCLUSIONES

Se logró identificar las covariables de importancia mayor en la formación de los 19 grupos principales de suelos en México, las cuales se presentan en orden descendente de importancia y se agrupan en factores formadores del suelo como el relieve, clima y material parental. Las primeras siete covariables identificadas son, índice de posición topográfica, índice de aridez, curvatura, radiación solar, temperatura, densidad de drenaje y geología, desempeñaron un papel fundamental en la formación y distribución de los suelos en el país.

El relieve, representado por el índice de posición topográfica, curvatura y densidad de drenaje, influye en la distribución del agua y la erosión, lo que a su vez afecta la formación y características de los suelos.

El clima, medido a través del índice de aridez, radiación solar y temperatura, determina los patrones de humedad y temperatura que influyen en los procesos de formación y desarrollo del suelo.

La inclusión de estas covariables en el análisis permite una caracterización más completa de los suelos en México, al considerar los factores formadores clave que influyen en su formación y distribución. Esto proporciona una base sólida para comprender la diversidad y variabilidad de los suelos en el país, así como para establecer relaciones entre las características del paisaje y los procesos pedogenéticos.

DECLARACIÓN DE ÉTICA

No aplicable.

CONSENTIMIENTO PARA PUBLICACIÓN

No aplicable.

DISPONIBILIDAD DE DATOS

Los datos originales se encuentran disponibles en la página de INEGI, es el Conjunto de datos de Perfiles de suelos. Escala 1:250 000. Serie II (Continuo Nacional) https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825266707

CONFLICTO DE INTERESES

Los autores declaran que no tienen conflicto de intereses en competencia.

FINANCIACIÓN

CONACYT.

CONTRIBUCIÓN DE LOS AUTORES

Idea principal de la investigación: D.S.F.R. Corrida del modelo Random Forest: B.G.C. Análisis de datos: B.G.C. y D.S.F.R. Redacción: C.B.G., D.S.F.R., L.C.B. y C.R.A.

AGRADECIMIENTOS

Al-CONACYT por el financiamiento de la beca de estudios de maestría de la autora principal.

LITERATURA CITADA

Andrade-Saltos, V. A., & Flores, P. (2018). Comparativa entre classification trees, random forest y gradient boosting; en la predicción de la satisfacción laboral en Ecuador. *Ciencia Digital, 2*(4-1), 42-54. https://doi.org/10.33262/cienciadigital.v2i4.1..189

Blanco-Sepúlveda, R., & Senciales-González, J. (2001). La influencia de los factores formadores en las variaciones de las características y propiedades de los suelos de los montes de Málaga. Baetica. *Estudios de Arte, Geografía e Historia, 23*, 9-24.

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

Buol, S. W., Hole, F.D., & McCraken, R. J. (1989). *Soil genesis and classification* (3ª ed.). USA: Iowa State University Press. ISBN: 0813814626

Cajuste-Bontemps, L., & Gutiérrez-Castorena M. C. (2011). El factor relieve en la distribución de los suelos en México. En P. Krasilnikov, F. J. Jiménez-Nava, T. Reyna-Trujillo, & N. E. García-Calderón (Eds.). *Geografía de suelos de México* (pp. 73-84) Ciudad de México, México: Universidad Nacional Autónoma de México.

Chawla, N. V., Bowyer, K.W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Artificial Intelligence Research, 16*, 341-378. https://doi.org/10.1613/jair.953

Colín-García, G., Fernández-Reynoso, D. S., Martínez-Menez, M. R., Ríos-Berber, J. D., Sánchez-Guzmán, P., Rubio-Granados, E., & Ibáñez-Castillo, L. A. (2017). Clasificación digital de suelos a través de covariables ambientales de la cuenca del río Mixteco. *Terra Latinoamericana, 35*(4), 281-291.

CONAGUA (Comisión Nacional del Agua). (2023). SMN: Normales climatológicas por estado. Consultado el 19 de abril, 2023, desde https://smn.conagua.gob.mx/es/climatologia/informacion-climatologica/normales-climatologicas-por-estado

Copernicus (2020). Copernicus Global Land Service. Consultado el 19 de abril, 2023, desde https://land.copernicus.vgt.vito.be/PDF/portal/Application.html#Browse;Root=513186;Collection=1000063;Time=NORMAL,NORMAL,-1,,,-1

Díaz-Padilla, G., Sánchez-Cohen, I., Guajardo-Panes, R. A., Del Ángel-Pérez, A. L., Ruíz-Corral, A., Medina-García, G., & Ibarra-Castillo, D. (2011). Mapeo del índice de aridez y su distribución poblacional en México. *Revista Chapingo Serie Ciencias Forestales y del Ambiente, 17*(Especial), 267-275. https://doi.org/10.5154/r.rchscfa.2010.09.069

Esri (2016). *ArcGis User´s Guide*. *Released 10.1*. Redlands, CA, USA: Environmental Systems Research Institute.
FAO (Organización de las Naciones Unidas para la Agricultura y la Alimentación). (1999). *Base Referencial Mundial del Recurso Suelo*. Informes sobre recursos mundiales de suelos No. 84. Roma, Italia: FAO-ISRIC-SICS. ISBN: 92-5-304141-2
Figueroa-Jáuregui, M. L., Martínez-Menez, M. R., Ortiz-Solorio, C. A., & Fernández-Reynoso, D. (2018). Influencia de los factores formadores en las propiedades de los suelos en la Mixteca, Oaxaca, México. *Terra Latinoamericana, 36*, 287-299. https://doi.org/10.28940/terra.v36i3.259
Fragoso-Servón P., Bautista, F., Pereira, A., & Frausto, O. (2016). Distribución de Suelos en ambientes tectokársticos en la porción este de la Península de Yucatán, México: *GEOS, 36*(2), 265-273.
García-Calderón, N. E. (2011). Los ecosistemas como factor geográfico de distribución de suelo. En P. Krasilnikov, F. J. Jiménez-Nava, T. R. Trujillo, & N. E. García-Calderón (Eds.). *Geografía de suelos de México* (pp. 99 -118.) Distrito Federal, México: Universidad Nacional Autónoma de México. ISBN: 9786070227042
Heung, B., Bulmer, C. E., & Schmidt, M. G. (2014). Predictive soil parent material mapping at a regional-scale: a random forest approach. *Geoderma, 214-215*, 141-154.
IMT (Instituto Mexicano del Transporte) (1998). *Génesis, identificación y uso de los suelos en México: distribución, propiedades, clasificación y manejo de suelos residuales y transportados con aplicaciones a la ingeniería civil*. Documento técnico No. 19. Sanfandila, Querétaro: IMT.
INEGI (Instituto Nacional de Estadística, Geografía e Informática). (2001). *Síntesis de Información geográfica del estado de Baja California*. Aguascalientes, Aguascalientes, México: INEGI,
INEGI (Instituto Nacional de Estadística, Geografía e Informática). (2007). *Conjunto de Datos Vectoriales Edafológico, Escala 1: 250,000, Serie II*. (Continuo Nacional). Aguascalientes, Aguascalientes, México: INEGI.
INEGI (Instituto Nacional de Estadística, Geografía e Informática), (2010). *Red Hidrográfica escala 1:50 000. Edición 2. RH01 - RH37*. Aguascalientes, Aguascalientes, México: INEGI.
INEGI (Instituto Nacional de Estadística, Geografía e Informática), (2011). *Continuo de Elevaciones Mexicano (CEM). 2.0.* Aguascalientes, Aguascalientes, México: INEGI.
INEGI (Instituto Nacional de Estadística, Geografía e Informática). (2013). *Conjunto de Datos de Perfiles de Suelos, Escala 1:250 000 Serie II (Continuo Nacional).* Aguascalientes, Aguascalientes, México: INEGI.
Jenness, J., Brost, B., & Beier, P. (2013). *Land Facet Corridor Designer: Extension for ArcGis*. Arizona, USA: Jenness Enterprises.
Jenny, H. (1961). Derivation of state factor equations of soil and ecosystems. *Soil Science Society of America Journal, 25*(5), 385-388. https://doi.org/10.2136/sssaj1961.03615995002500050023x
Jeune, W., Francelino, M. R., Souza, E. D., Fernandes-Filho, E. I., & Rocha, G. C. (2018). Multinomial Logistic Regression and Random Forest Classifiers in Digital Mapping of Soil Classes in Western Haiti. *Revista Brasileira de Ciência do Solo, 42*, e0170133. https://doi.org/10.1590/18069657rbcs20170133
Krasilnikov, P., García-Calderón, N. E., & Galicia-Palacios, M. S. (2007). Soils developed on different parent materials. *Terra Latinoamericana, 25*(4), 335-344.
Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310
Liaw, A., & Wiener, M. (2002). Classification and Regression by random Forest. *RNews, 2*(3), 18-22.
McBratney, A. B., Mendonca-Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma, 117*(1-2), 3-52. https://doi.org/10.1016/S0016-7061(03)00223-4
Medina-Castellanos, E., Sánchez-Espinosa, J. & Cely-Reyes, G. (2017). Génesis y evolución de los suelos del valle del Sibundoy. *Ciencia y Agricultura*, *14*(1), 95-105. https://doi.org/10.19053/01228420.v14.n1.2017.6092
Montgomery, D. R. (2007). Soil erosion and agricultural sustainability. *Proceedings of the National Academy of Sciences of the United States of America, 104*(33), 13268-13272. https://doi.org/10.1073/pnas.0611508104
Pinheiro, H. S. K., Owens, P. R., Anjos, L. H. C., Carvalho, W., & Chagas, C. S., (2017). Tree-based techniques to predict soil units. *Soil Research, 55*(8), 788-798. https://doi.org/10.1071/Sr16060
Porta, J., López-Acevedo, M., & Poch, R. M. (2014). *Edafología: uso y protección de suelos.* Madrid, España: Mundi-Prensa. ISBN: 978-8484767503
Riley, S. J., DeGloria, S. D., & Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, *5*(1-4), 23-27.
Shi, X., Zhu, A., Burt, J., Choi, W., Wang, R., Pei, T., ... & Qin, C. (2007). An experiment with circular neighborhood in the calculation of slope gradient from DEM. *Photogrammetric Engineering & Remote Sensing, 73*(2), 143-154.
Stum, A. K., Boettinger, J. L., White, M. A., & Ramsey, R. D. (2010). Random forests applied as a soil spatial predictive model in arid Utah. In A. E. Hartemink, & A. B. McBratney (Eds.). *Progress in Soil Science* (pp. 179-190). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-90-481-8863-5_15
Vásquez-Rasgado, P. S., & Rodríguez-Ortiz, G. (2018). Los suelos de los valles centrales de Oaxaca. *Revista Mexicana de Agroecosistemas, 5*(2), 156-167.