PREDICCIÓN DE CARBONO ORGÁNICO EN LOS SUELOS DE MÉXICO A 1M DE PROFUNDIDAD Y 90M DE RESOLUCIÓN ESPACIAL (1999-2009)

Soil Organic Carbon Predictions across Mexico at 1m of Soil Depth and 90m of Spatial Resolution (1999-2009)

Mario Guevara, Rodrigo Vargas

Department of Pland and Soil Sciences, University of Delaware, Newark DE, USA

Resumen

Estudios sobre la variación espacial del carbono orgánico del suelo (COS) son indispensables para mejorar el conocimiento sobre el ciclo global del carbono. Este trabajo documenta el desarrollo de un mapa digital de COS para México a 1m de profundidad y 90 m de resolución espacial representativo del periodo 1991-2009. Un ensamble de árboles de regresión con una eliminación recursiva de variables explica 54% de la variabilidad total empleando una técnica de validación cruzada de muestras independientes. El modelo predictivo produce un error promedio de 0.54 kg m^2 de COS a 1m de profundidad. Se discuten las limitaciones del mapa presentado y las oportunidades de investigación para mejorar la precisión en trabajos futuros. Se estima un total de 16.03 ± 4.24 Pg de COS en el primer metro de suelo mineral para el territorio mexicano. Este resultado es conservador comparado con otros trabajos previos (globales y nacionales). Con este trabajo se provee un marco de trabajo en mapeo digital de suelos útil para habilitar programas de monitoreo estatales y municipales de COS con bajo costo computacional.

Palabras clave: *mapeo digital de suelos, carbono orgánico del suelo 0-100 cm, predicción espacial, 90m de resolución espacial.*

Summary

Studies of the spatial variation of soil organic carbon (SOC) are essential to improve knowledge about the global carbon cycle. This work documents the development of a digital SOC map for Mexico at 1m of soil depth and at 90 m of spatial resolution representative of the period 1991-2009. A model ensemble of regression trees with a recursive elimination of variables explains 54% of the total variability using a cross-validation technique of independent samples. The predictive model produces an average error of 0.54 kg m^2 of SOC at 1m depth. The limitations of the proposed map and the research opportunities to improve the accuracy in future work are discussed. A total of 16.03 ± 4.24 Pg of SOC is estimated in the first meter of mineral soil for the Mexican territory. This result is conservative compared to previous works (global and national). In this study we provide a reference framework on digital soil mapping

useful for enabling state and municipal SOC monitoring programs with low computational cost.

Index words: *digital soil mapping, soil organic carbon 0-100cm, spatial prediction, 90m of spatial resolution.*

Introducción

Los mapas de distribución del carbono orgánico del suelo (COS) son herramientas de gestión requeridas para la formulación e implementación de políticas públicas relacionadas con el potencial natural de los suelos y su respuesta funcional al cambio ambiental global (Stockmann, et al., 2013). Información actual y precisa sobre los contenidos y distribución espacial del COS es constantemente requerida para la planeación de sistemas agrícolas y la identificación de suelos degradados (Powlson et al., 2016). Esto es porque el COS es un indicador directo de la capacidad del suelo para regular funciones de los ecosistemas, como la infiltración de agua a horizontes más profundos o su capacidad para almacenar, transportar y transformar nutrientes en formas disponibles para las plantas (Lal et al., 2018). El COS es entonces un indicador de funciones ecosistémicas e interacciones entre factores biofísicos del suelo, la vegetación y la atmósfera. El COS es tambien un indicador clave de procesos como la degradación de tierras o la productividad primaria, sin embargo, actualmente los estimados de COS a nivel global y nacional son una causa principal de incertidumbre en modelos globales de carbono. Por tanto, los estimados actuales de COS requieren constantemente mayor precisión, exactitud y resolución espacial y temporal.

Actualmente existe gran incertidumbre en los modelos predictivos del ciclo global del carbono que proviene de los estimados de COS (Tifafi et al., 2017). Diversos esfuerzos por caracterizar los contenidos de COS muestran discrepancias a diversas escalas espaciales y temporales (Lajtha et al., 2018). Esto es principalmente porque los diversos estimados de COS están basados en múltiples colecciones de datos patrimoniales y métodos de colecta de datos que representan indistintamente condiciones pasadas (en algunos casos >50 años) y condiciones actuales (e.g., Hengl et al., 2017, Guevara et al., 2018). México es un país pionero en cuanto al mapeo y documentación de sus recursos naturales ya que desde 1968, el Instituto Nacional de Estadística, Geografía e Informática (INEGI) trabaja en protocolos para generar información geográfica a nivel nacional sobre los recursos naturales (e.g., INEGI Serie 1 y 2, Krasilnikov et al., 2013). Con la información disponible para México, se han presentado estimaciones de carbono en el suelo (dentro de los primeros 30cm de profundidad) que varían entre seis y 18 Pg (Guevara et al., 2018). Síntesis recientes sugieren que el contenido actual de carbono en los primeros 30cm de profundidad en México es cercano a 9 Pg (Paz et al., 2016), pero existen discrepancias entre diversos estimados (Tifafi et al., 2017). En México se asume que el contenido a un metro de profundidad es de 18 Pg, el doble que a los 30 cm de profundidad (Lajtha et al., 2018). Sin embargo, no existe información detallada (e.g., en pixeles con resolución espacial < 100m) de cobertura nacional que nos permita saber cuál es el la distribución espacial y el contenido de carbono en el suelo a un metro de profundidad. Esta alta resolución espacial no solo es importante para la caracterización espacial del COS pero también para para proveer información a una escala relevante para generar planes de manejo y políticas públicas (FAO,

2017). El COS almacenado en la superficie del suelo (e.g., 0-15 o 0-30 cm) es más sensible a cambios de uso de suelo y transformaciones a la cobertura vegetal o a la acción directa del clima comparado con el carbono almacenado a mayores profundidades (e.g., 1 m de profundidad). El carbono almacenado a un metro de profundidad es por tanto más estable que el COS almacenado en la superficie del suelo.

La estimación del contenido de COS requiere dos variables edáficas adicionales: la densidad aparente del suelo (i.e., la relación entre peso y volumen) y el contenido de fragmentos rocosos (fragmentos > 2mm) (Nelson y Sommers, 1982). Estas variables tienen una disponibilidad limitada a nivel país y son la causa principal de errores en los estimados del COS (Poeplau et al., 2017). A pesar de los grandes esfuerzos realizados (Cruz-Cárdenas et al., 2014, Paz et al., 2016), México no es una excepción ya que existen grandes áreas sin información disponible sobre mediciones directas de COS, densidad aparente o fragmentos rocosos (Krasilnikov et al., 2013). Esto representa un gran reto para mejorar los estimados actuales de la distribución del COS al nivel nacional.

Existen varios métodos para predecir con datos de COS áreas sin información. Una posibilidad es asignar (e.g., con una ponderación) un valor de carbono a cada categoría de un mapa de suelos (e.g., tipos de suelo) disponible o a la intersección de varios mapas de variables relacionadas con la variabilidad espacial del COS (i.e., tipos de clima, tipos de rocas, tipos de geoformas, mapas de uso de suelo y tipos de vegetación) (Yigini et al., 2018). Estas capas posteriormente se generalizan (e.g., con el criterio de área mínima cartografiable) a una escala donde (idealmente) todas las categorías en el mapa están representadas con mediciones directas y datos de carbono. Otra manera es hacer un mapeo predictivo de COS construyendo un modelo estadístico (e.g., lineal, no lineal, basado en hipótesis, basado en datos) (Hengl y MacMillan, 2018). Este modelo depende de las relaciones entre los datos de COS y la información ambiental disponible representativa del área de interés, la cual se puede obtener a partir de productos satelitales, modelos digitales de elevación (i.e., Geomorfometría o análisis digital de terreno) y diversos tipos de mapas temáticos para representar el ambiente de formación de suelos (Reuter and Hengl, 2012). El ambiente de formación de suelos deriva de la información asociada a los factores de formación de suelos: clima, organismos vivos, topografía (relieve) y geología, que interactúan en un periodo determinado de tiempo (para el caso de interés) (Jenny, 1941). Combinando los datos disponibles y las capas ambientales es posible generar predicciones continuas de COS (y de otras de sus propiedades físicas, químicas y biológicas) en áreas sin datos disponibles y estimar un error asociado a estas predicciones (Lagacherie, et al., 2019). Las técnicas asociadas a estas estimaciones estadísticas para generar mapas digitales de COS pertenecen al área de estudio del mapeo digital de suelos (McBratney et al., 2003).

Avances recientes en mapeo digital de suelos y la disposición de datos relacionados al COS y el ambiente de formación de suelos han resultado en predicciones globales de COS a 250m de resolución espacial (Hengl et al., 2017). Las estimaciones globales de COS no necesariamente representan de manera precisa el COS para un país específico (Guevara et al., 2018). Esto se debe a que cada región (o país) tiene limitaciones particulares de información y el COS está asociado a condiciones de formación de suelo específicas para dicha región (o país). Con la disponibilidad de información específica de COS para México y el desarrollo de nuevas técnicas de Geomorfometría (i.e., análisis digital de terreno) es posible

generar predictores topográficos del COS a una escala espacial mucho más detallada (e.g., <100m) usando datos a nivel país (Amatulli et al., 2019). Una estimación a esta resolución espacial para los más de 2 millones de kilómetros cuadrados de México representa un reto computacional que hasta el momento ha sido reservado para instituciones con acceso a recursos de computación de alto rendimiento (*High performance Computing* o HPC en inglés). Por lo tanto, existe la necesidad de desarrollar técnicas para implementar estimaciones con gran resolución espacial de COS (y otras variables biofísicas del suelo) con recursos computacionales de bajo costo disponibles para múltiples usuarios (Beaudette y O'Geen, 2009).

El objetivo de este trabajo fue generar una predicción espacial del COS a 1 m de profundidad y a la resolución espacial de 90 m, empleando diversas capas de información ambiental como factores predictivos. Esta predicción se basó en un modelo estadístico seleccionado a partir de la comparación de diversas combinaciones de predictores ambientales y datos disponibles usando criterios de información como medidas de desempeño. La información generada será de utilidad para la validación y calibración de estimaciones nacionales y globales de carbono, y para la interpretación espacial de COS en México a escalas espaciales relevantes para el manejo de ecosistemas terrestres. Este objetivo se logró con recursos computacionales que generalmente están disponibles en la mayoría de las instituciones nacionales (e.g., agencias y centros de investigación) con intereses de predicción de COS y otros recursos naturales.

Métodos

El mapa digital de COS de México a un metro de profundidad y 90m de resolución espacial fue generado en R Core Team (2018) y el código se encuentra disponible en el material suplementario. La predicción de COS fue preparada a lo largo de una cuadrícula regular de píxeles de 90 m (250 901 811 pixeles de 90m) representativa del área dentro de los límites geográficos del país. La generación del modelo se basa en los datos disponibles a nivel nacional, y el proceso de predicción se genera por estado. En la predicción se aplican los coeficientes del modelo predictivo a toda el área de interés. Estos procesos (modelación y predicción) de México se realizan en una computadora portátil con 15.5 Gb de memoria, un procesador Intel® Core™ i7-7700HQ CPU @ 2.80GHz × 8 nodos y un sistema operativo Debian GNU/Linux 9 (stretch) 64-bit. El límite político para cada estado se obtuvo del proyecto *Global Administrative Areas* (2012).

Para el desarrollo de nuestro modelo predictivo, solamente se usaron los datos disponibles entre 1999 y 2009 provenientes de 2852 perfiles de suelo descritos a lo largo del territorio nacional (Figura 1) por el INEGI en su serie 2 de información edafológica (Krasilnikov, 2014). Para probar la capacidad predictiva de un modelo de aprendizaje automático aplicado a la variabilidad espacial del COS, fue seleccionada solo una decada de datos relativamente bien representada en la serie 2 de INEGI para generar una línea base de mapeo (relativamente 'reciente') basada en un solo sistema de organización de la información de suelos disponible (INEGI 2011, Figura 1). En este trabajo se asume un escenario de datos 'estático' pero representativo de la década analizada con la finalidad de evitar confusión en los resultados asociada a cambios en el carbono orgánico del suelo durante períodos más largos de tiempo. De esta manera emerge una fuente inevitable de incertidumbre asociada a posibles cambios, durante 1999 y 2009, en el ambiente

de formación de suelos (y consecuentemente cambios en el COS) por actividad humana o por patrones climáticos ocurridos durante este periodo de tiempo. Esta incertidumbre va más allá del enfoque principal de este trabajo, el cual está enfocado en la predicción de la variabilidad espacial del COS y en proponer un marco de trabajo que puede ser aplicado (en trabajos futuros) a múltiples series de colecta de datos en México. Trabajos previos han documentado las características principales de la bases de datos con información de suelos y series de colecta dirigidas por el INEGI (e.g., Krasilnikov, 2014) que pueden ser usadas con fines de monitoreo de COS.

Para obtener un valor específico de COS para un metro de profundidad fue usada la propuesta metodológica propuesta por Beaudette (2013) para estimar propiedades del suelo y relaciones con profundidad. Posteriormente aplicamos las funciones de suavizado de áreas equivalentes propuestas por Bishop et al., (1999) y por Malone et al., (2009) para su aplicación en mapeo digital de suelos siguiendo la implementación de Hengl et al. (2017).

Estimación de COS

Para el cálculo de COS se utilizó la ecuación (Eq.1) propuesta por Nelson y Sommers (1983). Esta ecuación relaciona de manera lineal la concentración de carbono orgánico (CO, g kg), la densidad aparente del suelo (BLD, gr cm^3), el contenido de fragmentos rocosos (CFR, %) y la profundidad del suelo representada en cm (H, 1 m).

$$COS = CO/1000 \times H/100 \times BLD \times (100 - CFR)/100 \quad \text{Eq. 1}$$

Donde COS representa el contenido total de carbono orgánico (g kg) a 1 m de profundidad de suelo mineral. La base de datos utilizada no cuenta con información de BLD. Por tanto, la BLD fue estimada de manera lineal a partir del contenido de materia orgánica (MO) empleando la ecuación (Eq. 2) propuesta por Saini (1996).

$$BLD = 1.53\ (\pm 0.1) - 0.05\ (\pm 0.01) * MO \quad \text{Eq.2}$$

Donde MO es igual a CO * 1.724. Este método fue generado por Saini y colaboradores (1966) para diversos suelos a nivel global, mostrando una mayor precisión para suelos escasamente drenados (r=0.85) que para suelos imperfectamente drenados (r=0.80) o suelos bien drenados (r=0.63). Los CFR (> 2 mm) fueron clasificados por su tamaño en piedras, guijarros o gravas y reportados en porcentaje en cada polígono de suelo analizado (INEGI, 2011). Para este trabajo estos valores fueron agrupados en 5 clases porcentuales (0, 20, 40, 60 y 80%) correspondientes con las 5 clases que usa INEGI para reportar la pedregosidad en cada perfil de suelo analizado (INEGI, 2011).

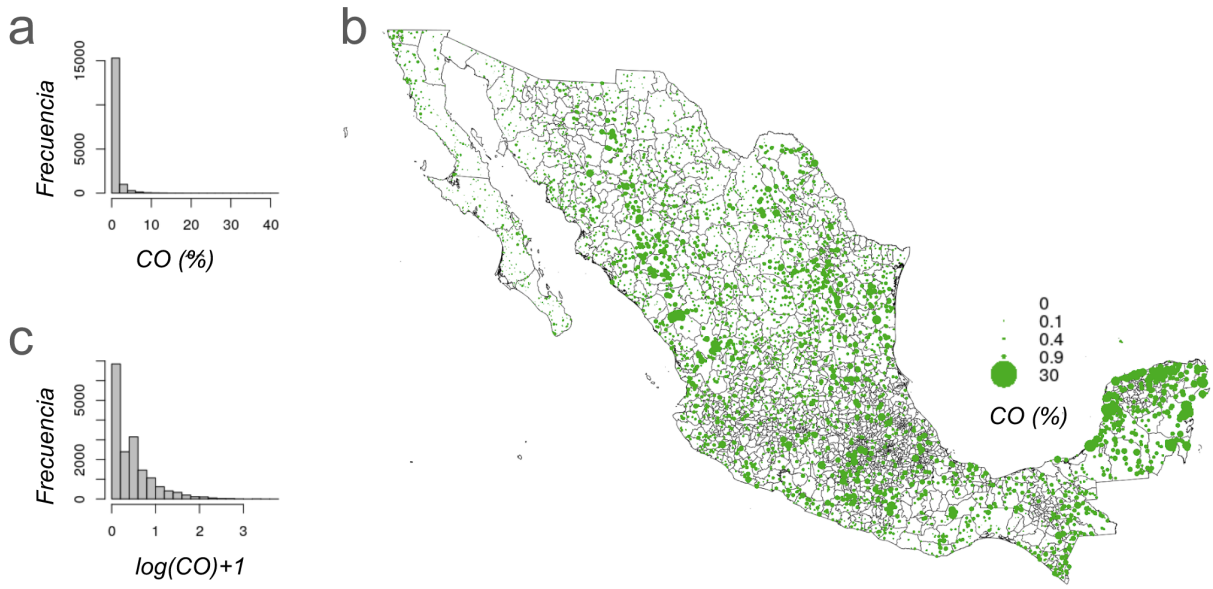


Figura 1 Distribución estadística de los datos disponibles de COS (a) y la transformación a su logaritmo natural + 1 (b). Distribución espacial de los datos de entrenamiento para el modelo predictivo (c). El tamaño del círculo está asociado con la concentración (%) de COS en cada punto.

Factores de predicción del COS

Para representar el ambiente de formación de suelos empleamos un conjunto armonizado de covariables ambientales en una cuadrícula regular de 90x90m a lo largo del territorio mexicano. La base de referencia fue el modelo digital de elevaciones (MDE) de INEGI a 90 m de resolución espacial. De este MDE fueron calculados atributos topográficos primarios como la pendiente del terreno y la exposición. También se calcularon atributos topográficos secundarios como el índice de rugosidad del terreno, el índice de posición topográfica, el índice analítico de sombra del terreno y el índice de escurrimiento de flujo superficial (Figura 3a). Estos atributos topográficos se calcularon en R empleando la implementación propuesta por Hijmans (2017), con base en los trabajos previos de Burrough y McDonnell (1998) y de Wilson et al., (2007). También fue utilizada la información climática (precipitación y temperatura) de WorldClim para condiciones 'recientes' (1970-2000, Fick y Hijmans, 2017). Estas capas de precipitación y temperatura fueron el insumo requerido para calcular el índice de aridez (evapotranspiración/precipitación) con base en el modelo de balance hídrico empírico propuesto por Thornthwaite (1948). Estas capas climáticas tienen una resolución espacial de 1km y fueron armonizadas en la cuadrícula base de 90m del MDE pero manteniendo el mismo valor en todos los píxeles de 90 m que corresponden a cada pixel de 1 km. Este procedimiento se repitió para armoniza en la cuadrícula regular de 90x90m, dos capas adicionales de información preparadas para el proyecto SoilGrids (Hengl et al., 2014, 2017): a) el índice de vegetación mejorada (EVI) derivado del sensor MODIS (Moderate Resolution Imaging Spectroradiometer); y b) un producto satelital generado por el USGS (United States Geological Survey) que representa la edad aproximada (en millones de años) de la geología y variabilidad litológica asociada (Reuter and Hengl, 2012).

Un problema común en el mapeo digital de suelos es la selección del tamaño del pixel (Hengl 2006). Para este trabajo decidimos utilizar un tamaño de píxel de 90 m de resolución espacial para seguir con las especificaciones de GlobalSoilMap.net para el mapa global de propiedades del suelo (Sánchez et al., 2009). En este trabajo se utilizaron capas ambientales con diversas resoluciones espaciales (clima y geología a 1 km, vegetación a 250m y topografía a 90m) y la armonización de esta información en un raster de 90 m de resolución espacial es otro factor que puede acarrear incertidumbre a nuestro modelo predictivo. Sin embargo, esta organización de covariables ambientales del suelo donde el clima y geología controlan la variación más general de suelos (e.g., a 1 km de resolución espacial) y el tipo de vegetación controlando una variación intermedia (e.g., a 250 m de resolución espacial) entre la topografía (e.g., a 90 m de resolución espacial) y el clima también es congruente con la interacción de factores formadores del suelo propuestos anteriormente (Jenny, 1941). La ventaja de este marco de trabajo es que tanto los datos como las covariables ambientales pueden ser actualizadas en el sistema con la finalidad de mejorar gradualmente (con nuevos datos y nuevas covariables) los resultados y mapas de COS.

Otro problema que también afecta la capacidad predictiva de los modelos tiene que ver con la multitemporalidad de los datos disponibles para representar el ambiente de formación de suelos. Para este trabajo, covariables de COS como la topografía o la geología provienen de productos relativamente actuales (Reuter and Hengl, 2012) pero estáticos, como el modelo digital de elevaciones (que no explican la dinámica temporal de la topografía y su influencia en la evolución del paisaje edáfico). Otras capas climáticas como la precipitación y temperatura representan provienen de productos como worldclim (Fick et al., 2017) que representan un periodo de tiempo (1970-2000) diferente a la década de los datos de COS seleccionada para este trabajo. Estas incertidumbres pueden propagarse a los estimados de COS.

En este trabajo ejecutamos la rutina de mapeo digital de suelos (descrita en los parrafos anteriores) en México con la finalidad de demostrar el potencial de un modelo conceptual dirigido por datos para la constante actualización de mapas de COS y el establecimiento de programas de monitoreo de COS en México. Los estimados de COS entonces, a partir de la metodología empleada en el presente trabajo, pueden mejorar a medida que se incorporen en futuros esfuerzos, a) estrategias automáticas para la selección apropiada del pixel (e.g., mayor detalle espacial y menor error predictivo), 2) estrategias dirigidas por datos para la identificación de las variables más importantes controlando la variabilidad espacial del carbono (y su sensibilidad a diversos escenarios de disponibilidad de datos) y 3) estrategias automáticas para identificar eficientemente cambios en los reservorios de COS a los largo de distintos periodos de tiempo y sus principales variables explicativas.

Forma del modelo

La media de estas capas de información se centró en 0 para reducir el riesgo de ruido en el modelo asociado a diversas dimensiones en los insumos empleados. De esta manera integramos un conjunto de variables ambientales para representar los diversos factores de formación de suelos (i.e., clima, relieve, topografía, geología y vegetación, Jenny, 1941) y generar predicciones de carbono en suelos sin información disponible (entre perfiles de suelo) de acuerdo con la ecuación (Eq. 3) y la formulación del marco de trabajo para el mapeo digital de suelos descrito previamente por McBratney et al., (2003).

$$COSxyz \sim f(MDExyz + ATxyz + Pxyz + Txyz + BHxyz + EVIxyz + Gxyz) + \varepsilon \quad \text{Eq.3}$$

Donde COS (representado por los datos disponibles entre 1999 y 2009) para un lugar determinado por las coordenadas (x,y) y a una profundidad específica de suelo (1m), puede ser representado por una función estadística (*f*) a partir de las relaciones que presenta con las capas ambientales que representan el ambiente de formación de suelos. Estas capas incluyen al MDE y sus atributos topográficos derivados (AT), precipitación y temperatura (P, T), el balance hídrico (BH), el EVI(1970-2000) y la capa de edades geológicas (G). La *f* en la Eq. 3 tomó forma de un ensamble de árboles de regresión conocido como bosques aleatorios (*Random Forests en inglés*, Breiman 2000) y Ɛ representa el error asociado a cada modelo (*f*). Random Forests es una técnica de minería de datos que permite modelar relaciones no-lineales entre la variable de respuesta (i.e., SOC) y sus variables explicativas (e.g., factores predictivos del COS). Esta técnica ha mostrado un elevado poder predictivo para el mapeo digital de suelos a escalas nacionales (Adhikari et al., 2014), regionales (Guevara et al., 2018) y globales (Hengl et al., 2017).

Selección del modelo

En este trabajo se asume que no existe un método que sea una bala de plata para generar los mejores resultados con todas las bases de datos de COS disponibles y que por tanto es necesario probar y compara la capacidad predictiva de diversos modelos o algoritmos para la predicción espacial del COS (Guevara et al., 2018). Existe una gran cantidad de métodos estadísticos para desarrollar predicciones de variables numéricas como el COS. Los modelos lineales (i.e,. regresión lineal múltiple) y los modelos basados en árboles de regresión como Random Forests son las formas estadísticas más comunes para el mapeo digital de COS (Lamichhane et al., 2019). Random Forests fue escogido para el desarrollo de este trabajo porque mostró un mejor desempeño estadístico modelando COS dado el escenario de datos disponibles en la colección de perfiles de INEGI serie 2 (INEGI, 2011), comparado con otros métodos de aprendizaje automático.

Métodos como *kknn (*vecinos cercanos ponderados por distancias y ‘kernels’ o funciones de forma de la distribuciones estadísticas de los datos de entrenamiento disponibles, Hechenbichler y Schliep, 2004) y maquinarias de soporte vectorial (Cortes y Vapnik, 1995), así como métodos basados en á modelos lineales generalizados (Gelman, 2009) y algorithmos impulsados por gradientes de aprendizaje automatico (e.g., G*radient Boosting*, Friedman, 2001) fueron sometidos a un proceso de riguroso de validación cruzada repetida (*V-fold cross validation*) y comparados con *Random Forests* (Breiman, 2000). Por tanto, para respaldar la selección del modelo predictivo empleado, en este trabajo se reporta el resultado del proceso de remuestreo *V-fold cross validation* aplicado a los modelos anteriormente mencionados (y sus posibles combinaciones) (Polley, 2011). Con este proceso se obtiene información sobre la sensibilidad de los modelos a variaciones en los datos disponibles e información sobre el riesgo (*V-fold cross validation risk estimate*) de incrementar el error de predicción (e.g., generalización de errores) al usar un modelo que no se ajusta de manera estable (precisa) a los datos disponibles cuando una porción de estos datos (i.e., 5 y 10% de los datos disponibles) es removida con el propósito de validar un modelo predictivo en ausencia de una muestra totalmente independiente. El *V-fold cross validation*

risk estimate fue estimado usando el paquete *Super Learner* de R (SL, Polley, 2011). Con este paquete podemos encontrar combinaciones optimas de diversos modelos predictivos basados en ponderaciones asociadas con el error promedio de cada modelo usando diversas formas de remuestreo estadistico y validación cruzada. *Random Forests* en el paquete *Super Learner* de R requiere del paquete *ranger*, una implementación rápida de *Random Forests* (en C++) variante del paquete original *randomForest* de R (Liaw and Wiener 2002, Wright and Ziegler, 2017).

Refinamiento y verificación del modelo predictivo

Para verificar/validar nuestros modelos predictivos empleamos una técnica de eliminación recursiva de variables donde el modelo se repite muchas veces y en cada realización utiliza una combinación diferente de predictores ambientales. Este metodo esta bien explicado en trabajos previos y es comunmente usado para selección de vatiables (Guyon et al., 2002). Con este método podemos obtener información acerca de las factores de predicción que son más importantes disminuyendo los errores en el modelo predictivo (Kohavi y John 1997). En cada modelo el predictor menos informativo (usando como indicador el error derivado de la validación cruzada de cada modelo *Random Forest*) queda afuera de la siguiente interacción hasta encontrar la combinación mínima de predictores ambientales que minimizan el error (e.g., error medio cuadrático, error medio absoluto) y maximizan la varianza explicada (r^2) de cada predicción. La estimación de medidas de desempeño (errores y r^2) se llevó a cabo empleando una técnica de validación cruzada con particiones de datos 80/20% para entrenar y validar los modelos predictivos.

Para obtener una medida espacialmente explícita de la incertidumbre fueron generadas predicciones a los datos de validacion (no se usaron en el modelo) y residuales independientes del modelo predictivo fueron calculados. Estos residuales fueron interpolados empleando el mismo algoritmo (*Random Forest*) y las mismas variables explicativas para generar un mapa de errores el cual expresamos en porcentaje promedio de error (e.g., COS / Ɛ*100) para facilitar su lectura e interpretación.

Diversas combinaciones de predictores fueron comparadas. Este proceso se repitió 5 veces (i.e, *repeated 5 fold cross validation* en inglés) para considerar la varianza en las predicciones asociada a combinaciones diferentes de datos y predictores ambientales. Una vez estimadas las medidas de desempeño del modelo (RMSE=raíz cuadrada del error medio cuadrático, r^2= varianza explicada, MAE=error medio absoluto), fue generada una predicción a 90 m de resolución espacial a lo largo de la República Mexicana, por estado, y usando en paralelo los recursos computacionales disponibles.

Resultados

Los datos de COS estimados muestran valores entre 0.03 a 135 kg m^2 a un metro de profundidad de suelo mineral, con un valor medio de 1.22 y una mediana de 0.67 kg m^2, indicando un sesgo hacia la derecha consistente con la distribución estadística de los datos disponibles (Figure 1a). Este sesgo a la derecha de la distribución estadística disminuye con la transformación logarítmica (Figure 1b). La mayoría de los valores más altos de carbono fueron encontrados en los primeros 30 cm de suelo mineral, hacia el sureste del país, mientras que las áreas áridas y semiáridas del centro norte del país mostraron los menores

valores de COSs (Figura 1c).

Valores de COS

Los valores de COS tienen una variación inversamente proporcional con la BLD a lo largo del primer metro del suelo mineral (Figure 2). La BLD muestra valores entre 0.68 y 1.6 gr/cm^3, un valor medio de 1.52 y una mediana de 1.55 gr/cm^3. Los CFR, que van de 0 a >80%, presentaron una media de 30 y una mediana de 20% y los valores disponibles muestran una clara disminución con la profundidad de manera abrupta después de los primeros 25 cm de suelo mineral (Figura 2).

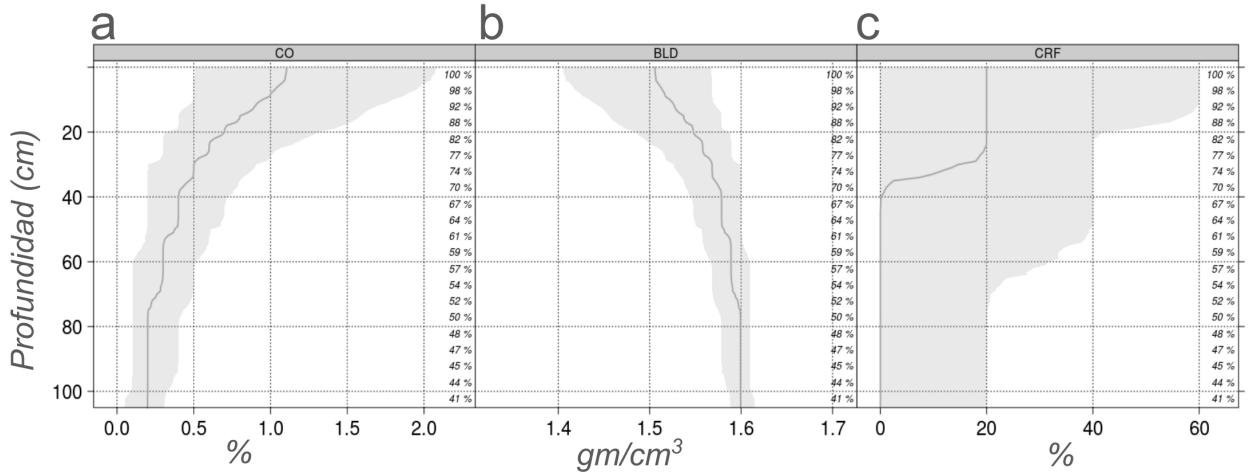
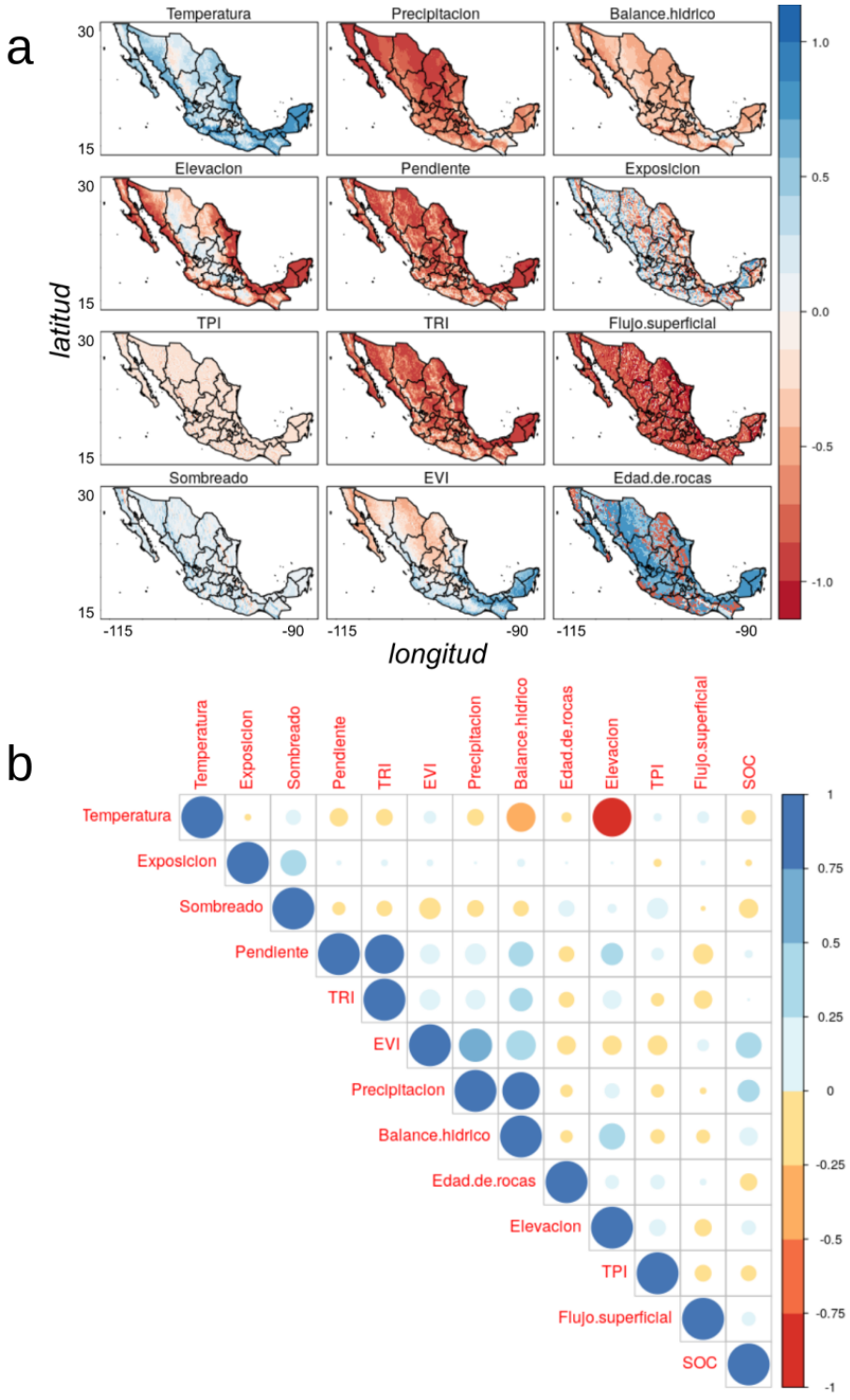


Figura 2 Variación de los datos de concentración de CO (eje x) con la profundidad del suelo mineral hasta 1.05 m (eje y) (a). Se muestran las relaciones entre la profundidad con los datos estimados de densidad aparente (BLD) (b) y con los datos disponibles de fragmentos rocosos (CFR) (c).

Correlacion entre COS y sus predictores

Los factores de predicción (Figura 3a) que mejor se correlacionaron positivamente con los datos de COS estimados fueron el EVI (r=0.37), la precipitación (r=0.27) y el BH (r=0.18). De manera opuesta (i.e., correlaciones negativas), las mejores correlaciones entre COS fueron con el índice de sombra (r=-0.20), el índice de posición topográfica (r=-0.13) y la temperatura (r=-0.11) (Figura 3b). Estas correlaciones son estadísticamente significativas y muestran valores de probabilidad cercanos al 99% ($p < 0.001$). En relación a su posición geográfica, los datos disponibles muestran correlaciones significativas con las coordenadas de los puntos muestreados ($p < 0.001$). Identificamos una correlación positiva con la longitud (r=0.47) y una correlación negativa con la latitud (r=-0.30), lo cual confirma que los datos estimados de COS muestran incrementos en sus valores de oeste a este y decrementos en sus valores del sur al norte de México (Figura 1c).



390 Figura 3 (a) Covariables ambientales usadas para predecir el COS en México a 90m de resolución espacial. TPI, índice de posición topográfica. TRI, índice de rugosidad del terreno. EVI, índice de vegetación (a). Escala de valores estandarizada entre -1 y 1. Correlograma (coeficiente de Pearson) entre

el COS y sus predictores ambientales (b).

Selección del modelo

El resultado de la validación cruzada aplicada a los distintos modelos predictivos usando *V-fold cross validation* y los datos de COS de la serie 2 de INEGI revela que *Random Forest(SL.ranger, Figura 4)* es el modelo que reduce al máximo el riesgo de errores (*V-fold cross validation risk estimate*) en la predicción de COS (Figura 4). El *V-fold cross validation risk estimate* muestra un error relativo mayor en otros modelos comparados con *Random Forests*. Las combinaciones de los algoritmos empleados (*Super Learner*, Figura 4) están ponderadas por el riesgo de errores de cada modelo y el mejor modelo (el que mas reduce el error, Discrete SL en Figura 4) no fue significativamente distinto al error generado por *Random Forests* (SL.ranger_All, Figura 4). Esto sugiere que esta técnica es predominantemente mejor prediciendo el COS que los otros algoritmos empleados (SL.bayesglm= modelos lineales generalizados, SL.kernelKnn_All=*kknn*, SL.ksvm_All=maquinarias de soporte vectorial, SL.xgboost_All= *Gradient Boosting*, Figura 4) dados los datos disponibles en la serie dos de INEGI (Figura 1).

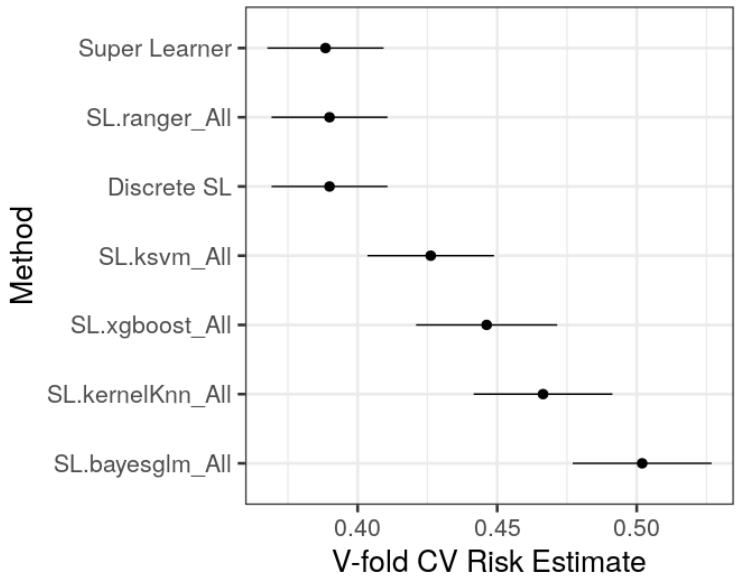


Figura 4 Gráfica nativa del paquete *Super learner* para la selección de modelos predictivos empleando el índice de riesgo relativo de error (V-fold cross validation estimate) derivado de la validación cruzada. *Super Learner* es una combinación de los modelos generadas usando el error relativo como factor de ponderación y diversas combinaciones convexas. Discrete SL representa diversas combinaciones (con diversos pesos) de multiples realizaciones del mejor modelo seleccionado por *Super learner* (*Random Forests*). *Random Forests* fue implementado con el paquete ranger de R = SL.ranger_All. Los otros algoritmos empleados mostraron un mayor error relativo (SL.bayesglm= modelos lineales generalizados, SL.kernelKnn_All=kknn,

SL.ksvm_All=maquinarias de soporte vectorial, SL.xgboost_All= Gradient Boosting.

Eliminación recursiva de variables

La capacidad predictiva de nuestros modelos alcanzó una varianza explicada (r^2) de 54±0.03 (por encima del 50 % usando solo los datos de la serie 2 de INEGI, 2011) y empleando todas las variables disponibles para representar el ambiente de formación de suelos (n=12; Tabla 1). Sin embargo, la eliminación recursiva de variables sugiere que esta capacidad predictiva puede lograrse empleando solamente la siguiente combinación de 6 variables explicativas: elevación + EVI + Precipitación + Temperatura + edad de rocas + TPI (Tabla 1).

Tabla 1 Resultados de la validación cruzada (usando conjuntos de 5% de datos para validar, 5 realizaciones de cada modelo) derivada de nuestra estrategia de modelación basada en la eliminación recursiva de variables. RMSE=raíz cuadrada del error medio cuadrático, r^2=varianza explicada, MAE=error medio absoluto, RMSESD=desviación estándar de la raíz cuadrada del error medio cuadrático, r^2SD=desviación estándar de varianza explicada y MAESD=desviación estándar del error medio absoluto. Las desviaciones (SD) representan las medidas de incertidumbre y la varianza de cada modelo asociado a variaciones en los datos para entrenar y validar.

Variables	RMSE	r^2	MAE	RMSESD	r^2SD	MAESD
1	0.81	0.09	0.63	0.03	0.03	0.02
2	0.67	0.31	0.50	0.02	0.03	0.01
3	0.62	0.40	0.46	0.02	0.03	0.01
4	0.58	0.48	0.43	0.01	0.03	0.01
5	0.56	0.51	0.42	0.01	0.02	0.01
6	0.54	0.54	0.40	0.01	0.02	0.01
12	0.54	0.54	0.40	0.01	0.03	0.01

Usando solamente las variables más importantes seleccionadas por la eliminación recursiva de predictores se obtuvo un error promedio de 0.54±0.01 kg m^2 a 1m de profundidad de suelo mineral (Tabla 1). Este valor se encuentra por debajo del primer cuartil de la distribución de los datos empleados para entrenar el modelo predictivo (3.39 kg m^2 a 1m de profundidad). La varianza explicada decrece hasta 31±0.03% y el error se incrementa hasta 0.67 kg m^2 a 1m de profundidad cuando sólo se usan las dos variables explicativas más informativas en el modelo predictivo después de la eliminación recursiva de variables (EVI + temperatura).

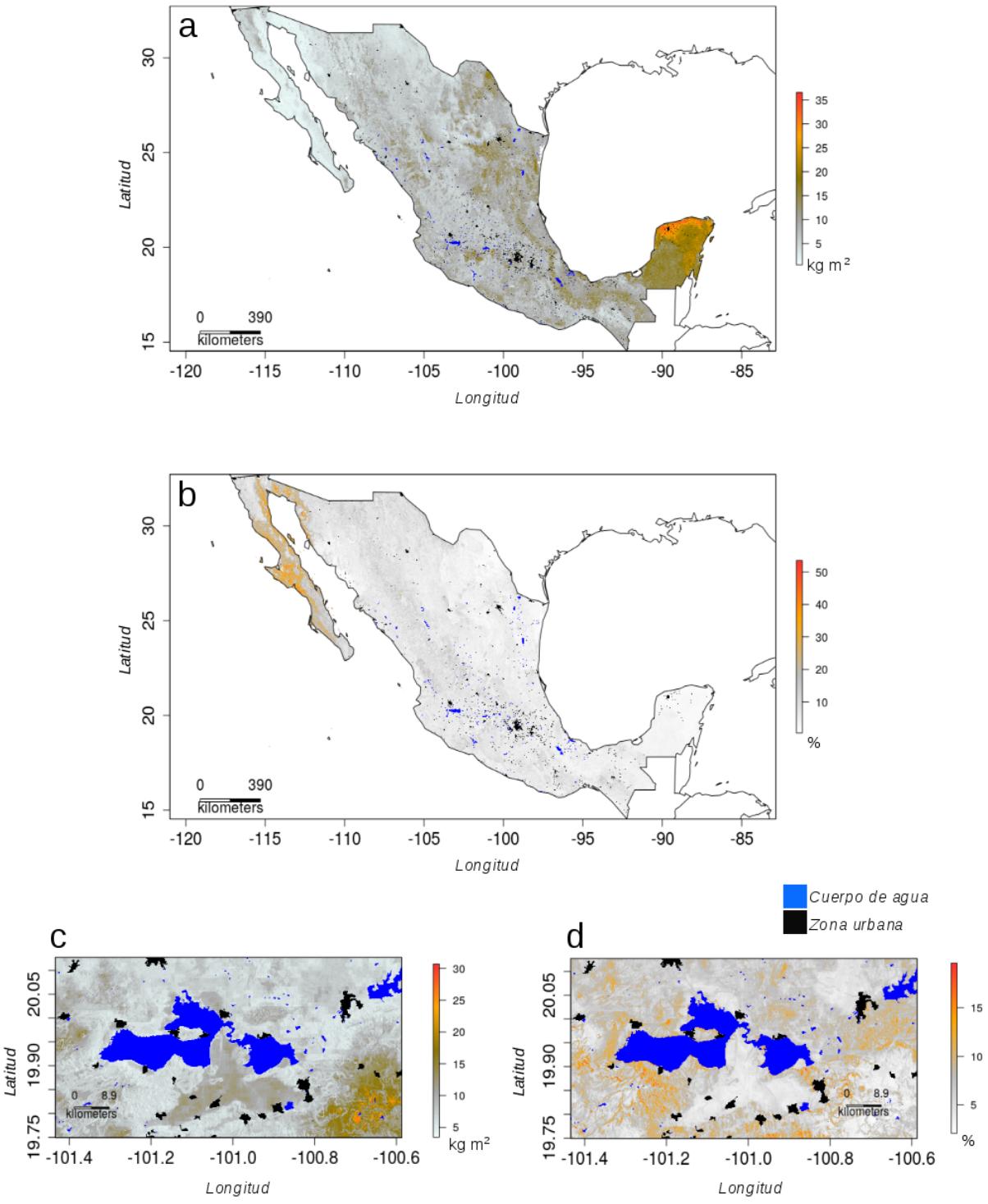


Figura 5 (a) Mapa digital de carbono orgánico en el suelo (kg m2) a 1m de profundidad y 90m de
462 resolución espacial, la línea negra representa límites entre países vecinos. (b) Ɛ , Porcentaje de errores en el modelo predictivo (mapa de incertidumbre). (c) Acercamiento, COS en la Cuenca del Lago de Cuitzeo,

al centro de México entre los límites de Michoacán y Guanajuato, para visualizar un ejemplo del nivel de detalle alcanzado mapeando el COS a 90 m de resolución espacial. (d) Acercamiento del mapa de errores de prediccion de COS en la Cuenca del Lago de Cuitzeo.

El mapa nacional de COS en México sugiere un total de 16.03 ± 4.24 Pg de COS (± 1 desviación estándar, Figura 5a). El mapa de errores interpolados (Ɛ) sugiere una incertidumbre de modelación de ± 1.68 Pg de COS, mostrando mayores incertidumbres (entre 50 y 100 %) en sitios aridos y semiaridos del noroeste del país, en la Península de Baja California y el desierto de Sonora (Figura 5b). Al nivel nacional los resultados muestran un valor promedio de 8.68 kg m^2 a 1 metro de profundidad con valores que varían de 0.27 a 38.77 kg m^2 (Tabla 2). A nivel estatal, los resultados muestran obvias relaciones entre el COS y el tamaño de los estados, pero los resultados muestran tambien variaciones importantes entre los contenidos de COS en estados con tamaños relativamente similares, pero bajo diversas condiciones geográficas (e.g., Yucatan-Sonora, Nayarit-Tabasco, Tlaxcala-Aguascalientes). La Ciudad de México, Aguascalientes y Morelos son los estados con reservorios de COS menores (Tabla 2). Gracias a su gran extensión territorial los estados Chihuahua, Coahuila y Durango mostraron los mayores reservorios de COS (Figura 6).

Tabla 2 Reporte de COS por estado. Total de COS en Pg y descripción estadística de los datos de COS modelados (media, mín, máx, desviación estándar (SD), el logaritmo del número de píxeles de 90 m modelados para cada estado (n) y la densidad del COS modelado en toneladas (Ton/Km^2).

Estado	Total (Pg)	Media (kg m^2)	Min (kg m^2)	Max (kg m^2)	SD (kg m^2)	n (píxeles)	Densidad (Ton km^2)
Aguascalientes	0.03	6.39	3.02	14.97	1.26	655329	7038.75
Baja California	0.29	3.57	0.63	18.23	2.20	9948743	3929.21
Baja California Sur	0.21	2.76	0.71	15.46	1.45	9249971	3040.89
Campeche	0.86	14.93	2.58	32.98	2.54	7095863	16453.99
Chiapas	0.63	8.69	1.44	21.99	2.79	8975054	9579.00
Chihuahua	1.51	5.84	0.98	19.12	1.71	31826863	6433.44
Coahuila	1.37	8.35	0.86	38.41	3.04	20280451	9196.86
Colima	0.05	8.28	3.30	18.93	1.67	690793	9129.61
Distrito Federal	0.01	8.56	2.30	32.87	3.10	198139	9435.31
Durango	0.97	7.46	1.61	32.86	2.35	15972961	8225.33

Guanajuato	0.21	6.62	3.09	19.89	1.30	3855479	7299.81
Guerrero	0.54	8.33	1.24	20.78	2.09	7966888	9177.39
Hidalgo	0.19	8.53	2.85	22.93	2.33	2797573	9402.95
Jalisco	0.56	7.33	1.63	18.59	1.65	9497597	8076.65
México	0.19	8.21	3.20	26.95	2.21	2871163	9046.71
Michoacán	0.49	8.26	1.65	30.70	2.61	7319411	9101.94
Morelos	0.04	7.81	3.41	16.29	1.45	626505	8607.18
Nayarit	0.21	7.49	2.31	23.47	1.95	3445650	8252.07
Nuevo León	0.63	9.29	2.13	21.55	2.58	8412272	10241.74
Oaxaca	0.77	8.62	1.84	25.08	2.75	11098742	9499.31
Puebla	0.29	8.74	2.33	25.51	2.07	4160136	9635.11
Querétaro	0.09	7.69	3.30	16.56	1.77	1439690	8471.49
Quintana Roo	0.83	18.92	2.74	32.02	3.61	5417598	20851.00
San Luis Potosí	0.62	9.65	3.14	21.46	2.33	7959124	10633.81
Sinaloa	0.37	6.67	1.34	21.18	1.81	6884320	7347.37
Sonora	0.94	4.73	0.27	22.64	2.34	24433438	5209.30
Tabasco	0.22	9.14	1.51	21.28	2.56	2923017	10074.28
Tamaulipas	0.77	9.39	1.81	23.64	2.61	10098778	10351.00
Tlaxcala	0.03	7.93	3.07	20.69	2.03	515708	8736.70
Veracruz	0.69	9.91	2.04	34.84	2.40	8539746	10919.12
Yucatán	0.87	22.74	3.17	38.77	5.24	4745119	25061.50
Zacatecas	0.55	7.00	2.27	18.96	1.72	9634081	7713.79
Nacional	16.03	8.68	0.27	38.77	4.24	250901811	8692.13

Las mayores densidades de COS a un metro de profundidad de acuerdo con el modelo predictivo generado son Campeche, Quintana Roo, y Veracruz. En contraste, estados como Baja California (norte y sur), Sonora y Chihuahua presentaron las menores densidades de COS por unidad de área (Tabla 2). Estas relaciones entre el área de los estados (representada por el número de píxeles modelados) y sus contenidos de COS son a nivel nacional significativas, con un valor r^2 igual a 0.71 (Figura 6).

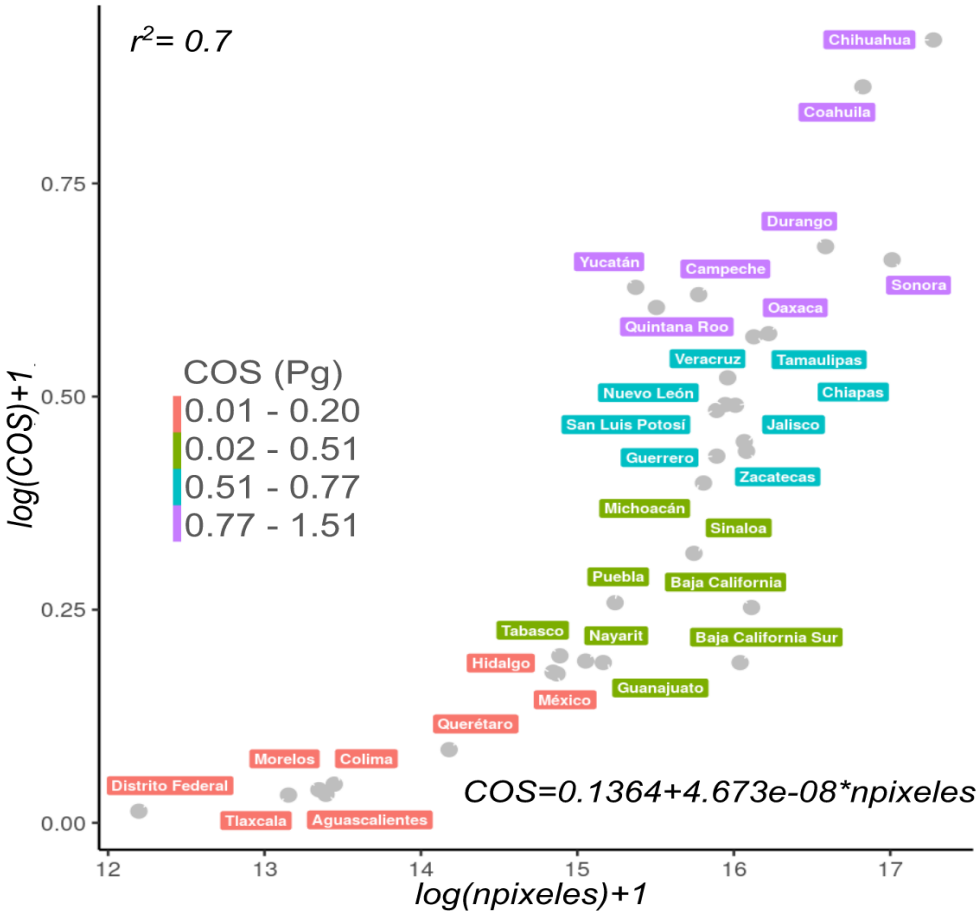


Figura 6 Relación entre el COS modelado y el área de cada estado representada por el número de píxeles (npíxeles) de 90x90 m. Los valores de COS y área se presentan en una transformación logarítmica (+1) que permitió reducir el sesgo entre las diferentes magnitudes de las variables en ambos ejes del plano. Los colores muestran una agrupación por cuantiles basada en la distribución estadística de los estimados de carbono por estado. Mostramos los coeficientes del ajuste lineal (70% de varianza explicada) entre el contenido de COS (en Petagramos) y el área (npíxeles) para cada estado.

Discusión

Fue generado un mapa digital del contenido de COS en México a un metro de profundidad y con una resolución espacial de 90 m usando datos recolectados por INEGI entre 1999 y 2009. Este mapa (Figura 5a) sigue una metodología basada en minería de datos implementada en una plataforma de código abierto (R, Core Team 2018) que permite obtener información espacialmente explícita sobre el error asociado a la predicción de COS también a 90m de resolución espacial (Figure 4b). Proponemos el uso de fuentes de información pública y metodologías reproducibles para la generación de mapas digitales y estimaciones de COS a escalas relevantes (píxeles <100m) para el manejo de ecosistemas y el desarrollo de políticas

públicas. El mapa de COS generado en este trabajo representa una herramienta de análisis que puede ayudar a resolver la actual discrepancia entre los distintos estimados de COS existente a diversas escalas espaciales (Tifafi et al., 2018, Guevara et al., 2018), así como reducir la incertidumbre en los pronósticos de COS en relación con el cambio ambiental global (Crowther et al., 2016, van Gestel et al., 2018).

Contenidos de COS

Para el periodo de análisis (1999-2009), el COS estimado a 1 metro de profundidad en este trabajo (16.03 ± 4.24 Pg) es un valor conservador comparado con reportes previos (Lajtha et al., 2018). La baja disponibilidad de datos de BLD o la naturaleza categórica de datos disponibles de CFR pueden ser la causa principal del sesgo en los diversos estimados de COS (Poeplau et al. 2017). Generar marcos de trabajo reproducibles para contar con estimados precisos de COS a 1m de profundidad es una prioridad (para México y otros países de Latinoamérica) ya que existe una discrepancia en estimados disponibles que va de 6 a 18 Pg de COS, sólo en la superficie del suelo (e.g., 0-30 cm, Guevara et al., 2018). Una síntesis para México, empleando diversas fuentes de datos (i.e., puntos, polígonos, imagenes) de carbono a 1 m de profundidad reporta 9 Pg para la superficie (0-30 cm de suelo mineral) y 14 Pg para el perfil de 0 a 100 cm (Paz et al., 2016). Esta varianza es un factor común en muchos países y a la escala global; por tanto, es imperante reducir la incertidumbre en los estimados actuales de COS para mejorar nuestro conocimiento sobre la contribución del COS al ciclo global del carbono.

El reto constante para reducir la incertidumbre en los estimados de COS es incrementar la resolución espacial y temporal de las predicciones e incrementar la precisión y exactitud de los modelos generados. Para esto es importante contar con sistemas de inferencia de suelos interoperables y replicables (Vargas et al., 2017). Una fuente de discrepancia que afecta a los modelos de variabilidad espacial del COS es la resolución espacial de los estimados disponibles (Hengl et al., 2006, Shangguan et al., 2017) . Otros estudios sugieren que el COS puede ser subestimado debido al uso de covariables (e.g., topográficos) derivadas a partir de resoluciones espaciales gruesas (e.g., píxeles de 1x1 km, Chen et al., 2018). Comparado con estimados previos a resoluciones más gruesas (Cruz-Cárdenas et al. 2014, FAO e ITPS, 2018) o de carácter global (Hengl et al., 2017), nuestros resultados proveen un nivel de detalle espacial de 90 m (Figura 5c) que puede contribuir con varios esfuerzos como: a) GlobalSoilMap.net para el mapa global de suelos (Sanchez et al., 2009, Arrouays et al., 2017); b) iniciativas de las Naciones Unidas para combatir la degradación de tierras (FAO, 2017); y c) reducir la incertidumbre en los pronósticos de cambio climático asociados al ciclo global del carbono (Crowther et al., 2016, Walsh et al., 2017).

Desempeño estadístico

Como línea base de incertidumbre, nuestro mapa digital de COS está asociado con un reporte de desempeño estadístico y un mapa de erroes (Ɛ) que captura la varianza de diversas combinaciones de datos y parámetros del modelo predictivo. Monitorear el desempeño estadístico de los modelos de variabilidad espacial del COS bajo distintos tratamientos de datos (i.e., logaritmo, no logaritmo) y combinaciones de factores predictivos representativos de múltiples periodos de tiempo o colectados con diversos métodos es requerido para reducir la incertidumbre de los múltiples estimados actuales de COS

(Lagacherie et al., 2019). Esto es porque las diversas medidas de desempeño estadístico (r^2, RMSE, MAE) proveen información útil para mejorar la selección de parámetros de los modelos predictivos del COS y consecuentemente, reducir los errores asociados a sus predicciones. Nuestro modelo predictivo sugiere errores promedio debajo del primer cuartil de la distribución estadística de los datos de COS disponibles (Tabla 1), lo cual sugiere un poder predictivo capaz de capturar la distribución estadística de los datos de entrenamiento con una varianza explicada por encima del 50%. Con 6 variables explicativas (elevación, EVI, Precipitación,Temperatura, edad de rocas y TPI), estos valores son comparables con reportes de otros países que documentan el desempeño estadístico de los modelos predictivos para desarrollo de los mapas digitales de COS a nivel global (250 m de resolución espacial, Hengl et al., 2017), continental (90 m de resolución espacial, Rossel et al., 2014) y nacional (30 m de resolución espacial, Adhikari et al., 2014).

Reto computacional

Generar mapas de alta resolución espacial a nivel región-país se convierte en un reto computacional a medida que el área de interés incrementa. Es común que los mapas digitales del suelo a escalas detalladas (i.e., pixeles <1x1km) de áreas de gran extensión territorial se desarrollen en sistemas de cómputo de alto rendimiento con infraestructura y mantenimiento costoso pero que facilitan el manejo de grandes bases de datos (Chaney et al., 2019). Sin embargo, las instituciones mexicanas que tienen el mandato de generar y actualizar información relacionada con el suelo y sus funciones (e.g, INEGI, Comisión Nacional Forestal) no necesariamente cuentan con estos sistemas computacionales. Sin embargo, estas instituciones generalmente albergan la mayor cantidad de información de campo para generar y validar modelos de variabilidad espacial de COS a escalas nacionales (Krasilnikov et al., 2013). Por lo tanto, es indispensable generar protocolos para análisis de datos y mapas digitales de COS con bajo costo computacional y con recursos computacionales disponibles en la mayoría de las instituciones interesadas en generar este tipo de información.

Por ejemplo, para analizar tendencias del COS asociadas a cambios de uso de suelo o clima a escalas detalladas (i.e., píxeles de 90 m) en México, es necesario resolver primero el reto computacional para la construcción de un modelo predictivo preciso y de un algoritmo de predicción eficiente (i.e., rápido) en los 2 millones de km^2 del país. Con esto podremos agilizar la obtención periódica de resultados y proveer una línea base para habilitar sistemas nacionales de monitoreo de COS. Hoy en día existen recursos computacionales de libre acceso para el análisis de datos geográficos como aquellos provistos por *Google Earth Engine* (Gorelick et al., 2017). Aunque existen retos asociados a la transferencia eficiente de datos (depende de una buena conexión a interne) y requiere un lenguaje de programación de alto nivel llamado *javascript*, esta plataforma ha demostrado ser eficiente para el mapeo digital de suelos con grandes bases de datos (Padarian et al., 2015). En este trabajo un reto principal fue manejar grandes bases de datos (i.e.., matrices de n píxeles por n predictores) usando R en computadoras convencionales. Por tanto nuestros resultados representan un ejemplo de otra posibilidad para analizar la compleja variabilidad del COS al nivel nacional usando una plataformas de código abierto y fuentes públicas de información ambiental.

En este trabajo fue empleada una estrategia de modelación a nivel nacional empleando 6 núcleos de

procesamiento de una computadora portátil en paralelo (i.e., núcleos trabajando al mismo tiempo), lo cual permite llevar a cabo la eliminación recursiva de variables y simplificar el tiempo de procesamiento sin agotar la memoria disponible para cada núcleo en el procesador de la computadora portátil. Además, el modelo predictivo se dividió por estados para maximizar el uso de la memoria (16 MB RAM), la cual se hubiera saturado si se intentara predecir para toda el área de interés (i.e., los 2 millones de Km^2 del país) al mismo tiempo. Los estados de Chihuahua, Sonora y Coahuila resultaron ser demasiado grandes para generar las predicciones a 90m en una sola pieza; por tanto, estos estados fueron subdivididos en dos archivos del mismo tamaño. Usando este protocolo es posible generar el mapa de COS para México a 90 m de resolución en 24 horas con el sistema de cómputo descrito previamente. El avance computacional para las predicciones de COS representa una herramienta que puede ayudar con aplicaciones para el manejo de ecosistemas terrestres (FAO, 2017) y para caracterizar la calidad del suelo (Bünemann et al., 2018). Una implicación directa de este trabajo es la posibilidad de habilitar en monitoreo del suelo a partir de la producción masiva de mapas digitales de propiedades biofísicas del suelo (como el COS) que puedan ser mejorados a medida que nuevos datos y covariables (representativas de los factores de formación del suelo a profundidades específicas) sean disponibles.

Retos e implicaciones

En este trabajo se propone que una forma de resolver el reto computacional requerido para hacer predicciones de COS a nivel nacional, a 90 m de resolución espacial, pero reconocemos que la información generada sobre COS no es libre de errores (Figura 5b). Es importante conocer la magnitud e ubicación de estos errores porque pueden propagarse en futuras aplicaciones de los productos de COS generados, sobre a escalas detalladas (Figura 5d). Nuestro mapa representa una línea base del periodo 1999-2009, determinado por los datos de entrenamiento disponibles durante este periodo de tiempo. Estos datos de entrenamiento, podrían ser no representativos de condiciones actuales en áreas sometidas a cambios de uso de suelo recientes (después de 2009), pero este es el caso para cualquier análisis usando datos de suelo patrimoniales (Mayr et al., 2010, Sulaeman et al., 2013, Karunaratne et al., 2014).

Con este trabajo se pretende incrementar la interoperabilidad en los diferentes sectores interesados en el COS. Las barreras de interoperabilidad se han descrito como conceptuales, organizacionales, tecnológicas y culturales (Vargas et al., 2017). Este trabajo busca incrementar la interoperabilidad para el entendimiento del COS en México al reducir barreras conceptuales y tecnológicas. Primero, propone un marco conceptual para estimar el COS en México (i.e., usando Geomorfometria y covariables relacionadas a los factores de formación del suelo). Segundo, se propone una metodología usando sistemas computacionales de bajo costo para incrementar la resolución espacial de las predicciones de COS y tratando de reducir el error asociado a los modelos predictivos. Cabe destacar que el error asociado a los modelos predictivos proviene de las imperfecciones en los datos de COS disponibles y en el uso de distintos tipos de insumos, con diversas resoluciones espaciales y temporales (Heuvelink, 2018). Así que un reto fundamental es reducir barreras culturales de interoperabilidad en México ya que se necesitan metodologías transparentes, fuentes públicas de bases de datos de fácil acceso y sistemas de códigos computacionales abiertos para avanzar en el cocimiento del COS y maximizar la información para mejorar el uso, manejo y conservación de los recursos naturales de México.

Conclusiones

Fue generando un mapa digital del COS en México a 1 m de profundidad de 90 m de resolución espacial y representativo del periodo 1999-2009. Estimamos un total de 16 Pg de carbono modelado en más de 250 millones de píxeles lo largo del territorio mexicano. Este estimado de COS fue generado en una computadora portátil con 16 Gb de memoria usando 6 núcleos (de 15) de procesamiento en paralelo. El tiempo de procesamiento para los casi dos millones de kilómetros cuadrados de México tarda aproximadamente 5-7 horas generando una predicción en una base estatal.

El modelo predictivo del COS a nivel nacional para el periodo de tiempo analizado (>50% de varianza explicada, 1999-2009) es reproducible a medida que nuevos datos o nuevas covariables estén disponibles. La metodología empleada permite obtener una medida de error de modelación que se puede ser monitoreada y asimilada en múltiples experimentos de mapeo digital de suelos (e.g., con diferentes combinaciones de datos y covariables para entrenar modelos representativos de distintos periodos de tiempo) con la finalidad principal de mejorar la calidad de la información generada y habilitar el monitoreo del COS a nivel nacional.

Por tanto, el mapa digital de COS generado en este trabajo representa una herramienta que puede ayudar en el desarrollo de información para la gestión, formulación e implementación de políticas públicas relacionadas con el potencial natural de los suelos y su respuesta funcional al cambio ambiental en México.

Los códigos de trabajo y archivos (i.e., datos de COS y covariables ambientales) o instrucciones para la reproducción total o parcial de este trabajo se encuentran disponibles y bajo constante actualización en: https://github.com/DSM-LAC/MEXICO/tree/master/soc_map_mx_90m_inegi_s2_bdSaini. Estos resultados pueden ser constantemente actualizados a medida que nuevos datos, covariables de COS, o nuevas preguntas de investigación emergen con el avance de la ciencia del ciclo del carbono en México.

Agradecimientos: *M.G. agradece una beca de Conacyt para estudios de doctorado (382790). RV agradece apoyo por parte de NASA Carbon Monitoring Systems (80NSSC18K0173).*

Referencias

Adhikari, K., Hartemink, A. E., Minasny, B., Kheir, R. B., Greve, M. B., & Greve, M. H. (2014). Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark. PLoS One, 9(8), e105519. doi: 10.1371/journal.pone.0105519

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., & Domisch, S. (2019). Geomorpho90m - Global high-resolution geomorphometry layers: empirical evaluation and accuracy assessment. PeerJ Preprints. doi: 10.7287/peerj.preprints.27595v1

Arrouays, D., Richer-de-Forges, A. C., Chen, S., Saby, N., Martin, M., Libohova, Z., ...Hempel, J. (2017). GlobalSoilMap history and main achievements. Taylor & Francis, 1–4. doi: 10.1201/9781351239707-1

Beaudette, D. E., Roudier, P., & O'Geen, A. T. (2013). Algorithms for quantitative pedology: A toolkit for soil scientists. Comput. Geosci., 52, 258–268. doi: 10.1016/j.cageo.2012.10.020

Beaudette, D. E. and O'Geen, A. T.: Soil-Web: An online soil survey for California, Arizona, and Nevada, Computers & Geosciences, 35(10), 2119–2128, doi:10.1016/j.cageo.2008.10.016, 2009.

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., (1999) Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma, 91(1-2): 27-45.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. doi: 0.1023/A:1010933404324

Burrough, P., and R.A. McDonnell, 1998. Principles of Geographical Information Systems. Oxford University Press.

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., ...Yimam, Y. (2019). POLARIS Soil Properties: 30-m Probabilistic Maps of Soil Properties Over the Contiguous United States. Water Resources Research, 0(0). doi: 10.1029/2018WR022797

Chen, S., & Arrouays, D. (2018). Soil carbon stocks are underestimated in mountainous regions. Geoderma, 320, 146–148. doi: 10.1016/j.geoderma.2018.01.029

Cortes, C. and Vapnik, V.: Machine Learning, 20(3), 273–297, doi:10.1023/a:1022627411411, 1995.

Crowther, T. W., Todd-Brown, K. E. O., Rowe, C. W., Wieder, W. R., Carey, J. C., Machmuller, M. B., ...Bradford, M. A. (2016). Quantifying global soil carbon losses in response to warming. Nature, 540(7631), 104. doi: 10.1038/nature20150

Cruz-Cárdenas, G., López-Mata, L., Ortiz-Solorio, C. A., Villaseñor, J. L., Ortiz, E., Silva, J. T., & Estrada-Godoy, F. (2014). Interpolation of Mexican soil properties at a scale of 1:1,000,000. Geoderma, 213, 29–35. doi: 10.1016/j.geoderma.2013.07.014

FAO 2017. Soil Organic Carbon: the hidden potential. Food and Agriculture Organization of the United Nations Rome, Italy

FAO and ITPS. 2018. Global Soil Organic Carbon Map (GSOCmap) Technical Report. Rome. 162 pp

Fick, S.E. and R.J. Hijmans, 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology

Friedman, J. H.: machine., The Annals of Statistics, 29(5), 1189–1232, doi:10.1214/aos/1013203451, 2001.

Gelman A., Aleks Jakulin, Maria Grazia Pittau and Yu-Sung Su. (2009). “A Weakly Informative Default Prior Distribution For Logistic And Other Regression Models.” The Annals of Applied Statistics 2 (4): 1360--1383. http://www.stat.columbia.edu/~gelman/research/published/priors11.pdf

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Global Administrative Areas (2012). GADM database of Global Administrative Areas, version 2.0. [online] URL: www.gadm.org.

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., ...Vargas, R. (2018). No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. SOIL, 4(3), 173–193. doi: 10.5194/soil-4-173-2018

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V.: Machine Learning, 46(1/3), 389–422, doi:10.1023/a:1012487302797, 2002.

Hechenbichler K. and Schliep K.P. (2004) Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, Discussion Paper 399, SFB 386, Ludwig-Maximilians University Munich (http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper399.ps)

Hengl, T. (2006). Finding the right pixel size. Computers & Geosciences, 32(9), 1283–1298. doi: 10.1016/j.cageo.2005.11.008

Hengl, T., MacMillan, R.A., (2019). Predictive Soil Mapping with R. OpenGeoHub foundation, Wageningen, the Netherlands, 370 pages, www.soilmapper.org, ISBN: 978-0-359-30635-0.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ...Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS One, 12(2), e0169748. doi: 10.1371/journal.pone.0169748

Heuvelink, G. B. M. (2018). Uncertainty and Uncertainty Propagation in Soil Mapping and Modelling. SpringerLink, 439–461. doi: 10.1007/978-3-319-63439-5_14

Hijmans, R. (2017). raster: Geographic Data Analysis and Modeling. R package version 2.6-7. https://CRAN.R-project.org/package=raster

Instituto Nacional de Estadística, Geografía e Informática (INEGI) Guia para la Interpretación Cartográfica, Edafologia escala 1:250 000 serie 2. Aguascalientes México, 2011, 32pp.

http://internet.contenidos.inegi.org.mx/contenidos/Productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825231606.pdf

Instituto Nacional de Estadística, Geografía e Informática (INEGI) -Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) - Instituto Nacional de Ecología (INE). (2008). 'Ecorregiones Terrestres de México'. Escala 1:1000000. México. De forma abreviada puede citarse así: INEGI, CONABIO e INE. 2008. 'Ecorregiones terrestres de México'. Escala 1:1000000. México.

Jenny, H. (1941) Factors of Soil Formation A System of Quantitative Pedology. Dover Publications, New York, 281 p.

Karunaratne, S., Bishop, T., Odeh, I., Baldock, J., & Marchant, B. (2014). Estimating change in soil organic carbon using legacy data as the baseline: issues, approaches and lessons to learn. CSIRO PUBLISHING. doi: 10.1071/SR13081

Kohavi, R. and John, G. H.: Wrappers for feature subset selection, Artificial Intelligence, 97(1–2), 273–324, doi:10.1016/s0004-3702(97)00043-x, 1997.

Krasilnikov, P., M. C. Gutiérrez-Castorena, R. J. Ahrens, C. O. Cruz-Gaistardo, S. Sedov, and E. Solleiro-Rebolledo. 2013. The soils of Mexico. Springer. Dordrecht, Netherlands.

Kuhn M. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret

Lamichhane, S., Kumar, L. and Wilson, B.: Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review, Geoderma, 352, 395–413, doi:10.1016/j.geoderma.2019.05.031, 2019.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., & Saby, N. P. A. (2019). How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma, 337, 1320–1328. doi: 10.1016/j.geoderma.2018.08.024

Lal, R., Smith, P., Jungkunst, H. F., Mitsch, W. J., Lehmann, J., Nair, P. K. R., ...Ravindranath, N. H. (2018). The carbon sequestration potential of terrestrial ecosystems. Journal of Soil and Water Conservation, 73(6), 145A–152A. doi: 10.2489/jswc.73.6.145A

Lajtha, K., V. L.Bailey, K. McFarlane, K. Paustian, D. Bachelet, R . Abramoff, D. Angers, S. A. Billings, D. Cerkowniak, Y. G. Dialynas, A. Finzi, N. H. F. French, S. Frey, N. P. Gurwick, J. Harden, J. M. F. Johnson, K. Johnson, J. Lehmann, S. Liu, B. McConkey, U. Mishra, S. Ollinger, D. Paré, F. Paz Pellat, D.

deB. Richter, S. M. Schaeffer, J. Schimel, C. Shaw,J. Tang, K. Todd-Brown, C. Trettin, M. Waldrop, T. Whitman, and K. Wickland, 2018: Chapter 12: Soils. In Second State of the Carbon Cycle Report (SOCCR2): A Sustained Assessment Report [Cavallaro, N., G. Shrestha, R . Birdsey, M. A. Mayes, R . G. Najjar, S. C. Reed, P. Romero-Lankao, and Z. Zhu (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, pp. 469-506, https://doi.org/10.7930/SOCCR2.2018.Ch12.

Liaw A.and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Mayr, T., Rivas-Casado, M., Bellamy, P., Palmer, R., Zawadzka, J., & Corstanje, R. (2010). Two Methods for Using Legacy Data in Digital Soil Mapping. SpringerLink, 191–202. doi: 10.1007/978-90-481-8863-5_16

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M. (2009) Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma, 154(1-2): 138-152.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1), 3–52. doi: 10.1016/S0016-7061(03)00223-4

Nelson, D.W., and L.E. Sommers (1982) Total carbon, organic carbon, and organic matter. p. 539-580. In A.L. Page et al. (ed.) Methods of soil Analysis. Part 2. 2nd ed. Agron. Monogr. 9. ASA and SSSA, Madison, WI.

Bünemann, E. K., Bongiorno, G., Bai, Z., Creamer, R. E., De Deyn, G., de Goede, R., ...Brussaard, L. (2018). Soil quality – A critical review. Soil Biology and Biochemistry, 120, 105–125. doi: 10.1016/j.soilbio.2018.01.030

Padarian, J., Minasny, B., & McBratney, A. B. (2015). Using Google's cloud-based platform for digital soil mapping. Computers & Geosciences, 83, 80–88. doi: 10.1016/j.cageo.2015.06.023

Paz Pellat, F., Argumedo Espinoza, J., Cruz Gaistardo, C. O., Etchevers B., J. D., & de Jong, B. (2016). Distribución espacial y temporal del carbono orgánico del suelo en los ecosistemas terrestres de México. Terra Latinoamericana, 34(3), 289–310. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S0187-57792016000300289&lng=es&nrm=iso

Polley, E. C. and Mark: Super Learner In Prediction, Collection of Biostatistics Research Archive [online] Available from: https://biostats.bepress.com/ucbbiostat/paper266/ (Accessed 29 September 2019), 2011.

Poeplau, C., Vos, C., & Don, A. (2017). Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content. SOIL, 3(1), 61–66. doi: 10.5194/soil-3-61-2017

Powlson, D. S., Stirling, C. M., Thierfelder, C., White, R. P., & Jat, M. L. (2016). Does conservation

agriculture deliver climate change mitigation through soil carbon sequestration in tropical agro-ecosystems? Agriculture, Ecosystems & Environment, 220, 164–174. doi: 10.1016/j.agee.2016.01.005

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reuter HI, Hengl T. Worldgrids—a public repository of global soil covariates. In: Miasny B, Malone BP, McBratney AB editors. Digital Soil Assessments and Beyond—Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney: CRC Press; 2012. pp. 287–292. Available: https://doi.org/10.1201/b12728-57

Rossel, R. A. V., Webster, R., Bui, E. N., & Baldock, J. A. (2014). Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Global Change Biology, 20(9), 2953–2970. doi: 10.1111/gcb.12569

Saini, G. R. (1966). Organic Matter as a Measure of Bulk Density of Soil. Nature, 210(5042), 1295. doi: 10.1038/2101295a0

Sanchez, P. A., Ahamed, S., Carré, F., Hartemink, A. E., Hempel, J., Huising, J., ...Zhang, G.-L. (2009). Digital Soil Map of the World. Science, 325(5941), 680–681. doi: 10.1126/science.1175084

Shangguan, W., Hengl, T., de Jesus, J. M., Yuan, H., & Dai, Y. (2017). Mapping the global depth to bedrock for land surface modeling. Journal of Advances in Modeling Earth Systems, 9(1), 65–88. doi: 10.1002/2016MS000686

Stockmann, U., Adams, M. A., Crawford, J. W., Field, D. J., Henakaarchchi, N., Jenkins, M., ...Zimmermann, M. (2013). The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agriculture, Ecosystems & Environment, 164, 80–99. doi: 10.1016/j.agee.2012.10.001

Sulaeman, Y., Minasny, B., McBratney, A. B., Sarwani, M., & Sutandi, A. (2013). Harmonizing legacy soil data for digital soil mapping in Indonesia. Geoderma, 192, 77–85. doi: 10.1016/j.geoderma.2012.08.005

Thornthwaite, C. W. (1948). An Approach toward a Rational Classification of Climate. Geographical Review, 38(1), 55–94. doi: 10.2307/210739

Tifafi, M., Guenet, B., & Hatté, C. (2018). Large Differences in Global and Regional Total Soil Carbon Stock Estimates Based on SoilGrids, HWSD, and NCSCD: Intercomparison and Evaluation Based on Field Data From USA, England, Wales, and France. Global Biogeochemical Cycles, 32(1), 42–56. doi: 10.1002/2017GB005678

van Gestel, N., Shi, Z., van Groenigen, K. J., Osenberg, C. W., Andresen, L. C., Dukes, J. S., ...Hungate,

B. A. (2018). Predicting soil carbon loss with warming. Nature, 554(7693), E4. doi: 10.1038/nature25745

Vargas, R., Alcaraz-Segura, D., Birdsey, R., Brunsell, N. A., Cruz-Gaistardo, C. O., de Jong, B., ...Toledo-Gutierrez, K. P. (2017). Enhancing interoperability to facilitate implementation of REDD+: case study of Mexico. Carbon Management, 8(1), 57–65. doi: 10.1080/17583004.2017.1285177

Walsh, B., Ciais, P., Janssens, I. A., Peñuelas, J., Riahi, K., Rydzak, F., ...Obersteiner, M. (2017). Pathways for balancing CO2 emissions and sinks. Nature Communications, 8, 14856. doi: 10.1038/ncomms14856

Wilson, M.F.J., O'Connell, B., Brown, C., Guinan, J.C., Grehan, A.J., 2007. Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. Marine Geodesy 30: 3-35.

Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1-17. https://doi.org/10.18637/jss.v077.i01.

Yigini, Y., Olmedo, G.F., Reiter, S., Baritz, R., Viatkin, K. and Vargas, R. (eds). 2018. Soil Organic Carbon Mapping Cookbook 2nd edition. Rome, FAO. 220 pp.