Anexo 2:
Mapeo digital de datos públicos de pH del suelo en el centro de México empleando co-variables de libre acceso a la resolución espacial de 1km.

Mario Guevara, Carlos Arroyo y Julián Equihua

Introducción

El pH del suelo indica su nivel de acidez y su variabilidad depende de varios factores, como el tipo de roca y la intensidad de la precipitación. Entender la variabilidad del pH del suelo es importante para la planeación apropiada de cultivos, entre otras actividades productivas. De acuerdo con Cruz-Cárdenas et al., (2014), el pH del suelo en México dominante es ligeramente alcalino y neutro, estos suelos se encuentran en zonas áridas, y suelos con menor pH dominan áreas forestales templadas y tropicales. Siguiendo el modelo conceptual *SCORPANe* (Suelos, Clima, Organismos Relieve, Material parental, Edad, Coordenadas de observaciones disponibles y Error, McBratney, et al., 2003) y un marco de trabajo para combinar la regresión con el kriging para el mapeo de suelos propuesto por Hengl, (2004), el objetivo de este ejercicio es relacionar datos de observaciones de pH en la superficie del suelo en México con co-variables ambientales de fuente pública y generar un modelo predictivo que permita predecir el pH superficial en un mapa continuo de cobertura nacional a la resolución espacial de 1 km.

El modelo de regresión tomará forma de un ensamble de árboles de regresión bajo la técnica de modelación *random forest* (Breiman, 2001). En esta técnica se construyen muchos árboles de regresión, los cuales se ajustan a partir de la división sucesiva de los datos con o sin pruebas de hipótesis para generar sub grupos de datos estadísticamente homogéneos, los candidatos para cada división en cada árbol están restringidos a una muestra aleatoria del total de observaciones y al final los mejores resultados de cada árbol son promediados.

Posteriormente serán estimados los residuales (los errores) y serán modelados con Kriging ordinario para generar un mapa de residuales que será sumado a la predicción espacial *de random forest*. Además de considerar con el componente *e* de nuestro modelo conceptual, la finalidad de esto es que el usuario tenga las herramientas (el código de R) para que de acuerdo con sus necesidades pueda implementar ambos métodos (uno prácticamente sin supuestos y otro con supuestos sobre la estructura espacial y estacionalidad de la variable de interés) de manera combinada o independiente.

Entre las co-variables encontraremos variables climáticas de UNIATMOS (*Unidad de Informática para las Ciencias Atmosféricas y Ambientales, Centro de Ciencias de la Atmósfera-UNAM*), precipitación anual acumulada y temperatura media anual (Fernandez-Eguiarte et al., 2014), y una capa de balance hídrico generada con estos datos en CONABIO, basada en el método empírico de Thornwaite (1948). También se incluyen co-variables topográficas, geológicas y relacionadas con la vegetación (índices de vegetación) provenientes de percepción remota, sistematizadas por el proyecto *WorldGrids.org*, una iniciativa de ISRIC (*International Soil Resource Information Center*) para compilar información relacionada al ambiente de formación de suelos con el objetivo de proveer mapas de libre acceso y cobertura global. Lea documentación de capas en http://worldgrids.org/doku.php?id=wiki:layers.

Tabla 1 Co-variables del suelo empleadas para mapear el pH del suelo superficial

Co-variable	**Descripción**	**Unidades**
UNIATMOS		
Precipitación anual acumulada	Precipitación anual acumulada	mm
Temperatura media anual	Temperatura media anual	Grados Celsius
Balance hídrico	Indica el déficit o superávit de agua de acuerdo a la evapotranspiración potencial	mm
WorldGrids.org		
DEMSRE2	Modelo global del relieve basado en SRTM 30m	Metros
EVMMOD3	Media de los valores mensuales de las series de tiempo de EVI basado en MODIS	Sin unidades
EVSMOD3	Desviación estándar de los valores mensuales de las series de tiempo de EVI basado en MODIS	Sin unidades
GEAISG3	Edad de la geología basada en la geología superficial	Sin unidades
INMSRE3	Radiación solar media potencial derivada con SAGA GIS	kWh/m^2
L3POBI3	Unidades fisiográficas	Sin unidades
LAMMOD3	Valor medio de la serie de tiempo de LAI basado en MODIS	Porcentaje de cobertura
SLPSRT2	Mapa de pendientes derivado de DEMSRE2	Porcentaje
TDHMOD3	Valor máximo de 8 días de las series diurnas LST de MODIS	Grados Celsius
TDLMOD3	Valor mínimo de 8 días de las series diurnas LST de MODIS	Grados Celsius
TDMMOD3	Valor medio de 8 días de las series diurnas LST de MODIS	Grados Celsius
TDSMOD3	Desviación estándar de 8 días de las series diurnas LST de MODIS	Grados Celsius
TNMMOD3	Valor medio de 8 días de las series nocturnas LST de MODIS	Grados Celsius
TWISRE3	Índice topográfico de humedad derivado en SAGA GIS usando DEMSRE2	Sin unidades

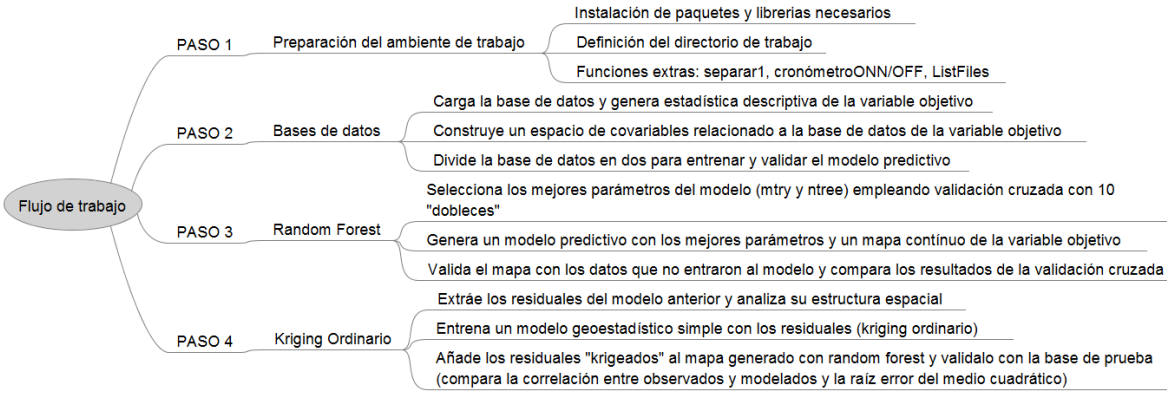
Los datos de pH son parte de la Información Nacional sobre Perfiles de Suelo series 1 y 2 de INEGI. Esta base de datos es de libre acceso en cualquier ventanilla de atención a clientes de INEGI, integra datos de variables cualitativas y cuantitativas del suelo y tiene más de 4,000 perfiles con datos de laboratorio.
http://www.inegi.org.mx/geo/contenidos/recnat/edafologia/vectorial_serieI.aspx
http://www.inegi.org.mx/geo/contenidos/recnat/edafologia/vectorial_serieII.aspx
Para el presente ejercicio serán empleados 467 observaciones de pH contenidas en estas colecciones de datos para un fragmento del centro de México (figura 1).

Figura 1, zona de estudio, gradiente de pH abarcando zonas ácidas del eje neovolcánico y ambientes sedimentarios y calizos del altiplano mexicano.

La zona donde el pH será modelado se incluye dentro de los límites del recuadro en línea negra, muestra un área aproximada de 25, 480 000 hectáreas. Las co-variables empleadas se incluyen como material adicional. Esta información fue transformada a una proyección Conforme Cónica de Lambert. Para R es necesario que el usuario se asegure que todas las bases de datos espaciales se encuentren en el mismo sistema de coordenadas, así como la misma extensión geográfica (número de pixeles). Esto puede ser verificado de manera simple en SAGA GIS y en R (lea la documentación de los paquetes *rgdal, raster, sp y maptools*). El código está preparado para un archivo tipo *shapefile* que contiene la información de las observaciones de INEGI como los entrega. Sin embargo puede adaptarse a cualquier conjunto de datos y co-variables.

Flujo de trabajo:

Figura 2 Esta metodología sigue 4 pasos principales y se comparten datos, covariables y códigos de fuente libre para su procesamiento y reproducibilidad.



Abra el código *RutinaDocumentadaRF_RK.r* en R o en Tinn R o RStudio. Note como este código está organizado en 4 pasos principales como se muestra en la figura 2.

Paso 1: Dirija la dirección o ruta de su directorio de trabajo con el comando setwd(), como se especifica en el código (línea 7). Una vez señalado el directorio o ruta de trabajo (el cual

es el único cambio que hay que hacer al código en este paso). Copie y pegue todo el paso 1 (línea 1 a línea 48) en la consola de R. La función install.packages () solamente es necesaria correrla una vez, y cada paquete de R necesario se activa en cada sesión con la función library (). Todo aquello que siga después del símbolo # es nota y R hará caso omiso.

Paso 2: La base de datos de campo se encuentra en un formato tipo shape file, del cual leeremos en R su componente *.dbf, que contiene la tabla de información. Esta base de datos debe de contener, al menos, dos columnas de coordenadas y una columna con la variable objetivo (aquella que se quiere modelar, en este caso el pH superficial del suelo). En esta base de datos se cuenta con información por horizonte a distintas profundidades asociadas a las mismas coordenadas. Serán seleccionados a continuación aquellos datos correspondientes solamente al horizonte 1 del suelo y serán eliminadas aquellas variables que no sean el pH y sus coordenadas espaciales (líneas 56-59). Posteriormente será generado un histograma de distribución de datos y la estadística descriptiva de la variable a modelar. Los valores de los predictores a cada observación de campo serán extraídos a continuación y cuatro nuevas bases de datos serán generadas, data y net que corresponden a los datos de campo y al espacio de covariables, respectivamente; entrenamiento y prueba que corresponden al 80 y 20 % de la base de campo data. El nombre de uno de los predictores incluyendo su extensión es necesario como referencia de extracción (en este caso “BalHid.tif", línea 69), el cual genera una columna (net$BalHid<-NULL) que tiene que ser eliminada posteriormente (línea 75). Note como los valores no asignados (NA´s) fueron reemplazados por 0.001 (líneas 92 y 93)

Paso 3 Copie y pegue en su consola de R todo este paso (111-162). En este paso no tiene que cambiar nada a menos que prefiera otro valor para los dobleces de la validación cruzada (cross=10 en línea 119, k=10 en línea 125). La raíz del error medio cuadrático y la correlación entre observados y modelados (rmse y cor) son el resultado de la validación cruzada mientras que rmseTest y corTest son el resultado de la validación externa. El parámetro ntree=1000 (en las líneas 118 y 137) indican el número de árboles que se van a generar en el ensamble, la función plot(bestRF) permitirá saber si son suficientes, si se alcanza una tendencia estable de error con el incremento del número de árboles. Debido a que los mismos nombres de predictores en la base de campo (incluyendo x y y) deben aparecer en la base de covariables para la predicción (función predict), note como se genera una nueva base de datos llamada cov (línea 141) que pone en la base de covariables, los nombres de las coordenadas tal cual y como vienen en la base de campo. La línea plot(r) generará un gráfico que muestra los datos interpolados con la téncica random forest como una función de las covariables mencionadas.

Paso 4: copie y pegue el paso 4 a su consola de R, la primera parte se trata de extraer los residuales del modelo random forest y analizar su estructura espacial. Esto significa identificar si la variable objetivo (en este caso los residuales del modelo random forest) presenta autorrelación espacial y de ser así, en qué medida. Para esto será generado un variograma experimental que y los valores asociados a él serán estimados (nugget, silt y range, pepita, tope y rango) a ojo (línea 188) y luego serán ajustados automáticamente (líneas 190-192). corCVkrig y rmseCVkrige muestran la validación cruzada del kriging de los residuales. La función plot(r1) muestra el mapa de los residuales “krigeados”, el cual será sumado al primer mapa generado con random forest y por último será validado con la

muestra de datos que no entro a ninguno de los modelos (rmseTestRFRK y corTestRFRK). Compare resultado e interprete los errores haciendo uso de conocimiento experto sobre la variable de interés.

***Resultados**:*

Figura 3 Histograma de frecuencias y distribución estadística de los datos de interés:

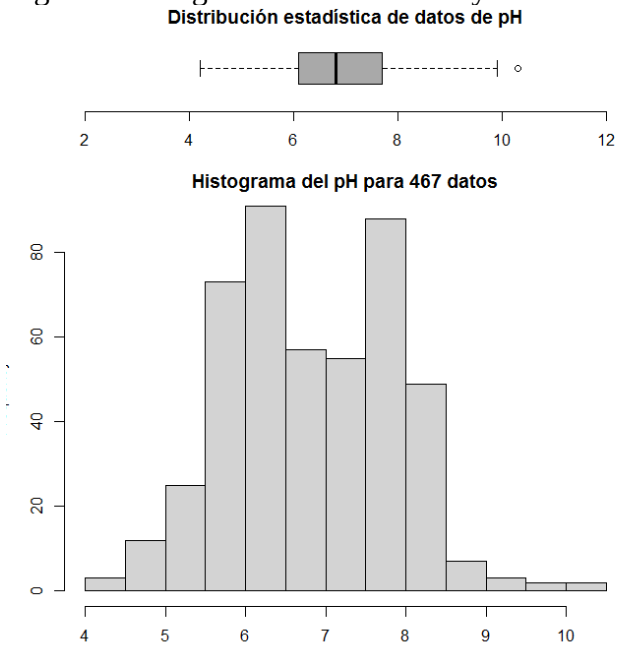


Figura 4 Distribución especial de datos para entrenar y validar el modelo predictivo

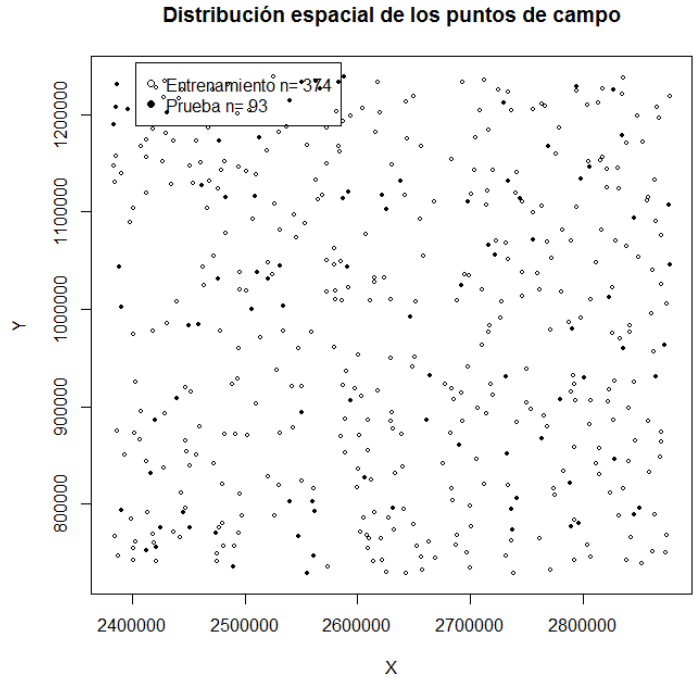


Figura 5 mapa de pH del suelo superficial 1km de resolución espacial

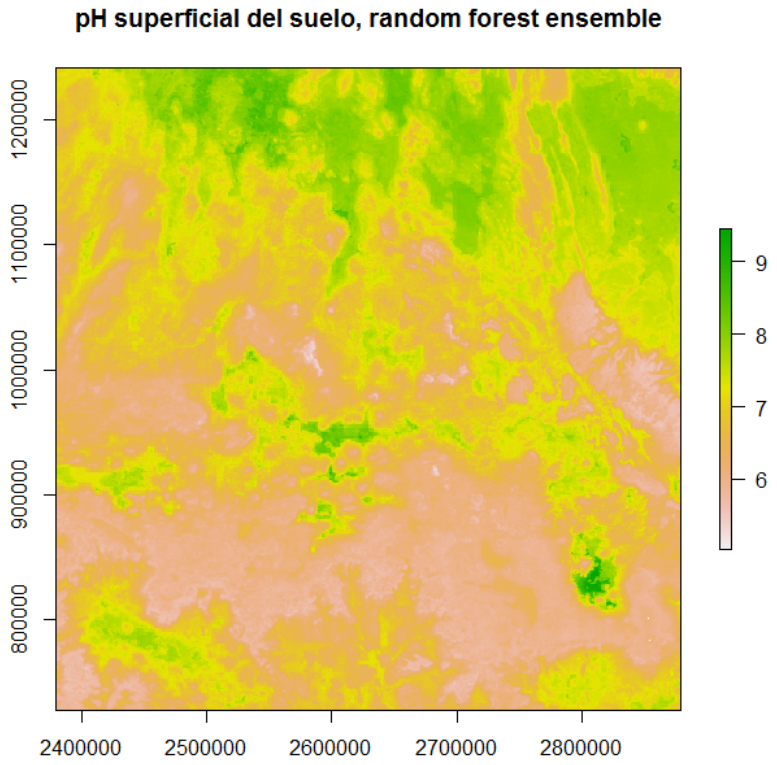


Figura 6 errores de random forest (validación cruzada con 10 dobleces)

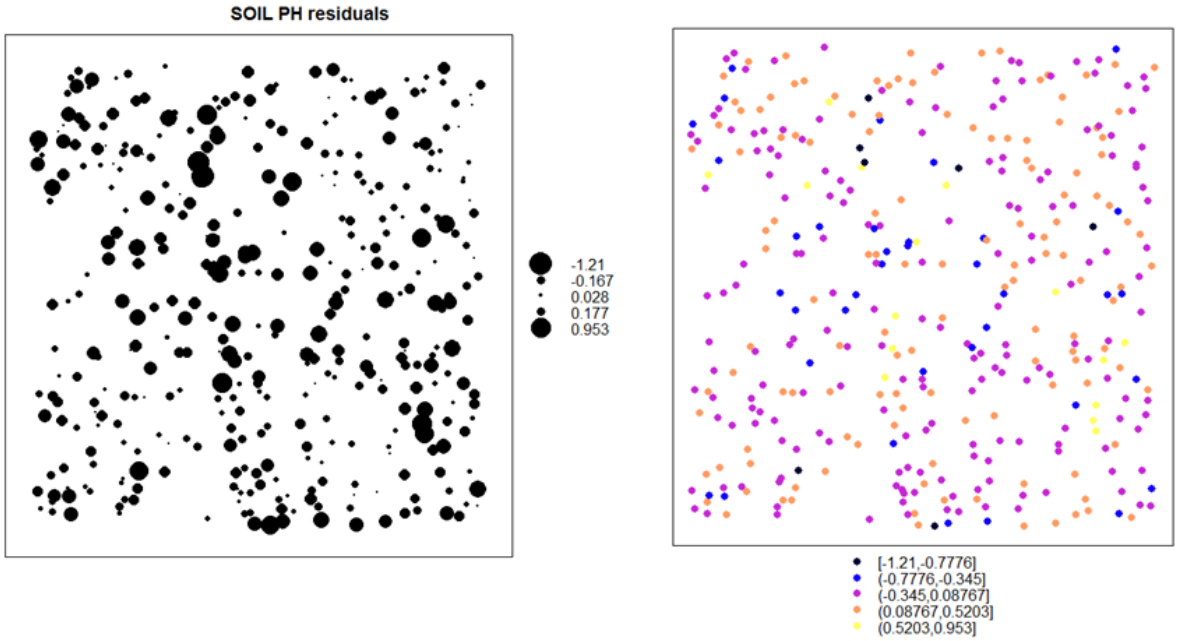


Figura 7 variograma experimental de residuales de random forest modelo exponencial

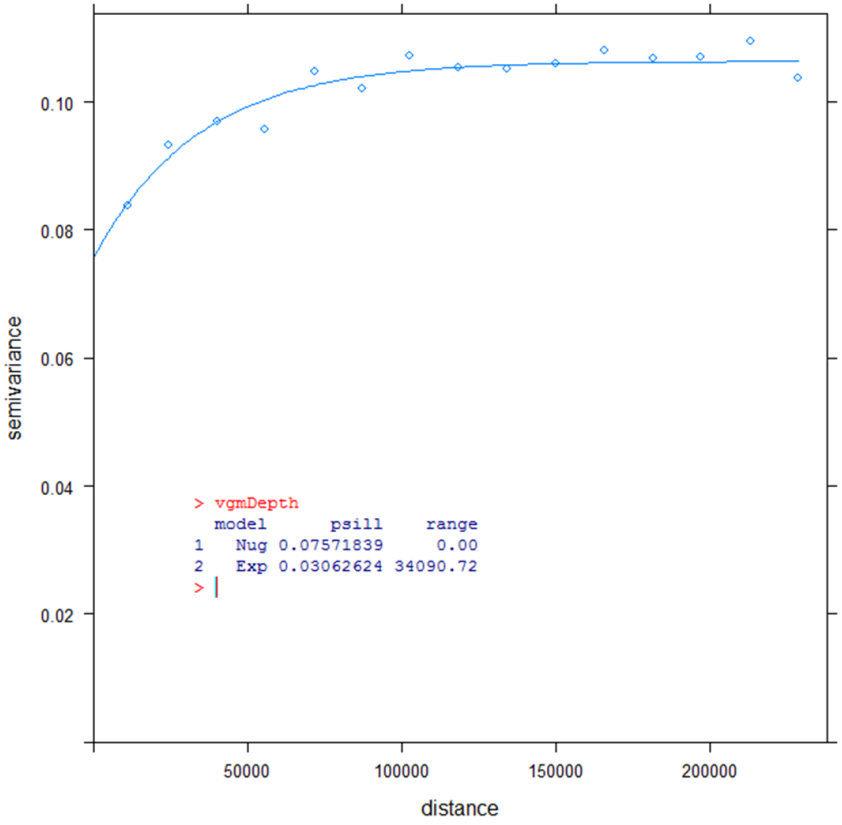


Figura 8 residuales de random forest “krigeados”

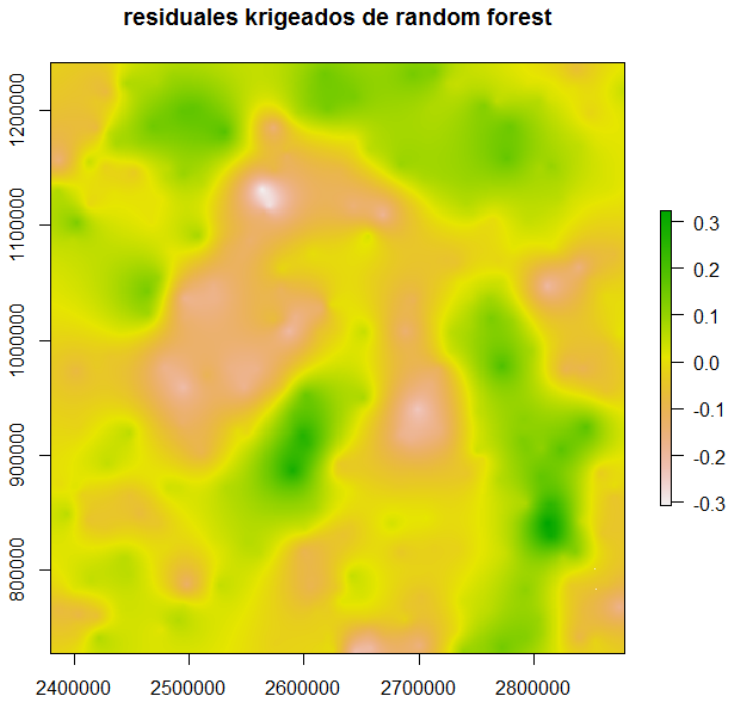


Figura 9 Mapa de pH basado en regression-kriging

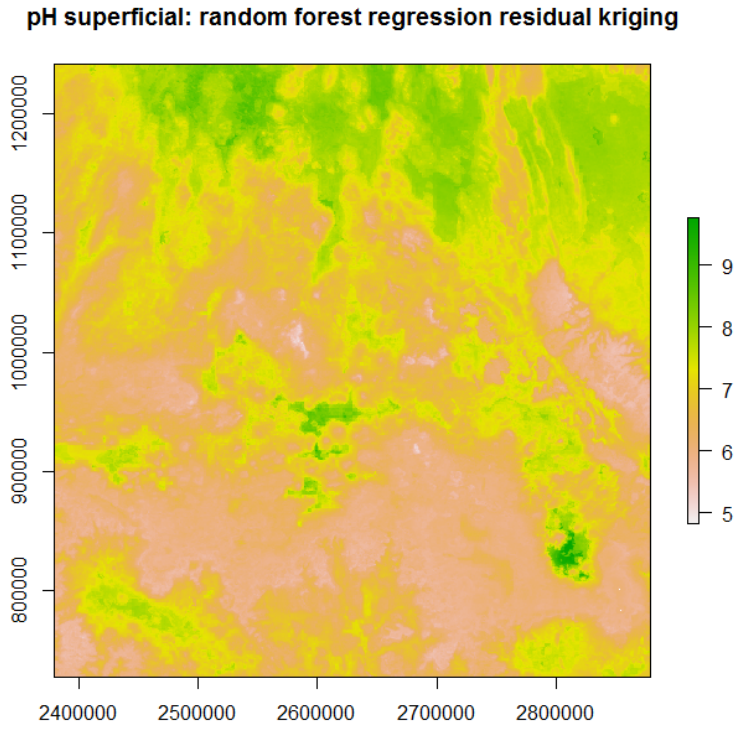


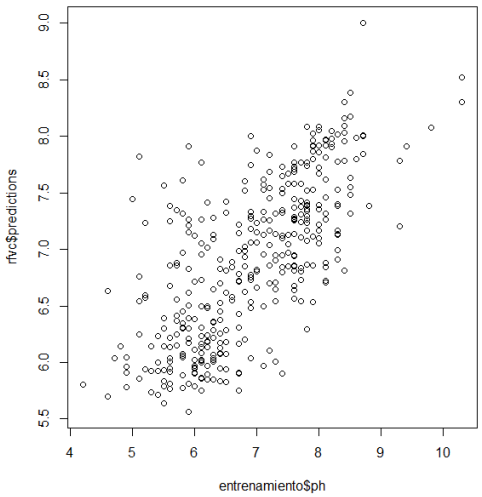
Tabla 2 validación de resultados

```
> rmse
[1] 0.7495035
> corr
[1] 0.7052123
> corTest
[1] 0.6417049
> rmseTest
[1] 0.8030157
> corCVkrig
[1] 0.1151205
> rmseCVkrig
[1] 0.3167793
> corTestRFRK
[1] 0.6554031
> rmseTestRFRK
[1] 0.7935602
```

rmse= raíz de error medio cuadrático, validación cruzada 10 folds modelo random forest, corr= correlación entre observados y modelados, validación cruzada 10 folds modelo random forest, corTest = correlación entre observados y modelados, rmseTest = raíz de error medio cuadrático validación externa validación externa, corCVkrig = correlación entre observados, validación cruzada de kriging de residuales, rmseCVkrig = raíz de error medio cuadrático, validación de kriging de residuales, corTestRFRK = correlación entre observados y modelados, validación externa de random forest mas kriging de residuales,

rmseTestRFRK = raíz de error medio cuadrático, validación externa de random forest mas kriging de residuales.

Figura 10 correlación entre observados y modelados, validación cruzada de modelo random forest (0.80).



Discusión:

Contrario a las expectativas, la suma de los residuales “krigeados” al modelo random forest no mejoró sino empeoró marginalmente el error y la correlación entre observados y modelados, quizá por la falta de autocorrelación espacial y el elevado “efecto pepita” que el variograma de residuales muestra. Es importante recordar que la falta de congruencia entre la escala 1:1 a la que se tomó el dato de campo y la gruesa resolución de las covariables (pixeles de 1km) influye en la calidad de los modelos predictivos. Otro factor importante es la baja densidad de puntos de muestreo para la gran extensión territorial del área a mapear. Aun así los resultados generados son confiables y comparables con los reportados previamente en la literatura para México (Cruz-Cárdenas et al., 2014) y para otros sitios de estudio, en Latinoamérica (Gonzalez et al., 2007), en Australia (Henderson et al., 2005) y en Europa (Reuter et al., 2008).

Conclusion:

En el presente documento (y su material adicional) se comparte una metodología para la cartografía digital de suelos en México tomando como referencia para su ejemplificación los datos patrimoniales de pH de suelo generados por INEGI y un conjunto de co-variables ambientales de dominio público.
Los métodos de inferencia espacial que se comparten son kriging ordinario, con supuestos relacionados a la estructura espacial de la variable objetivo, y random forest, un ensamble de árboles de decisión prácticamente sin supuestos y que funciona para ambos, clasificación y regresión.

Este marco de trabajo permite la inclusión de nuevos datos y nuevas covariables para mejorar la calidad de los modelos predictivos. Está diseñado para su reproducción en plataformas de código abierto. Las rutinas demuestran que es posible obtener resultados confiables con estos datos disponibles, lo cual aplica para otras propiedades contenidas en la base de datos de INEGI.

El sistema es lo suficientemente flexible como para que el usuario incluya sus propias bases de datos. En resumen, este sistema permite la armonización de bases de datos, la estimación de estadística descriptiva y autocorrelación espacial, la generación de modelos predictivos por la vía de random forest (haciendo uso de covariables) o kriging ordinario (sin covariables) y la validación estadística de resultados empleando validación cruzada y un porcentaje de datos que no tuvo que ver en la generación de los modelos predictivos.

Referencias:

Breiman 2001b. “Random Forests.” *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Cruz-Cárdenas, Gustavo, Lauro López-Mata, Carlos Alberto Ortiz-Solorio, José Luis Villaseñor, Enrique Ortiz, José Teodoro Silva, and Francisco Estrada-Godoy. 2014. 29–35. doi:10.1016/j.geoderma.2013.07.014.

Gonzalez J., A. Jarvis, S. Cook, T. Oberthur, M. Rincon-Romero, J. Bagnell y M. Bernardine 2008 Cap. 33 Digital Soil Mapping of Soil Properties in Honduras Using Readily Available Biophysical Datasets and Gaussian Processes en E. Hartemink et al. (Eds.), Digital Soil Mapping with Limited Data, Springer

Hengl, T., G. Heuvelink, and A. Stein. 2004. “A Generic Framework for Spatial Prediction of Soil Variables Based on Regression-Kriging.” *Geoderma* 120 (1): 75–93.

Henderson B., Bui E., Moren C., and Simon D 2005 “Australia-wide predictions of soil properties using decision trees” Geoderma 124, 383-398.

Fernandez-Eguiarte, A., R. Zavala-Hidalgo, and R Romero-Centeno. 2014. “Atlas Climático Digital de México”. Centro de Ciencias de la Atmósfera. Universidad Nacional Autónoma de México. http://uniatmos.atmosfera.unam.mx/ACDM/

McBratney, A. B, M. L Mendonça Santos, and B Minasny. 2003. “On Digital Soil Mapping.” *Geoderma* 117 (1–2): 3–52. doi:10.1016/S0016-7061(03)00223-4.

Reuter H., Rodriguez L., Hengl T., and Montanarella L., 2008, Continental scale digital soil mapping using European soil profile data: soil pH Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie – Heft 19 / 2008

Thornwaite C.W. 1949, An approach toward a rational classification of climate. Geographical review 38, 1, 55-94.