Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Trabajo de Especialización

Carbono en Biomasa Microbiana

Autor: Ing. Pablo Facundo Andreoni

Supervisor: Dr. Marcelo Soria

Índice

Introducción

Acerca de la problemática de estudio

Los datos a analizar se refieren principalmente a concentraciones de carbono de la biomasa microbiana (C), nitrógeno (N) y fósforo (P) del suelo, el carbono orgánico del suelo, nitrógeno total, y fósforo total a nivel de bioma y global. Esta información fue compilada en base a un amplio relevamiento de publicaciones desde finales de la década de 1970 hasta 2012 e incluye 3422 instancias provenientes de 315 trabajos de investigación científica (*papers*). Los datos corresponden a muestras de suelo recolectadas primariamente a 0-15cm de profundidad, y algunas a 0-30cm. Además, los mismos fueron compilados para concentraciones de biomasa microbiana para muestras de perfiles de suelo a profundidades de 100cm. Las coordenadas de latitud y longitud de los sitios donde se tomaron las muestras estuvieron disponibles para la mayoría de las muestras, lo que permitió ensamblar propiedades del suelo adicionales, características del sitio, distribuciones de la vegetación, biomas y datos climáticos de largo plazo desde varias fuentes globales de suelo, cobertura de la tierra, y datos del clima en general. Estos atributos del sitio se incluyeron junto con los datos de biomasa microbiana [1].

La imagen debajo (Fig. 1) muestra el área delimitada para el muestreo con un rectángulo azul.

Fig. 1. *Área correspondiente a los datos disponibles.*

Se entiende por biomasa la "materia orgánica originada en un proceso biológico, espontáneo o provocado, utilizable como fuente de energía" [2]. "La biomasa microbiana constituye el componente vivo de la materia orgánica del suelo (…). La cantidad de biomasa microbiana contenida en el suelo y los cambios estacionales sufridos por ella, van a estar influidos por la cantidad de materia orgánica del suelo, por factores climáticos, uso de la tierra y por las características físico-químicas del suelo (...) lo que la convierte en un indicador altamente sensible de los cambios sucedidos en el suelo (…)" [3].

En particular el carbono de la biomasa microbiana guarda un especial interés desde el punto de vista ecológico por cuanto es un agente potencial para funcionar como depósito de dióxido de carbono en los suelos. "Los científicos aseguran que reside más carbono en el suelo que en la atmósfera y en toda la vida vegetal combinadas; existen 2500 billones de toneladas de carbono en el suelo, comparadas con 800 billones de toneladas en la atmósfera y 560 billones de toneladas en la vida vegetal y animal. A su vez, comparado con muchas soluciones desde el punto de vista de la geoingeniería, el almacenamiento de carbono en el suelo es simple: implica restaurar el carbono a donde pertenece (…). A través de la fotosíntesis la planta toma el carbono del aire para formar compuestos de carbono. El remanente que la planta no necesita para crecer es exudado a través de las raíces para alimentar organismos del suelo, por medio de los cuales el carbono es humificado o pasado a ser estable. El carbono orgánico es uno de los componentes principales del suelo y ayuda a darle al suelo su capacidad de retención de agua, su estructura y su fertilidad (…). Existe la necesidad de encontrar oportunidades para incrementar el carbono en el suelo en todos los ecosistemas – desde bosques tropicales a pastizales y humedales – reforestando áreas degradadas, incrementando el cubrimiento (*mulching*) de biomasa en lugar de su combustión, mediante el uso a gran escala de carbón vegetal, gestión de praderas mejorada, y restauración de manglares, marismas salinas y pedreras marinas (…).Muchos científicos afirman que las prácticas de agricultura regenerativa pueden volver atrás el reloj en cuanto a pérdida de carbono, reduciendo el dióxido de carbono atmosférico impulsando a su vez la productividad del suelo e incrementando la capacidad de resistencia a inundaciones y sequías (…)" [4].

De acuerdo a lo dicho anteriormente queda de manifiesto la importancia de poder determinar los niveles de biomasa microbiana en el suelo para un correcto diagnóstico de la efectividad del mismo como agente de retención de carbono y como indicador de fertilidad, y en este sentido se dirigirá el presente estudio.

Objetivos

El propósito de este trabajo es (1) analizar la relación existente entre las variables geográficas, medioambientales y físico-químicas del suelo con la biomasa microbiana contenida en el mismo y (2) indagar acerca de la factibilidad del modelado predictivo en este sentido.

Fuente de los datos

El conjunto de datos original (de acuerdo a cita [1]) fue obtenido de la Administración Nacional de Aeronáutica y del Espacio (*NASA)* de los Estados Unidos, a través de su Centro de Archivos Activos Distribuidos para Dinámicas Bioquímicas (*ORLN DAAC*). El mismo consta, como se adelantó, de 3422 registros y unas 20 variables. A efectos de este análisis, ciertas variables fueron desestimadas (referencias, comentarios, entre otras). De esta manera, la lista completa de variables resultantes[i] es la que se muestra en la tabla debajo (en adelante serán citadas

[i] Algunas de estas variables fueron sujeto de ajuste o se crearon a partir de otras, como se detalle en la sección siguiente.

indistintamente con nombre en inglés en las figuras correspondientes a salidas de software o en español, en el cuerpo del texto).

Tabla 1 *Variables del conjunto de datos*

Nombre	**Nombre original**	**Tipo**
Bioma	*Biome*	Categórica
País	*Country*	Categórica
Latitud	*Latitude*	Continua
Longitud	*Longitude*	Continua
Elevación	*Elevation*	Continua
Temperatura Media Anual	*MAT: Mean Annual Temperature*	Continua
Precipitación Media Anual	*MAP: Mean Annual Precipitation*	Continua
Carbono en Biomasa Microbiana del Suelo	*SMBC: Soil Microbial Biomass – Carbon*	Continua
Nitrógeno en Biomasa Microbiana del Suelo	*SMBN:* Soil Microbial Biomass – *Nitrogen*	Continua
Fósforo en Biomasa Microbiana del Suelo	*SMBP: Soil Microbial Biomass – Phosphorus*	Continua
Carbono Orgánico del Suelo	*SOC: Soil Organic – Carbon*	Continua
Nitrógeno Total	*TN: Total – Nitrogen*	Continua
Fósforo Orgánico Total	*TOP: Total Organic – Phosphorus*	Continua
pH	*pH*	Continua
Profundidad – Límite Superior	*UD: Upper Depth*	Continua
Profundidad – Límite Inferior	*LD: Lower Depth*	Continua

Pre procesamiento de los datos

A efectos de garantizar la consistencia del conjunto de datos, se llevaron a cabo las tareas de revisión de rangos de valores de las variables continuas y categorías de las variables nominales.

Tratamiento de categorías

1. Se unificaron nombres de los países en mayúsculas y minúsculas (ej.: "United Kingdom" y "United kingdom") y se completaron nombres truncos (ej.: "United States of" por "United States of America").
2. Se establece como premisa que, a efectos de este análisis, "Norway" (*Noruega*) no incluye "Svalvard" y Russia (*Rusia*) no incluye Russia in Europe (*Rusia europea*).
3. Se completaron los nombres de biomas truncos (ej.: "Natural Wet") de Natural Wetland, Temperate Broadleaf Forest, Temperate Coniferous Forest y Tropical/Subtropical Forest.

Transformaciones en rangos continuos

1. Se aplicó una transformación para la elevación del terreno, de manera de obtener el promedio en los casos en que se especificaba un rango (por ejemplo "200-250m" se re expresó como "225m"), lo cual representa aproximadamente el 3% del total. Asimismo, se modificaron valores acompañados de texto aclaratorio para que fueran exclusivamente numéricos.
2. Se aplicó la misma transformación promedio del punto 1 en el caso de la precipitación media anual, siendo en este caso menos del 1% del total.
3. En lo relativo a los límites de profundidad de excavación superior e inferior, se quitó texto innecesario que acompaña las cifras, y se eliminaron valores de sólo texto que no tienen una interpretación en términos del concepto. Asimismo, en el caso del límite máximo, se eliminaron las leyendas aclaratorias en casos donde no se especificaba valor numérico, es decir el valor resultante es nulo lo cual representa aproximadamente el 2% del total. Se reemplazó la leyenda "Surface" (Superficie) por el valor 0 y el mismo valor se indicó para el límite superior.
4. En el caso del pH se elimina el valor 63,83 en el desierto israelí, el cual se considera inválido en términos de la escala de 0-14 para esta variable [5].

Tratamiento de datos faltantes

Teniendo en cuenta que la mayor parte de las variables presentan datos faltantes previo al análisis se llevó a cabo imputación en dos etapas: por búsqueda de los datos faltantes en fuentes alternativas y por medio de algoritmos automáticos basados en los datos completos disponibles.

En primer término, en lo que respecta a la variable elevación del terreno (originalmente con 791 casos con valores sobre 3422 en total) se realizó una búsqueda basada en coordenadas (Latitud y Longitud, las cuales estaban completas en alrededor del 95% de los casos) a partir de un módulo específico que disponibiliza google en forma de API: *Google Maps Elevation* [6], para completar los valores faltantes. En los casos en que el valor resultante era negativo se buscó el dato en la base de GpsVisualizer [7], para reemplazar el valor cuando fuera incorrecto (se considera que no puede existir elevación menor a los 423m por debajo del nivel del mar y al no haber trabajado con biomas acuáticos, el umbral se incrementa a 258m [8]).

En la misma línea, se operó con las variables de temperatura media anual (originalmente con 1128 casos con valores sobre 3422 en total) y precipitación media anual (con 1285 casos con valores). Para completar los valores de estas dos variables, se utilizaron los datos de *WorldClim – Global Climate Data* [9]. Los mismos consisten de un conjunto de 19 variables bioclimáticas (derivadas de valores mensuales de temperatura y precipitación, y que resultan más interesantes desde el punto de vista biológico), que representan promedios del rango temporal 1960-1990, entre las cuales se utilizaron *BIO1 = Annual Mean Temperature* y *BIO12 = Annual Precipitation*. Esta información se obtuvo con el grado de resolución más alto disponible correspondiente a 30 arco segundos equivalentes a aproximadamente 1km, que los autores generaron a partir de un método de interpolación de promedios mensuales de datos meteorológicos de varias estaciones distribuidas a lo largo y ancho del globo [10]. De esta manera el objetivo fue encontrar el dato de temperatura y precipitación para cada una de las coordenadas correspondientes a casos con datos faltantes. En este sentido, se obtuvieron los datos por partes para hacer posible su posterior procesamiento (ver Fig. 2 para el detalle de las regiones).

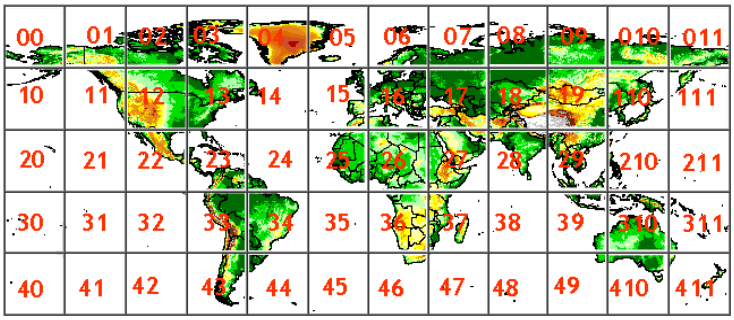


Fig. 2 *Representación gráfica de las "celdas" que agrupan datos de los espacios geográficos asociados.*

Los datos están organizados con formato grilla (*raster*), por lo que se utilizó la biblioteca raster de R para su procesamiento, consistente en hallar los valores objetivo por medio del método bilinear (el resultado se calcula como interpolación de los valores de las cuatro celdas más cercanas del raster, para un par de coordenadas dado) [11] y aplicar una sencilla transformación a los datos de temperatura (expresados en ºC x 10). De esta manera, se consiguió completar los datos faltantes de ambas variables meteorológicas, en un 91% en el caso de temperatura y un 90% en el caso de precipitación.

Los siguientes gráficos representan (Fig. 3.a) la proporción de valores perdidos de las distintas variables y (Fig. 3.b) un mapa con las combinaciones observadas de valores faltantes (en orden de mayor frecuencia de aparición de abajo hacia arriba). El color amarillo se utilizó para representar valores faltantes, y el violeta para valores definidos.

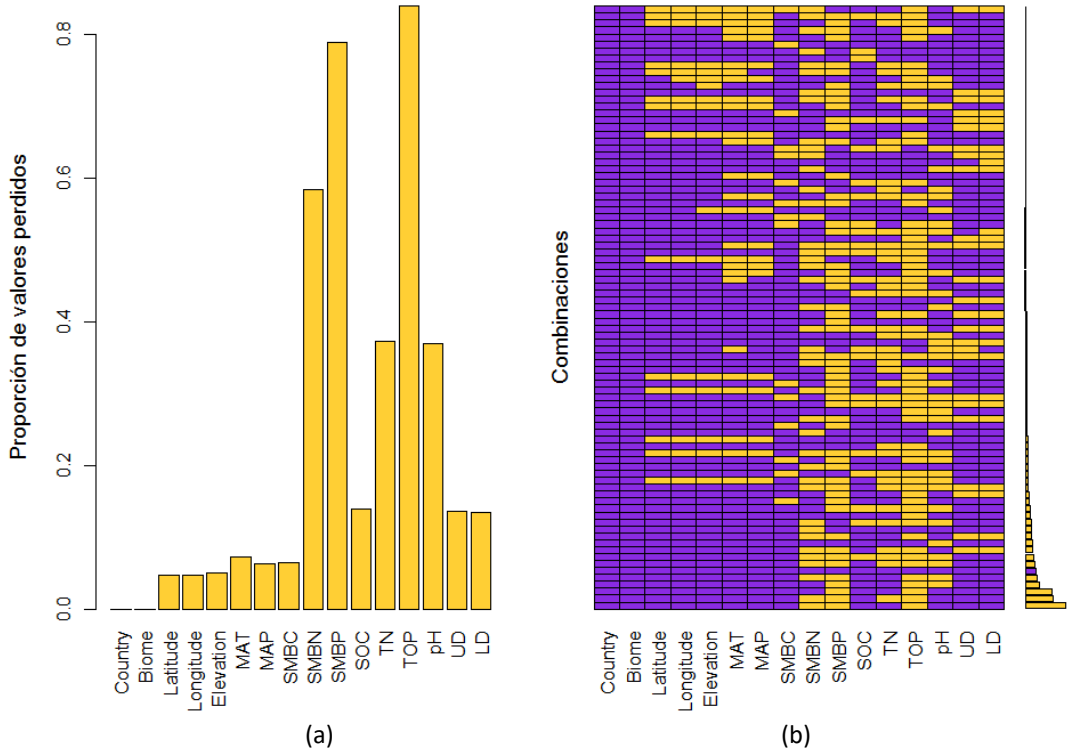


Fig. 3 *Proporción de valores perdidos (a) y combinaciones observadas en valores perdidos (b) iniciales.*

Se observa que las variables geográficas y climáticas, presentan una proporción baja de valores faltantes; mientras que las frecuencias más altas se registran para las variables de biomasa, en particular para el fósforo y el nitrógeno, asimismo para el pH.

El gráfico de la derecha muestra que la combinación de variables para la que se observa mayor frecuencia de datos faltantes. Así, SMBN, SMBP y TOP, seguida de cerca por SMBN, TN y TOP, y por SMBP y TOP son las combinaciones que más se repiten, siendo todas ellas subconjuntos de las variables con mayor proporción de datos faltantes.

En base a lo antedicho, se decide eliminar del conjunto de datos a analizar las 4 variables relacionadas a fósforo y nitrógeno.

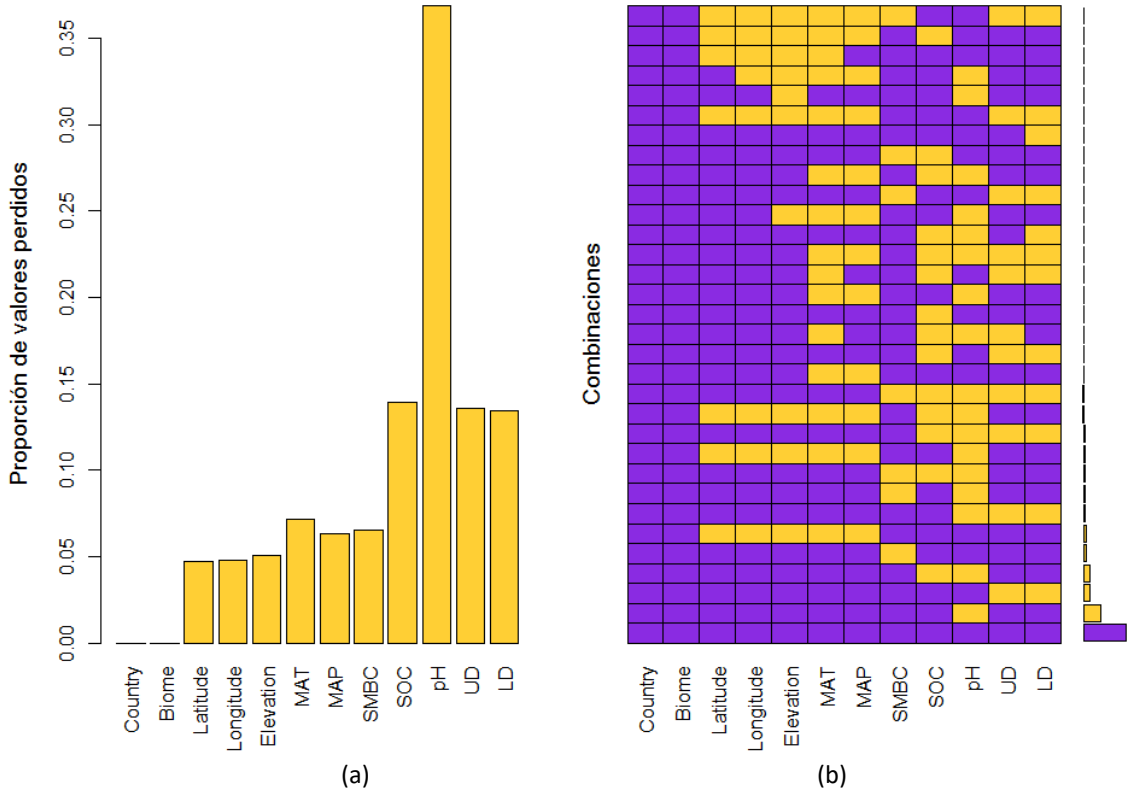


Fig. 4 *Proporción de valores perdidos (a) y combinaciones observadas en valores perdidos (b) resultantes.*

Una vez quitadas las variables, se puede notar (ver Fig. 4) que la proporción de faltantes más alta se registra ahora para el pH con alrededor del 35%, mientras que para las demás variables no se supera el 15%.

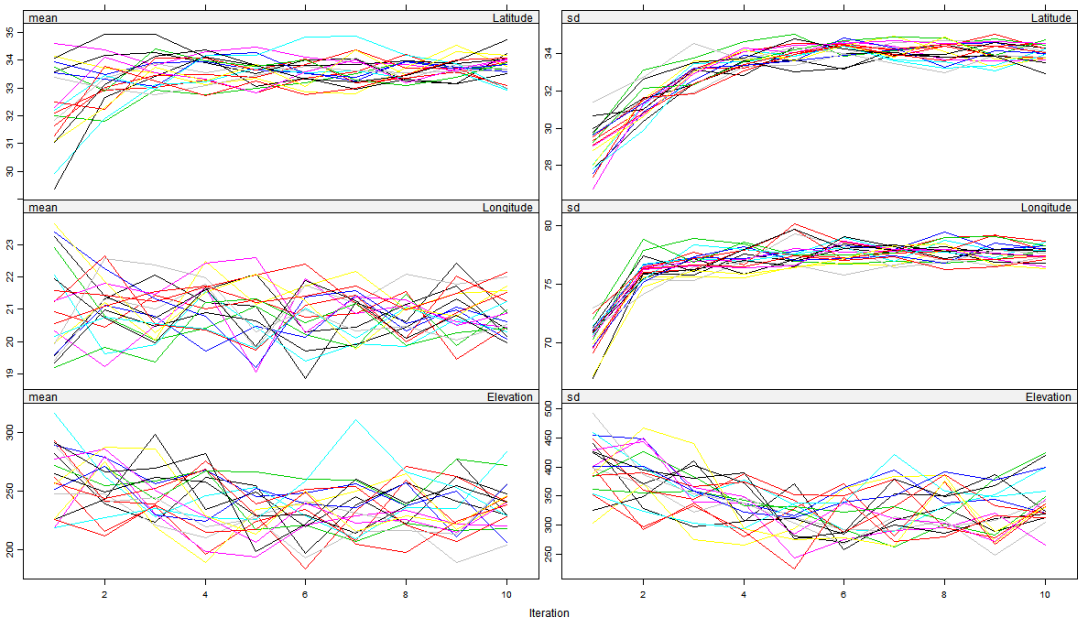
En lo que se refiere a las combinaciones, resalta el hecho que ahora el patrón más frecuente es el de datos completos en todas las variables, seguido por datos faltantes sólo en pH, en los límites de profundidad, y en SOC y pH como los más representativos. Se observa además en este gráfico, que las combinaciones observadas son de naturaleza heterogénea, y por ende no podrían identificarse patrones fuertes en términos de la asociación de variables con datos faltantes. En base a lo anterior, no habría razones para presumir que la ocurrencia de datos faltantes no sea aleatoria.

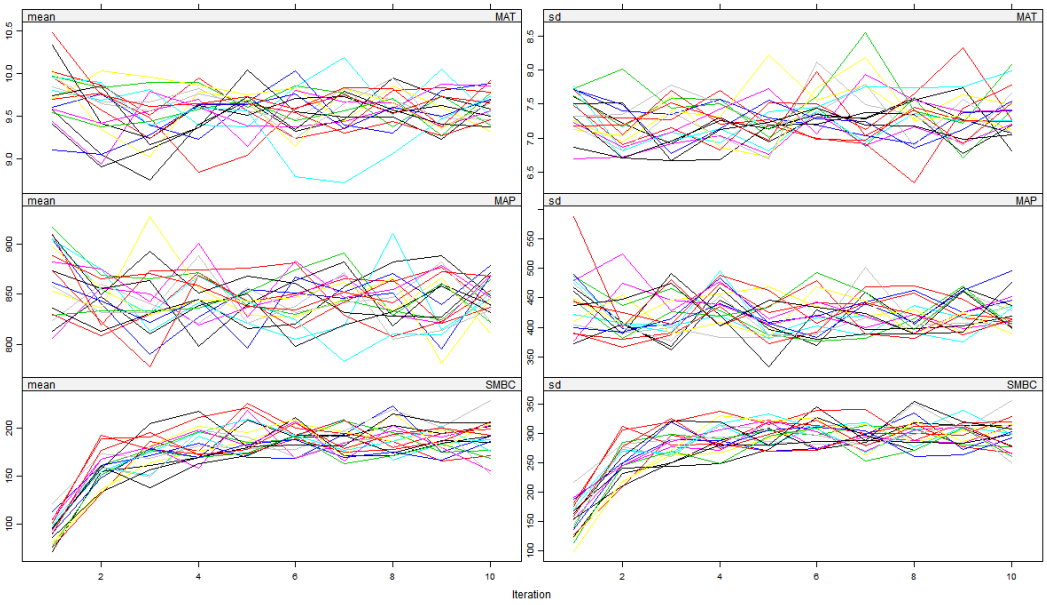
Como segundo paso, a efectos de completar los valores faltantes para permitir un análisis más completo de los datos, se eligió utilizar dos de los métodos más avanzados para problemas de datos incompletos complejos: imputación múltiple basada en ecuaciones encadenadas (*MICE* por sus siglas en inglés) [12] y *Miss Forest* [13], basado en el algoritmo de clasificación *Random Forest* [14], compararlos y determinar la opción más conveniente.

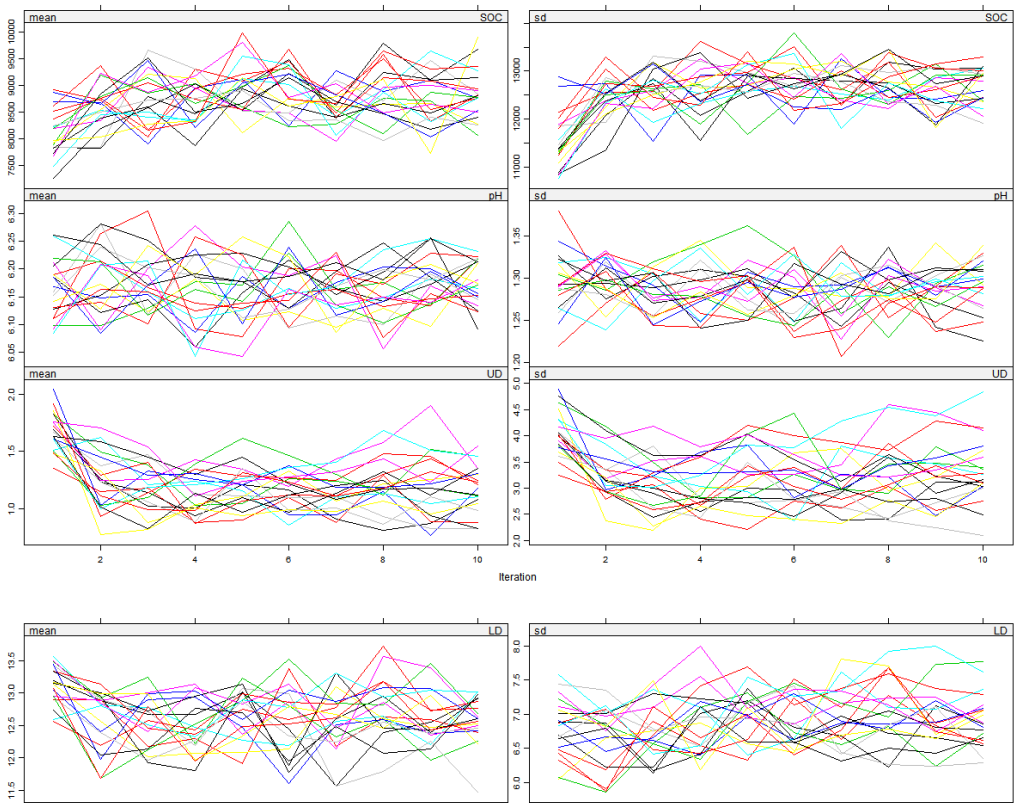
El primer método consiste en generar un modelo de regresión lineal para cada variable con valores perdidos, de manera tal que las demás variables con valores completos funciones como regresaras en cada caso. Así, el valor estimado de la regresión se imputa como valor de la variable en cuestión en base al método *predictive mean matching* (concordancia de media predictiva). Este procedimiento (estocástico) se repite n veces con un máximo de iteraciones, para generar varias

versiones de datos imputados (bajo el supuesto que esa diversidad de opciones aproxima mejor la naturaleza del dato faltante a efectos de su uso posterior para propósitos de clasificación, por ejemplo). En este caso, n se fijó en 20 y el máximo de iteraciones en 10.

El modelo generado, consistió en la utilización de todas las 12 variables previamente seleccionadas. A continuación, se presenta el gráfico que muestra la media (Fig. 5.a) y el desvío estándar (Fig. 5.b) de cada una de las variables conforme se incrementa el número de iteraciones, donde cada curva representa una de las versiones de datos imputados generadas por el modelo.







(a) (b)

Fig. 5 *Media (mean) (a) y desvío estándar (SD) (b) a lo largo de las iteraciones (iterations) de cada uno de los conjuntos generados para cada variable.*

Como se puede observar (ver Fig. 5) en todos los casos se presenta convergencia de las variables (todas las líneas se entremezclan entre sí) en términos de los dos parámetros analizados, con mayor o menor variabilidad. En el caso de elevación, MAT, MAP, pH, UD y LD las curvas se observan medianamente estables a lo largo de las iteraciones; mientras que, en el caso de latitud, longitud, SMBC y SOC, se observa una clara tendencia inicial, que conduce la convergencia hacia valores más altos que los iniciales (como resultado de la elección de valores iniciales demasiado bajos por parte del algoritmo, lo cual es particularmente notorio en SMBC).

El gráfico siguiente (Fig. 6), por otra parte, muestra las curvas de densidad para los distintos valores de las variables. Los conjuntos de datos imputados se representan con líneas magenta, y el conjunto de datos sin imputar con línea celeste.

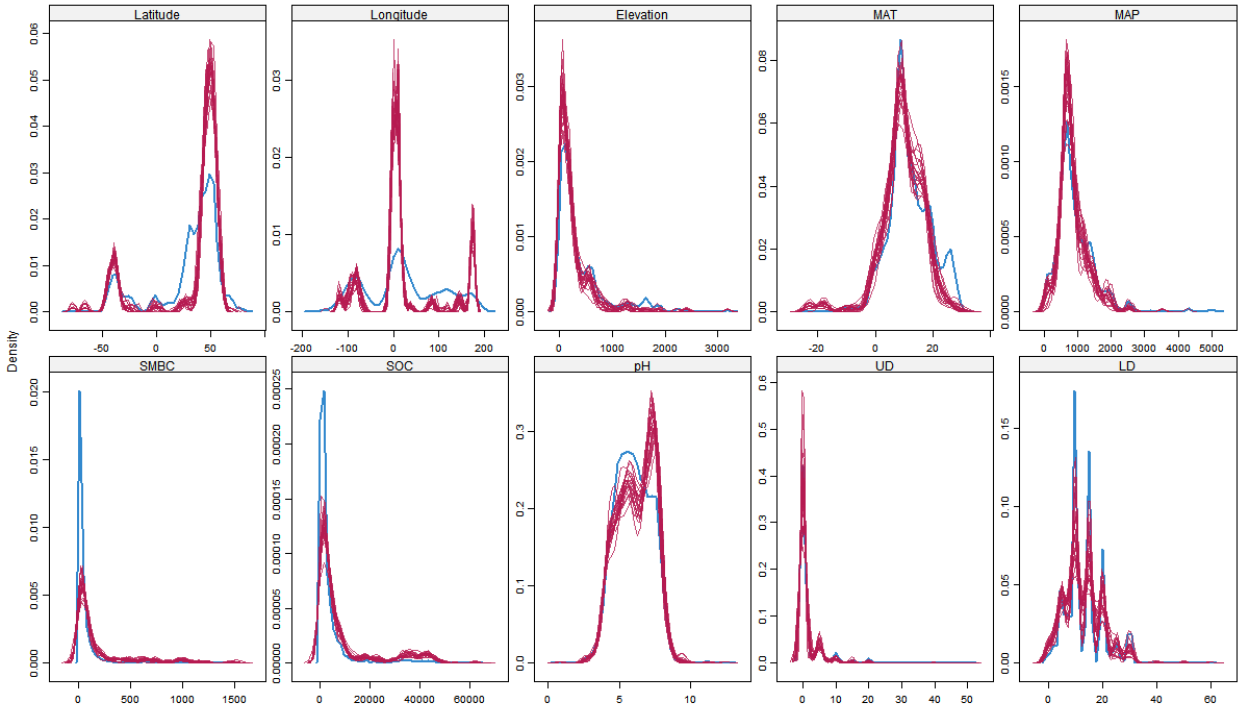


Fig. 6 *Densidad (density) para el rango de valores de las variables para los conjuntos de datos imputados con MICE (magenta) y el conjunto de datos original (celeste).*

Como se puede observar, las curvas de datos imputadas siguen mayormente la naturaleza de los datos sin imputar, exacerbando la frecuencia de los valores más frecuentes en éstos. Un caso llamativo es pH, donde este comportamiento esperado pareciera no cumplirse, existiendo un sesgo marcado hacia la mayor ocurrencia de valores en la zona de los neutros, mientras que la distribución original muestra una mayor preponderancia de los ácidos.

El segundo algoritmo evaluado, *Miss Forest* (Selvas de Perdidos), es un método de imputación no paramétrico. Para cada variable con datos perdidos, el algoritmo ajusta un modelo *Random Forest* (Selvas Aleatorias) en base a la parte observada, y predice la parte faltante. El algoritmo repite estos dos pasos hasta que se cumple el criterio de paro o se alcanza la cantidad máxima de iteraciones especificada. Este criterio de paro, consiste en evaluar las diferencias entre el resultado de las imputaciones de la iteración previa, y el correspondiente a la iteración actual, deteniendo el procesamiento tan pronto como estas diferencias se incrementen [15].

En este caso, al igual que con el algoritmo anterior, también se consideraron la totalidad de las variables preseleccionadas, y la cantidad de iteraciones se fijó en 10. El conjunto de datos imputado a partir del modelo, se acompaña de una estimación de error basada en el método *out-of-bag* (OOB por sus siglas, fuera de la bolsa) de *Random Forest*. En este caso, el error es de 0,28 (28%).

El siguiente gráfico (Fig. 7), muestra las curvas de densidad para los distintos valores de las variables. Como antes, el color magenta corresponde a los datos imputados y el color celeste a los datos originales.

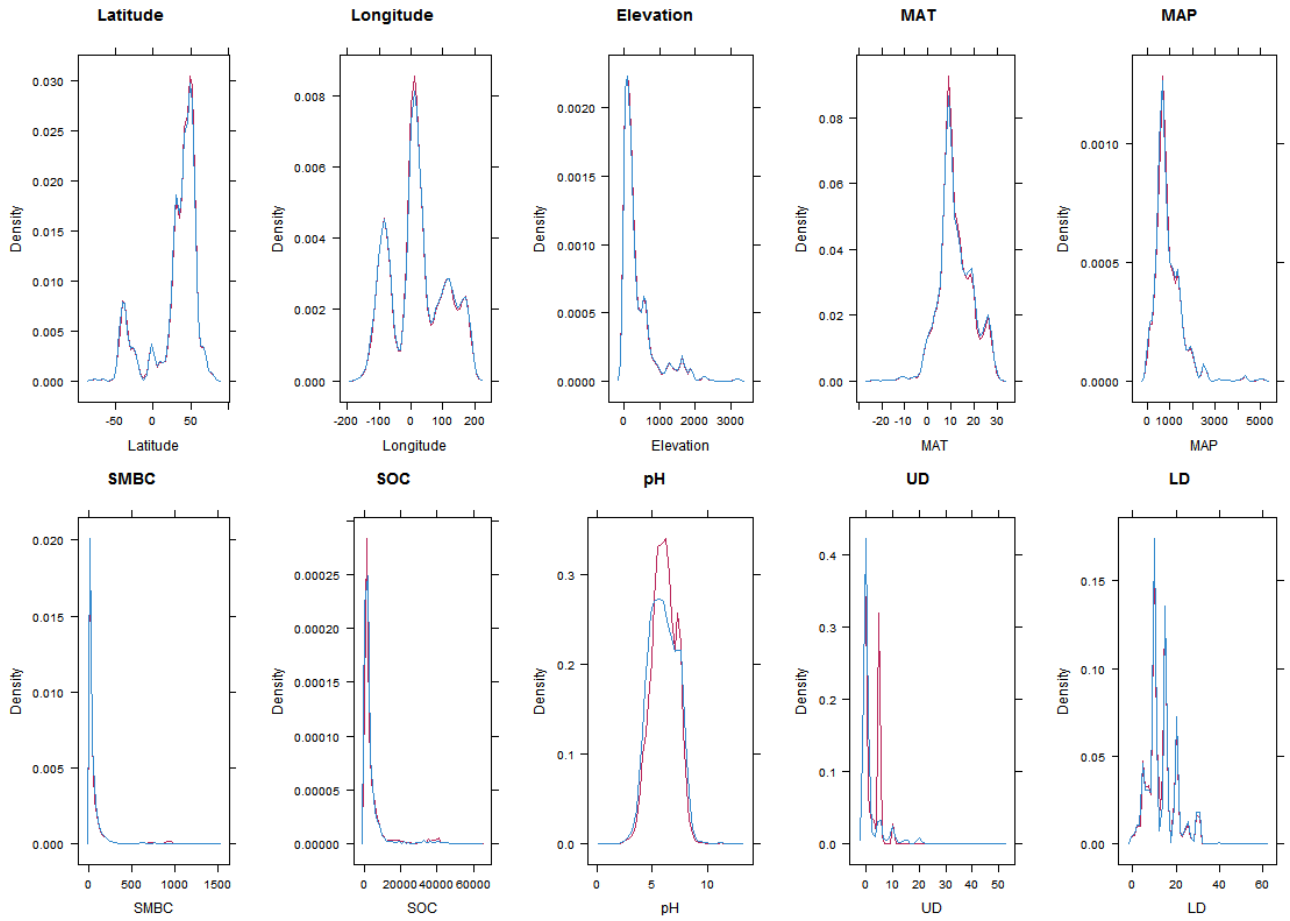


Fig. 7 *Densidad para el rango de valores de las variables para el conjunto de datos imputado con Miss Forest (magenta) y el conjunto de datos original (celeste).*

Como se puede apreciar, la curva de densidad de los datos imputados se ajusta bien a la correspondiente a datos originales con valores faltantes. En algunos casos, las frecuencias son más altas para los datos imputados, por ejemplo, para SOC, pH y UD, pero en todos ellos el comportamiento de la curva de datos imputados sigue al de la curva de datos originales (los valores altos en una son más altos en la otra). Es clara la diferencia respecto del método de imputación anterior en términos de la variable pH, donde la distribución no se ve modificada en *Miss Forest* a partir de la imputación. Asimismo, para las variables SMBC y SOC, la curva de valores imputados generada con este último método ajusta mucho mejor a la curva de valores originales en las zonas de alta frecuencia. En el caso de Latitud y Longitud, las curvas generadas con *Miss Forest* son muy parecidas, mientras que en *MICE* existe una marcada diferencia, en particular para valores altos de ambas variables.

A efectos de comparar los dos métodos analíticamente, se siguió la siguiente estrategia:

a. Se preparó un conjunto de datos reducido para el que se conozcan los valores verdaderos de combinaciones faltantes simuladas.

1. Se calculó la cantidad de registros incompletos, entendiendo por ello la cantidad de casos donde para al menos una de las variables se tenía valor faltante (1784/3422).
2. Se seleccionó una muestra aleatoria de casos incompletos, igual a la cantidad de casos completos (1638).
3. Se replicó el mismo patrón de datos faltantes de la muestra aleatoria de casos, sobre el conjunto de los registros completos, para generar un conjunto lo más representativo posible de la distribución original.

b. Se realizó la imputación de valores faltantes con cada uno de los algoritmos.
c. Se computó en cada caso la métrica: raíz cuadrada del error cuadrático medio normalizada (*NRSME* por sus siglas en ingles)[ii].

En el caso de *MICE*, el error promedio de los 20 conjuntos de datos imputados fue de 0,58 (58%), en tanto que para *Miss Forest* el error fue de 0,33 (33%).

Debido a la amplia diferencia sobre la prueba efectuada, sumada a las características de la distribución observadas, se decide utilizar la imputación realizada con Miss Forest. Para el conjunto de datos completo, el error estimado es del 0,29 (29%).

[ii] $$\sqrt{\frac{mean((X_{true} - X_{imp})^2)}{var(X_{true})}}$$

Donde Xtrue es la matriz de datos completa, Ximp, la matriz de datos imputados, y 'mean'/'var' usados como notación breve de la media empírica y la varianza calculadas sobre los valores perdidos continuos solamente.

Análisis de los datos

Descripción de las variables

1) Bioma / *Biome*

Tabla 2 *Frecuencia de valores de la variable bioma*

Bioma	Frecuencia	%
Tierra Cultivable / Cropland	1621	47.37
Pastizales / Grassland	336	9.82
Bosque Templado Caducifolio / Temperate Broadleaf Forest	263	7.69
Bosque Tropical/Subtropical / Tropical/Subtropical Forestt	215	6.28
Bosque de Coníferas Templado / Temperate Coniferous Forest	210	6.14
Desierto / Desert	206	6.02
Prado / Pasture	167	4.88
Taiga / Boreal Forest	126	3.68
Arbustos / Shrub	78	2.28
Humedal Natural / Natural Wetland	70	2.05
No Mencionado / Not mentioned	46	1.34
Tundra / Tundra	39	1.14
Glaciar / Glacier	17	0.5
Sabana / Savanna	15	0.44
Suelo Descubierto / Bare soil	13	0.38

2) País / Country

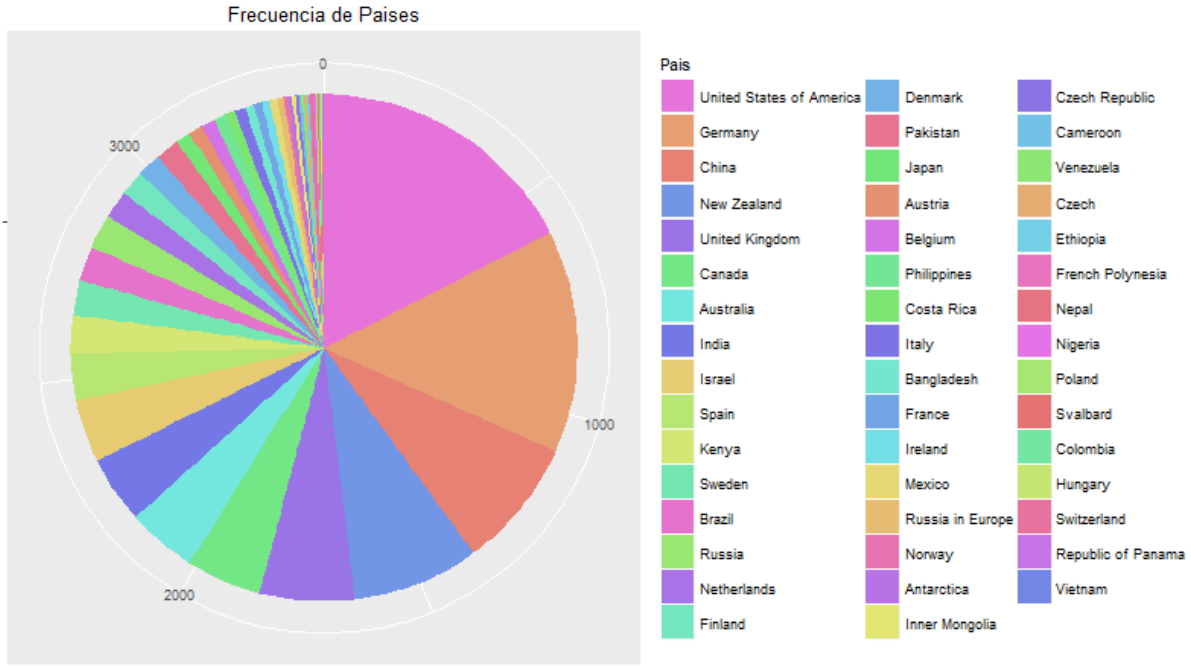


Fig. 8 *Representación gráfica de frecuencia de países*

3) Latitud / Latitude (grados decimales)

Tabla 3 *Estadísticos de la variable latitud*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
-77.63	26.82	41.83	30.62	50.81	79	30.08

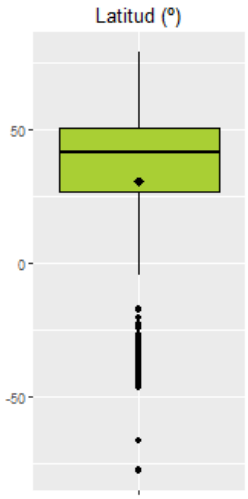


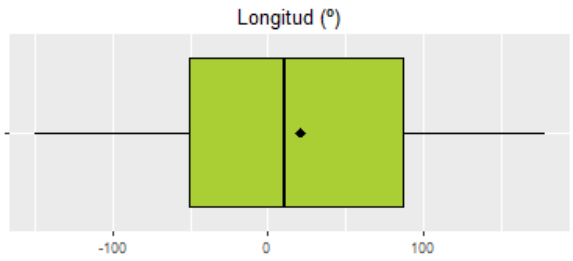
Fig. 9 *Boxplot de la variable latitud*

Se observa que esta variable se distribuye con asimetría negativa. Esta característica responde a que la mayor parte de las muestras corresponden al hemisferio norte. Pero aun así muchas se han tomado del hemisferio sur (correspondiente a los valores atípicos en el gráfico).

4) Longitud / *Longitude* (grados decimales)

Tabla 4 *Estadísticos de la variable longitud*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
-149.8	-50.87	10.5	20.88	87.15	177.9	83.52



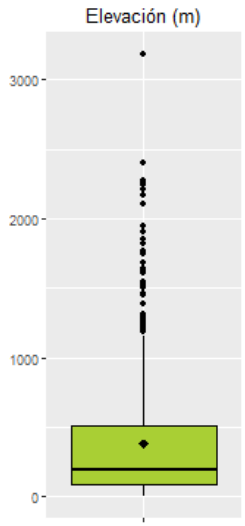
En el caso de Longitud, a diferencia de la variable anterior, se observa un criterio mucho más homogéneo de muestreo a lo ancho del planeta, dando como resultado una distribución aproximadamente simétrica.

Fig. 10 *Boxplot de la variable longitud*

5) Elevación / *Elevation* (m)

Tabla 5 *Estadísticos de la variable elevación*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
-2	80.86	192	377.5	512.9	3189	474



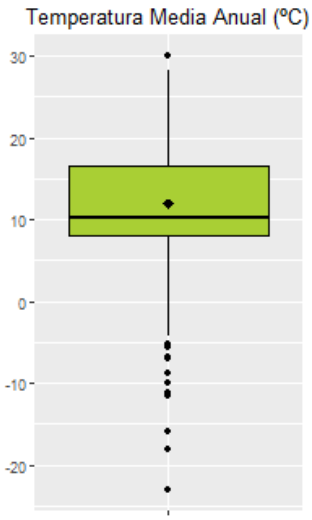
En el caso de la elevación se observa mayor preponderancia de las mediciones a baja altura o en depreciones respecto del nivel del mar. La distribuciòn es sesgada a derecha, donde alturas superiores a los 1000m se corresponden con valores atípicos.

Fig. 11 *Boxplot de la variable elevación*

6) Temperatura Media Anual / *Mean Annual Temperature* (ºC)

Tabla 6 *Estadísticos de la variable temperatura media anual*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
-23	8	10.2	11.9	16.5	30	7.14



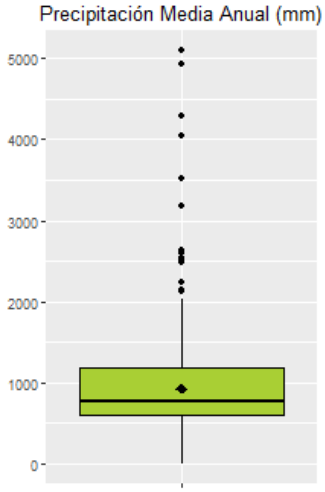
La temperatura media anual sigue una distribución asimétrica con cola a izquierda, con algunos valores atípicos en la zona de las temperaturas bajo cero. Las mismas corresponden a dos casos de -23 y dos casos de -18 en el desierto polar antártico. No se consideran valores anómalos.

Fig. 12 *Boxplot de la variable temperatura meda anual*

7) Precipitación Media Anual / *Mean Annual Precipitation* (mm)

Tabla 7 *Estadísticos de la variable precipitación media anual*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
0.1	600.3	775	926.4	1190	5100	583.91



En el caso de esta variable, se observa una distribución positiva, con cola a derecha, con algunos valores atípicos severos de precipitaciones altas. Estos valores corresponden a la pluviselva costarricense y filipina, y los valores atípicos moderados, a ciertas áreas de pastizales en Costa Rica y el Reino Unido. No se consideran puntos anómalos.

Fig. 13 *Boxplot de la variable precipitación meda anual*

8) Carbono en Biomasa Microbiana del Suelo / *Soil Microbial Biomass - Carbon* (mmol Kg-1)

Tabla 8 *Estadísticos de la variable carbono en biomasa microbiana del suelo*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
0.04	15.12	28.06	76.84	60.58	1508	166.54

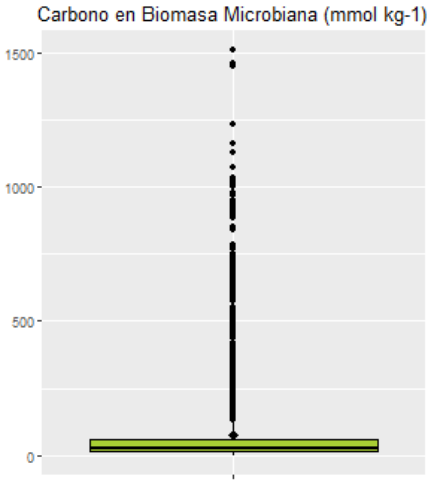


Fig. 14 *Boxplot de la variable carbono en biomasa microbiana del suelo*

Las dos variables referidas a la concentración de carbono siguen distribuciones similares (ver Fig. 14 y Fig. 15). Esto se debe a que la composición de la biomasa es relativamente constante. La distribución es aproximadamente positiva, con gran cantidad de valores atípicos a derecha, evidenciando que las concentraciones altas son frecuentes y a distintos niveles.

9) Carbono Orgánico del Suelo / *Soil Organic - Carbon* (mmol Kg-1)

Tabla 9 *Estadísticos de la variable carbono orgánico del suelo*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
8.333	941.7	1658	4587	4164	63920	8285.66

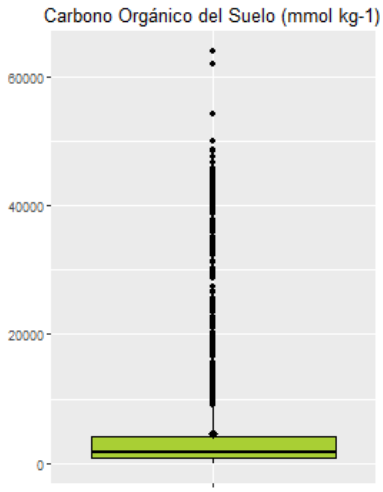
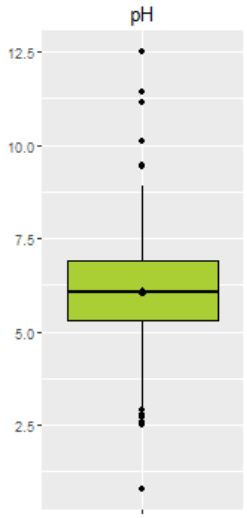


Fig.15 *Boxplot de la variable carbono orgánico del suelo*

10) pH

Tabla 10 *Estadísticos de la variable pH*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
0.8	5.3	6.049	6.073	6.89	12.5	1.1



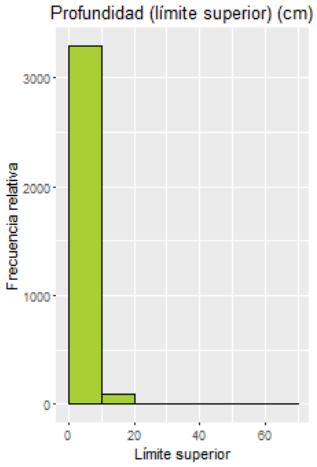
En el caso del potencial Hidrógeno, los datos siguen una distribución aproximadamente simétrica, centrada en un valor ligeramente ácido, con algunos valores atípicos en la zona de los valores altos ácidos y alcalinos.

Fig.16 *Boxplot de la variable pH*

11) Profundidad (límite superior) / *Upper Depth* (cm)

Tabla 11 *Estadísticos de la variable límite superior de profundidad*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
-6.275e-15	0	0	1.472	0.305	50	3.92



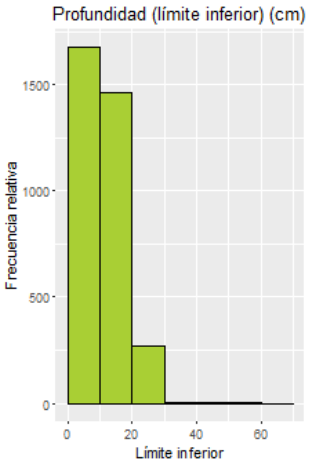
Se observa como en el caso de esta variable las muestras se concentran en zonas bajas de profundidad, mientas que una parte ínfima se corresponde con mediciones a profundidades iniciales mayores. Esto se debe a que, a mayor profundidad, las variables biológicas pierden relevancia.

Fig.17 *Histograma de la variable límite superior de profundidad*

12) Profundidad (límite inferior) / *Lower Depth* (cm)

Tabla 12 *Estadísticos de la variable límite inferior de profundidad*

Min.	1º Q	Mediana	Media	3º Q	Max.	DE
0	10	10.85	13.05	15	60	6.57



En el caso del límite inferior de profundidad, se observa una distribución más pareja pero centrada en valores igualmente bajos, en comparación con el límite superior.

Fig.18 *Histograma de la variable límite inferior de profundidad*

Reducción dimensional

Como primer paso en vistas a analizar la posibilidad de reducir la cantidad de variables del problema, se analiza correlación[iii] entre las variables en estudio, según la siguiente tabla (Tabla 13).

[iii] Se utilizó el método de Pearson.

Tabla 13 *Correlación entre las variables en estudio.*

	Latitude	Longitude	Elevation	MAT	MAP	SMBC	SOC	pH	UD	LD
Latitude	1.00000000	-0.5506416262	-0.08619441	-0.43288632	-0.34087238	0.10504889	0.085464151	0.0845258403	-0.03317057	0.15400418
Longitude	-0.55064163	1.0000000000	-0.02986302	0.14303846	0.05590561	0.01175224	0.006799817	-0.0002514405	0.07418590	-0.05266194
Elevation	-0.08619441	-0.0298630234	1.00000000	-0.09341429	-0.08228253	0.01904867	0.012971678	0.0125228454	0.00154546	-0.09530805
MAT	-0.43288632	0.1430384640	-0.09341429	1.00000000	0.36742192	-0.28235565	-0.294172576	0.0902247901	0.03647579	0.16727086
MAP	-0.34087238	0.0559056130	-0.08228253	0.36742192	1.00000000	-0.03295108	0.050210799	-0.3849368153	-0.05840948	-0.01415992
SMBC	0.10504889	0.0117522375	0.01904867	-0.28235565	-0.03295108	1.00000000	0.827923713	-0.2423673096	-0.09998031	-0.27988279
SOC	0.08546415	0.0067998168	0.01297168	-0.29417258	0.05021080	0.82792371	1.000000000	-0.3621153387	-0.10494595	-0.29100207
pH	0.08452584	-0.0002514405	0.01252285	0.09022479	-0.38493682	-0.24236731	-0.362115339	1.0000000000	0.04220659	0.13408060
UD	-0.03317057	0.0741858974	0.00154546	0.03647579	-0.05840948	-0.09998031	-0.104945946	0.0422065919	1.00000000	0.47895593
LD	0.15400418	-0.0526619420	-0.09530805	0.16727086	-0.01415992	-0.27988279	-0.291002067	0.1340805954	0.47895593	1.00000000

Se observa correlación alta las variables de Carbono: SMBC y SOC, mientras que, entre las demás variables, y para con estas dos, es más bien baja.

Se decide indagar en la reducción dimensional a partir de la determinación de componentes principales, no a efectos de simplificar la cantidad de variables en estudio, sino más bien para comprender la naturaleza de sus relaciones. Esta técnica exploratoria, que carece de supuestos, permite transformar un conjunto de variables correlacionadas en un conjunto menor de variables no correlacionadas.

A continuación, se muestra el gráfico de sedimentación (Fig. 19) correspondiente a las componentes principales halladas utilizando el método de Pearson. Basado en el criterio del bastón roto, se observa un punto de corte con 3 componentes.

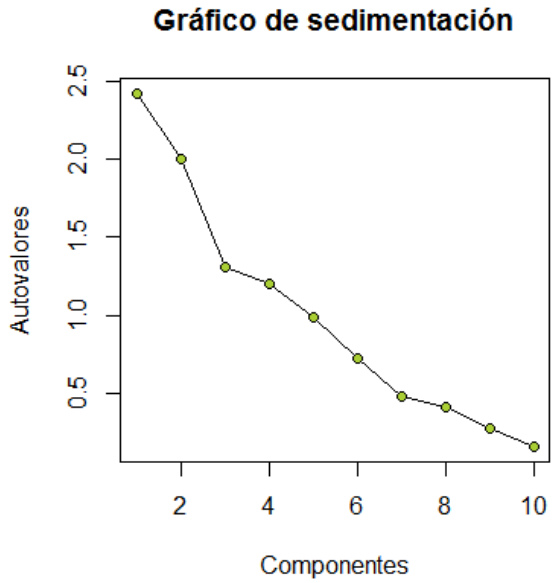


Fig 19. *Gráfico de sedimentación de componentes principales.*

Utilizando el criterio de Kaiser (ver Tabla 14), debieran sin embargo considerarse 4 componentes (con autovalor mayor o igual a 1).

Tabla 14 *Porcentaje de la varianza total explicada por cada componente*

	Autovalor	%	% (acum)
comp 1	2.4220504	24.220504	24.22050
comp 2	1.9951319	19.951319	44.17182
comp 3	1.3112095	13.112095	57.28392
comp 4	1.2077534	12.077534	69.36145
comp 5	0.9909393	9.909393	79.27084
comp 6	0.7297544	7.297544	86.56839
comp 7	0.4797366	4.797366	91.36575
comp 8	0.4170689	4.170689	95.53644
comp 9	0.2850967	2.850967	98.38741
comp 10	0.1612591	1.612591	100.00000

Se observa en la tabla superior que, para explicar un porcentaje de varianza razonablemente alta, cercano al 80% harían falta 5 componentes. Considerando que se tienen 10 variables, explicar más del 80% de la variabilidad de los datos con la mitad de las variables indica cierto grado de asociación entre las mismas.

A continuación (Tabla 15), se analizarán la primera y segunda componentes por ser las más significativas en términos de aporte a la variabilidad.

Tabla 15 *Coordenadas de las dos primeras componentes para cada variable.*

	Coord. comp 1	Coord. comp 2
Latitude	-0.29152546	-0.79770768
Longitude	0.16239815	0.55764116
Elevation	-0.07817654	-0.02073148
MAT	0.57001143	0.48768866
MAP	0.08651400	0.67277344
SMBC	-0.82966558	0.12900506
SOC	-0.85235408	0.19439414
pH	0.41250056	-0.43355213
UD	0.33610035	-0.15108850
LD	0.52364900	-0.30283653

Se observa que ambas componentes son de forma, dado que presentan coeficientes tanto de signo positivo como negativo.

Para la primera componente, las variables de Carbono, son altamente influyentes en términos negativos, es decir que en la medida en que los valores de estas variables son más altos, los valores de esta componente se reducen. Por otro lado, la temperatura media anual y el límite inferior de profundidad de la excavación, actúan en sentido contrario.

En lo que respecta a la segunda componente, el peso preponderante lo tiene latitud haciendo que los puntos norte del planeta presenten baja esta componente. Asimismo, la precipitación media anual y longitud incrementan esta componente.

A continuación, se muestra el gráfico biplot (Fig. 20) que representa la relación entre casos, variables y componentes.

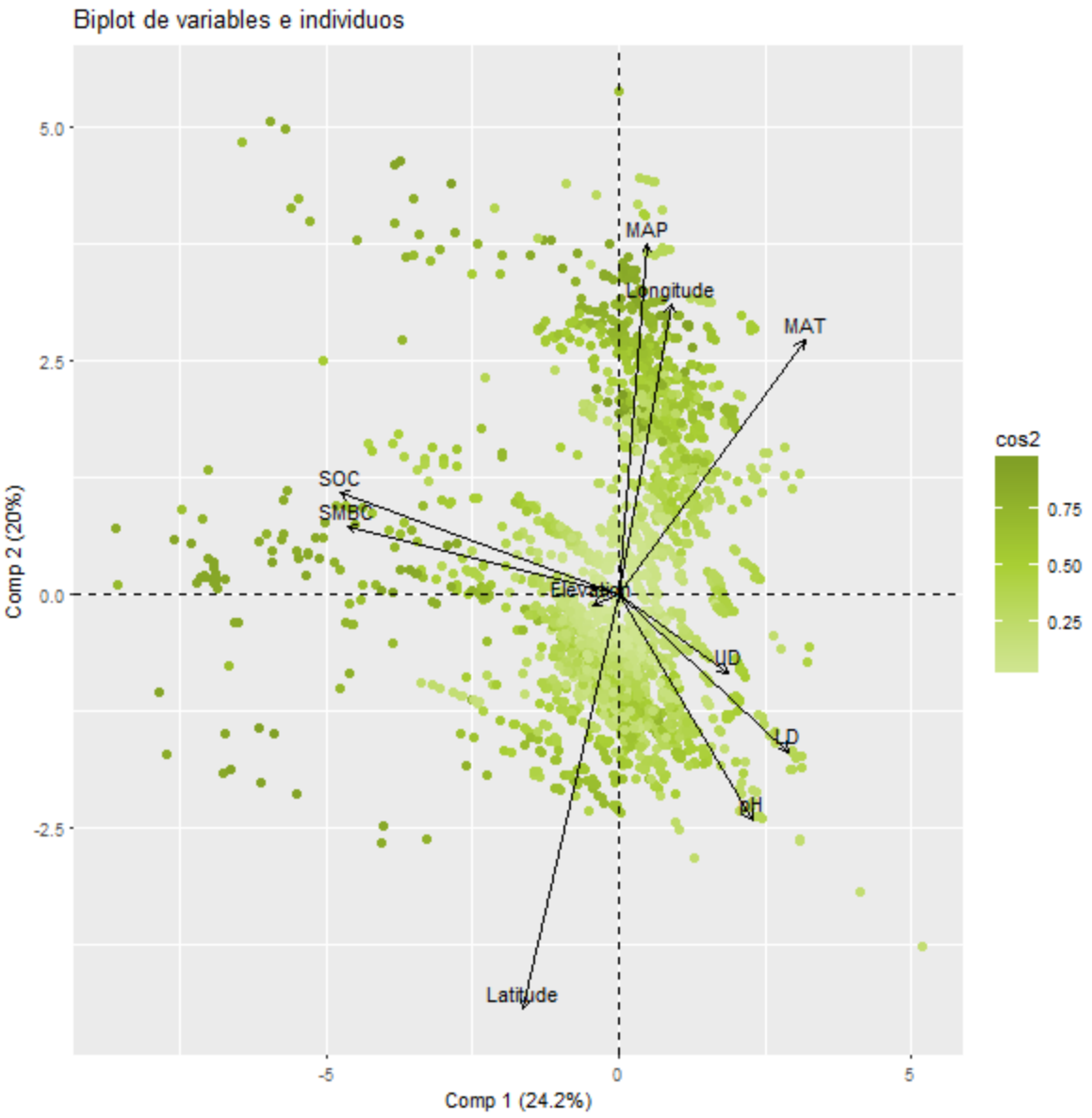


Fig. 20 *Biplot para las dos primeras componentes principales*

Considerando el eje de la *Componente 1*, se observa como el individuo más a la izquierda, correspondiente a un bioma de Bosque Boreal en Finlandia, presenta por ejemplo un valor de biomasa de Carbono de 1233,3 mmol Kg-1, muy por encima de la media de 76,84 mmol Kg-1. Asimismo, este resultado corresponde a una excavación de entre los 0 y 2,8 metros por debajo del nivel del mar (límite inferior no muy bajo), con temperatura media anual de -0,5ºC, por debajo de la media de 11,9ºC. Entre tanto, el individuo en el extremo opuesto a la izquierda, correspondiente a un bioma de Tierra Cultivable en Alemania, presenta un valor de biomasa de Carbono de apenas 1,08 mmol Kg-1, bastante por debajo de la media, y temperatura media anual de 8ºC. Este resultado corresponde a una excavación de entre 50 y 60 metros por debajo del nivel del mar (límite inferior marcadamente bajo), y a una temperatura media anual de 8,5ºC mucho más cercana a la media que el caso anterior.

De manera similar, analizando ahora la *Componente 2* se observa como el caso con esta componente más alta corresponde a un bioma de Pastizales en Nueva Zelanda, habiéndose relevado por tanto en latitud sur y hacia el extremo este del planeta, con abundantes lluvias, donde la precipitación media anual alcanza los 4293,2mm, muy por encima de la media de 926,4mm. Mientras que, en contrapartida, el punto con esta componente más baja corresponde al mismo caso visto anteriormente de Tierra Cultivable en Alemania, por tanto, en el hemisferio norte y con precipitaciones que se ubican en los 580mm, por debajo de la media.

Se puede observar, además, analizando el gráfico, como todas las variables relacionadas a los indicadores de concentración de componentes del suelo funcionan en la misma dirección, hacia los valores bajos de la *Componente 1*, se encuentran altamente correlacionadas. Mientras que, contrarresta este efecto el límite inferior (y superior, en menor medida) de la profundidad de la excavación, como se había observado a partir de los signos de los coeficientes. El largo de este vector sin embargo es inferior al correspondiente a las variables mencionadas anteriormente siendo mayor la influencia de estas últimas en el balance.

De igual manera, puede hacerse la analogía para la *Componente 2*, en términos de latitud (y precipitación media anual) versus longitud. El indicador de pH asimismo ejerce una leve influencia en sentido negativo.

Por otra parte, se observa como la longitud y la concentración de carbono en biomasa microbiana no se encuentran correlacionadas en absoluto (se representan con vectores ortogonales), de manera tal que los indicadores de concentración de componentes en el suelo no parecen estar asociados con el punto a lo ancho del planeta.

Por último, puede analizarse la calidad de representación de las observaciones en términos de los dos ejes asociados con las componentes. Esto se puede medir por medio de la métrica coseno cuadrado ($\cos^2$)[iv], donde valores altos (próximos a 1) indican que el punto está bien representado por las componentes, y valores bajos (próximos a 0) indican que las componentes no representan tan bien el punto. Esta métrica se muestra en el gráfico biplot por medio de la escala de verdes (0-1). Así, puede observarse como los puntos de mayor calidad, se ubican en la zona periférica de la nube, con mayor preponderancia en el primer y cuarto cuadrante, es decir, para valores positivos de la segunda componente. Es más, puede observarse como se aglutinan en torno a la dirección y sentido de los vectores que representan las variables de carbono, por un lado, y precipitación y longitud, por otro; y en las zonas en el arco entre los dos grupos de vectores, en la región más

[iv] $\cos^2$ indica la contribución de una componente a la distancia cuadrada de la observación al origen. Corresponde al cuadrado del coseno del ángulo del triángulo rectángulo conformado por el origen, la observación (en el espacio original), y su proyección sobre la componente, y se calcula como:

$$\cos^2_{i,\ell} = \frac{f^2_{i,\ell}}{\sum_{\ell} f^2_{i,\ell}} = \frac{f^2_{i,\ell}}{d^2_{i,g}}$$

Donde *i* representa la observación y *l* la componente, y $d^2_{i,g}$ es la distancia al cuadrado de una observación dada al origen. La misma se calcula, en base al teorema de Pitágoras, como la suma de los valores al cuadrado de las puntuaciones de la observación en términos de cada una de las componentes $f^2_{i,\ell}$ (observación "valuada" en las componentes). [16]

Por último, el hecho de utilizar coseno cuadrado, permite sumar las contribuciones con respecto a cada componente para obtener la ponderación de una observación en el (hiper)plano reducido por las componentes.

alejada del origen. En menor proporción de casos, esto último se verifica en relación a la dirección y sentido del vector de latitud. En este sentido, podría decirse que observaciones con valores altos en las variables de carbono (cobra sentido en relación a su preponderancia en términos de correlación), longitud alta (en el extremo derecho respecto de Greenwich) y con valores de precipitación altos, son los que se encuentran bien representados por este modelo. El mapa que se muestra a continuación (Fig. 21) [17] captura la región del planeta descripta.

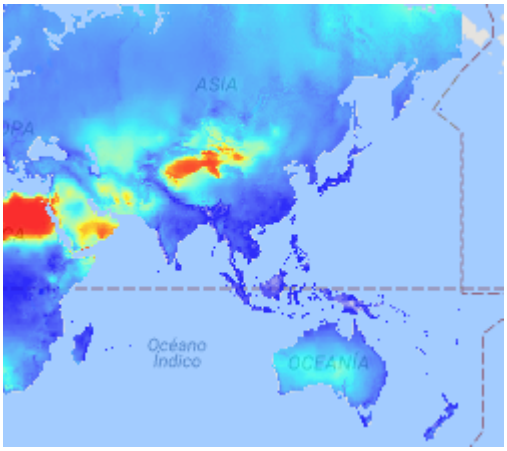


Fig. 21 *Precipitaciones en el extremo oriente del planeta (escala azul oscuro a rojo de mayor a menor precipitación respectivamente).*

En términos de los países para los que se observan datos, podría decirse que Nueva Zelanda, Australia, Finlandia, China, India, Indonesia, Pakistán, Japón, y Filipinas son los lugares en que es más probable que se den las condiciones del modelo, correspondiendo a Biomas de Bosque Boreal (al norte), prado (al sur).
Cabe destacar finalmente que el bioma es una variable ecológica resultado de la interacción de factores geográficos (latitud, longitud, elevación) con variables climáticas (temperatura, precipitaciones) y variables biológicas, que a largo plazo, junto con las variables geológicas no registradas, determinan las características de los suelos.

Análisis de conglomerados

Continuando con la línea investigativa centrada en la naturaleza de la variable de carbono en biomasa microbiana cabe la pregunta acerca de si es factible agrupar los datos a partir de las demás variables en estudio y si es posible encontrar una interpretación de estos grupos en términos de la variable de interés.

Comenzando con el análisis primeramente es necesario establecer si los datos, en las variables de latitud, longitud, elevación, temperatura media anual, precipitación media anual, pH y profundidad (límites superior e inferior de la excavación) tienen tendencia a la agrupación. A tal efecto se emplea un test basado en el estadístico de Hopkins[v] [18], que mide la probabilidad de que un conjunto de datos dado esté generado por una distribución uniforme. La hipótesis nula de la prueba establece que los datos se distribuyen uniformemente mientras que la hipótesis alternativa establece que los datos no están distribuidos uniformemente.

El resultado del test para los datos de interés es 0,015[vi] lo que lleva a rechazar la hipótesis nula (con un 95% de confianza) y por tanto se concluye que existe tendencia de los datos al agrupamiento.

Como segundo paso se calculan las distancias euclídeas entre los puntos para las variables mencionadas, y luego se modela el agrupamiento jerárquico de los datos utilizando el método promedio. Se observa que el coeficiente de correlación cofenética es de 0,83, indicado que el dendograma asociado preserva las distancias de los datos sin modelar tomados de a pares. Sin embargo, es posible incorporar una variable más al modelo: Bioma, que resulta ser una variable categórica. En este sentido se recalculan las distancias, esta vez utilizando el coeficiente de Gower, y se vuelve a modelar el agrupamiento jerárquico. Como resultado se observa que con el agregado de la variable mejora el coeficiente de correlación cofenética, que resulta ser de 0,84.

A continuación, se muestra el dendograma (Fig. 22) correspondiente a 8 grupos. La elección de la cantidad se basó en un corte que generara la menor cantidad de grupos de tamaño intermedio entre el bloque central, por un lado, y los bloques muy pequeños a izquierda y derecha, por otro.

$$H_{ind} = \frac{\sum_{i=1}^{n} q_i}{\sum_{i=1}^{n} q_i + \sum_{i=1}^{n} w_i}$$

[v] El estadístico compara la distancia w_i entre los objetos reales y sus vecinos más cercanos, con las distancias q_i entre los objetos artificiales uniformemente generados sobre los datos y sus vecinos más cercanos reales. Este proceso se repite varias veces para fracciones del total de la población. Si los objetos están uniformemente distribuidos, q_i y w_i serán parecidos y el estadístico será cercano a 0,5. Si existe tendencia al agrupamiento, las distancias para los objetos artificiales serán mayores que para los objetos reales, porque estos objetos artificiales están distribuidos homogéneamente mientras que los objetos reales están agrupados, y por ende el valor del estadístico será más grande.

[vi] Se emplearon la totalidad de los datos para la prueba.

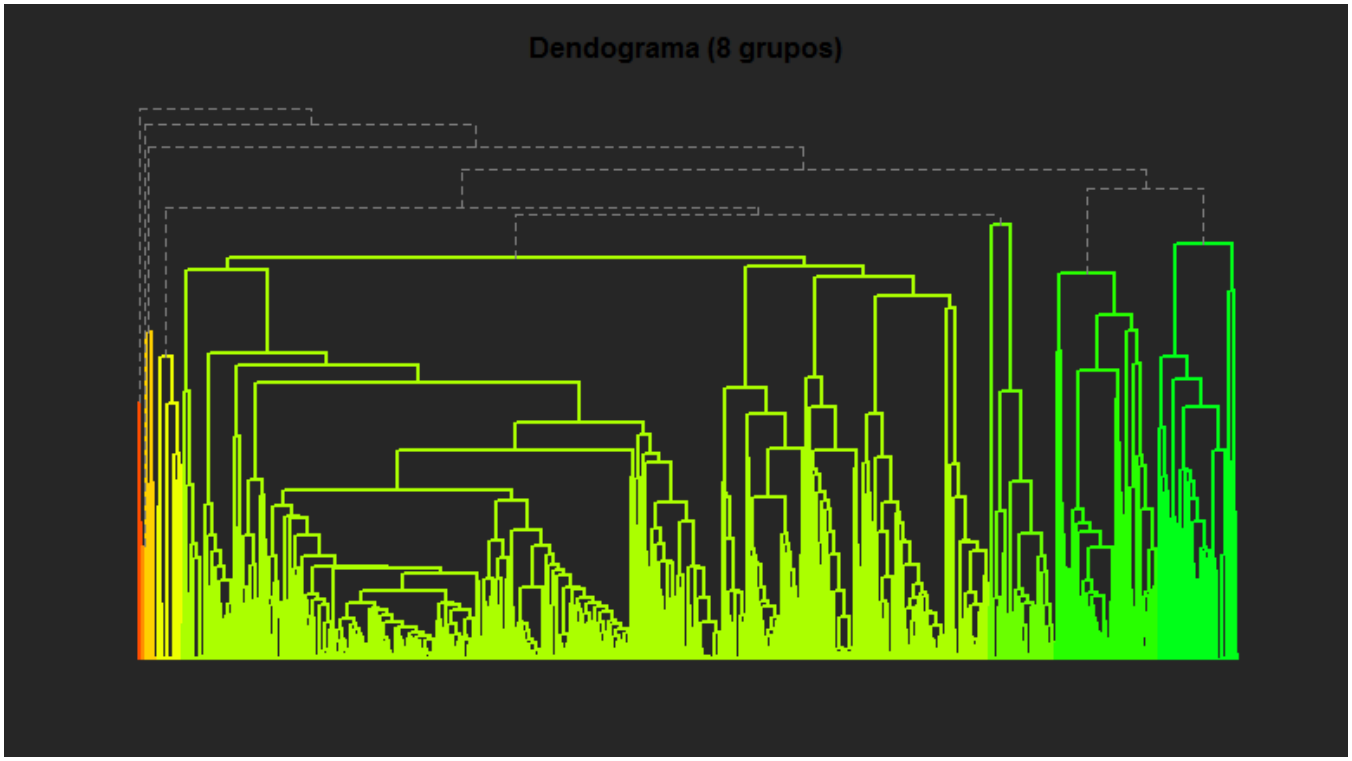
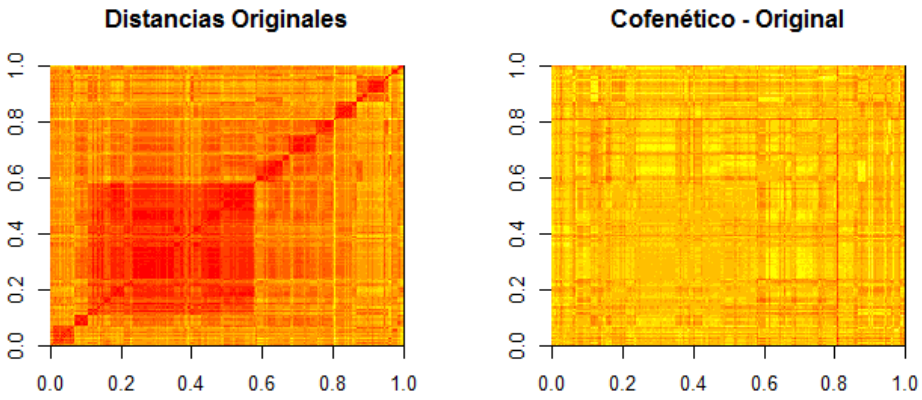
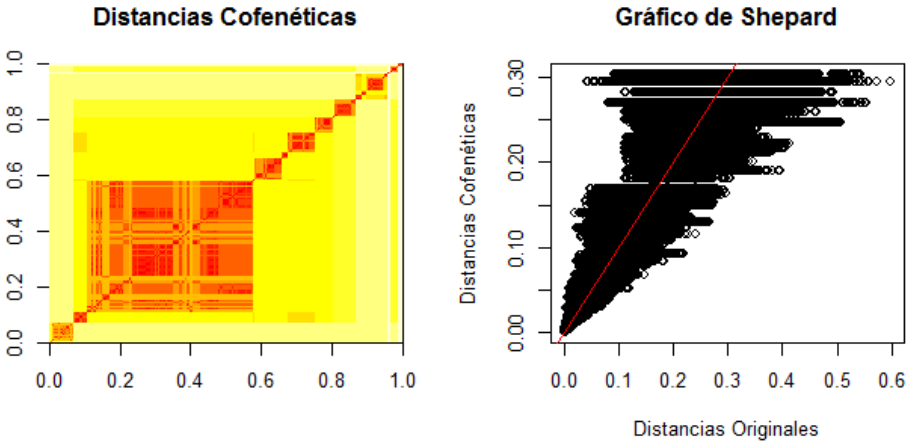


Fig. 22 *Dendograma correspondiente a agrupación jerárquica con corte de 8 grupos.*

Asimismo, se analiza el efecto de aplicar el agrupamiento en términos de disimilitud y distancias.



(a) (b)



(c) (d)

Fig. 23 *Diagrama de distancias originales (a), diagrama cofenético original (b), diagrama de distancias cofenéticas (c) y gráfico de Shepard (d).*

El diagrama de Distancias Originales (Fig. 23.a) muestra los datos de la matriz de disimilitud original, ordenados en base a las 8 agrupaciones seleccionadas. Se observan rectángulos rojos en la diagonal principal asociados con cada una de las agrupaciones. Una de ellas representa la mayor para te de los casos, y existen varias otras agrupaciones de tamaño más o menos similar. El diagrama de Distancias Cofenéticas (Fig. 23.c) muestra las distancias cofenéticas ordenadas en base a los grupos. El gráfico Cofenético – Original (Fig. 23.b) muestra la diferencia entre las distancias cofenéticas y originales, presentando una distribución pareja que indica la correspondencia entre los términos de la mencionada diferencia. Por último, el Gráfico de Shepard (Fig. 21.d) compara ambas distancias mencionadas, donde mientras mejor sea la captura de las distancias por parte de las agrupaciones, más cerca de la recta de pendiente unitaria se concentrarán los puntos. En este caso se observa una buena distribución en este sentido.

En segundo término, se revisarán alternativas de agrupamiento por partición. En este sentido se evalúan dos algoritmos: K-medios y PAM, y en ambos casos la medida de ajuste elegida para determinar la cantidad de agrupamientos a seleccionar es el ancho promedio del coeficiente de *Silhouette* (Silueta).

En el caso de K-medios a diferencia del agrupamiento jerárquico, se utilizan sólo las variables numéricas mencionadas anteriormente (no es factible utilizar la matriz de distancia que incorpora la variable nominal *Bioma*). Para este algoritmo, se observa por comparaciones sucesivas, que el valor adecuado de K es 3[vii] . A continuación, el gráfico con los valores de Silhouette (Fig. 24).

[vii] Para superar el valor del ancho promedio de Silhouette es preciso que K sea al menos 20, lo cual dificulta sensiblemente la posterior interpretación de los agrupamientos.

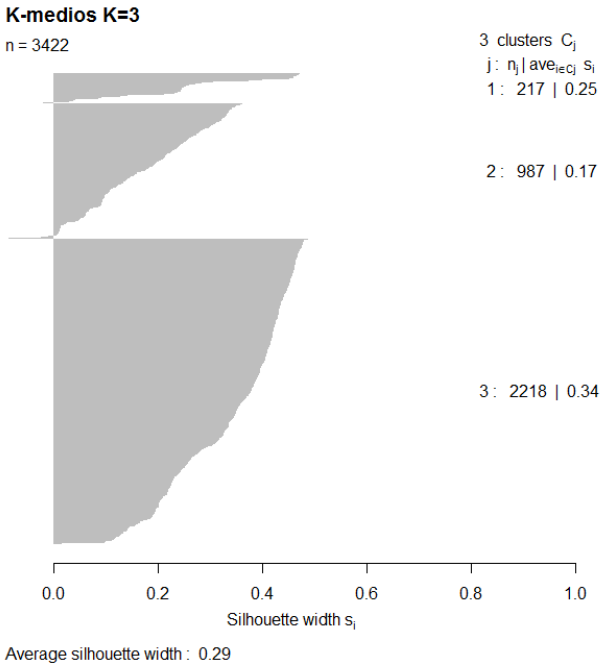


Fig. 24 *Representación de Silhouette para K-medios (K = 3).*

En el caso de PAM (ver Fig. 25), si se contemplan todas las mismas variables que se utilizaron para el agrupamiento jerárquico. Bajo este algoritmo, también por comparaciones sucesivas se encuentra que la cantidad adecuada de medioides es 7[viii].

[viii] A valores mayores no se obtiene una mejora significativa en el ancho promedio de Silhouette.

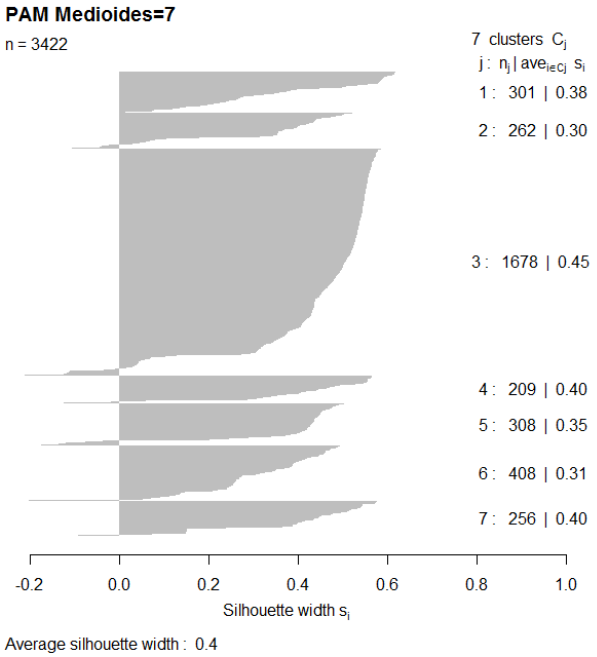


Fig. 25 *Representación de Silhouette para PAM (7 grupos).*

Una vez descriptas las tres metodologías de agrupación elegidas, ahora se procede a interpretar las agrupaciones en los tres casos, denominados en los sucesivo: *clustK* (para K-medios con K=3), *clustH* (para el agrupamiento jerárquico de corte 8) y *clustP* (para PAM con 7 grupos). Los gráficos que se muestran a continuación representan en la parte central la dispersión respecto de dos variables, y en la parte superior y a la derecha (para cada variable respectivamente) se muestran los diagramas de densidad.

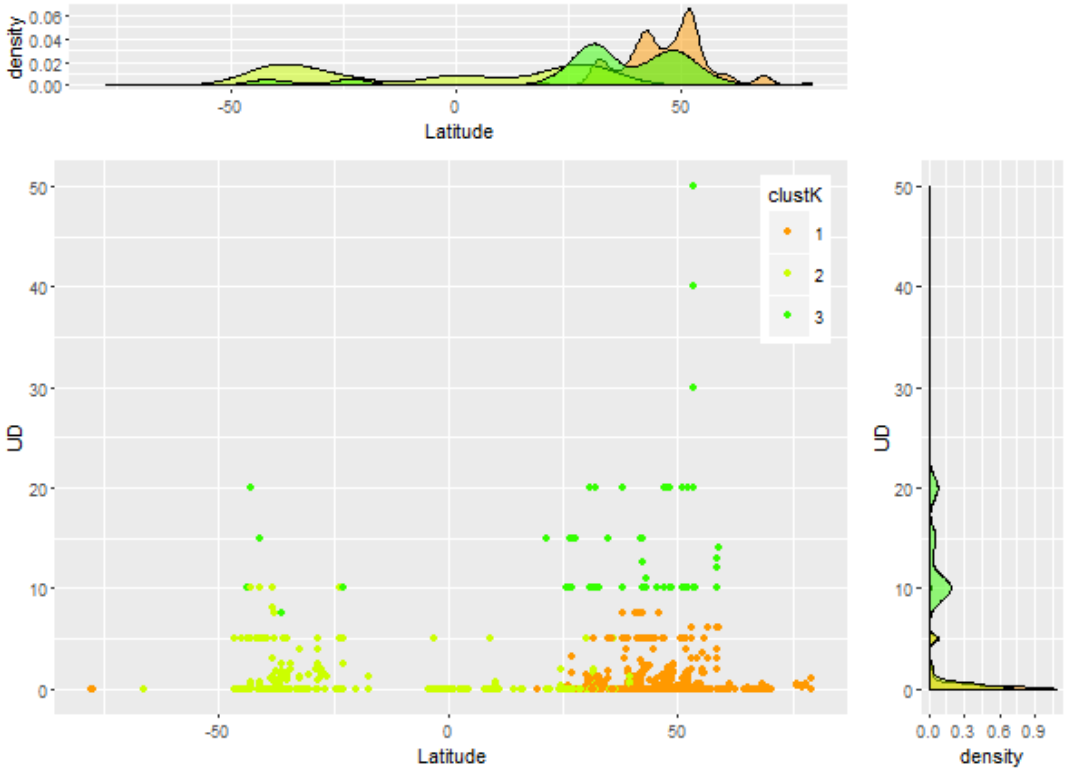


Fig. 26 *Diagramas de densidad y dispersión para las variables latitud y límite superior de profundidad en clustK.*

En el primer gráfico correspondiente a clustK (Fig. 26), se observa que, datos correspondientes a latitudes bajas (hemisferio sur y cierto margen del hemisferio norte próximo al ecuador) quedan comprendidos dentro del grupo 2 con límite inferior próximo a 0, mientras que, a la vez límites superiores de profundidad de excavación altos resultan bien delimitados dentro de la agrupación 3 a lo largo de diferentes valores de latitud.

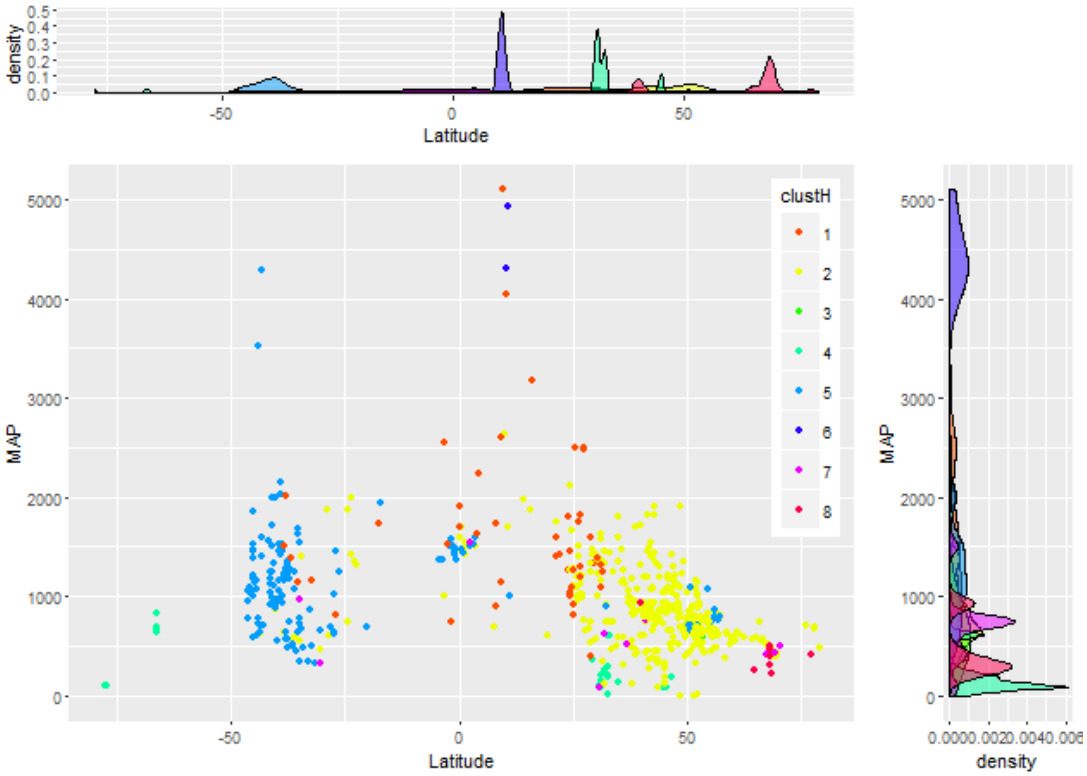


Fig. 27 *Diagramas de densidad y dispersión para las variables latitud y precipitación media anual en clustH.*

De manera similar para el caso de clustH (ver Fig. 27) las agrupaciones permiten separar los datos por precipitación, observándose en este caso como el grupo 4 encierra puntos bien al sur con precipitaciones escasas, mientras que el 6 hace lo propio con puntos de latitud norte próximos al ecuador y con abundantes lluvias.

En el caso de clutH se observa una buena clasificación de los biomas (ver Fig. 28).

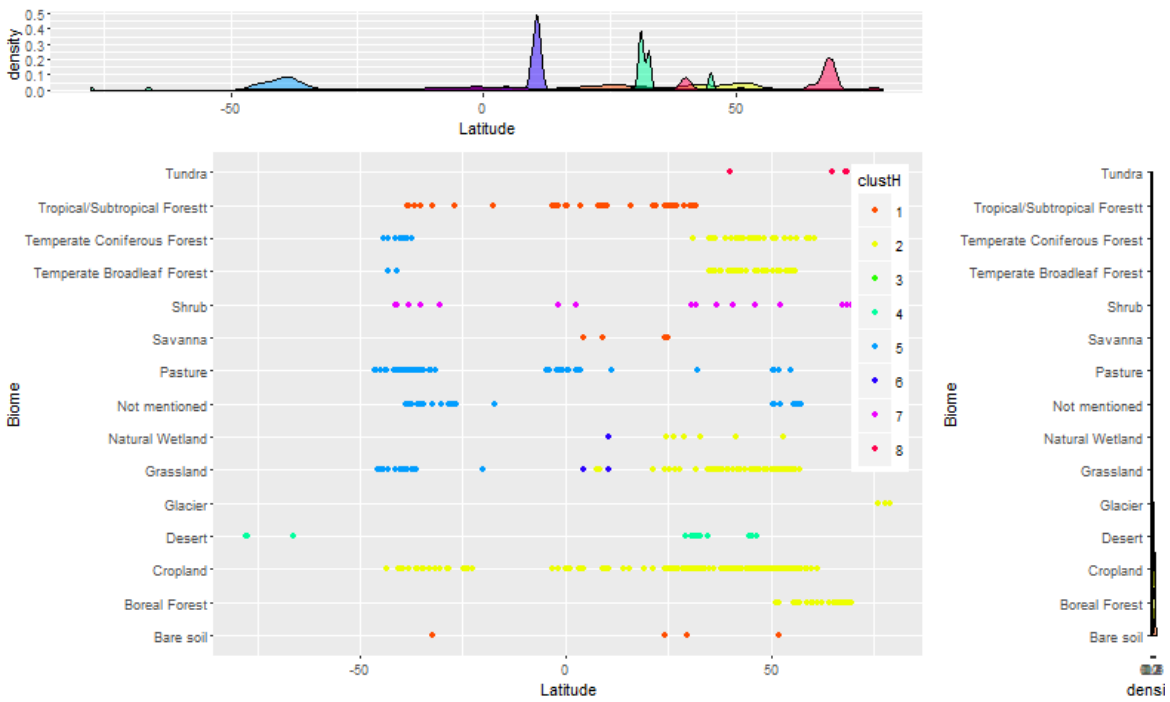


Fig. 28 *Diagramas de densidad y dispersión para las variables latitud y bioma en clustH.*

Existen biomas enteramente contenidos en grupos. En particular Bosque Tropical/Subtropical, Sabana y Suelo desnudo, quedan enteramente contenidos en el grupo 1. Desierto queda enteramente contenido en el grupo 4, Tundra enteramente contenido en el grupo 8 y Arbustos enteramente contenido en el grupo 7. En tanto Tierra cultivable queda contenido en el grupo 2, al igual que Bosque Boreal; mientras que Bosque Templado de Coníferas y Bosque Templado Caducifolio, quedan contenido en el mismo grupo para latitud norte, y en el grupo 5 para latitud sur, al igual que Pastizales. Por último, el bioma Prado queda enteramente contenido en el grupo 2.

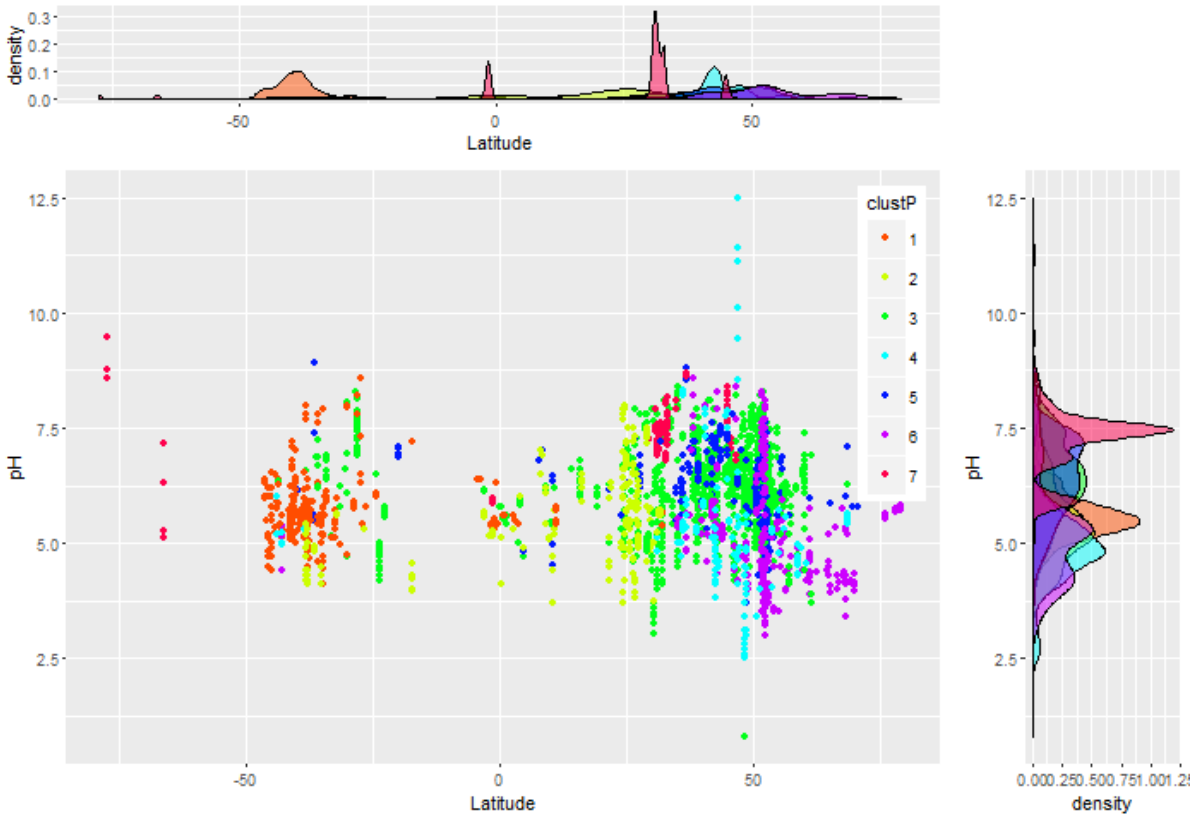


Fig. 29 *Diagramas de densidad y dispersión para las variables latitud y bioma en clustP.*

Por último, en el caso de clustP (ver Fig. 29) la determinación en términos de latitud no es tan buena, como en el caso de clustH para discriminar en este sentido, sin embargo, sí se observa que el grupo 7 captura suelos alcalinos en latitud sur, mientras que el grupo 2 captura suelos ácidos en latitud próxima al ecuador, ligeramente hacia el norte.

Vale la pena mencionar que en el caso de clustP la diferenciación de biomas no es tan notable pese a haberse utilizado la matriz de disimilitud que contiene a la variable biomas al igual que en clustH.

Ahora bien, retornando al interés inicial en la variable carbono en la biomasa microbiana (SMBC), vale la pena hacer analizar su distribución en función de los tres criterios de agrupación. A continuación, el detalle.

Tabla 16 *Estadísticos de clustK.*

Clst.	n	Min.	1º Q	Mediana	Media	3º Q	Max.	DE
1	2218	0.0400	15.790	29.92	86.97	64.37	1508.0	185.04725
2	987	1.6700	15.980	28.33	66.57	61.43	1460.0	134.01233
3	217	0.1917	4.392	13.33	20.02	24.25	383.8	32.16566

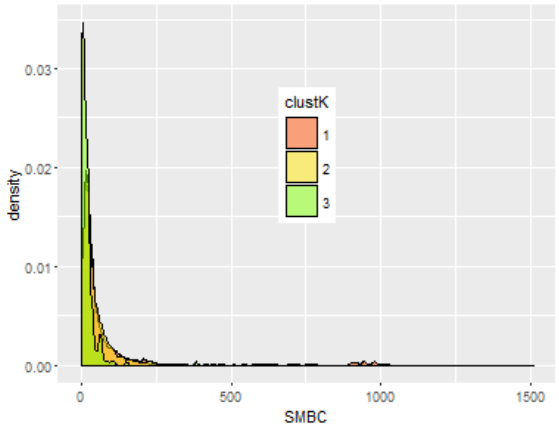


Fig. 30 *Diagrama de densidad de la variable carbono en biomasa microbiana del suelo para los agrupamientos de clustK.*

El método de clustK (ver Tabla 16) se muestra medianamente efectivo para discriminar casos con baja SMBC en el grupo 3, sin embargo, los valores altos de la variable se reparten más o menos equitativamente entre los otros dos grupos.

Tabla 17 *Estadísticos de clustH.*

Clst.	n	Min.	1º Q	Mediana	Media	3º Q	Max.	DE
1	243	2.230	18.340	30.000	79.320	57.520	1460.0	185.1458472
2	2513	0.360	15.000	26.830	68.840	52.000	1449.0	153.4193507
3	6	1.083	1.229	1.458	1.889	2.375	3.5	0.9742043
4	206	0.040	4.200	13.900	46.640	29.620	604.2	91.0418962
5	320	1.670	27.290	56.880	74.480	96.480	610.9	69.5556660
6	17	9.170	49.670	92.920	102.200	160.200	226.8	66.1151586
7	78	4.670	17.690	24.210	126.000	98.450	1158.0	237.3110080
8	39	80.830	233.300	739.200	657.500	979.400	1508.0	399.7516554

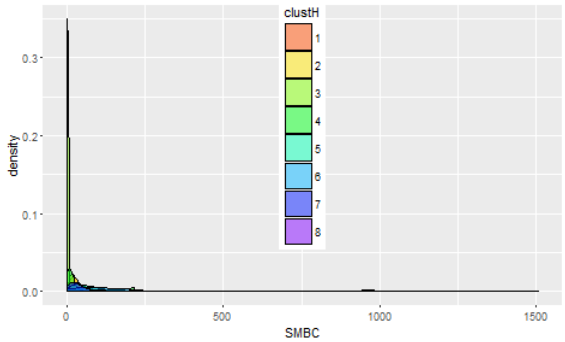


Fig. 31 *Diagrama de densidad de la variable carbono en biomasa microbiana del suelo para los agrupamientos de clustH.*

Por su parte, el método clustH (ver Tabla 17) reúne algunos casos de alto valor en la componente en los grupos 1, 7 y 8. En el último caso, la cantidad de registros es baja, y el desvió estándar es alto, sin embargo, media y mediana no se encuentran tan alejadas. Mientras tanto, comparando los casos de los grupos 1 y 7, se observa que en el caso del grupo 1 la media es algo más alta que en el caso del grupo 7, y está más próxima a la mediana que en el otro grupo. Asimismo, el desvío estándar es más bajo en el grupo 1 que en los otros dos y la cantidad de casos significativamente más grande. En conclusión, podría decirse que los datos del grupo1 son los que tienen la componente comparativamente alta desde el punto de vista estadístico. Valores bajos se registran en el grupo 4 y 5.

Tabla 18 *Estadísticos de clustP.*

Clst.	n	Min.	1º Q	Mediana	Media	3º Q	Max.	DE
1	301	1.670	26.250	56.08	74.10	93.33	647.2	72.36458
2	262	2.230	19.000	32.33	106.00	71.82	1460.0	216.82761
3	1678	1.083	12.840	22.11	32.59	34.48	782.8	51.44793
4	209	0.640	27.000	49.00	123.00	99.42	1158.0	212.80969
5	308	1.250	19.580	45.87	67.46	94.29	1158.0	84.94699
6	408	0.360	32.440	65.92	246.40	286.20	1508.0	340.60549
7	256	0.040	5.127	15.00	43.57	28.44	604.2	82.78634

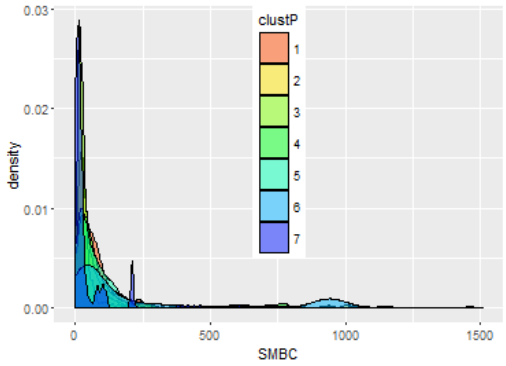


Fig. 32 *Diagrama de densidad de la variable carbono en biomasa microbiana del suelo para los agrupamientos de clustP.*

En cuanto al método clustP (ver Tabla 18), los valores altos se concentran en el grupo 2 con alto desvío estándar, media y mediana que difieren más que en el caso del grupo 1 del método anterior, pero el número de casos es significativamente más alto. Los casos con valores de la variable bajos, se encuentran reunidos en el grupo 1, 3 y 7, destacando el grupo 3 por la cantidad de casos y el desvío estándar bajo (sin bien el valor máximo es superior al de los otros dos grupos). Por lo demás existe una distribución medianamente pareja de los casos, a diferencia de los otros dos métodos donde una mayoría se aglutinaba en una menor cantidad de grupos con relativamente poca cantidad de casos.

Adicionalmente, pueden compararse los gráficos de densidad en Fig.30, Fig. 31 y Fig. 32., que permiten visualizar gráficamente las anteriores consideraciones.

Análisis de la Varianza

Vale la pena analizar, si los grupos considerados en la sección anterior presentan características diferencias en términos de media de la variable de interés SMBC. Dado que dicha variable no reúne los requisitos de normalidad ni homocedasticidad, se recurre al análisis no paramétrico. En este sentido, y teniendo en cuenta que las distribuciones por bioma se muestran relativamente homogéneas, se realiza la prueba de *Kruskall-Wallis* por comparaciones de a pares.

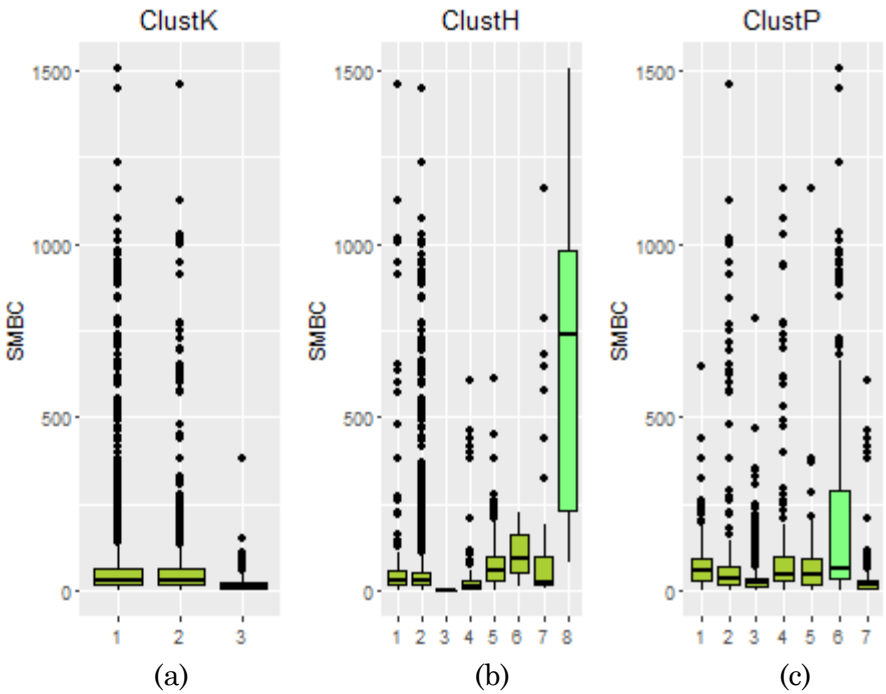


Fig. 33 *Boxplot de la variable carbono en biomasa microbiana para las agrupaciones de clustK (a), clustH (b) y clustP (c).*

La única excepción se refiere a los grupos 8 de clustH y 6 de clustP donde la varianza es notablemente distinta al de las otras distribuciones (ver Fig. 33) y por tanto cabe esperar anticipadamente que constituyan un grupo en sí mimos, por lo demás se espera que a grupos distintos exista una diferencia significativa de media.

A continuación, los resultados del test[ix].

[ix] En todos los casos el p valor asociado a chi cuadrado es significativo, con confianza del 95%.

Tabla 19 *Resultados del test para clustK (a), clustH (b) y clustP (c)*

clustK (a)

Tratam.	Media	Clase
1	1775.4689	a
2	1743.4894	a
3	912.1613	b

clustH (b)

Tratam.	Media	Clase
8	3279.85897	a
6	2543.41176	b
5	2234.41563	b
7	1798.36538	c
1	1776.60082	c
2	1661.36053	c
4	1084.04369	d
3	48.16667	e

clustP (c)

Tratam.	Media	Clase
6	2428.630	a
1	2200.246	b
4	2138.502	bc
5	2027.688	cd
2	1907.191	d
3	1393.343	e
7	1150.039	f

En el caso de clustK (Tabla 19.a) podría decirse que los grupos 1 y 2 que concentran la mayoría de los casos, conforman una misma clase. Se trata, sin embargo, de grupos heterogéneos que abarcan valores grandes y pequeños de la variable SMBC. Sin embargo, el grupo 3, resulta separado en una clase distinta y con media sensiblemente inferior, por lo que podría decirse que efectivamente reúne los valores pequeños de la variable SMBC.

Para clustH (Tabla 19.b) se observa que el grupo 3 separa bien los valores bajos, mientras que los grupos 4, 2, 1 y 7 corresponden a valores medios, y los grupos 6 y 5, a valores altos. El grupo 8, como se mencionó con anterioridad, por seguir una distribución marcadamente distinta al resto, constituye una clase en sí mismo, pero nada puede decirse acerca de su media en relación con el resto. Sin embargo, se observa como el agrupamiento jerárquico es más efectivo que el agrupamiento K-medios para separar los valores más bajos y más altos, que son precisamente (en particular los segundos) los que revisten interés.

Por último, en clustP (Tabla 19.c) hay un solapamiento de grupos y las categorías se hacen más difusas, todas ellas de naturaleza similar. De cualquier manera, podría decirse que se tiene un grupo de valores medios en 7 y 3, y luego valores altos en 2, 5, 4 y 1, lo cual resulta una clasificación que más bien es efectiva en la discriminación de varios niveles con poca variación entre los altos. La salvedad es el grupo 6 que presenta una distribución distinta al resto y por ende no puede decirse que separe bien valores altos en una clase propia. En base a lo anterior, se observa que el agrupamiento jerárquico es también preferible al agrupamiento PAM.

Clasificación

Por último, luego de realizado cierto análisis exploratorio y estadístico, vale la pena preguntarse si existe un método que permita clasificar distintas muestras evaluadas en las variables numéricas de interés y los biomas, en términos de la cantidad de carbono en la biomasa microbiana del suelo. Como se hizo notar al comienzo de este estudio, poder determinar en qué medida el carbono está presente en el suelo puede significar un indicio de la calidad de ese suelo como agente de retención del carbono. Así, podría conjeturarse que es posible conseguir un modelo, en tanto simplificación de la realidad, que permitiese predecir carbono en biomasa microbiana, partiendo de las variables longitud, latitud, elevación del terreno, precipitación media anual, temperatura media anual, pH, profundidad de la excavación que se realizó para tomar la muestra (en base a límite superior e inferior) y bioma. Como se mencionó cuando se consideraron componentes principales, la otra variable de carbono: carbono total del suelo, se encuentra en relación directa con la variable objetivo y por ende no amerita considerarla como parte del modelo, amén que en caso de incorporarla la determinación de su valor no justificaría el esfuerzo de predicción (se asume que si fuera preciso realizar las determinaciones de estas magnitudes en laboratorio también podría hacerse lo propio con carbono sin mayor esfuerzo). En términos técnicos, la variable de carbono en biomasa presentaba menor cantidad de valores perdidos que la variable de carbono total por lo que se la prefiere.

Para abordar esta problemática existe una primera consideración en relación al tipo de variable a predecir. Por un lado, se trata de una variable continua (valuada real) y al mismo tiempo, para la muestra de datos que se tiene, sigue una distribución no normal con varianza alta, una media centrada en la zona de los valores bajos de la variable, pero no representativa en términos del intervalo de valores posibles. En vistas a lo anterior, se decide discretizar la variable objetivo tomando en cuenta un criterio fijo. El tercer cuartil de la variable corresponde al valor 60,58, y se considera que a partir de ese valor el componente de carbono de biomasa microbiana del suelo es alto, por lo tanto, se divide el rango en dos intervalos consecutivos: hasta 60,58 y a partir de 60,59, componiendo una clase binaria: *verdadero* para alto y *falso* para bajo.

En primer término, se utiliza el algoritmo Máquina de Vector Soporte (*SVM* por sus siglas en inglés). Por tratarse de una clase desbalanceada, donde sólo el 19% de los registros corresponden a los valores altos de carbono, entonces se aplica una restricción de costo, penalizando en tres veces cada falso negativo. En lo que respecta a la configuración del algoritmo, el parámetro c de complejidad se fijó en 1, y el parámetro épsilon (error de redondeo) en 1^{-12}. Por otra parte, se eligió una función *Kernel* (núcleo) de tipo radial[x] con parámetro gamma 30 (determinado experimentalmente). A partir del procesado de los datos en esquema de validación cruzada de 10 pliegues se obtuvieron los resultados de clasificación que se muestran debajo (Tabla 20 y Tabla 21).

SVM: Instancias clasificadas correctamente 87,2%
Exactitud por clase

[x] K(x, y) = e^-(gamma * <x-y, x-y>) El algoritmo evalúa categorías de a pares.

Tabla 20 *Métricas de evaluación para SVM*

Clase	Razón.VP	Razón.FP	Precisión	Exhaustividad	F	MCC	Area.ROC	Area.PRC
Falso	0.882	0.158	0.944	0.882	0.912	0.684	0.862	0.921
Verdadero	0.842	0.118	0.703	0.842	0.766	0.684	0.862	0.632
Prom. Ponderado	0.872	0.148	0.884	0.872	0.875	0.684	0.862	0.849

Tabla 21 *Matriz de confusión para SVM*

Ref.	Clasificado.a	Clasificado.b
a = True	2263	304
b = False	135	720

Como se puede apreciar en la tabla superior (Tabla 20), el modelo obtenido se muestra eficaz a la hora de clasificar los casos verdaderos con un porcentaje de área bajo la curva ROC del 86%.

Como segundo algoritmo se considera el árbol de decisión C4.5 [19] (en su implementación J48), considerando el mismo esquema de costo. Como parámetros del modelo se establece un factor de confianza de 0,25 y un valor mínimo de objetos por nodo hoja de 2.

J48: Instancias correctamente clasificadas 88,52%
Exactitud por clase

Tabla 22 *Métricas de evaluación para J48*

Class	Razón.VP	Razón.FP	Precisión	Exhaustividad	F	MCC	Area.ROC	Area.PRC
Falso	0.892	0.137	0.951	0.892	0.921	0.716	0.904	0.945
Verdadero	0.863	0.108	0.728	0.863	0.790	0.716	0.904	0.804
Prom. Ponderado	0.885	0.130	0.896	0.885	0.888	0.716	0.904	0.909

Tabla 23 *Matriz de confusión para SVM*

Ref.	Clasificado.a	Clasificado.b
a = True	2291	276
b = False	117	738

En este caso se observan parámetros similares de exactitud (Tabla 22 y Tabla 23), con resultado levemente superior para el porcentaje bajo la curva ROC, siendo del 90%. A su vez, en este caso se dispone de una estructura resultante de árbol que permite analizar la lógica de clasificación.

Partiendo de la raíz del árbol se observa como la variable correspondiente al límite inferior de profundad presenta la mayor ganancia de información normalizada de entre todas las variables, generando dos ramas a partir de un valor de corte de aproximadamente 10m. En el caso de la rama que contiene los valores menores o iguales al punto de corte, pudiéndose catalogar como de las excavaciones poco profundas, la segunda variable en importancia de acuerdo al criterio de separación es pH. En este caso, el punto de corte es de aproximadamente 7, con lo cual podría hablarse de suelos ácidos, por un lado, y suelos alcalinos por otro. Tanto para la rama de suelo ácido, como en el caso de la rama correspondiente a excavaciones profundas, la siguiente variable en orden de relevancia es bioma. Así, se observan las siguientes reglas más relevantes (en términos de proporción de casos con alto contenido de carbono en biomasa microbiana del suelo):

- Para excavaciones poco profundas y suelos ácidos:
    - El bosque boreal, el desierto, los humedales naturales, el bosque caducifolio templado, el bosque de coníferas y la tundra, presentan contenido alto de carbono en biomasa microbiana.
    - El bosque tropical/subtropical en el sur del planeta.
    - En el caso de pastizales y prado, cuando la temperatura es relativamente baja (menor a 12ºC y 16ºC, respectivamente).

- Para excavaciones profundas:
    - El bosque boreal y los humedales naturales, en presencia de suelo no ácido (pH mayor a 5 y 6 respectivamente).
    - El bioma de arbustos con temperatura relativamente baja (menor a 18ºC) y suelo elevado (a más de 260 m).
    - El bosque caducifolio templado considerando el hemisferio este hasta América del Sur y el área este de América del Norte.
    - El bosque tropical/subtropical donde las precipitaciones son abundantes (mayores a 1056mm), el suelo es ligeramente ácido (pH entre 4 y 5) y el terreno se encuentra ligeramente elevado (menor a 35m).
    - El bioma tundra.
    - El prado con precipitaciones altas (mayor a 1024mm), en la norte y centro del planeta, con límite superior de profundidad de excavación relativamente bajo (menor a 6).
    - El bioma pastizales con suelos ácidos, límite inferior de profundidad no demasiado alto (menor a 20), relativamente baja temperatura (menor a 12ºC) y alta elevación (mayor a 114m).
    - El bioma tierras cultivables con límite inferior y superior de profundidad relativamente bajos (menor a 11 y 3 respectivamente), temperatura y precipitación más bien bajas (menor a 19ºC y entre 491 y 648mm) en el hemisferio este y en el sur y centro del planeta, al igual que los puntos del mismo bioma, pero en excavaciones bien profundas (entre 24 y 27m) en el hemisferio este.

Por último, se aplica un tercer método, pero esta vez con el objetivo de estimar el valor numérico de carbono en biomasa microbiana del suelo. En este sentido, se utiliza el algoritmo de árbol de

regresión (en su implementación *REPTree*), en ensamble, a partir de la técnica de *Bagging (Boostrap Aggregating)*. Esta técnica consiste en crear muestras separadas del conjunto de datos utilizado para entrenamiento, y construir un clasificador por cada uno. Los resultados de estos múltiples clasificadores se combinan (por promedio). La ventaja radica en que cada muestra del conjunto de entrenamiento es diferente (de tamaño fijo, generado por muestreo aleatorio con reemplazo), dando a cada clasificador que se entrena, un foco y perspectiva diferente del problema, reduciendo la varianza y el sobreajuste (*overfitting*). En este caso, el tamaño de la muestra se elige igual a la del conjunto utilizado para entrenamiento, y la cantidad de iteraciones se fija en 100.

En lo que se refiere al algoritmo base, la estructuración del árbol se rige por el criterio de minimización de varianza total. Luego, en cada hoja el valor de la variable a predecir se determina en función de la media de los casos de entrenamiento reunidos en esa hoja. En términos de parámetros, la profundidad máxima del árbol se fijó en 7, la cantidad mínima de instancias por hoja en 2, y la proporción mínima de varianza sobre todo los datos de un nodo que se requiere para realizar una división en 0,00005.

Como resultado, se obtuvo un árbol de 212 hojas, donde la variable seleccionada como raíz es longitud, y separa los casos en dos partes: los que se encuentran al este de los montes Urales (Europa y América) y los que se encuentran al oeste (Asia y Oceanía). En el primer caso, la segunda variable de corte es el Bioma, mientras que en el segundo lo es la elevación, en sitios elevados (a más de 155m) y no elevados. Todas las demás variables se utilizan para clasificar en lo sucesivo.

A continuación, se muestran los resultados de la validación cruzada de 10 pliegues (Tabla 24).

Tabla 24 *Métricas de evaluación para REPTree*

Métrica	Valor
Coeficiente de Correlación	0.8934
Error Absoluto Medio	30.1309
Raíz Cuadrada del Error Cuadrático Medio	74.8561
Error Absoluto Relativo	37.1503 %
Raíz Cuadrada del Error Cuadrático Relativo	44.9464 %

Se observa como la correlación entre el valor calculado para el modelo y el valor real de carbono en biomasa microbiana del suelo es alta. Incluso comparativamente hablando es más alto que el porcentaje de instancias clasificadas correctamente a partir de los modelos generados con los dos algoritmos anteriores. Asimismo, supera al modelo del algoritmo de árbol de regresión M5 [20] (en su implementación M5P), que presenta una correlación más baja (en el orden de 0,87) y errores más altos. Sin embargo, en términos generales los errores obtenidos son más bien altos, sobre todo si se los compara con las métricas derivadas de los modelos anteriores (donde por caso, la tasa de falsos positivos no supera el 15%). En contrapartida, la ventaja de este modelo es que permite estimar un valor numérico de la cantidad de carbono, y no solo una determinación de si este valor es alto o no.

Conclusión

El desarrollo del estudio permitió, por medio del análisis exploratorio y estadístico de las variables, probar la conjetura inicial acerca de la posibilidad de encontrar relaciones entre los biomas, variables geográficas y físico-químicas, con características del suelo. Si bien no se observa una alta correlación entre las variables pudieron determinarse algunas relaciones entre ellas en términos de su contribución a la varianza global del conjunto: tipo de variables de indicadores de concentración correlacionadas, variables de posicionamiento opuestas e influencia de profundidad de la excavación. Por otra parte, se ensayaron agrupaciones de los datos en base a distintos criterios y se pudieron establecer relaciones en cuanto a las variables intervinientes para con la variable de carbono en biomasa microbiana, destacando el método de agrupamiento jerárquico.

Por último, se observó que es factible determinar, con cierto grado de exactitud inherente a la simplificación del modelo propuesto, que tan probable es encontrar concentraciones altas de carbono, en función de las distintas variables que caracterizan a las muestras y, por otra parte, estimar un valor para dicha concentración.

Citas

[1] XU, X.; THORNTON, P.E.; POST, W.M. *A Compilation of Global Soil Microbial Biomass Carbon, Nitrogen, and Phosphorus Data* [en línea]. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, EEUU. 2014. Disponible en: DOI: 10.3334/ORNLDAAC/1264. Disponible en https://daac.ornl.gov/SOILS/guides/Global_Microbial_Biomass_C_N_P.html.
[2] BIOMASA. En *Diccionario de la lengua española*. 22.ª edición. Espasa. Madrid. 2001. p. 320.
[3] IGLESIAS, M.T. *Estudio del carbono de la biomasa microbiana en suelos alterados*. Lazaroa 29: 117-123. 2008.
[4] SCHWARTZ J.D. Soil as Carbon Storehouse: New Weapon in Climate Fight?. *Yale environment 360*. 2014. [fecha de consulta: 8 de mayo de 2016]. Disponible en <http://e360.yale.edu/feature/soil_as_carbon_storehouse_new_weapon_in_climate_fight/2744/. (traducción del original en inglés)>.
[5] PH [en línea]. En *Wikipedia, La enciclopedia libre, 2016*. Colaboradores de Wikipedia [fecha de consulta: 7 de febrero del 2016]. Disponible en <https://es.wikipedia.org/w/index.php?title=PH&oldid=88873566>.
[6] API DE GOOGLE MAPS. [fecha de consulta: 8 de mayo de 2016]. Disponible en <https://developers.google.com/maps/documentation/elevation/intro?hl=es-419>.
[7] GPS VISUALIZER. [fecha de consulta: 8 de mayo de 2016]. Disponible en <http://www.gpsvisualizer.com/elevation>.
[8] ANEXO: TIERRA BAJO EL NIVEL DEL MAR [en línea]. En *Wikipedia, La enciclopedia libre, 2016*. Colaboradores de Wikipedia [fecha de consulta: 23 de abril del 2016]. Disponible en <https://es.wikipedia.org/w/index.php?title=Anexo:Tierra_bajo_el_nivel_del_mar&oldid=90577871>.
[9] HIJMANS, R.J.; CAMERON S.E.; PARRA J.L. at all. *WorldClim versión 1*. Museum of Vertebrate Zoology, Universidad de California, Berkeley, California, EEUU.
[10] HIJMANS, R.J., CAMERON S.E.; J.L. PARRA; JONES P.G.; JARVIS A. *Very high resolution interpolated climate surfaces for global land areas*. International Journal of Climatology 25: 1965-1978. 2005.
[11] HIJMANS, R.J. at all. R Package 'raster': *Geographic Data Analysis and Modeling*. CRAN. 2016. p. 78.
[12] VAN BUUREN S.; GROOTHUIS-OUDSHOOM K. *Mice: Multivariate Imputation by Chained Equations in R*. Journal of Statistical Software. 2011. 45(3), pp. 1-67.
[13] STEKHOVEN. D.J.; BÜHLMANN P. *MissForest - nonparametric missing value imputation for mixed-type data*. 2011.
[14] BREIMAN L. *Random Forests*. Statistics Department, University of California, Berkeley, California, EEUU. 2001.
[15] STEKHOVEN. D.J. *Using the missForest Package*. 2011.
[16] ABDI, H.; WILLIAMS, L.J. *Principal Component Analysis*. Vol 2. John Wiley & Sons, Inc. 2010.
[17] THE WORLD BANK – CLIMATE PORTAL – ANNUAL PRECIPITATION 1960-1990 [en línea]. Banco Mundial [fecha de consulta 8 de mayo de 2016]. Disponible en <http://sdwebx.worldbank.org/climateportal/index.cfm?page=global_map>.
[18] HOPKIS B. *A New Method for determining the Type of Distribution of Plant IndividualsB*. Ann. Bot. 1954. 18 (2) p 213.

[19] QUINLAN R.J. *C4.5: Programs for Machine Learning*. San Francisco, California, EEUU. Morgan Kaufmann. 1993.
[20] QUINLAN R.J. *Learning with Continuous Classes*. 5th Australian Joint Conference on Artificial Intelligence. Singapore. 1992. pp 343-348.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. 2016.
Versión de R utilizada: 3.2.3.

FRANK E.; HALL M.A.; WITTEN I.H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Cuarta Edición. Morgan Kaufmann. 2016.
Versión de Weka utilizada: 3.7.