GUÍA SESIÓN HACKATHON DÍA 2

"Descubre los patrones secretos de la diversidad vegetal en Andalucía"

Resumen

Como base de datos partimos de los inventarios de vegetación en zonas naturales de Andalucía con coordenadas y el número de especies de plantas leñosas que tienen (Consejería de Medio Ambiente 2006). La variable de diversidad a explicar con vuestros conocimientos de ciencia de datos y ecología forestal sería el **número de especies de plantas leñosas**. Esta base de datos es única en Andalucía ya que se hicieron aproximadamente 50000 inventarios de vegetación para describir toda la variabilidad ambiental de Andalucía (Fig. 1). Para cada uno de estos inventarios, se identificaron todas las especies de plantas presentes, en un área acorde con el tipo de formación forestal (e.g. más grande para bosques y más pequeñas para pastizales).

Fig 1: muestreo de vegetación

Para este hackathon partiremos de **15591 inventarios** (Fig. 2) de los cuales, un 80% los usaréis para crear el modelo y el 20% restante para una evaluación objetiva. Os hemos facilitado el trabajo aportando un set de variables explicativas (también llamadas independientes) relacionadas con clima, topografía y suelo que podrían explicar los patrones de diversidad en Andalucía. Sin embargo, puede que no todas las variables sean relevantes a priori ni que estén todas las necesarias. Aprovechad vuestro conocimiento colectivo del equipo para seleccionar e incorporar variables que mejoren el ajuste y calidad del modelo.

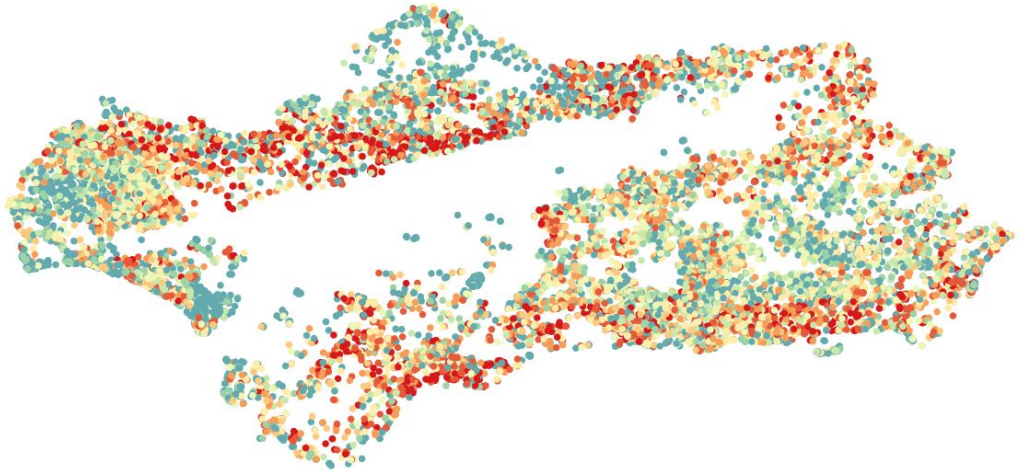


Fig. 2: Inventarios vegetación de Andalucía (Consejería de Medio Ambiente 2006).

Objetivo

En este reto tendréis que aunar esfuerzos para conseguir los el mejor modelo predictivo de diversidad vegetal en Andalucía basado en variables ambientales de acuerdo a los siguientes criterios:

1. Teoría ecológica e interpretación los resultados
2. Ajuste y precisión del modelo

¿Qué hay que entregar (jueves 4) ?

1. Dos diapositivas sobre vuestro modelo que demuestren el uso de la teoría ecológica y la interpretación de los resultados. Trabajad en GoogleSlides para luego pasar el enlace en el día de presentaciones.
2. Vuestras predicciones sobre los datos de test (ver instrucciones abajo).

Metodología

Para desarrollar este reto os recomendamos que trabajéis de forma colaborativa en GoogleColab. Como primer paso, podéis crear un enlace único para vuestro grupo y compartidlo únicamente con los organizadores (por si surgen dudas). Tenéis más información sobre esta herramienta en los tutoriales del hackathon (ver enlace abajo)

Google Colab en Python:
https://colab.research.google.com/notebook#create=true
Google Colab en R:
https://colab.research.google.com/notebook#create=true&language=r.

Fuente de datos principales

Archivo	Descripción	Formato	Origen de datos
inventario_hack.csv Descargar aquí (también podéis cargarlo directamente en GoogleColab, ver aquí)	Nuestra tabla maestra que incluye la variable a explicar y las predictoras.	csv delimitado por ;	https://laboratoriorediam.cica.es/geonetwork/srv/spa/catalog.search#/metadata/0a212501-dc46-4e9a-beac-04e14e137565

Contenido de la tabla - variables generales

Nombre variable	Descripción
cod_punt	Identificador del inventario
modelo	**test**: punto de inventario que se usará para evaluar objetivamente el ajuste de vuestro modelo. Para estos inventarios no tenéis la riqueza de especies, sólo las variables ambientales. **train**: punto de inventario que usaréis para calibrar vuestro modelo. Para estos inventarios tenéis tanto la riqueza como las variables ambientales. Son los puntos que tenéis que empezar a trabajar
riqueza_wood	Número de especies leñosas identificadas en el inventario
longitud	Longitud (X) en WGS84 del inventario
latitud	Latitud (Y) en WGS84 del inventario

Contenido de la tabla - variables ambientales (predictoras potenciales)
Estas variables tienen su origen en el proyecto Biomasa Forestal en Andalucía. Si quieres ver más información sobre cómo se calcularon las variables puedes ver el siguiente capítulo del proyecto

VARIABLES OROGRÁFICA		
1	mde	Elevación del terreno (m)
2	pte	Pendiente del relieve (grados)
3	cur_md	Curvatura media (1/m)
4	orien	Orientación del relieve (grados)
5	ins	Insolación (adimensional)
VARIABLES CLIMÁTICAS		
6	ptt	Precipitación anual (mm)
7	pin	Precipitación de invierno (mm)
8	pp	Precipitación de primavera (mm)
9	pv	Precipitación de verano (mm)
10	po	Precipitación de otoño (mm)
11	ta	Temperatura media anual (ºC)
12	tminf	Temperatura media de las mínimas del mes más frío (ºC)
13	tmaxc	Temperatura media de las máximas del mes más cálido (ºC)
14	osc1	Oscilación térmica media (ºC)
15	tmf	Temperatura media del mes más frío (ºC)
16	tmc	Temperatura media del mes más cálido (ºC)
17	osc2	Oscilación térmica total (ºC)
18	etott	Evapotranspiración de referencia anual (mm)
19	ssup	Suma de superavits (mm)
20	sdef	Suma de déficits (mm)
21	dseq	Duración de la sequía (meses)
22	iha	Índice hídrico anual (adimensional)
VARIABLES EDÁFICAS		
23	are	Media ponderada del contenido de arena en todo el perfil del suelo (%)
24	lim	Media ponderada del contenido de limo en todo el perfil del suelo (%)
25	arc	Media ponderada del contenido de arcilla en todo el perfil del suelo (%)
26	ps	Profundidad del suelo hasta el horizonte R (cm)
27	cod_hid	Media ponderada de la conductividad hidráulica saturada en todo el perfil del suelo (cm/día)
28	mo	Media ponderada de Materia Orgánica en el perfil del suelo (%)
29	mo_sup	Contenido de materia orgánica en el horizonte superficial del suelo (%)
30	ph	Media ponderada del pH en todo el perfil del suelo
31	tf	Media ponderada del contenido de tierra fina en todo el perfil del suelo (%)
32	ca	Caliza activa (%)
33	cic	Capacidad de intercambio catiónico (meq/100 gr)
34	psb	Porcentaje de saturación de bases (%)
35	n_sup	Contenido de nitrógeno en el horizonte superficial del suelo (%)
36	crad	Media ponderada de la capacidad de retención del agua en todo el perfil del suelo (mm/m)

Tenéis libertad para seleccionar y buscar nuevas variables de otras fuentes de datos. En la REDIAM tenéis muchísima información extra. Pensad en la variable que os falta que pueda mejorar el modelo de biodiversidad en Andalucía y buscad el ráster. Aquí tenéis el repositorio REDIAM.

Valoración de resultados

Mejor modelo ecológico

No siempre el mejor modelo (mayor ajuste) es el más apropiado para generalizar los datos. Para mejorar este aspecto es fundamental que el diseño del modelo, es decir las variables que usaremos como explicativas, sean relevantes y tengan un sentido ecológico. De esta forma garantizamos que no se generen relaciones falsas que no puedan ser generalizadas en otras zonas.

Por otro lado, cada tipo de modelo tiene ciertas asunciones y limitaciones. Así por ejemplo, para los modelos de regresión lineales hay que tener en cuenta entre otros los criterios de co-linealilidad y homocedasticidad. Es relevante que consideréis estos aspectos si bien es cierto que siempre habrá que tener un compromiso entre estas limitaciones y el objetivo del modelo (la predicción).

Valoraremos este **criterio de acuerdo a vuestro diseño del modelo y cómo se ajusta a la realidad ecológica de Andalucía**. También consideraremos las asunciones y aspectos que habéis tenido en cuenta a la hora del tipo de modelo. En este aspecto no tenéis que ser perfectos, pero cualquier aspecto que incluyáis se valorará positivamente.

Mejor ajuste del modelo sobre los datos de validación.

Aquí se trata de generar el mejor modelo predictivo. Es decir, el que predice mejor unos datos de diversidad no conocidos. En este caso usaremos la métrica de Error cuadrático medio (*RMSE* en inglés).

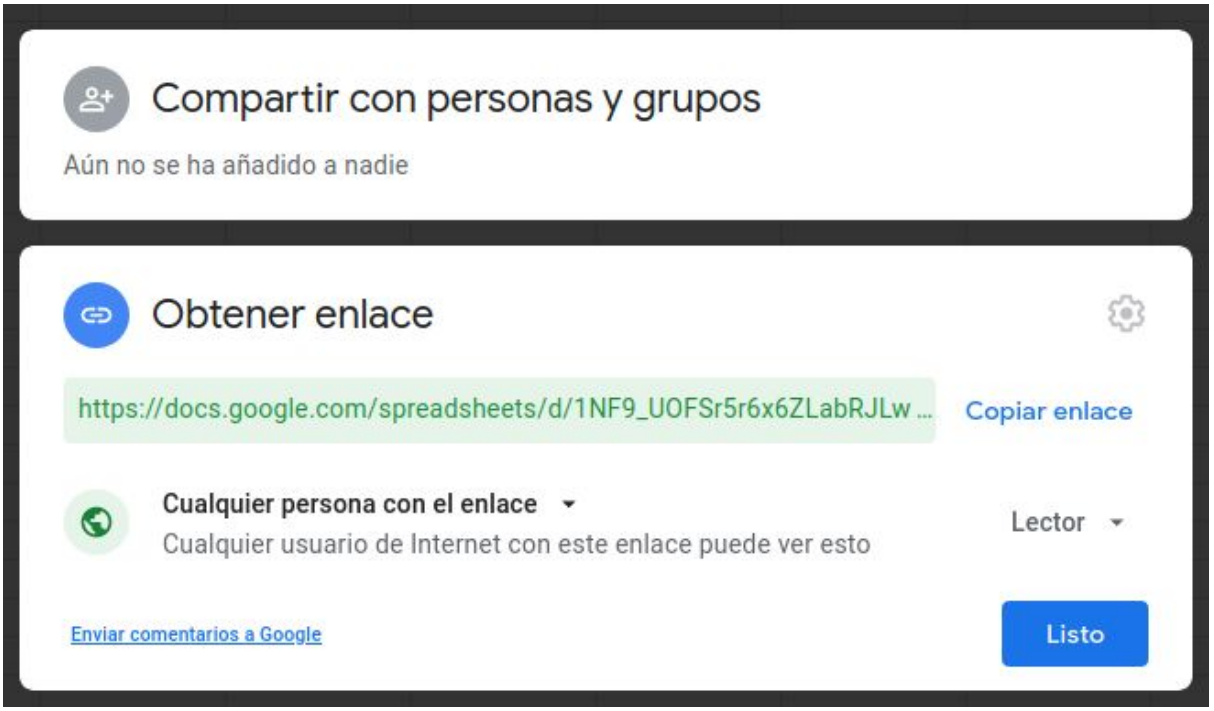
¿Cómo evaluaremos esta precisión?

En la tabla que os proporcionamos, los registros están divididos entre tipo “test” y “train”. Usareis los de tipo “train” para generar vuestro modelo. Una vez que habéis generado vuestro mejor modelo pasaréis a hacer una predicción sobre los datos de “test” para estimar la riqueza de especies. Una vez que tenemos las predicciones, copiamos **la columna “cod_punt” y la columna estimada de riqueza** en un documento de hoja de cálculo de Google. Cuando hayáis copiado la columna tendréis una hoja de cálculo como esta.

Para copiar y ser eficiente, haced click en el primer elemento y pulsáis Ctrl + Mayus + Flecha hacia abajo. Este comando selecciona toda la columna. Pulsad después Crtl + C y se copia. Para pegarlo en la hoja de Google, haced Ctrl + V.

La competición empieza ahora. Cada uno, cuando saque la hoja de Google con sus predicciones, tiene que sacar el enlace personal para que pueda ser evaluado. Esto se hace de la siguiente forma:

Estamos en la hoja de cálculo con las predicciones y pulsamos > Buscamos el botón verde de *Compartir* > Pulsamos en *Cambiar* o *Cambiar para cualquier persona con el enlace* y dejamos la configuración tal como se muestra en la siguiente imagen:



Y pulsamos *Copiar Enlace* y posteriormente *Listo*.

Una vez que tenemos este enlace con nuestras predicciones, nos vamos al formulario que se muestra aquí y lo rellenamos.

IMPORTANTE: Este paso solo hay que hacerlo una vez, ya que se actualiza automáticamente cuando nosotros peguemos una predicción diferente en la hoja creada en nuestro Google Drive.

Para ver los resultados que tenemos, nos vamos a esta hoja. Y aquí se muestra tanto nuestro RMSE (calidad de los modelos, se muestra junto a nuestro nombre) como el orden en el ranking final (de mejor a peor). Recordad, el RMSE, mientras más pequeño, mejor.

Metodologías que podrías necesitar

a) Atacar directamente repositorios de google drive
   - R:
     1) Para archivos GoogleSheet https://googlesheets4.tidyverse.org/
     2) Para archivos en GoogleDrive https://googledrive.tidyverse.org/
   - Python: https://pypi.org/project/gdown/

b) Análisis espacial:
   - QGIS: Extracción de datos ráster a punto (point sampling tool):
   - R: paquete raster función extract(). Ver este tutorial

c) Modelización
   - Guía de ciencia de datos en Python
   - Guía de ciencia de datos en R