

# IMM con R. Lab4: Análisis de Clasificación

Dra. Maite Mascaro y Dr. Edlin Guerra Castro

04/03/2022

Caso real: Contaminación en bentos por plataformas petroleras.

El siguiente caso de estudio fue publicado por Gray JS, Clarke KR, Warwick RM, Hobbs G (1990). Detection of initial effects of pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. Marine Ecology Progress Series 66, 285-299. Este estudio consiste en la evaluación del macrobentos y varios contaminantes del sedimento en 39 sitios dispuestos en un diseño radial alrededor de una plataforma de perforación petrolera en el mar del Norte, donde se espera que los contaminantes asociados a la actividad petrolera afecten la estructura del ecosistema. La disposición de los sitios es circular, alejándose cada ciertos kilómetros del centro de perforación. Para esta actividad ignoraremos la distancia de cada sitio respecto a la plataforma. Intentaremos clasificar las muestras según sus descriptores, es decir, dejaremos que los datos cuenten su propia historia en vez de forzarlos a un modelo específico.

##1. Análisis de clasificación o Cluster analysis. Evaluación de los métodos de agrupación

- a) Importe la matriz de datos “sedimentos”
- b) ¿Cuáles son las dimensiones del objeto?
- c) ¿Cuáles son las variables que allí aparecen y cuáles deberían incluirse en el análisis?
- d) Seleccione sólo las variables que ambientalmente describan la muestra de sedimento, incluyendo contaminantes. Llame a esa selección “sed”.
- e) Evalúe las escalas de magnitud de las variables, así como sus promedios. ¿Qué función de R podría ser útil para esto?
- f) Considerando la escala de magnitud de las variables, ¿qué tipo de pretratamiento debe ser aplicado?
- g) Aplique el pretratamiento y llame a la matriz resultante “sed.stand”
- h) Aplique distancias Euclidianas a la matriz pretratada y llame a ese objeto “DE.sed”
- i) Genere un dendrograma jerárquico con el método de agrupamiento simple (“single linkage”). Para esto será necesario aplicar la función `hclust`. Le sugerimos leer la documentación en R sobre `hclust` del paquete “Stats”. Por ahora le adelantamos que los argumentos necesarios son la matriz de distancias y el método de vinculación, que en este caso será el simple. Llame al cluster “single.c” y gráfiquelo. ¿Cuántos clusters (grupos) identifica? ¿Es capaz de reconocer algún patrón lógico? ¿Cuáles son los sitios que más se parecen? ¿Este gráfico evidencia de contaminación alrededor de la plataforma petrolera?
- j) Genere un nuevo dendrograma jerárquico, pero ahora con el método completo (“complete linkage”). Llame al cluster “comp.c” y gráfiquelo. Responda: ¿cuántos grupos identifica? ¿hay algún patrón más lógico? ¿cambió la clasificación respecto al método anterior?
- k) Genere un tercer dendrograma pero ahora con el método promedio (“average linkage”). Llame al cluster “ave.c” y gráfiquelo. Responda: ¿cuántos grupos identifica? ¿hay algún patrón más lógico? ¿cambió la clasificación respecto a los dos métodos anteriores? Para tener una mejor perspectiva, puede proyectar los tres dendogramas usando la función `par` con una fila y tres columnas, y llamando nuevamente a los tres dendogramas. Es recomendable asignar un título a cada plot para poder hacer comparaciones efectivas.

- l) Modifique la posición de las etiquetas agregando como atributo en cada plot “hang = -0.1”. ¿Prefiere esta salida o la anterior?

```
par(mfrow=c(1,3),mar=c(2,4,4,2))
plot(ave.c, main="average linkage", hang = -0.1)
plot(comp.c, main="Complete linkage", hang = -0.1)
plot(single.c, main="Single linkage", hang = -0.1)
```

- m) Hasta ahora, la identificación de grupos es completamente arbitraria. Usaremos un método computacionalmente intensivo para “podar” el dendrograma. Esta función es la `simprof` del paquete `{clustsig}`. Copie en su consola `??simprof` para obtener ayuda de cómo ejecutar `simprof`. Asigne un nombre al resultado `simprof` e identifique cuántos grupos realmente distinguibles hay explorando ese objeto.

```
sig.clust <-
  simprof(
    sed.stand,
    num.expected = 500,
    num.simulated = 99,
    method.cluster = "average",
    method.distance = "euclidean",
    method.transform = "identity",
    alpha = 0.05,
    sample.orientation = "row",
    const = 0,
    silent = TRUE,
    increment = 100
  )
```

Grafique su resultado con la función `simprof.plot`

```
par(mfrow = c(1, 1))
sig.clust.dend <- simprof.plot(sig.clust, leaflab = "perpendicular")
```

## ##2. Métodos de interpretación: Heat Maps para a variables biológicas

- Importe la matriz de datos “macrofauna.csv” y renómbrela “matriz.macrofauna”
- Evalúe el contenido de la matriz: unidades, escalas y dimensiones.
- Genere una nueva matriz pero filtrando las primeras dos columnas. Llámelas “macrofauna”
- Calcule descriptores univariados por cada sitio, use la riqueza de especies, la abundancia de individuos y el índice de diversidad de Simpson. Para ello genere un data.frame con las columnas como las variables, y filas como los sitios. Llame a la tabla generada “uni”. Trate de identificar patrones de variación en estos estimadores según el sitio
- Para usar una aproximación multivariada, aplique el índice de similitud Bray-Curtis luego de transformar las abundancias a raíz cuadrada. Llame a la matriz de abundancias transformadas “r.c” y a la matriz Bray-Curtis resultante “bray1”. Trate de identificar un patrón de asociación entre las muestras viendo a “bray1”.
- Genere un dendrograma jerárquico con el método de agrupamiento “average linkage” o UPGMA. Para esto será necesario aplicar la función `hclust`. Llame al cluster “cluster.bray” y conviértalo en dendrograma con nuevo nombre “dend.macro”. ¿Encuentra familiar esa representación? compárela con el dendrograma `ave.c` de las variables ambientales en un solo plot. Identifique un patrón respecto a grupos y muestras que lo conforman. ¿Hay correspondencia en ambos análisis?
- Identifique la relación entre las matrices con la función Mantel. Esto nos dará una idea de cuán similares son las relaciones espaciales según sus descriptores. En teoría, si los contaminantes afectan el macrobentos, debe haber correlación entre las matrices. Ejecute la prueba Mantel usando el coeficiente

de correlación Spearman y construya una hipótesis nula con 9999 permutaciones. `mantel(bray1, DE.sed, method="spearman", permutations=9999)`

- h) ¿Cuán importante cree usted que es la similitud entre ambas matrices? Sirve esta aproximación para asociar matrices de igual dimensión?
- i) Ahora apliquemos la prueba estadística “simprof” para podar el dendograma del macrobentos usando un método cuantitativo. ¿Cuántos grupos estadísticamente significativos identifica el análisis jerárquico usando el método promedio?

```
library(clustsig)
sig.clust.macro <-
  simprof(
    r.c,
    num.expected = 1000,
    num.simulated = 999,
    method.cluster = "average",
    method.distance = "braycurtis",
    method.transform = "identity",
    alpha = 0.05,
    sample.orientation = "row",
    const = 0,
    silent = TRUE,
    increment = 100
  )

sig.clust.dend.m <- simprof.plot(sig.clust.macro, leaflab = "perpendicular")
```

- j) Utilicemos ahora la función “heatmap” para apreciar la contribución de las especies a la estructuración de los grupos. Primero, como práctica, generemos un factor, con base en el grupo “distancia” en la matriz.macrofana.
- k) Identifique cuáles son los niveles del factor distancia generado llamando a ese objeto distancia.
- l) Lea en el menú de ayuda sobre la función “heatmap”. Esta función requiere definir varios argumentos, los más importantes son: la matriz con los valores de abundancia de especies y el dendograma final. Básicamente, la función reordena las filas en la matriz de abundancias, y asigna un color a las abundancias, que aumenta según la intensidad. Las columnas las asocia con base en un dendograma basado en correlaciones entre las especies. Luego genera un gráfico que facilita interpretar las especies que generan los grupos. Lo primero que debe hacer es convertir el dataframe “r.c” a una matriz de nombre “r.c.m”

Luego debe llamar a la función “heatmap”, incluyendo la matrix r.c.m y al dendograma “dend.macro”

- n) ¿Puede interpretar algo del HeatMap generado?
- o) Mejoresmos el heatmap incluyendo solo a las 10 especies que mejor contribuyen a diferenciar los grupos en las distancias 1 y 4. Para esto usaremos la herramienta SIMPER desarrollada por Clarke (1993), original del software PRIMER, pero incluida en la librería Vegan. Esta rutina descompone las similitudes entre cada par de sitios y relativiza el peso de cada especie en contribuir a la disimilitud promedio entre dos grupos. Para más información:

Apliquemos “simper” a la matriz “r.c” usando el factor distancia. Para poder acceder al resultado, llamaremos a este “simper\_1v4” y al resumen de los resultados sum1v4

- o) El objeto “sum1v4” es una lista, y como tal, podemos observar sus elementos uno a uno. Identifiquemos las especies que mayor contribución a las diferencias entre los grupos 1 y 4 con el siguiente comando
- p) Generemos un vector de nombre “sp” con los nombres de las 20 especies con mayor contribución a las diferencias entre la distancia 1 y 4 y usémoslo como subsetting en la matriz “r.c”, que simultáneamente

podemos convertir de data.frame a matriz con el nombre “r.c.sub”

- q) Genere un nuevo heatmap, pero usando la matriz de 20 especies. ¿Qué logra apreciar?
- r) Trate de mejorar el heatmap pero ahora permitiendo que las filas se agrupen según la distancia a la que pertenecen. ¿cómo haría esto?