IMM con R. Lab2: Medidas de Asociación: Distancias

Dra. Maite Mascaro y Dr. Edlin Guerra Castro

01/03/2021

Parte 1: Análisis en modo R

Existe el interés de explorar como cambian las condiciones abióticas del agua en distintas localidades de la costa nor-occidental de México, con la finalidad de explorar potenciales relaciones entre dichas variables, así como la similitud entre distintas localidades con base en sus condiciones abióticas. Las variables fueron: temperatura (°C), salinidad (ups), pH, oxígeno disuelto (mg/l), clorofila (mg/m^3), nitrito (mg/l). Los datos se encuentran en la pestaña 'abiot' del archivo 'datosIAM.xlsx'.

1. Importa los datos registrados, copiando el siguiente código. Identifica qué hace cada línea de comando. Examina el objeto que creaste y responde a las siguientes preguntas:

```
library(readxl)
dat <- read_excel("datosIAM.xlsx", sheet = "abiot")
dat
```

   a) ¿Cuántos descriptores están siendo usados para caracterizar las 6 localidades?
   b) ¿Son similares las escalas (y unidades) en las que están siendo medidas las variables abióticas? Explica.
   c) Aplica la función `summary` sobre el objeto *dat* y describe lo que obtienes. ¿Sirve esta información resumida para comprender cómo se correlacionan unas variables con otras?
   d) Copia el siguiente código para trasponer la matriz *dat*, y aplica la función `summary` sobre la nueva matriz *dat.t*. ¿Sirve esta informaccción para comparar las localidades?

```
dat.t <- t(as.matrix(dat))
summary(dat.t)
```

   e) A partir de la observación detenida de 'dat', ¿puedes saber si algunas localidades están caracterizadas por ciertos valores de las variables (por ejemplo, salinidad alta, temp baja)?

2. Usa la función `stack` para apilar la columnas (valores de variables) unas encima de otras. Esto te permitirá elaborar gráficas usando la función `qplot` (del paquete `ggplot2`) con la información contenida en *dat*. Para ello corre el código y sigue estas intrucciones:

```
library(ggplot2)
dat.s <- data.frame(stack(dat), loc = row.names(dat))
```

   a) Examina el objeto *dat.s* y describe a qué se refiere **ind**, **values** y **loc**.
   b) Copia el siguiente código e interpreta la gráfica de salida ¿Qué puedes decir sobre las diferencias en los valores de las variables para las distintas localidades? ¿Hay variables que varían conjuntamente?

```
qplot(
  data = dat.s,
  x = loc,
  y = values,
  group = ind,
  col = ind,
  shape = ind,
```

```
  geom = c("point", "line"),
  xlab = "Localidad",
  ylab = "Unidades de Medición"
)
```

c) Copia ahora el siguiente código e interpreta esta segunda gráfica. ¿Qué puedes decir sobre las diferencias entre localidades?¿Hay localidades que se parecen entre si por los valores de sus variables?

```
qplot(
  data = dat.s,
  x = ind,
  y = values,
  group = loc,
  col = loc,
  shape = loc,
  geom = c("point", "line"),
  xlab = "Variables",
  ylab = "Localidades"
)
```

d) ¿Cómo harías un resúmen de la información numérica y gráfica obtenida hasta ahora? ¿Puedes responder a alguna de las preguntas de los investigadores?

3. Si obtuvieras una matriz con los valores de covariación entre todas las variables abióticas, considerando los valores en las 6 localidades:

a) ¿Cuántas columnas y filas tendría dicha matriz? ¿Qué forma tendría?
b) ¿Qué habría en la diagonal?
c) Obtén ésta matriz de covariación usando la función `cov` y verifica tus respuestas.
d) ¿Cuáles son las variables que se correlacionan directamente y cuáles los hacen inversamente? ¿Cuáles son las que están más fuertemente asociadas? Para ello usa la función `cor`.
e) Aplica la función `pairs` a la matriz de correlaciones, y examina el resultado gráfico. ¿Cuáles son las unidades de los ejes en los gráficos obtenidos?

Medidas de distancias

4. Obtén una matriz con las medidas de distancia Euclidiana entre todas las localidades considerando los valores de las 6 variables abióticas. Usa la función `dist` del paquete `stats` e identifica qué ocurre en cada una de las tres línea de comando.

```
dist(dat, method = "euclidean")
dist(dat, method = "euclidean", diag = TRUE)
dist(dat, method = "euclidean", diag = TRUE, upper = TRUE)
```

a) ¿qué ocurre en cada una de las tres línea de comando?
b) ¿Cuántas columnas y filas tiene la matriz? ¿Qué forma tienen?
c) ¿Qué hay en la diagonal?
d) ¿Puedes decir qué localidades son las más y las menos similares entre si?
e) Si aplicas la función `sort` al objeto resultante de `dist` puedes ordenar estos datos en una escala de distancia euclideana ¿Sirve esto a tu propósito de describir resumidamente las diferencias entre localidades?

5. Copia el siguiente código para elaborar una gráfica tipo raster de la matriz de distancias euclideanas

```
DE <- dist(dat, method = "euclidean")
DE.s <- data.frame(stack(data.frame(as.matrix(DE))), loc = row.names(dat))
qplot(
```

```
  data = DE.s,
  x = ind,
  y = loc,
  fill = values,
  geom = c("raster", "text"),
  label = round(values, digits = 2),
  xlab = "Localidad",
  ylab = "Localidad",
  main = "Euclideana",
  size = 1
)
```

a) ¿Qué ganaste con esta gráfica en términos de la descripción de los datos abióticos?
b) Tomando en cuenta la información del objeto *dat*, identifica la variable que tiene más preponderancia para hacer que dos localidades se parezcan (o distingan). Explica tu respuesta.
c) ¿Consideras que la distancia euclideana representa con fidelidad qué tanto se parecen 2 localidades por sus condiciones abióticas?

6. Para ver el efecto de una estandarización de las variables sobre la distancia euclideana, aplica la función `decostand` de `vegan` siguiendo el código a continuación. El argumento **standardize** transforma cada medición en z-scores (ésta centra y divide entre desviación estandar) para volver las medidas comparables.

```
library(vegan)
dat.stan<-decostand(dat, method = "standardize")
DE.stan <- dist(dat.stan, method = "euclidean")
DE.stan
```

a) Elabora el gráfico raster con la matriz de DE transformada, sustituyendo los nombres de los objetos en el código del numeral 5, y compara ambas gráficas. ¿Qué cambió?

b) ¿Cual de las dos formas de resumir la información te parece más realista?

7. Para ver el efecto de distintas medidas de asociación sobre estos datos ambientales, obtén las distancias de **Hellinger** y **Manhattan**, a partir de la matriz *dat* y compara las matrices triangulares con la de **DE.stan**. Nota: para obtener la transformación de **Hellinger**, se aplica `decostand` a la matriz *dat* previo a obtener la euclideana. Si tienes dificultades usa la función `help`.

a) Copia el siguiente código para comparar los valores de las medidas usadas en esta actividad en términos de sus escalas relativas. ¿Cuáles fueron las diferencias y similitudes en la escala entre estas medidas de asociación?

```
par(mfrow = c(2, 2), mar = c(2, 4, 2, 2))
dotchart(as.vector(DE), main = "Euclideana")
dotchart(as.vector(DE.stan), main = "Euclideana (estandarizada)")
dotchart(as.vector(DH), main = "Hellinger")
dotchart(as.vector(DM), main = "Distancia Manhattan")
par(mfrow = c(1, 1))
```

b) Obtén las gráficas raster con base en las 4 medidas de asociación usadas en esta actividad para apoyar tu exploración. ¿Cuáles fueron las diferencias y similitudes en el agrupamiento de las localidades usando las distintas medidas de asociación?

```
DE.s <- data.frame(stack(data.frame(as.matrix(DE))), loc = row.names(dat))

p1<- qplot(
  data = DE.s,
  x = ind,
```

```
  y = loc,
  fill = values,
  geom = c("raster", "text"),
  label = round(values, digits = 2),
  xlab = "Localidad",
  ylab = "Localidad",
  main = "Euclideana",
  size = 1
)

DE.stan.s <- data.frame(stack(data.frame(as.matrix(DE.stan))), loc = row.names(dat))
p2 <- qplot(
  data = DE.stan.s,
  x = ind,
  y = loc,
  fill = values,
  geom = c("raster", "text"),
  label = round(values, digits = 2),
  xlab = "Localidad",
  ylab = "Localidad",
  main = "Euclideana (estandarizada)",
  size = 1
)
DE.stan.s

DH.s <- data.frame(stack(data.frame(as.matrix(DH))), loc = row.names(dat))
p3 <- qplot(
  data = DH.s,
  x = ind,
  y = loc,
  fill = values,
  geom = c("raster", "text"),
  label = round(values, digits = 2),
  xlab = "Localidad",
  ylab = "Localidad",
  main = "Hellinger",
  size = 1
)
DH.s

DM.s <- data.frame(stack(data.frame(as.matrix(DM))), loc = row.names(dat))
p4 <- qplot(
  data = DM.s,
  x = ind,
  y = loc,
  fill = values,
  geom = c("raster", "text"),
  label = round(values, digits = 2),
  xlab = "Localidad",
  ylab = "Localidad",
  main = "Manhattan",
  size = 1
)
```

```
DM.s

library(patchwork) #Con este paquete podemos acomodar gráficos generados con ggplot2

(p1 + p2) / (p3 + p4)
```

c) Con base en los datos de la matriz dat ¿cuál medida escogerías para describir estos datos?

Recuerde cargar las respuestas a estas preguntas en el archivo **mi_solucion_lab2.Rmd** y convertirlo con `Knit` en documento Word. Descargue el archivo y súbalo en el Google Classroom como entregable de la tarea.