Análisis multivariante de la comunidad

Carlos Iván Espinosa

Octubre 2016

Contents

Prefacio

La comunidad biológica se refiere a una agrupación de poblaciones de especies que se presentan juntas en el espacio y el tiempo (Begon et al. 1999). Este concepto plantea que las comunidades tienen unos límites en el espacio y el tiempo, y que estos límites están dados por la distribución de las poblaciones. Sin embargo, la distribución de las poblaciones no es homogénea y cada población responde diferente en el espacio y el tiempo.

De esta forma la caracterización de una comunidad biológica se constituye en un reto ya que implica poder rescatar los efectos que se dan a varios niveles en la comunidad. El definir por ejemplo ¿Dónde inicia y termina una comunidad? o ¿Cómo difieren las comunidades entre localidades? o ¿Cómo la comunidad responde a las condiciones ambientales o disturbios? representan algunas de las principales preguntas que necesitamos responder. Una de las formas de responder estas preguntas puede ser intentar cuantificar las similitudes entre localidades.

Objetivos

- Comprender las bases teóricas para el cálculo de similitudes de la estructura de la comunidad entre localidades.
- Utilizar herramientas de análisis para calcular índices de similitud y distancias entre comunidades.

Figure 1: Stenocercus iridicens

Chapter 1

Medidas de similitud

La caracterización de una comunidad biológica presenta varios retos a los ecólogos. ¿Dónde inicia y termina una comunidad? ¿Cómo difieren las comunidades entre localidades? ¿Cómo la comunidad responde a las condiciones ambientales o disturbios? ¿Cómo se mantiene la diversidad en un área determinada? son algunas de las temáticas con mayor desarrollo científico dentro de la ecología de comunidades. En el presente capítulo se introduce a los estudiantes en los conceptos que los ecólogos utilizan para comparar comunidades y se presenta una guía de algunos de los análisis básicos de análisis de la estructura de las comunidades.

1.1 Medidas de abundancia

> "La abundancia se refiere al número de individuos de una especie en una determinada área"
>
> — (Smith and Smith 2010)

Cuando hablamos de la composición de especies de una comunidad nos referimos al conjunto de especies que habitan una determinada localidad. Típicamente, esto incluye cierto grado de abundancia de cada especie, pero puede también ser simplemente un listado de especies en esa localidad, donde se registra la presencia o ausencia de cada especie. Ahora, imaginemos que tenemos cuatro localidades (A, B, C, D) donde recogemos los datos de densidad de dos especies; *Tabebuia billbergii* y *Geofroea spinosa*, especies características de bosques secos tropicales. Podemos introducir datos hipotéticos de abundancia para cada especie en cada una de las localidades.

```
dens <- data.frame(T.bil = c(1, 1, 2, 3), G.spi = c(21, 8, 13, 5))
row.names(dens) <- LETTERS[1:4]
dens
```

```
##   T.bil G.spi
## A     1    21
## B     1     8
## C     2    13
## D     3     5
```

Generamos un gráfico para ver cuánto se parece cada sitio (Figura 1.1)

```
par(mar=c(4,4,1,1), mgp=c(1,0.3,0), tcl= -0.2)
plot(dens, type = "n", cex.axis=0.8)
text(dens, row.names(dens), col ="blue")
```

En la Figura 1.1 vemos que la composición de especies en el sitio A es diferente de la composición del sitio D. Es decir, la distancia entre el sitio A y D es mayor que entre los otros sitios. Lo siguiente que nos deberíamos

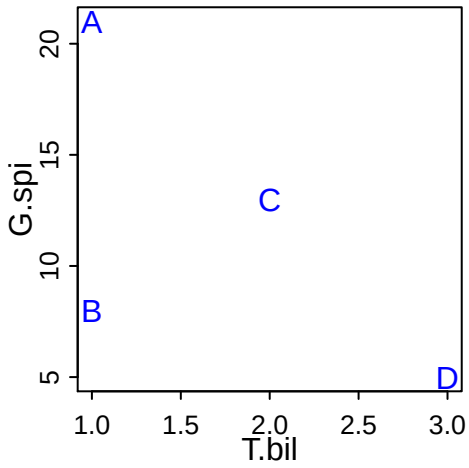


Figure 1.1: Distancias de cuatro localidades hipotéticas

preguntar es; ¿qué tan distantes están los dos sitios? Claramente, esto depende de la escala de medición (los valores de los ejes), y sobre cómo medimos la distancia a través del espacio multivariado (Stevens, 2009).

Estas diferencias entre sitios son dependientes de la abundancia de cada especie. En el caso de *G. spinosa* su eje varía entre 5 y 21, mientras que para *T. billbergii* varía entre 1 y 3. Una forma de corregir esta distorsión es calcular la densidad relativa de cada especie, de esta forma cada especie variará entre 0 y 1 (Stevens, 2009). Para ello dividimos la abundancia de cada especie para la suma total de los individuos de las especies en esa muestra.

> "La distancia se refiere a la diferencia en un espacio multidimensional (dado por las especies) entre dos comunidades. Esta distancia puede ser medida por múltiples vías"

```
dens[1, ]/sum(dens[1, ])
```

```
##        T.bil     G.spi
## A 0.04545455 0.9545455
```

Este resultado nos muestra que el sitio A esta constituido en un 95% por *G. spinosa*, mientras que *T. billbergii* aporta únicamente el 5%. Cuando nos referimos a densidad relativa hablamos de la densidad de una especie con referencia a algo, en el caso anterior con referencia a otras especies en el mismo sitio, pero también podríamos calcular en relación a otros sitios la misma especie.

```
dens[,1]/sum(dens[,1])
```

```
## [1] 0.1428571 0.1428571 0.2857143 0.4285714
```

Ahora podemos ver cómo *T. billbergii* varía en su abundancia en los cuatro sitios. El sitio A y B tienen el 14% de individuos mientras que el D tiene el 42% de los individuos de esta especie.

Ya sea que nuestras medidas de abundancia son absoluta o relativa, nos interesa conocer cuan diferente es la comunidad de una muestra (o sitio) con relación a la otra. En el ejemplo ha sido fácil entender la diferencia entre las dos comunidades debido a que teníamos únicamente dos especies, pero con más de tres especies es complicado observar estas diferencias gráficamente. Tal vez la forma más sencilla de describir la diferencia entre los sitios es calcular las *distancias* entre cada par de sitios.

1.2 Distancias entre sitios

La *distancia* entre dos muestras está dada por la diferencia entre la abundancia y la composición de especies, como lo hemos visto esto genera una distancia, en el caso del ejemplo la comunidad A esta más alejada de la comunidad D que de las otras dos.

Existen muchas formas de poder calcular las distancias entre estos puntos una de las más sencillas es la distancia *Euclidiana*. La distancia euclidiana entre dos sitios es simplemente la longitud del vector que conecta los sitios y la podemos obtener como $\sqrt{x^2 + y^2}$, donde *"x"* y *"y"* son las coordenadas (x, y) de distancia entre un par de sitios.

En nuestro caso si queremos comparar B y C tenemos que la distancia en el eje x es la diferencia de la abundancia de *T. bilbergii* entre el sitio B y C.

```
x <- dens[2, 1] - dens[3, 1]
```

Mientras que la distancia en el eje y es la diferencia en la abundancia de *G. spinosa* entre el sitio B y C.

```
y <- dens[2, 2] - dens[3, 2]
```

Ahora obtenemos las distancias entre los dos sitios

```
sqrt(x^2 + y^2)
```

```
## [1] 5.09902
```

Pero como en *R* todo es sencillo podemos utilizar la función *dist*

```
dist(dens)
```

```
##           A         B         C
## B 13.000000
## C  8.062258  5.099020
## D 16.124515  3.605551  8.062258
```

Si bien este cálculo es sencillo con dos especies, si tenemos que calcular la distancia para una comunidad con más de tres especies los cálculos son tediosos y largos. Para calcular la distancia *Euclidiana* entre pares de sitios con R especies utilizamos la siguiente ecuación:

$$D_E = \sqrt{\sum_{i=l}^{R} (x_{ai} - x_{bi})^2}$$

Distancia Euclidiana

Existen otras formas de medir distancias entre dos localidades. En ecología una de las distancias más utilizada es la distancia de *Bray-Curtis*, conocida también como *Sorensen*. Esta distancia es calculada como:

$$D_{BC} = \sum_{i=l}^{R} \frac{(x_{ai} - x_{bi})}{(x_{ai} + x_{bi})}$$

Distancia de Bray-Curtis

La distancia *Bray-Curtis* no es más que la diferencia total en la abundancia de especies entre dos sitios, dividido para la abundancia total en cada sitio. La distancia Bray-Curtis tiende a resultar más intuitiva debido a que las especies comunes y raras tienen pesos relativamente similares, mientras que la distancia euclidia depende en mayor medida de las especies más abundantes. Esto sucede porque las distancias euclidianas se basan en diferencias al cuadrado, mientras que Bray-Curtis utiliza diferencias absolutas. El elevar un número al cuadrado siempre amplifica la importancia de los valores más grandes. En la figura 1.2 se compara gráficos basados en distancias euclidianas y Bray-Curtis de los mismos datos.

Como se había comentado es virtualmente imposible representar una distancia en más de tres dimensiones (cada especie es una dimensión). Una forma sencilla de mostrar distancias para tres o más especies es crear un gráfico de dos dimensiones, intentando organizar todos los sitios para que las distancias sean aproximadamente las correctas. Está claro que esto es una aproximación nunca estas serán exactas. Una técnica que intenta crear un arreglo aproximado es escalamiento multidimensional no métrico (NMDS). Vamos a calcular las distancias para nuestra comunidad, primero vamos a añadir dos especies más a nuestra comunidad, *Ceiba trichistandra* y *Colicodendron scabridum*.

```
dens$C.tri<- c(11, 3, 7, 5)
dens$C.sca<- c(16, 0, 9, 4)
```

La función de escalamiento multidimensional no-métrico está en el paquete `vegan`. Aquí mostramos las distancias euclidianas entre sitios (Figura 1.2a) y las distancias de Bray-Curtis (Figura 1.2b).

```
library(vegan)

#Distancia Euclidiana
mdsE <- metaMDS(dens, distance = "euc", autotransform = FALSE, trace = 0)
#Distancia de Bray-Curtis
mdsB <- metaMDS(dens, distance = "bray", autotransform = FALSE, trace = 0)
```

```
par(mfcol=c(1,2), oma=c(1,1,1,1), mar=c(4,4,1,1),
    mgp=c(1,0.3,0), tcl= -0.2)

plot(mdsE, display = "sites",
     type = "text",main="a)Euclidiana",
     cex.axis= 0.7, cex.main=0.75, cex.lab=0.7)

plot(mdsB, display = "sites", type = "text",
     main="b)Bray-Curtis",
     cex.axis= 0.7, cex.main=0.75, cex.lab=0.7)
```

1.3 Similitud

Ahora que sabemos cuan distantes son los diferentes sitios, muchas veces nos podría interesar cuan similares son cada uno de los sitios a continuación se describen dos medidas de similitud; *Porcentaje de Similitud* e *Índice de Sorensen*.

El *porcentaje de similitud* puede ser simplemente la suma de los porcentajes mínimos de cada especie en la comunidad. Lo primero que debemos hacer es convertir la abundancia de cada especie a su abundancia relativa dentro de cada sitio. Para ello dividimos la abundancia de cada especie por la suma de las abundancias en cada sitio.

```
dens.RA <- t(apply(dens, 1, function(sp.abun) sp.abun/sum(sp.abun)))
dens.RA
```

```
##        T.bil     G.spi     C.tri     C.sca
## A 0.02040816 0.4285714 0.2244898 0.3265306
## B 0.08333333 0.6666667 0.2500000 0.0000000
## C 0.06451613 0.4193548 0.2258065 0.2903226
## D 0.17647059 0.2941176 0.2941176 0.2352941
```

El siguiente paso para comparar entre sitios, es encontrar el valor mínimo para cada especie entre los sitios que debemos comparar. Vamos a comparar los sitios A y B, para esto utilizamos la función `aplly`, la cual nos permite encontrar el valor mínimo entre las filas 1 y 2 (sitio A y B respectivamente). Para *T. billbergi* en el sitio A la abundancia relativa es 0.02 que es menor a la abundancia en el sitio B que es de 0.08.

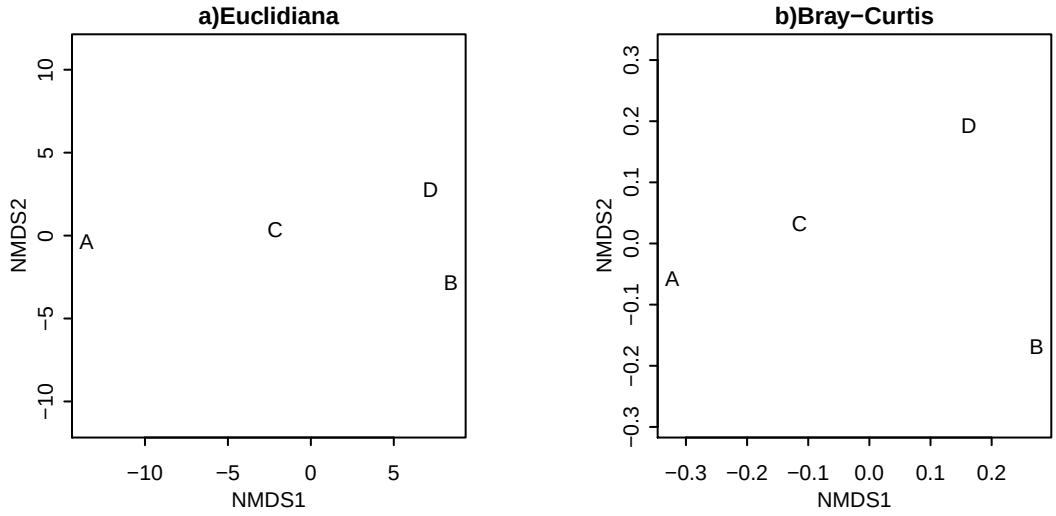


Figure 1.2: Arreglo de las parcelas en distancias multidimensionales no métricas (NMDS). Estas dos figuras muestran los mismos datos en bruto, pero las distancias euclidianas tienden a enfatizar las diferencias debidas a las especies más abundantes, mientras que Bray-Curtis no lo hace.

```
mins <- apply(dens.RA[1:2, ], 2, min)
mins
```

```
##      T.bil      G.spi      C.tri      C.sca
## 0.02040816 0.42857143 0.22448980 0.00000000
```

Finalmente para conocer el porcentaje de similitud entre los dos sitios sumamos estos valores y multiplicamos por 100.

```
sum(mins)*100
```

```
## [1] 67.34694
```

Esto significa que la comunidad A y B tienen un porcentaje de similitud del 67%.

El índice de Sorensen es la segunda medida de similitud que vamos a estudiar, este índice es medido como:

$$S_s = \frac{(2C)}{(A+B)}$$

Índice de Sorensen

Donde C es el número de especies en común entre los dos sitios, y A y B son el número de especies en cada sitio. Esto es equivalente a dividir las especies compartidas por la riqueza media.

Para calcular el índice de Sorensen entre los sitios A y B necesitamos definir el número de especies compartidas y luego la riqueza de cada uno de los dos sitios.

Definimos si alguna de las especies en uno de los sitios la abundancia no es igual a cero, eso nos dirá en qué casos se comparten especies. Finalmente, sumamos todas las especies que su abundancia es mayor a cero.

Table 1.1: Comunidades hipotéticas

	sp1	sp2	sp3	sp4	sp5	sp6	sp7	sp8
A	26	17	16	1995	159	0	362	0
B	0	35	14	236	54	0	496	57
C	24	0	26	17	88	18	907	20
D	35	18	24	2033	175	15	376	16
E	105	129	40	18	191	53	964	134

```
comp<- apply(dens[1:2, ], 2, function(abuns) all(abuns != 0))
comp
```

```
## T.bil G.spi C.tri C.sca
##  TRUE  TRUE  TRUE FALSE
```

```
Rs <- apply(dens[1:2, ], 1, function(x) sum(x > 0))
Rs
```

```
## A B
## 4 3
```

Como vemos, la abundancia de *C. scabridum* en uno de los dos sitios es igual a Cero, lo confirmamos al tener la riqueza por sitio. El sitio B tenemos únicamente 3 especies.

Ahora aplicamos la formula, dividimos las especies compartidas (*comp*) para la riqueza total de los dos sitios y lo multiplicamos por 2.

```
(2*sum(comp))/sum(Rs)
```

```
## [1] 0.8571429
```

Según el índice de Sorensen estos dos sitios son parecidos en un 86%. Los datos de los dos índices utilizados difieren entre sí, el porcentaje de similitud utiliza no solamente la presencia ausencia sino también la abundancia lo que podría estar reduciendo la similitud entre sitios.

1.4 Ejercicio: Análisis de similitud

Una de las preguntas básicas de un ecólogo es saber ¿Cómo de diferentes son dos comunidades?, en el presente ejercicio nos interesa entender comprender la similitud y distancias entre estas cinco comunidades hipotéticas (tabla 1.1)

Con los datos anteriores:

a. Convierta los datos en abundancia relativa por sitio (la suma en cada sitio debe ser igual a 1). Dibuje dos gráficas para representar; i) la abundancia total y ii) abundancia relativa de cada localidad. ¿Qué diferencias puede ver en la gráfica i y en la ii?

b. Calcule la distancia Euclideana y de Bray Curtis para cada sitio con las dos medidas de abundancia y grafíquelas utilizando el NMDS. ¿Cómo cambia entre distancias y abundancias? Explique las diferencias.

c. Evalúe la similitud (Sorensen) y el porcentaje de similitud entre pares de sitios. ¿Cuáles son los sitios más similares? ¿Cuál es la razón de las diferencias entre los índices utilizados?

Chapter 2

Análisis multivariado de la composición de la comunidad

Los índices de similitud nos permiten comparar las comunidades entre dos sitios, pero claramente cuando estudiamos las comunidades nuestros datos no son tan sencillos como lo que hemos utilizado hasta el momento. El organizar los datos de composición de la comunidad y poder interpretarlos en relación a otras comunidades, entender que comunidades son más similares entre sí, y saber si esta similitud o distancia es el resultado de unas respuestas al entorno pueden ser algunas de las cosas que podremos responder utilizando las técnicas de análisis multivariado de la comunidad. A continuación vamos a describir algunas técnicas de clasificación y ordenación que nos permitirán abordar estas temáticas.

Las técnicas de ordenación y clasificación son estrategias alternativas para simplificar los datos. La ordenación intenta simplificar los datos en un mapa que muestra las similitudes entre los puntos. La clasificación simplifica datos colocando los puntos similares en una misma clase o grupo Oksanen 2014[1].

Utilizaremos el paquete *Vegan* para los análisis de ordenación y clasificación, para mayor información puede referirse a Oksanen 2013[2].

2.1 Agrupamiento Jerárquico (Hierarchic Cluster)

A continuación vamos a realizar un análisis Cluster (análisis de conglomerados) utilizando la función *hclust* del paquete *vegan*. La función hclust necesita una matriz de disimilitudes como entrada. El Análisis de conglomerados intenta generar conglomerados que tengan la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

Aunque la función *dist* nos permite calcular disimilitudes, para el análisis de comunidades biológicas utilizaremos la función *vegdist* del paquete *vegan*. Esta función nos permite calcular varios índices de disimilitud. El método de cálculo de la disimilitud por defecto es Bray-Curtis (*"bray"*).

Una de las características importantes del método Bray-Curtis es que varía entre 0 y 1, dos comunidades que no comparten ninguna especie tendrían 1 como resultado.

Calculemos una matriz de disimilitudes usando el método Bray-Curtis, utilizaremos los datos de Barro Colorado Island (BCI) cargados en el paquete *vegan*. Para eso necesitamos cargar el paquete y los datos de BCI, únicamente utilizaremos los datos de los primeros 10 sitios.

[1] http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf
[2] http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

```
library(vegan)
data(BCI)

dist<- vegdist(BCI[1:10,], method="bray")
dist[1:10]
```

```
##  [1] 0.2706682 0.3501647 0.3682008 0.3725079 0.3744186 0.3518519 0.3424346
##  [8] 0.4235706 0.3770140 0.2873051
```

Podemos ver que el sitio 1 es 27% diferente al sitio 2, 35% al sitio 3, 36% al sitio 4 y así sucesivamente con los 10 sitios.

Con la matriz de disimilitudes calculada se puede analizar los puntos que conforman una agrupación. Utilizaremos los métodos de agrupación de la función *hclust* que nos propone tres métodos de agrupamiento: agrupación simple, agrupación completa y agrupación promedio.

Todos los métodos inician con el agrupamiento de las dos comunidades (dos sitios) más similares y a partir de esta primera comparación se continúa con el resto de puntos.

A continuación ejemplificaremos el cálculo de las distancias usando los tres métodos. Extraemos los cinco primeros sitios de la matriz de BCI y generamos un nuevo objeto (S_BCI). Con este nuevo objeto calculamos la distancia entre los cinco sitios.

```
S_BCI<- BCI[1:5,]
dist1<- vegdist(S_BCI, method="bray")
dist1
```

```
##           1         2         3         4
## 2 0.2706682
## 3 0.3501647 0.2873051
## 4 0.3682008 0.3149523 0.3244078
## 5 0.3725079 0.3851064 0.3595041 0.3721619
```

1. En base de la matriz de disimilitudes se busca el par de puntos que se encuentren más cercanos (menos disimiles). En nuestro caso el punto 1 y 2 tienen la distancia más baja 0.27. Una vez identificado, inicia el proceso de agrupación y es donde se diferencian los tres métodos.
2. Con el primer grupo generado debemos comenzar la construcción del resto de grupos, para esto construimos una nueva matriz de disimilitud calculando las distancias desde este primer grupo (1-2) al resto de sitios. El cálculo de esta distancia es dependiente del método.

   Recuerde, para los sitios del 3 al 5 tendremos dos distancias, la distancia desde el sitio 1 y del sitio 2 a cada uno de estos sitios. Por tanto utilizaremos estas dos distancias para calcular la distancia desde el grupo.

- En el método de agrupación simple la distancia entre el grupo y el sitio 3 será igual a la distancia más baja comparando entre la distancia del sitio 1 y el sitio 2. En el caso de la distancia al sitio 3 el valor mínimo es 0.287.

- En el método completo el nuevo valor de distancia será el valor más alto, en este caso 0.350, y
- En el método de agrupación promedio, obtenemos el valor promedio entre las distancias primer grupo y el sitio 3 en este caso 0.318 (Tabla 1).

Tabla 1. Cálculo de nuevas distancias entre el grupo 1 (sitio 1 y 2) y los sitios restantes. ***A. simple***: cálculo de distancia mediante el método de agrupación simple. ***A. completa***: cálculo de distancia mediante el método de agrupación completa. ***A. promedio***: cálculo de distancia mediante el método de agrupación promedio.

Sitios	Sitio 1	Sitio 2	A. Simple	A. Completa	A. Media
Sitio 3	0.3501647	0.2873051	0.2873051	0.3501647	0.3187349

Sitios	Sitio 1	Sitio 2	A. Simple	A. Completa	A. Media
Sitio 4	0.3682008	0.3149523	0.3149523	0.3682008	0.3415765
Sitio 5	0.3725079	0.3851064	0.3725079	0.3851064	0.3788071

3. A partir de estos cálculos se construye nuevamente la matriz de distancia. Mostramos las nuevas matrices de distancias según el método de agrupación utilizado.

Para el método de agrupación **simple**

Sitio	Grupo1-2	Sitio 3	Sitio 4
3	0.2873051		
4	0.3149523	0.3244078	
5	0.3725079	0.3595041	0.3721619

Para el método de agrupación **completo**

Sitio	Grupo1-2	Sitio 3	Sitio 4
3	0.3501647		
4	0.3682008	0.3244078	
5	0.3851064	0.3595041	0.3721619

Para el método de agrupación **promedio**

Sitio	Grupo1-2	Sitio 3	Sitio 4
3	0.3187349		
4	0.3415765	0.3244078	
5	0.3788071	0.3595041	0.3721619

4. Se repite el procedimiento, se busca los puntos que tienen la menor disimilitud en la nueva matriz y se vuelve a calcular las distancias desde este nuevo grupo al resto de grupos, esto se repite tantas veces hasta que todos los sitios están asociados.

Podemos calcular directamente la agrupación utilizando la función *hclust*, y graficarlo con la función *plot*.

```
par(mfcol=c(1,3))

csim <- hclust(dist1, method="single")
ccom <- hclust(dist1, method="complete")
cpro <- hclust(dist1, method="average")

plot(csim, cex.axis=0.7)
plot(ccom, cex.axis=0.7)
plot(cpro, cex.axis=0.7)
```

Como podemos ver en la figura 2.1 en todos los casos el primer grupo es el mismo, el grupo entre los sitios 1 y 2 con una disimilitud de 0.27, a partir de este punto los dendrogramas varían según el método utilizado. En el caso del método simple la disimilitud más baja es entre el grupo 1-2 y el sitio 3, con una disimilitud del 0.287. En el caso del método completo la disimilitud más baja se da entre el sitio 3 y 4 que conforman un segundo grupo con una disimilitud de 0.32. Finalmente, en el caso del método de promedio la menor disimilitud se da entre el grupo 1-2 y el sitio 3 con una disimilitud de 0.31 (Figura 2.1)

Los métodos de agrupamiento jerárquico (cluster) producen clasificaciones donde todas las observaciones se

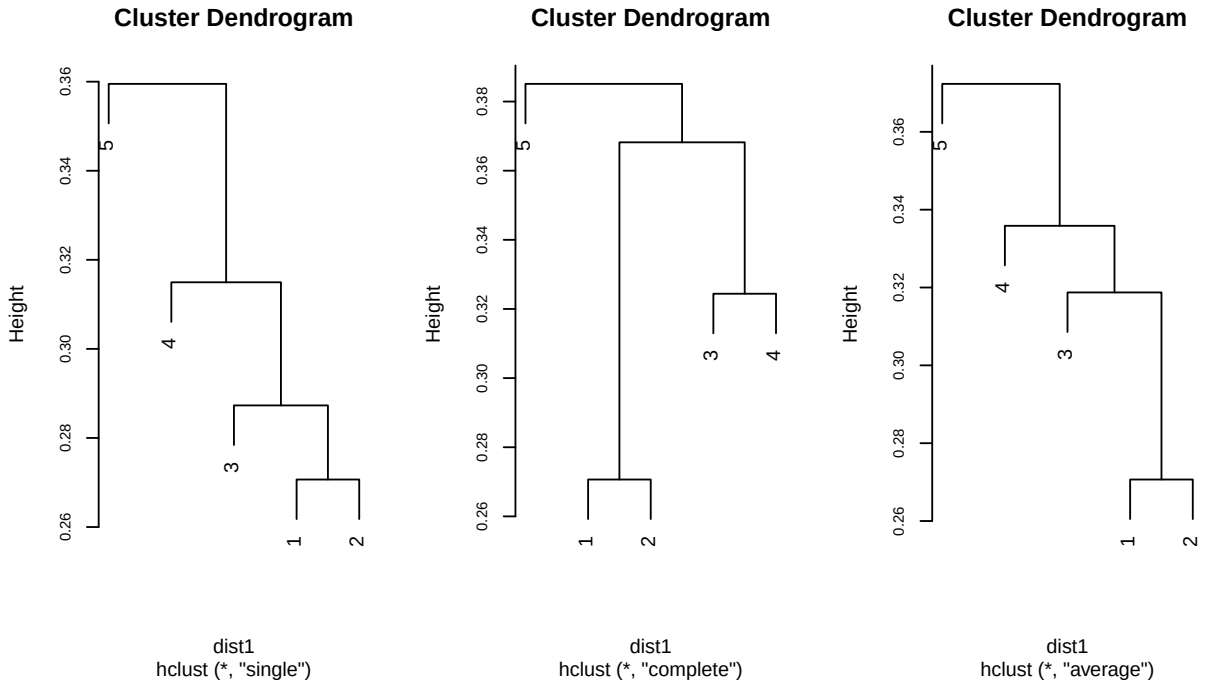


Figure 2.1: Dendrograma construido a partir de los 3 métodos de agrupación

encuentran agrupadas de diferente forma. En los extremos todas las observaciones se encuentran agrupadas en una sola clase o cada observación conforma su clase privada, entre estos extremos las observaciones forman diferentes agrupamientos con niveles de disimilitud variables. Normalmente nos interesa tener un cierto número de clases con niveles de disimilitud establecido. La conformación de estos grupos se puede mostrar visualmente con función **rect.hclust** (Figura 2.2)

```
par(mar=c(2,3,4,2))
plot(ccom, hang=-0.1, cex.axis=0.7, cex.lab=0.8, cex.main=0.8)
rect.hclust(ccom, 3)
```

Ahora podríamos obtener la pertenencia a un grupo y relacionarlo con otra variable explicativa, y analizar si la genración del grupo responde a algún factor.

```
grupo <- cutree(ccom, 3)
grupo
```

```
## 1 2 3 4 5
## 1 1 2 2 3
```

2.2 Interpretando el cluster

El análisis de conglomerados (cluster) no es un test estadístico, y como vimos hay varios factores que pueden afectar la generación de los grupos (Borcard et al., 2011), por lo que debemos ser consientes de lo que obtenemos como resultado. POdemos usar la función `summary()` para ver la información que tenemos luego de haber utilizado el `hclust`, estos datos pueden ser utilizados para interpretar el agrupamiento (Borcard et al., 2011).

Como vimos anteriormente el investigador puede decidir, en función de su experiencia y de los arboles generados, cuantos grupos se generan dentro del árbol y que metodo de agrupamiento utilizar, sin embargo,

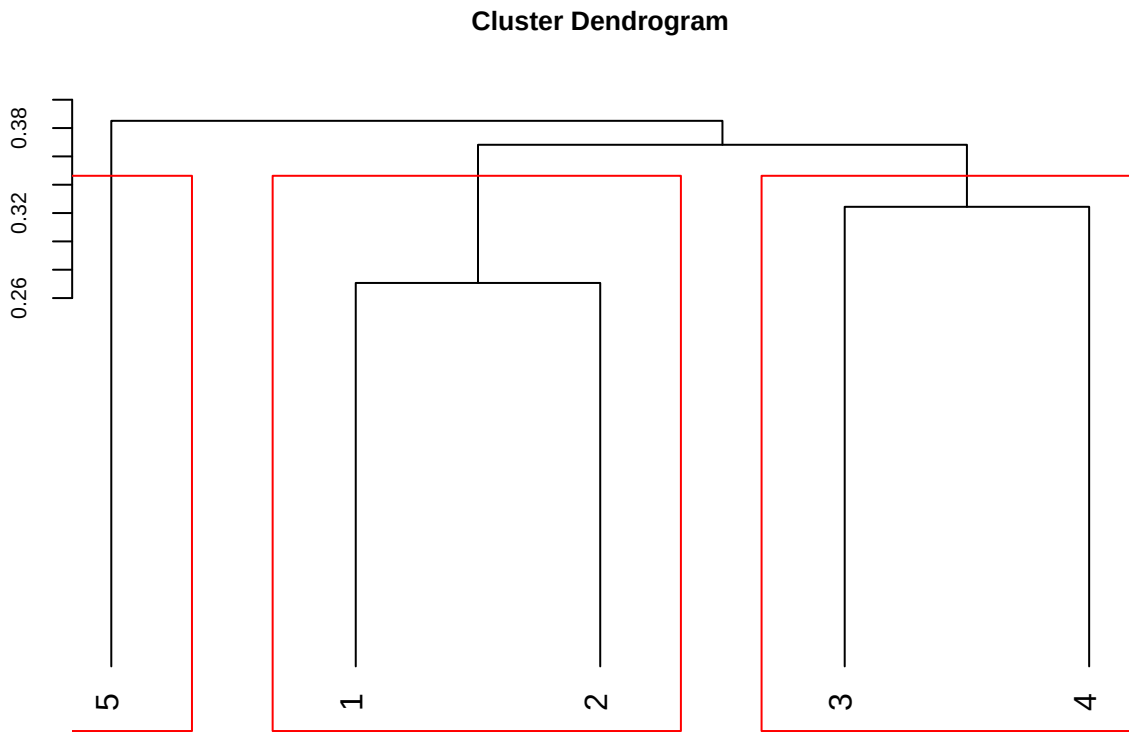


Figure 2.2: Dendrograma con número de grupos

podemos utilizar algunas funciones que nos permitan determinar grupos consistentes.

2.2.1 Elegir la función de enlace

Una forma que podemos utilizar para definir los grupos es la distancia Cofenética. Esta distancia es calculada como la distancia entre dos objetos de un mismo grupo en el dendrograma, la distancia desde el primer objeto al segundo objeto pasando por el nodo de unión de los dos objetos es la distancia Cofenética. Una matriz cofenética es una matriz que representa las distancias cophenéticas entre todos los pares de objetos. Con esta matriz podemos correlacionar con la matriz de disimilitud original. El método con la correlación cofenética más alta puede ser vista como la que produjo el mejor modelo de agrupación para la matriz de distancia.

```
#Calculamos la matriz cofenética para cada método de
#agrupamiento

csim_coph <- cophenetic(csim)
cpro_coph <- cophenetic(cpro)
ccom_coph <- cophenetic(ccom)

#Calculamos la correlación
cor(csim_coph, dist1); cor(cpro_coph, dist1);cor(ccom_coph, dist1)
```

```
## [1] 0.8143114
```

```
## [1] 0.846916
```

```
## [1] 0.7487461
```

Según estos datos el método promedio es el método que produce un mejor agrupamiento.

Otra forma de evaluar el mejor método es calcular la distancia de Gower, calculado como la suma de los cuadrados de la diferencia entre la matriz de distancia y la distancia Cofenética, el menor valor significa que es el mejor método de agrupamiento.

```
sim_gow <- sum((dist1-csim_coph)^2)
pro_gow <- sum((dist1-cpro_coph)^2)
com_gow <- sum((dist1-ccom_coph)^2)

sim_gow; pro_gow; com_gow
```

```
## [1] 0.007860928
```

```
## [1] 0.003917673
```

```
## [1] 0.01068659
```

En este caso vemos que la decisión usando la distancia de Gower y la Cofenética es la misma, el método promedio produce el mejor agrupamiento. Sin embargo, no siempre el resultado es consistente entre los dos métodos.

> Este proceso nos ha permitido obtener la mejor función de enlace, sin embargo, para definir cuales son los subconjuntos de datos (tener un punto de corte) se puede utilizar algunas otras herramientas.

2.2.2 Elegir el punto de corte

Como vimos anteriormente yo puedo definir un punto de corte para generar los grupos o puedo decidir cuantos grupos, sin embargo, este procedimiento es subjetivo. Podemos utilizar alguna información que nos permita tomar decisiones fundamentadas.

Podemos utilizar la **silhouette width (anchura de la silueta)** para medir el grado de pertenencia de un objeto a su agrupación, basado en la distancia media entre este objeto y todos los objetos de la agrupación a la que se pertenece, en comparación con la misma medida calculada para el siguiente grupo más cercano (Borcard et al., 2011). Utilizaremos la función `siluette` del paquete **cluster**. La salida de esta función varía entre 1 y -1. Los valores negativos significan que los objetos correspondientes probablemente se han colocado en un grupo erróneo.

A continuación el proceso utilizado:

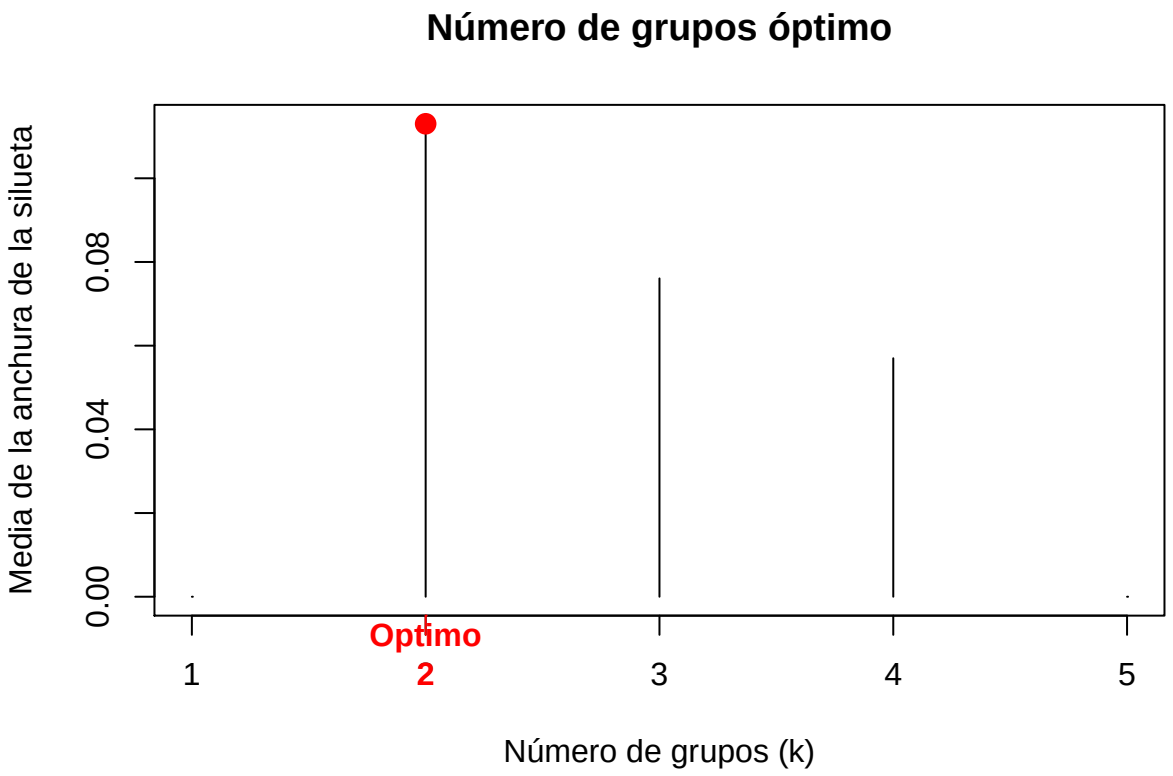
```
library(cluster)

#Generamos un vector vacío para colocar los valores
# medios de la anchura de la silueta (mas)
mas <- numeric(nrow(S_BCI))

#Calculamos y ponemos el <mas> en el vector generado
for( k in 2: (nrow(S_BCI)-1)){
    sil <- silhouette(cutree(ccom, k=k), dist1)
  mas[k] <- summary(sil)$avg.width
}

# Analizamos cual es el mejor punto de corte
k.best <- which.max(mas)

# Graficamos
plot(1:nrow(S_BCI), mas, type = "h", main="Número de grupos óptimo",
     xlab = "Número de grupos (k)", ylab="Media de la anchura de la silueta")
axis(1, k.best, paste("Optimo", k.best, sep="\n"  ),
     col="red", font=2, col.axis="red")
points(k.best, max(mas), pch=16, col="red", cex=1.5)
```




```
cat("", "Número óptimo de grupos k=", k.best, "\n",
    "Con un valor medio de anchura de la silueta de", max(mas), "\n")
```

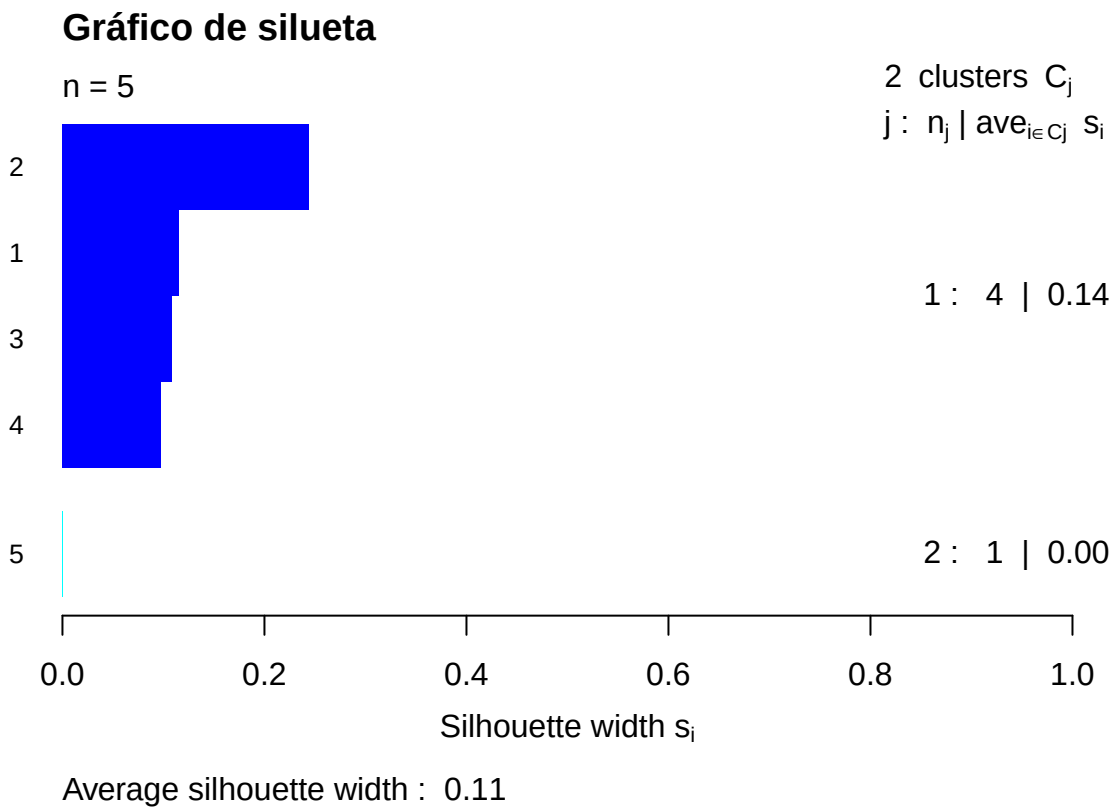
```
##  Número óptimo de grupos k= 2
##  Con un valor medio de anchura de la silueta de 0.1130222
```

A partir de este punto podría utilizar otras herramientas para definir el número de grupos. Ahora nos interesa saber si los grupos están balanceados y bien delimitados. Podemos utilizar el gráfico de la silueta

```
k<- 2
cutg <- cutree(ccom, k=k)
sil <- silhouette(cutg, dist1)
sil.o <- sortSilhouette(sil)

rownames(sil.o) <- row.names(S_BCI)[attr(sil.o, "iOrd")]

plot(sil.o, main= "Gráfico de silueta", cex.names = 0.8,
     col = cutg+3, nmax.lab=100)
```



Al parecer no ha sido el mejor ejemplo, sin embargo, podemos ver que los 2 grupos han sido consistentes. Vamos a probar con nuevos datos.

2.3 Ejercicio 2: Análisis de clasificación

Con el fin de determinar si existen agrupamientos de herbaceas dentro de una parcela permanente de 9ha en la Reserva Ecológica Arenillas realizaremos un análisis de Agrupamiento (Cluster).

Para esto disponemos de una matriz con datos de la composición de la comunidad que puede ser descargado aquí.

Los datos corresponden a un levantamiento de la vegetación de herbáceas en 4 tiempos distintos; final de invierno (abril 2012), estación seca (noviembre 2012), inicio del invierno (diciembre 2012), invierno (enero 2013). Se levantaron 4 cuadrantes de 0.5x0.5 m en cada vértice y centro de la parcela permanente de 9 hectáreas (113 muestras).

Con estos datos:

1. Calcular una matriz de disimilitud utilizando la distancia de Bray-Curtis.
2. Definir la mejor función de enlace para los tres métodos.
3. Definir usando la función silhouette cuantos grupos deberían generarse.
4. Realizar un gráfico del cluster y mostrar los grupos con la función rect.hclust
5. Evaluar si los grupos obtenidos responden a alguna de las variables de especies leñosas
6. Graficar las coordenadas **"x"** y **"y"** de las parcelas y colorear cada punto de acuerdo al grupo al que pertenece. Esto nos permitirá identificar si existe un patrón espacial en la generación de los grupos.

Chapter 3

Ordenaciónes Indirectas

Chapter 4

Ordenaciones Directas o Constreñidas

Bibliography

Borcard, D., Gillet, F., and Legendre, P. (2011). *Numerical Ecology with R.*

Stevens, M. H. H. (2009). *A Primer of Ecology with R.*