COMPOSICIÓN DE COMUNIDADES: Medidas de similitud

Carlos Iván Espinosa

Octubre, 2015

Introducción

La caracterización de una comunidad biológica presenta varios retos a los ecólogos. ¿Donde 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 documento 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 comunidades.

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 puede ser simplemente una lista de especies en esa localidad, donde se registra la presencia o ausencia de cada espécie. Ahora, imaginemos que tenemos cuatro localidades (A:D) donde recogemos los datos de densidad de dos especies; *Tabebuia billbergii* y *Geofroea spinosa*, dos 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 plot para ver cuanto se parece cada sitio (Figura 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")
```

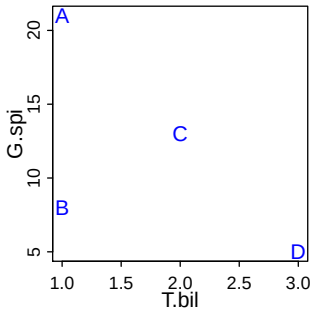

Figure 1: Distancias de cuatro localidades hipotéticas

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

En la Figura 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 otros sitios. Lo siguiente que nos deberiamos pregunta 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. Para ello dividimos la abundancia de cada especie para la suma total de los individuos de las especies en esa muestra.

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

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

Esto implica 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 como *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.

Distancias entre sitios

La *distancia* entre dos muestras esta 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 *Euclideana*. 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 A y D tenemos que la distancia en el eje *x* es la diferencia de la abundancia de *T. bilbergii* entre el sitio A y D.

```
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 A y B.

```
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 largo. Para calcular la distancia *Euclideana* entre pares de sitios con *R* especies utilizamos la siguiente ecuación (Distancia Euclideana):

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

Distancia Euclideana

Existen muchas otras formas de medir distancias entre dos localidades. En ecología posiblemente una de las distancias más utilizada es la distancia de *Bray-Curtis*, la cual muchas veces es conocida como *Sorensen*. Esta distancia es calculada como (Distancia de Bray-Curtis):

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

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 euclidea 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 5 se compara gráficos basados en distancias euclidianas y Bray-Curtis de los mismos datos en bruto.

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, pero las distancias pueden ser aproximadamente correctas. 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 2a) y las distancias de Bray-Curtis (Figura 2b).

```
library(vegan)

# Distancia Euclideana
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(2, 1), 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)Euclideana",
    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)
```

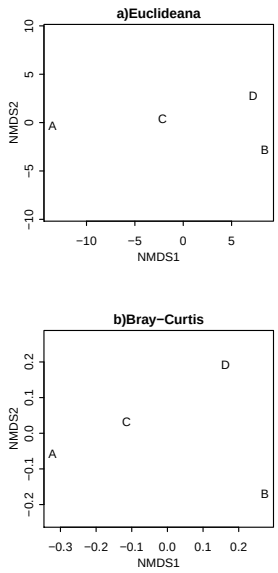

Figure 2: Arreglo de las parcelas en distancias multidimensionales no métricas (NMDS). Estas dos figuras muestran los mismos datos en bruto, pero las distancias euclidianas tienden a enfatizar las diferencias debidas a las especies más abundantes, mientras que Bray-Curtis no lo hace.

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* y *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
```

Lo siguiente para comparar entre sitios, lo que hacemos 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.

```
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 que casos se comparten especies. Finalmente, sumamos todas las especies que su abundancia es mayor a cero.

```
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 si, el porcentaje de similitud utiliza no solamente la presencia ausencia sino también la abundancia lo que podría estar reduciendo la similitud entre sitios.

Análisis multivariado de la composición de la comunidad

Los índices de similitud nos permite 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].

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

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

Análisis de Clasificación. 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.

```
library(vegan)
```

```
data(BCI)

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

```
##  [1] 0.2706682 0.3501647 0.3682008 0.3725079
##  [5] 0.3744186 0.3518519 0.3424346 0.4235706
##  [9] 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 3 métodos de agrupamiento: agrupación simple, agrupación completa y agrupación promedio.

Todos los métodos inician comparando dos comunidades y a partir de esta primera comparación se continúa con el resto de puntos.

A continuación ejemplificaremos el cálculo de las distancias usando los 3 métodos. Para esto extraemos los 5 primeros sitios de la matriz y generamos un nuevo objeto (S_BCI). Con este nuevo objeto calculamos la distancia entre los 5 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. En nuestro caso el punto 1 y 2 tienen la distancia más baja 0.27. Una vez identificado inicia el proceso de generación de la agrupación y es donde se diferencian los métodos de agrupación.

2. Construimos una nueva matriz de disimilitud calculando las distancias desde este primer grupo (1-2) al resto de sitios. El cálculo de la nueva distancia depende del método utilizado, en el método de agrupación simple, definimos cual es la distancia mínima desde los sitios del primer grupo (1-2) a cada uno de los sitios, así en el caso de la distancia al sitio 3 el valor mínimo es 0.287. En el caso de utilizar el método completo el nuevo valor de distancia será

el valor más alto, en este caso 0.350 y si utilizamos el método de agrupación promedio entre las distancias entre sitios del 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
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(3, 1))

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)
```

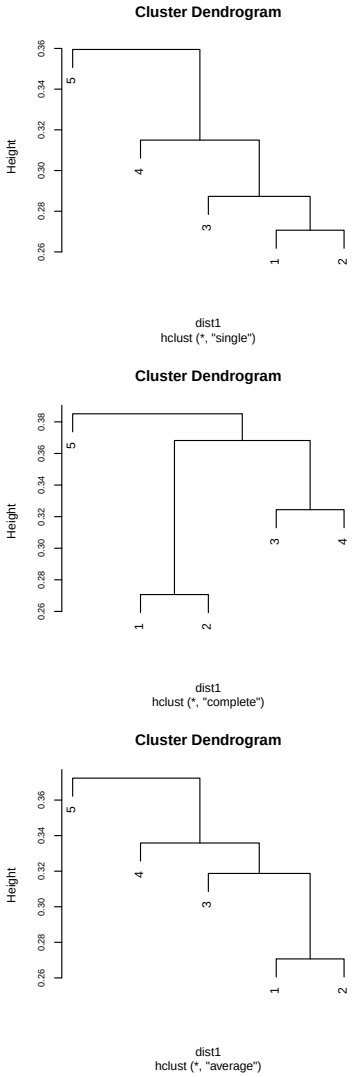

Figure 3: Dendrograma construido a partir de los 3 métodos de agrupación

Como podemos ver en todos los casos el primer grupo en todos los dendrogramas 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 (Figura 6). 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 6)

Los métodos de agrupamiento jerárquico (cluster) producen clasificaciones donde todas las observaciones se encuentran agrupados de diferente forma. En los extremos todas las observaciones se encuentran agrupados 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**

```
par(mar = c(2, 3, 4, 2))
plot(ccom, hang = -0.5, cex.axis = 0.7, cex.lab = 0.8,
    cex.main = 0.8)
rect.hclust(ccom, 3)
```

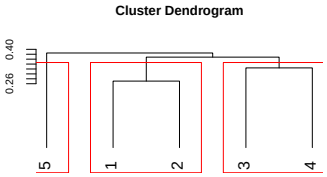

Figure 4: Dendrograma con número de grupos

Ejercicio 2. *Análisis de Clasificación de la comunidad*

Uno de los temas más interesantes en la ecología de comunidades es definir los límites de una comunidad en el tiempo y el espacio.

La definición de los límites de la comunidad presenta algunos retos: ¿Cómo definimos el límite? El límite puede ser bastante difuso puesto que la comunidad es la suma de poblaciones y estas a su vez tienen respuestas variables al ambiente, por tanto las ocurrencias no estan sincronizadas (no todas las poblaciones ocurren exactamente en los mismos puntos). El análisis de clasificación si bien no permite tener información de cuál es el límite exacto de la comunidad, si permite definir cuan similares son los puntos en los cuales se hace el muestreo.

El presente trabajo realizaremos un análisis cluster de una comunidad de matorral seco tropical en el cantón Catamayo. Los datos que se proporcionan corresponden a la cobertura promedio de cada especie.

Se establecieron 4 niveles alitudinales en dos localidades distintas Alamala y las Chinchas en cada altitud se establecieron 2 parcelas separadas entre 5 y 50 metros de altitud. Nos interesa entender como la comunidad cambia en el gradiente altitudinal y si esta tendencia se mantiene entre las localidades.

A continuación proponemos algunas preguntas que permiten canalizar los análisis para responder el objetivo planteado.

1. ¿Cuál es la similitud entre las dos localidades estudiadas? (Utilizar el índice de similitud y el índice de Sorensen)
2. La similitud entre los niveles altitudinales es igual al que mantienen las localidades o es dependiente de la altitud (Utilizar el índice de similitud y el índice de Sorensen). Para este análisis utilizaremos solo las altitudes comparables por lo que las parcelas de las Chinchas a 2070 y 2090, y de Alamala de 1670 y 1680 serán excluidas de la comparación.
3. Utilice los diferentes métodos de agregación con el fin de establecer agrupamiento entre las diferentes parcelas. Es importante tener presente que nos interesa saber si las diferencias en la comunidad se mantienen a lo largo del gradiente y estas son consistentes entre comunidades.
4. ¿Cuál es el efecto de las especies dominantes sobre los resultados obtenidos?. Podemos utilizar el método de Bray Curtis pero solo con datos de presencia ausencia (incluimos en el cálculo de distancia el parámetro [binary=TRUE])

Nota: En el caso de este estudio las parcelas tienen correlación espacial por lo que se debería usar un método cluster constreñido, el cual permite mantener esta correlación espacial. Sin embargo, para este ejercicio utilizaremos el método jerárquico