# Similitud de Comunidades biológicas

*Carlos Iván Espinosa*

*Noviembre 2018*

# 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 espaciales y temporales. Estos límites están dados por la distribución de las poblaciones a lo largo de un gradiente espacial o temporal. De esta forma los cambios en abundancia de las especies a lo largo de gradientes espaciales o temporales generan la zonación y la sucesión respectivamente.

La identificación de formaciones biológicas en el espacio (**zonación**) o las etapas seriales a lo largo del tiempo (**sucesión**) implica que tenemos la capacidad de establecer en que momento una comunidad cambia. Parece una tarea sencilla, pero realmente no lo es, ¿cuanto deberia cambiar una comunidad para poder hablar de etapas seriales o zonas distintas? y ¿cómo podemos calcular ese cambio? Una de las formas de responder estas preguntas puede ser intentar cuantificar las similitudes entre localidades.

# Objetivos

---

En este ejercicio mostramos las bases del cálculo de similitud y distancia entre comunidades, el cual se convierte en la base de los análisis multivariantes de la comunidad.

Específicamente nos interesa;

- Comprender las bases teóricas para el cálculo de similitudes y distancias en la comunidad entre localidades.
- Desarrollar mediciones de similitud entre localidades e interpretar los resultados.

Figure 1: *Stenocercus iridicens*

# Chapter 1

# Introducción

La composición y estructura de la comunidad varía a lo largo de un gradiente ambiental o a lo largo del tiempo. De esta manera, si un ecólogo realiza un muestreo de ese gradiente, tendrá cambios en la composición de la comunidad (las especies que la constituyen) y en la estructura (las abundancias de las especies) (figura 1.1).

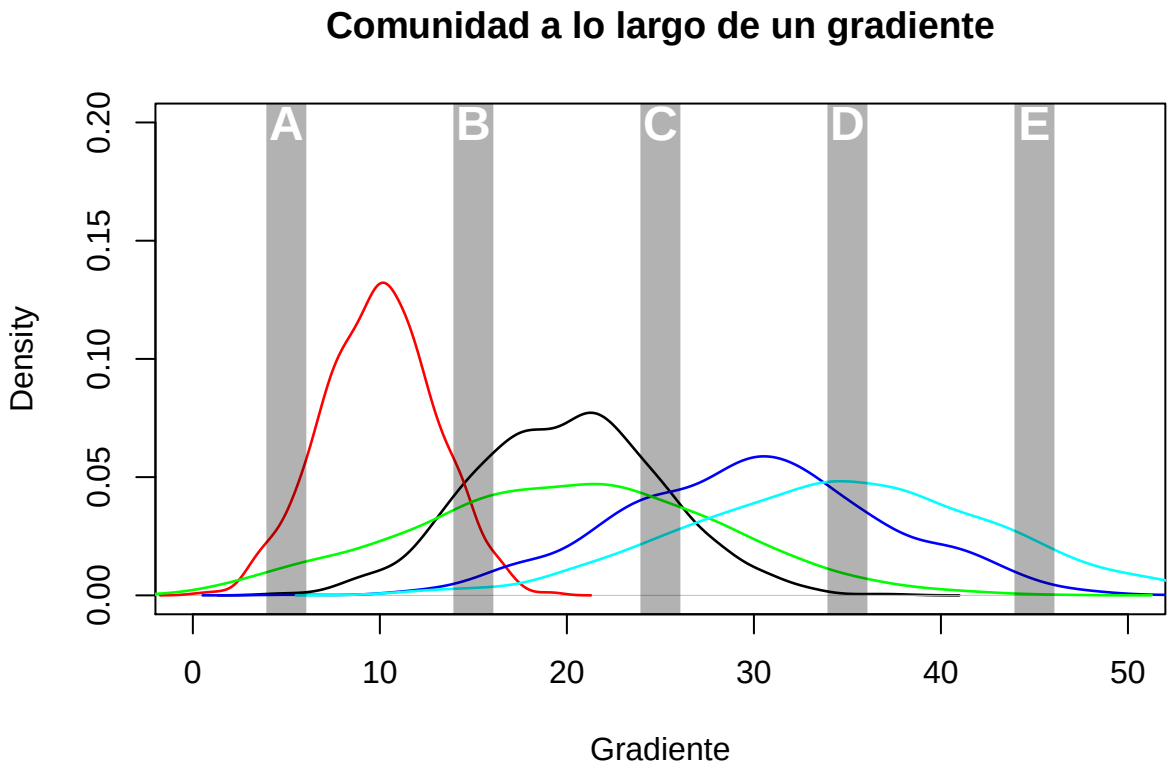


Figure 1.1: Ejemplo de la variación de una comunidad

Como podemos ver en la figura 1.1, en cada uno de los transectos (representados por la línea gris), la ocurrencia de las especies y su abundancia cambia. Por ejemplo, en la comunidad "B" tenemos cinco especies, la especie representada por la línea verde alcanza su máxima abundancia, mientras que la representada por

la línea azul se encuentra con una abundancia muy baja. En la comunidad “C” tenemos cuatro especies pero las abundancias son diferentes a las encontradas en la comunidad “B”. En el caso de la comunidad “C”, no solo las abundancias varían sino también la ocurrencia de especies.

Los cambios en la estructura y composición de la comunidad hacen que ciertos transectos sean más parecidos y que otros sean más distintos, en el caso del ejemplo es clara o más o menos claras las similitudes en cuanto a estructura y composición de las especies, sin embargo, en la realidad es más complejo determinar a simple vista estas similitudes. A continuación veremos que métodos se usan para cuantificar estas similitudes.

# Chapter 2

# Similitud, disimilitud y distancia

Pensemos que dos elementos se parecen más, cuando sus propiedades son más parecidas, en este dos comunidades se parecerán más si su composición y estructura es parecida. La *similitud* nos permite entonces tener un valor que define en qué medida dos comunidades se parecen. Aunque esta información es interesante, cuando se analizan muchas comunidades, el apreciar estas diferencias en cada par sería complejo, por lo que interesa poder representar estas comunidades en un plano. La graficación de las comunidades en un plano es posible si disponemos de medidas de **distancias** entre las comunidades. Las distancias pueden ser medidas a través de distancias simétricas (ejemplo Euclideana, Hellinger), o a través de medidas asimétricas (medidas de **disimilitud**), la otra cara de la similitud.

# Chapter 3

# Índices de Similitud

¿Cuán similares son dos localidades?, vamos a calcular dos tipos de similitudes una basada en incidencia (presencia-ausencia de especies) (ej. Índices de *Sorensen*, *Jaccard* y *Simpson*), y otra basada en la abundancia *Porcentaje de Similitud*. Imaginemos que tenemos cuatro localidades (A, B, C, D) donde recogemos los datos de densidad de cuatro especies; *Tabebuia billbergii*, *Geofroea spinosa*,*Ceiba trichistandra* y *Colicodendron scabridum*, 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),
                   C.tri = c(11, 3, 7, 5), C.sca = c(16, 0, 9, 4))
row.names(dens) <- LETTERS[1:4]
dens
```

```
##   T.bil G.spi C.tri C.sca
## A     1    21    11    16
## B     1     8     3     0
## C     2    13     7     9
## D     3     5     5     4
```

Generamos un gráfico para ver cuánto se parece cada sitio (Figura 3.1) basado en las dos primeras especies.

```
par(mar=c(4,4,1,1), mgp=c(1,0.3,0), tcl= -0.2)
plot(dens[,1:2], type = "n", cex.axis=0.8, xlim=c(0,20), ylim=c(0,25))
text(dens[,1:2], row.names(dens), col ="blue")
```

En la figura 3.1 vemos que la composición de especies en el sitio A es diferente de la composición del sitio D. Es decir, la similitud entre el sitio A y D es menor que entre los otros sitios. Lo siguiente que nos deberíamos preguntar es; ¿qué tan similares son los dos sitios?

## 3.1 Índices cualitativos

Los índices cualitativos son una medida que nos permite evaluar la similitud entre comunidades basados en presencia-ausencia de especies. A continuación veremos tres diferentes índices:

$$S_s = \frac{(2c)}{(a+b+2c)}$$

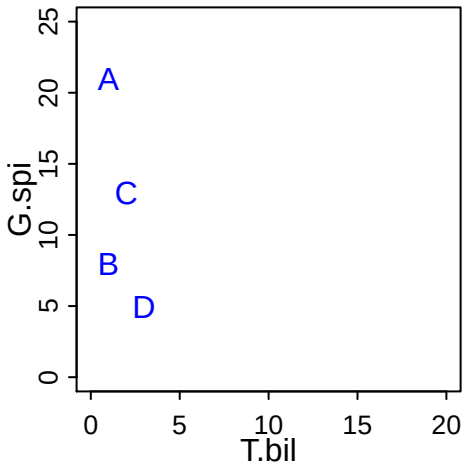


Figure 3.1: Similitud de cuatro localidades hipotéticas

Índice de Sorensen

$$S_s = \frac{(c)}{(a+b+c)}$$

Índice de Jaccard

$$S_s = \frac{(c)}{c + min(a+b)}$$

Índice de Simpson

Donde $c$ es el número de especies en común entre los dos sitios, $a$ y $b$ son el número de especies únicas en cada sitio. Las diferencias entre estos índices radica en la importancia que se le da a cada componente, en el caso del índice de Sorensen las especies compartidas tienen una gran importancia, por eso es multiplicada por dos. En el caso del índice de Simpson, es un índice usado cuando hay diferencias muy altas entre pares de comunidades, así restamos el peso obteniendo el valor mínimo de entre a y b.

Para calcular estos índices entre los sitios A y B necesitamos definir el número de especies compartidas y luego el número de especies únicas de los dos sitios.

```
comp <- dens
comp[comp>0] <- 1 #Generamos una matriz de presencia ausencia
comp
```

```
##   T.bil G.spi C.tri C.sca
## A     1     1     1     1
## B     1     1     1     0
## C     1     1     1     1
## D     1     1     1     1
```

```
a <- sum(colSums(comp[1:2,])==1&comp[2,]==0)#Ocurren en A pero no en B
b <- sum(colSums(comp[1:2,])==1&comp[1,]==FALSE)#Ocurren en B pero no en A
c <- sum(colSums(comp[1:2,])==2) #ocurren en A y B

a;b;c
```

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

```
## [1] 0
```

```
## [1] 3
```

Ahora obtenemos el valor de similitud entre los dos primeros sitios (A y B).

```
Sor <- (2*c)/(a+b+(2*c))
Jac <- c/(a+b+c)
Sim <- c/c+min(a,b)

Sor; Jac; Sim
```

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

```
## [1] 0.75
```

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

Según el índice de Sorensen estos dos sitios son parecidos en un 86%, mientras que para el índice de Jaccard es el 75% y para Simpson estos dos sitios son iguales (100%).

## 3.2 Índices cuantitativos

El *porcentaje de similitud* es la versión cuantitativa del índice de Sorensen este índice está basado en datos de abundancia y es calculado como:

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

Porcentaje de Similitud

Donde; $W$ es la sumatoria del valor mínimo de la abundancia entre las comunidades comparadas para cada especie. $A$ y $B$ es la suma de las abundancias de todas las especies en cada sitio.

```
library(knitr)

MatPS <- rbind(dens[1:2,], apply(dens[1:2, ], 2, min))#Obtenemos el valor mínimo de cada especie
MatPS <- data.frame(MatPS, Medidas=rowSums(MatPS), Tipo= c("A", "B", "W")) #Obtenemos W,A,B

kable(MatPS, caption = "Medidas para obtener el porcentaje de Similitud")
```

Table 3.1: Medidas para obtener el porcentaje de Similitud

| | T.bil | G.spi | C.tri | C.sca | Medidas | Tipo |
|---|---|---|---|---|---|---|
| A | 1 | 21 | 11 | 16 | 49 | A |
| B | 1 | 8 | 3 | 0 | 12 | B |
| 3 | 1 | 8 | 3 | 0 | 12 | W |

```
PS <- (2*MatPS[3,5])/(MatPS[1,5]+MatPS[2,5])
PS
```

```
## [1] 0.3934426
```

Esto significa que la comunidad A y B tienen un porcentaje de similitud del 39%. Los datos de los dos tipos de 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.

# Chapter 4

# Distancias entre sitios

Cuando tenemos dos comunidades muy parecidas entre sí tendremos valores altos de similitud, en contraposición los índices de distancia nos mostrarán valores altos cuando dos comunidades se parecen poco. Como habíamos mencionado anteriormente existes dos tipos de medidas de distancia;

- aquellas calculadas a partir de los índices de similitud usualmente como D= 1-Similitud. Así, para los índices de incidencia (presencia - ausencia) se pueden usar los índices de Jacard, Simpson o Sorensen, mientras que para los índices cuantitativos se puede usar el porcentaje de similitud, este último conocido como distancia de Bray Curtis.
- aquellas que no tienen medidas de similitud análogas, algunos de estos índices son; Euclidiana, Chord, Hellinger.

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 (figura 3.1).

## 4.1 Distancia Euclidiana

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

Table 4.1: Efecto del doble cero

| spp1 | spp2 | spp3 | sp4 | spp5 | spp6 |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 | 1 |

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

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

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

```
dist(dens[,1:2])
```

```
##           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

### 4.1.1 Efecto de doble-ceros y abundancia

Aunque la distancia Euclidiana es fácilmente interpretable, es poco usado en análisis biológicos. Normalmente los datos de comunidad están caracterizados por una gran cantidad de ceros (especies no encontradas en determinados sitios), el cálculo de la distancia Euclidiana incrementa la similitud entre comunidades que presentan ceros en la misma especie.

Según los datos mostrados en la tabla tendríamos que hay un gradiente, la primera comunidad comparte una especie con la comunidad dos y la comunidad dos comparte una especie con la comunidad tres. Los índices deberían permitir recuperar ese gradiente, veamos lo que pasa.

```
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-2
```

Table 4.2: Efecto de la abundancia

| spp1 | spp2 | spp3 |
|---|---|---|
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 0 | 8 | 7 |

```
vegdist(dcMat, "euclidean")
```

```
##   1 2
## 2 2
## 3 2 2
```

Como vemos, en el caso del ejemplo el doble cero de la comunidad uno y tres generan una mayor similitud, por lo que las tres comunidades son mostradas a igual distancia. Esto no debería ser un problema, si el cero realmente nos da información, pero en el caso de datos biológicos la razón de tener ese cero puede deberse a varias razones, por lo que realmente el cero no es informativo. En otros casos, normalmente en datos abióticos, el cero implica la ausencia de algo, por ejemplo tener cero mg de un contaminante es una información. De esta forma la distancia Euclidiana es usada sobre todo para interpretar datos ambientales.

Otra característica importante de la distancia euclidiana es que está fuertemente impactada por la diferencia de abundancias, recordemos que la diferencia de abundancias esta elevada al cuadrado, de esta forma la distancia entre dos comunidades puede estar marcada por la diferencia en abundancias más que por la diferencia en presencia de especies.

```
dcMat2 <- data.frame(spp1=c(0,1,0),spp2=c(1,0,8),
                     spp3=c(1,0,7))

kable(dcMat2, caption = "Efecto de la abundancia")
```

```
vegdist(dcMat2, "euclidean")
```

```
##           1         2
## 2  1.732051
## 3  9.219544 10.677078
```

Como vemos la comunidad uno se encuentra más cercana a la comunidad dos que a la tres. La distancia de la comunidad uno a la tres es de 9.21, aunque comparten dos especies la diferencia en abundancias es muy marcada. Por otro lado, la comunidad uno tiene una distancia de 1.73 a la comunidad dos, esta menor distancia se da aunque no comparten ninguna especie.

## 4.2 Distancia Bray-Curtis

Existen otras formas de medir distancias entre dos localidades. En ecología una de las distancias más utilizada es la distancia de *Bray-Curtis*, como mencionamos anteriormente esta distancia es el opuesto del porcentaje de similitud, que a su vez es la versión de abundancia del índice de 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* se refiere a 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 euclidiana 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 4.1 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).

La función de escalamiento multidimensional no-métrico está en el paquete `vegan`. Aquí mostramos las distancias euclidianas entre sitios (Figura 4.1a) y las distancias de Bray-Curtis (Figura 4.1b).

```
library(vegan)

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

Como podemos apreciar en el caso del ejemplo, la distancia de Bray-Curtis recupera la idea de un gradiente entre las comunidades, desde la comunidad uno a la tres. En el caso de la distancia Euclidiana la comunidad dos y tres se encuentran a igual distancia de la comunidad uno, como un efecto del doble cero.

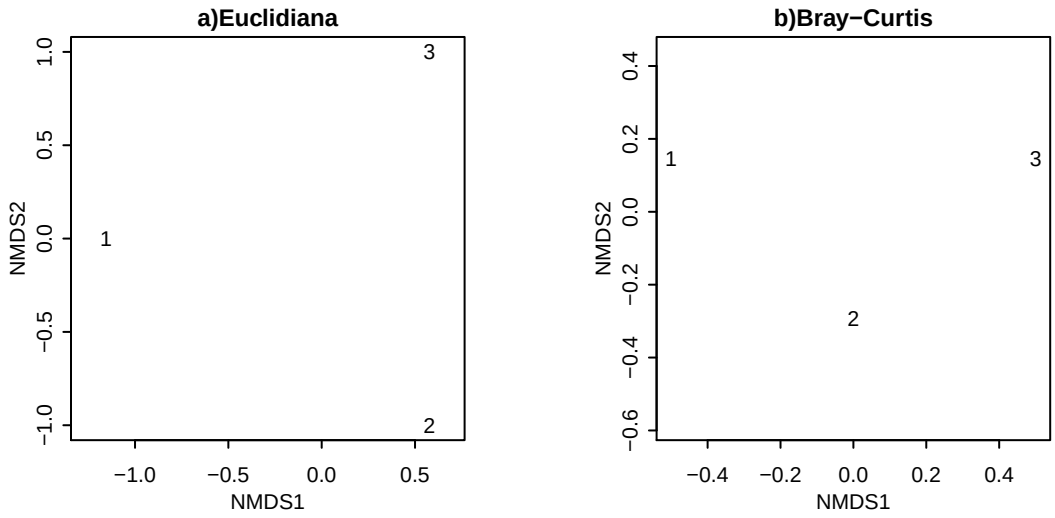


Figure 4.1: Arreglo de las parcelas en distancias multidimensionales no métricas (NMDS). Estas dos figuras muestran los mismos datos en bruto, pero las distancias euclidianas tienden a enfatizar las diferencias debidas a las especies más abundantes, mientras que Bray-Curtis no lo hace.

# Chapter 5

# Transformación y Estandarización de datos

Cuando trabajamos con datos multivariantes cabe la posibilidad de que los datos dentro de esta matriz tengan diferencias de magnitud importantes. Como vimos antes el cálculo de distancia entre los sitios puede verse fuertemente afectado por la magnitud de sus diferencias.

En el ejemplo que mostramos en el inicio, las similitudes entre comunidades basadas en las dos primeras especies, las diferencias entre las comunidades depende de la escala de medición (los valores de los ejes), y sobre cómo medimos la distancia a través del espacio multivariado (Stevens, 2009).

De esta forma, las 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 (Figura 5.1a). Veamos ahora que sucede con las similitud si incremento la abundancia de *T. billbergii* (Figura 5.1b).

```
par(mar=c(4,4,1,1), mgp=c(1,0.3,0), mfcol=c(1,2), tcl= -0.2)
dens1 <- dens
dens1$T.bil <- dens1$T.bil*100
plot(dens[,1:2], type = "n", cex.axis=0.8, xlim=c(0,30), ylim=c(0,30), main="a.")
text(dens[,1:2], row.names(dens), col ="blue")

plot(dens1[,1:2], type = "n", cex.axis=0.8, ylim=c(0,300), main="b.")
text(dens1[,1:2], row.names(dens1), col ="blue")
```

Como vemos en la figura 5.1 las distancias entre cada uno de los sitios cambio cuando incremento la abundancia de *T. bilbergi*, aunque este incremento fue proporcional. 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). Cuando nos referimos a densidad relativa hablamos de la densidad de una especie con referencia a algo, en este caso con relación a la abundancia de individuos de la misma especie en otros sitios.

Para calcular la densidad relativa dividimos la abundancia de cada especie para la suma total de los individuos de las especies en esa muestra.

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

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

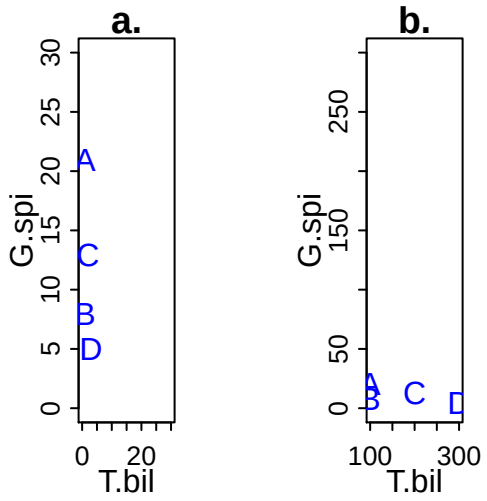


Figure 5.1: Distancias de cuatro localidades hipotéticas

```
dens1[,1]/sum(dens1[,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. Interesantemente, no hay diferencias en las proporciones entre las dos medidas que tenemos. ¿Qué pasó con las distancias?

En la figura 5.2 podemos apreciar que no hay diferencias entre las dos densidades cuando estoy usando la densidad relativa. Pero ¿Qué implicaciones biológicas tiene el usar las densidades relativas para calcular la distancia entre sitios?

Cuando usamos las densidades relativas estamos dando el mismo peso a todas las especies. En un ecosistema con una especie dominante y varias subordinadas, al usar la densidad relativa estoy eliminando esa dominancia. Es importante tener claro este punto, ya que las interpretaciones que puedo hacer con los datos de densidad y densidad relativa son distintos.

## 5.1 Transformación de datos

Como vemos la magnitud de las diferencias entre las variables tiene un impacto sobre el cálculo de la distancia, por lo que nos interesa poder manejar el efecto de esas diferencias, para lo cual desarrollamos una transformación.

La transformación de datos implica una modificación de los datos brutos a través de una ecuación algebraica. La transformación de datos implica una modificación independientemente para cada dato, no existe influencia del resto de datos.

En la tabla anterior podemos ver que la comunidad está compuesta por un par de especies dominantes y varias especies raras.

Al transformar los datos evitamos que las especies dominantes definan el cálculo de la distancia.

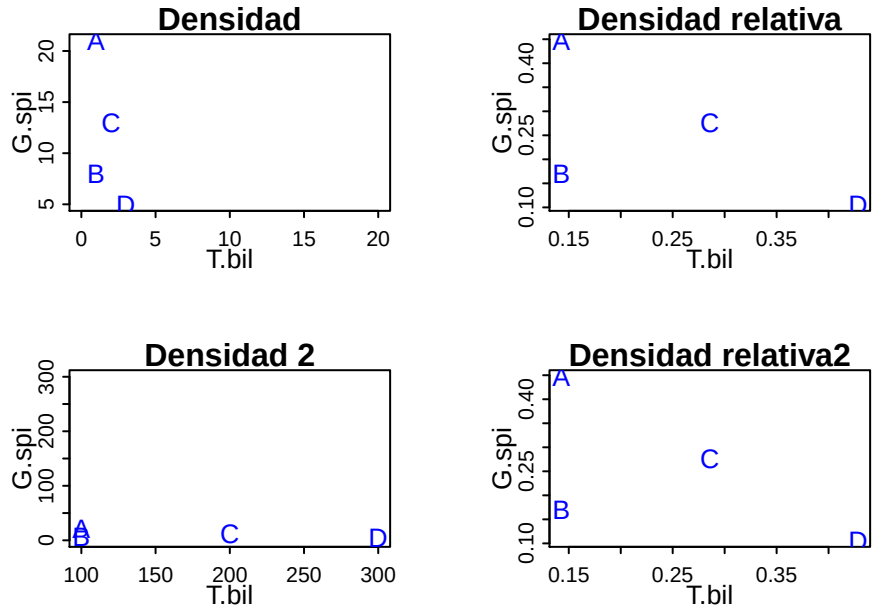


Figure 5.2: Distancias de cuatro localidades hipotéticas

Table 5.1: Comunidad de macroinvertebrados acuáticos

| LOCALIDAD | Allu | Atop | Atri | Baet | Bezz | Blep | Cera | Chel | Chim | Chir | Cole |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bo-1 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 1 | 3 | 18 | 4 |
| Bo-2 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 1 | 9 | 0 |
| Bo-3 | 0 | 0 | 0 | 6 | 0 | 0 | 1 | 1 | 1 | 9 | 0 |
| BP-1 | 0 | 3 | 0 | 81 | 0 | 0 | 0 | 0 | 0 | 27 | 0 |
| BP-2 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| BP-3 | 0 | 0 | 0 | 54 | 0 | 0 | 1 | 0 | 0 | 9 | 0 |
| Pa-1 | 1 | 0 | 0 | 984 | 0 | 0 | 0 | 0 | 0 | 81 | 0 |
| Pa-2 | 0 | 0 | 0 | 15 | 0 | 0 | 0 | 0 | 1 | 9 | 0 |
| Pa-3 | 0 | 0 | 0 | 93 | 1 | 0 | 0 | 0 | 0 | 18 | 0 |
| Ur-1 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 855 | 0 |
| Ur-2 | 0 | 0 | 1 | 12 | 0 | 0 | 0 | 1 | 0 | 9 | 0 |
| Ur-3 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 27 | 0 |

Table 5.2: Efecto de la transformación. Pequeñas diferencias

| sp1. | sp2. | sp1.sqrt | sp2.sqrt | sp1.log | sp2.log |
|---|---|---|---|---|---|
| 53 | 213 | 7.28 | 14.59 | 3.97 | 5.36 |
| 1 | 142 | 1.00 | 11.92 | 0.00 | 4.96 |
| 26 | 114 | 5.10 | 10.68 | 3.26 | 4.74 |
| 25 | 241 | 5.00 | 15.52 | 3.22 | 5.48 |
| 70 | 161 | 8.37 | 12.69 | 4.25 | 5.08 |
| 23 | 166 | 4.80 | 12.88 | 3.14 | 5.11 |
| 61 | 240 | 7.81 | 15.49 | 4.11 | 5.48 |
| 76 | 184 | 8.72 | 13.56 | 4.33 | 5.21 |
| 78 | 237 | 8.83 | 15.39 | 4.36 | 5.47 |
| 6 | 208 | 2.45 | 14.42 | 1.79 | 5.34 |

Existen varias posibilidades para transformar los datos, por lo que definir que función utilizar es importante. Cada tipo de transformación produce resultados distintos por lo que debemos utilizarlas con precaución.

La transformación más sencilla o menos intensa es la raíz cuadrada, mientras que el logaritmo es la transformación más intensa, podríamos utilizar la raíz cuarta como una función intermedia. La raíz cuadrada la utilizaríamos cuando tenemos diferencias con variaciones de una magnitud de diferencia (entre decenas y centenas), mientras que la transformación logarítmica la haríamos con comunidades donde hay más de una magnitud de diferencia (entre decenas y miles).

Aunque hay muchos autores que aconsejan realizar transformaciones hay que ser conscientes de lo que estamos haciendo, transformaciones muy fuertes en una matriz con pocas diferencias pueden hacer que, por ejemplo, las especies raras tengan igual peso que las dominantes, ¿esto es lo que queremos?

**Recuerde: las diferentes transformaciones tienen interpretaciones biológicas distintas. Debemos ser conscientes de lo que estamos haciendo y de su posterior interpretación biológica.**

Veamos un ejemplo:

```
set.seed(4)
aves<- data.frame(sp1= sample(1:90, 10), sp2= sample(100:250, 10))

insectos<- data.frame(sp1= sample(5:99, 10), sp2= sample(1000:2500, 10))

##¿Qué pasa cuando transformamos?
aveT <- round(cbind(aves, sqrt(aves),log(aves)),2)
colnames(aveT) <- paste(rep(c("sp1", "sp2"), 3), c("","","sqrt", "sqrt", "log", "log"), sep=".")

kable(aveT, caption = "Efecto de la transformación. Pequeñas diferencias")
```

```
insT <- round(cbind(insectos, sqrt(insectos),log(insectos)),2)
colnames(insT) <- paste(rep(c("sp1", "sp2"), 3), c("","","sqrt", "sqrt", "log", "log"), sep=".")

kable(insT, caption = "Efecto de la transformación, Grandes diferencias")
```

## 5.2 Estandarización de los datos

La estandarización de los datos permite modificar las variables transformándolas en unidades de desviación típica, lo que nos permite comparar entre valores de distribuciones normales diferentes, o de valores diferentes.

Table 5.3: Efecto de la transformación, Grandes diferencias

| sp1. | sp2. | sp1.sqrt | sp2.sqrt | sp1.log | sp2.log |
|---|---|---|---|---|---|
| 72 | 1851 | 8.49 | 43.02 | 4.28 | 7.52 |
| 98 | 1358 | 9.90 | 36.85 | 4.58 | 7.21 |
| 52 | 2316 | 7.21 | 48.12 | 3.95 | 7.75 |
| 50 | 1980 | 7.07 | 44.50 | 3.91 | 7.59 |
| 64 | 1722 | 8.00 | 41.50 | 4.16 | 7.45 |
| 79 | 2452 | 8.89 | 49.52 | 4.37 | 7.80 |
| 47 | 1687 | 6.86 | 41.07 | 3.85 | 7.43 |
| 94 | 1929 | 9.70 | 43.92 | 4.54 | 7.56 |
| 49 | 1579 | 7.00 | 39.74 | 3.89 | 7.36 |
| 96 | 1009 | 9.80 | 31.76 | 4.56 | 6.92 |

La estandarización o tipificación se lo realiza restando a cada valor el valor medio de la variable y dividiendo para la desviación estándar.

```
avesE <- (aves[,1]-mean(aves[,1]))/sd(aves[,1])
avesE
```

```
##  [1]  0.3819546 -1.4073822 -0.5471241 -0.5815345  0.9669301 -0.6503551
##  [7]  0.6572372  1.1733920  1.2422126 -1.2353306
```

```
round(mean(avesE),1);sd(avesE)
```

```
## [1] 0
```

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

Como vemos las variables estandarizadas tienen como propiedad que la desviación estándar es 1 y la media es 0.

# Chapter 6

# Ejercicio práctico

---

Una de las preguntas básicas de un ecólogo es saber ¿Cómo de diferentes son dos comunidades?. Como hemos visto existen varias decisiones que los investigadores debemos tomar, estas decisiones afectan directamente a los resultados que podemos obtener y por ende a las conclusiones biológicas que obtenemos de este análisis.

El presente ejercicio evaluaremos como las diferentes decisiones que tomamos entorno al procesamiento de datos afectan nuestras medidas de similitud, y cuáles son las conclusiones biológicas que obtenemos con uno u otro procedimiento. En la tabla 6.1 mostramos cinco comunidades hipotéticas.

Con los datos anteriores:

a. Convierta los datos en abundancia relativa por especie (la suma en cada especie 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?¿Qué implicaciones biológicas podría tener si utilizamos la primera o la segunda matriz para calcular las similitudes?

b. Calcule la distancia Euclidiana y de Bray Curtis para cada sitio con las dos medidas de abundancia y grafíquelas utilizando el NMDS. ¿Cómo cambia entre distancias y abundancias? ¿Por qué se dan estas diferencias? ¿Puede darle una explicación biológica a los diferentes resultados?

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? ¿De una interpretación biológica a estos resultados?

Table 6.1: 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 |

# Bibliography

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