ISSN 0373-5680 *Rev. Soc. Entomol. Argent. 63 (3-4): 1-10, 2004*

FORO

La aplicación de Técnicas de Ordenación Multivariadas en la Entomología

MANGEAUD, Arnaldo
Centro de Investigaciones Entomológicas de Córdoba. Cátedra de Estadística y Biometría.
Facultad de Ciencias Exactas Físicas y Naturales. Universidad Nacional de Córdoba.
Av. Vélez Sársfield 299. 5000 Córdoba. Argentina;
e-mail: amangeaud@com.uncor.edu

■ **RESUMEN.** Los métodos de ordenación son herramientas multivariadas muy utilizadas en la Entomología. En este foro se presenta una introducción a éstos y breves explicaciones sobre distintas Técnicas: Análisis de Componentes Principales (ACP), Análisis de Redundancia (AR), Análisis de Correspondencia. (AC), de Correspondencia Canónica (ACC) y de Correspondencia Detendenciada (ACD), Análisis de Coordenadas Principales (AcoP), Análisis Factoriales (AF), Modelos de Ecuaciones Estructurales (MEE) y Método de Procrustes.

PALABRAS CLAVE. Análisis de Componentes Principales. Redundancia. Correspondencia. Correspondencia Canónica. Correspondencia Detendenciada. Coordenadas Principales.

■ **ABSTRACT. Application of Multivariate ordenation methods in Entomology.** Ordenation methods are multivariate technics used a lot in Entomology. At this forum is presented an introduction to ordination methods and a short explanation over different technics: Principal Component Analysis (PCA), Redundancy Analysis (RA). Correspondence Analysis (CA). Canonical Correspondence Analysis (CCA). Detrended Correspondence Analysis (DCA) Principal Coordinate Analysis (PcoA). Factor Analysis (FA), Structural Equation Models (SEM) and Procrustes method.

KEY WORDS. Principal Component Analysis. Redundancy. Correspondence. Canonical Correspondence. Detrended Correspondence. Principal Coordinates.

INTRODUCCIÓN

Puede resultar muy poco novedoso comenzar este artículo con una sentencia como: "la naturaleza es muy compleja". Pero aunque esta frase tan utilizada diera lugar a una sonrisa, podemos convenir que es cierta. Cuando un investigador observa o intenta explicar los eventos que ocurren en la naturaleza, comienza por tomar características aisladas, es decir de a una por vez. Además, en el afán de cumplir con el principio de parsimonia y que las explicaciones sean biológicamente coherentes, utiliza muy pocas características o variables independientes que expliquen a otras, dependientes. Esto es apoyado por toda una batería de análisis de la Estadística, que podemos denominar Estadística univariada.

Qué ocurre cuando el investigador ávido de conocimientos quiere más y no se conforma con sólo una variable. Si ésto pasara por vuestras mentes, entonces la solución a vuestras inquietudes tiene nombre y apellido: Análisis Multivariados, un conjunto de técnicas y análisis que utilizan muchas (dos o más) variables a la vez.

Las variables se definen como las características que les poseen las unidades de observación

(el objeto de estudio). Cada unidad posee un sinnúmero de variables, pero el investigador sólo toma o mide las que le serán útiles a sus objetivos. Las variables pueden ser de varios tipos: Cualitativa Nominal, como Presencia-Ausencia, sexo, coloración, forma de la mina de un minador (Agromyzidae), etc. Cualitativa Ordinal, como la especificidad alimentaria de una categoría taxonómica (monófagas, oligófagas y polífagas, estadios larvales, etc.) Cuantitativa Discreta, como número de huevos, número de antenómeros, días hasta la pupación, número de parasitoides, de huéspedes, de especies, etc. Cuantitativa Continua, como largo del ala, distancia recorrida, peso de las ovariolas, etc.

Algunos conocimientos preliminares. En primer término cabe acotar que en los análisis de Ordenación (más adelante daremos su definición) se le dará a todas las variables el mismo peso, por lo que si una variable fue medida en metros, otra en cm, otra en mg x l^{-1}, los análisis tomarán sus respectivos valores y la escala en que están medidas algunas, llevará la atención de todo el análisis. Por otra parte si algunas poseen mucha variabilidad y otras muy poca, toda la atención se la llevará la variable con mayor dispersión en los datos. Para observar esto, antes de realizar los análisis pertinentes, los investigadores deben realizar la mayor cantidad posible de análisis descriptivos y gráficos preliminares. Nadie mejor que el " dueño" de los datos para conocer la posición y dispersión, simetría y distribución de cada una de las variables en estudio. A ésta información se la complementa con gráficos de cajas (*box-plot*) en busca de datos anómalos (*outliers*) que deben ser identificados. Nótese la palabra identificar y no eliminar. Un dato anómalo NO se debe eliminar sólo por el hecho de identificarlo, debe existir una razón biológica para su eliminación. Porqué vamos a discriminar a un dato "rarito"?, porque no es como la mayoría?.

Por otra parte resulta interesante realizar gráficos de dispersión (tipo xy) de a dos variables a la vez para "ver" como se van relacionando de a pares. Por último existen herramientas descriptivas multivariadas como perfiles multivariados, gráficos de estrellas, etc. que muestran patrones generales tomando todas las variables a la vez (Johnson & Wichern, 1998).

Con el objeto de quitarle peso relativo a algunas variables o para que se cumplan ciertos supuestos, el investigador puede realizar transformaciones a los datos originales, con ello se trata de aplicar un cambio de escala en cada elemento de la variable, independientemente de otro.

Existen varios tipos de transformaciones, cada una con objetivos diferentes, siendo las más comunes:

Logaritmos: $y_{ij}= \text{Log } (x_{ij}+1)$, donde y_{ij} es el valor transformado, x_{ij} es el valor original.

Se utiliza normalmente para conseguir distribuciones simétricas en aquellas asimétricas a la derecha. También se utiliza para "acercar" los datos que se hallan lejanos hacia la derecha del eje cuando se presenta un alto grado de variación entre las variables o cuando dentro de una de ellas hay mucha variación. La forma mas común de realizar esta transformación es sumar previamente al dato original una constante (uno) ya que si hay datos ceros, el logaritmo de cero no está definido.

Transformaciones de Potencia (*Power transformations*): $y_{ij}= (x_{ij}+1)^p$, donde $0<p<9$.

Podemos diferenciar la transformación cuando $p<1$, de aquella donde $p>1$. En la primera estamos hablando de una raíz que se utiliza, al igual que el logaritmo, para "traer" datos lejanos. En el segundo caso es la potencia tal como la conocemos, y se la utiliza para hacer simétrica distribuciones que son asimétricas a la izquierda.

Arcoseno: $y_{ij}=\sqrt[2]{\text{Asen (pij)}}$, donde p_{ij} es la proporción

Se utilizan para normalizar las proporciones.

Relativizaciones:

Se utilizan para cambiar la escala de un valor de la variable, en referencia a otros valores.

Relativizada al máximo o al total: $y_{ij}= x_{ij} / \max x_j$

Es la proporción de un valor con respecto al máximo valor observado o a la suma de éstos.

Ajustada a la media: $y_{ij}= x_{ij} - \overline{X}_i$

Se utiliza para quitarle el peso relativo de cada variable, ya que se consigue una variable con media cero, pero conserva la varianza y la forma de la distribución.

Estandarización (ajustada a la media y al desvío):

Tabla 1: Ejemplo de la estructura de la Matriz de datos, de Varianzas-Covarianzas y Correlación. Se observa una matriz original con tres variables tomadas a cinco unidades de observación.

Matriz original de datos X_{nxp} donde n=5 y p=3	**Matriz de Varianzas-Covarianzas S_{pxp} donde p=3**	**Matriz de Correlación R_{pxp} donde p=3**
$X=\begin{bmatrix} 1 & 2 & 5 \\ 4 & 7 & 1 \\ 2 & 1 & 6 \\ 5 & 6 & 2 \\ 3 & 4 & 6 \end{bmatrix}$	$S=\begin{bmatrix} 2,5 & 3,5 & -2,75 \\ 3,5 & 6,5 & -5,25 \\ -2,75 & -5,25 & 5,5 \end{bmatrix}$	$R=\begin{bmatrix} 1 & 0,87 & -0,74 \\ 0,87 & 1 & -0,88 \\ -0,74 & -0,88 & 1 \end{bmatrix}$

$y_{ij}=\frac{(x_{ij}-\bar{X}_i)}{S_i}$, donde $\bar{X}_i$ es la media de la variable y S_i es el desvío.

Se le quita el peso relativo y la variabilidad, todas las variables presentarán media cero y varianza uno. Pero se conservará la forma de la distribución.

Es necesario acotar que la transformación de los datos no es un dogma de fe, hacerla no necesariamente soluciona todos los problemas. El transformar las variables tiene que perseguir un objetivo (Afifi & Clark, 1996). Después de realizar estos cambios en las variables se debe proceder a hacer nuevamente todos los análisis descriptivos para observar si se cumplieron los objetivos de este proceso.

Algo de Matrices y Estadística. En textos de matemática se puede leer que una matriz es un arreglo de números en la forma de filas y columnas, o sea una tabla de doble entrada. Para realizar los análisis que nos convocan construiremos matrices de la siguiente forma: en cada fila se colocarán cada una de las unidades de observación desde 1 a n. En cada columna, las variables, desde la 1 a la p. Por lo tanto tendremos una matriz X_{nxp} donde n es el número de unidades de observación y p el número de variables a considerar (Tabla 1).

Recordemos que la varianza es una medida de dispersión de todos los datos con respecto a su propia media. La covarianza, por su parte, es la variación conjunta de dos variables a la vez. Se puede considerar a la covarianza como una varianza entre dos variables, y a la varianza como una covarianza de una variable consigo misma. Con la matriz **X** de datos se puede construir una matriz $\mathbf{S}_{pxp}$ (matriz de varianzas-covarianzas) donde en su diagonal principal se observan las varianzas de cada una de las variables y en los triángulos de la matriz hacia arriba y hacia abajo se presentan las covarianzas entre las distintas variables (Tabla 1). La suma de las varianzas de la diagonal principal es la varianza total de la matriz.

El índice de correlación de Pearson entre dos variables se define como el cociente entre la covarianza y el producto de los desvíos. Si se divide a cada uno de los números de la matriz **S** por el producto de los desvíos correspondientes, se obtendrá una nueva matriz $\mathbf{R}_{pxp}$ llamada matriz de correlación, donde en la diagonal principal se presentan todos los unos (correlación entre una variable y ella misma) y en los triángulos, las correlaciones entre cada par de variables (Tabla 1).

DE LO PARTICULAR A LO GENERAL

Ejemplo 1: Análisis de Componentes Principales (ACP).

Supongamos que un investigador está estudiando un género de mosquitos, y toma dos variables a cada individuo (unidad de observación): largo del ala y largo del cuerpo (ruego a los especialistas no sonreír por la simplicidad del ejemplo). Si la combinación de estas dos variables pudiera separar a algunas especies, entonces resultaría interesante dibujar todas las unidades de observación en un espacio de dos dimensiones. Entonces se realizaría un gráfico de dispersión, en dos dimensiones, como el de la Figura 1. Ahora, si el investigador toma tres variables, las unidades de observación son dibujadas en un espacio de tres dimensiones, si fueran 10 variables estarían en 10 dimensiones y generalizando: en p variables, corresponderían p dimensiones. Resulta prácticamente imposible encontrar patrones o grupos de individuos en más de tres dimensiones. Por ello se deben utilizar herramientas para ex-

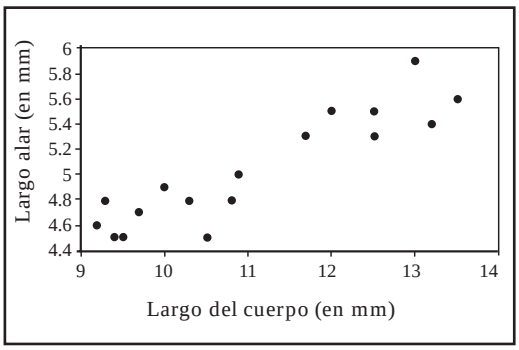

Figura 1: Gráfico de dispersión entre el largo del cuerpo y largo del ala largo de un género de mosquito.

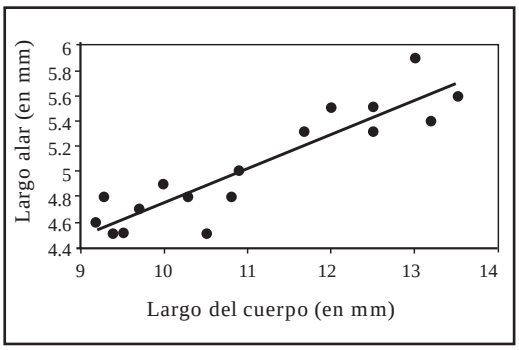

Figura 2: Gráfico de dispersión entre el largo del cuerpo y largo del ala largo de un género de mosquito donde figura el eje que absorve la mayor variabilidad de los datos.

traer conclusiones a partir de muchas dimensiones, es decir con todas las variables. Si volvemos a dos dimensiones veremos que una manera sencilla de encontrar un patrón es ver en qué sentido se presenta la máxima dispersión de datos. Este es el eje de mayor variabilidad de los datos y se construye pasando una recta de modo tal que se maximice la dispersión de datos en un sentido y se minimice la distancia de todos los puntos a ésta (Figura 2). Esto es lo que conocemos como autovector, vector propio o *eigenvector*. Asociado a este autovector se presenta un autovalor, valor propio o *eigenvalue*, que da una idea de la magnitud (el largo) del autovector. Esa es una medida de la variabilidad que absorbe ese vector. Después de ese eje de máxima variabilidad se tiene un segundo autovector, perpendicular (ortogonal) al primero y así sucesivamente hasta conseguir tantos p autovectores y autovalores como p variables posee la matriz.

Los autovalores y autovectores no surgen a partir de un dibujo, sino que se calculan a partir de matrices. Para ello se debe partir de una matriz original de datos $\mathbf{X}_{nxp}$ a partir de la cual se calcula la matriz $\mathbf{S}_{pxp}$. Si los datos originales han sido previamente estandarizados, lo que se obtiene es una matriz de correlación (**R**) donde se presentan los valores de los índices de correlación de Pearson. Así como en un análisis de regresión ordinaria (con método de mínimos cuadrados) se construye una recta (eje), aquí ocurre lo mismo, pero la forma de construírlo es distinta. La fórmula del eje no tiene ordenada al origen y consta de una pendiente por cada una de las p variables. Entonces ese eje de mayor variabilidad está compuesto por una porción de la información de cada variable, es lo que se llama una combinación lineal de las variables:

$$CP_1 = a_{11} X_1 + a_{12} X_2 + ... + a_{1p} X_p$$
$$CP_2 = a_{21} X_1 + a_{22} X_2 + ... + a_{2p} X_p$$
$$CP_p = a_{p1} X_1 + a_{p2} X_2 + ... + a_{pp} X_p$$

Donde CP_i: eje o componente principal, X_j: variables, a_{ij}: pendientes.

Como se vio anteriormente la varianza total del análisis es la suma de las varianzas de cada variable. Entonces se tiene que la suma de los autovalores es igual a la suma de todas las varianzas, es decir la varianza total. Por definición, los autovectores se presentan desde el mayor autovalor al menor de éstos. El primer autovector va a absorber mucha variabilidad, el segundo menos y así sucesivamente, pero cuántos autovectores son suficientes para explicar gran parte de la variabilidad de la matriz?. Una regla práctica sencilla dice: Sólo sirven los autovectores que absorben más variabilidad que el promedio de las varianzas (si se trabaja con una matriz de varianza-covarianza) o si son mayores que 1 (si se trabaja con una matriz de correlación). Este criterio de decisión se denomina **criterio de la raíz latente**. (Ver otros criterios en Hair *et al.*, 1995 y McGarigal *et al.*, 2000).

A partir de los autovalores escogidos se realizará un nuevo gráfico donde se ordenarán las unidades de observación. Allí se colocarán en los ejes x e y a los autovectores 1 y 2 y se observarán los patrones que presenta el gráfico: quiénes están en los extremos?, se observan grupos?, cuán cerca están las unidades de observación?. Se pue-

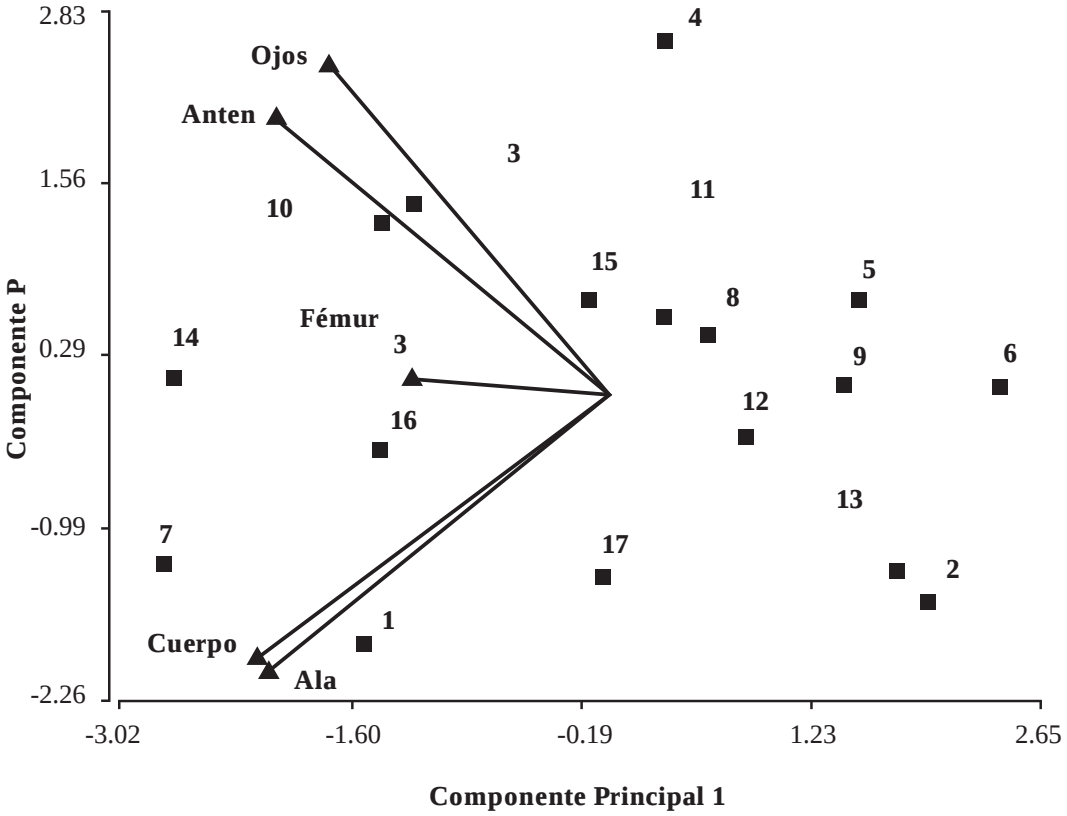

Figura 3: Gráfico (*Biplot*) de un Análisis de Componentes Principales, donde figuran las unidades de observación y las variables.

de realizar, además, otro gráfico solapado (*biplot*) donde superpone las variables. Estas se unen al centro del gráfico y dan una idea de cuán cerca están algunas variables de las unidades y cuán correlacionadas están las variables entre sí. Cuanto más agudo sea el ángulo entre variables la correlación es mayor, si el ángulo es de 90 grados la correlación nula y si es muy obtuso es una correlación inversa (Figura 3).

Lo que acabamos de ver es el conocido Análisis de Componentes Principales (ACP). Esta técnica forma parte de un grupo de métodos denominados de ordenación o de reducción de la dimensionalidad. Ordenan a las unidades de observación y reducen de p dimensiones a pocos vectores que llevan gran parte de la carga de la información original. Los objetivos de estos análisis apuntan a generar hipótesis y no a probarlas, por lo tanto no forman parte de la Estadística Inferencial.

Este Análisis es altamente recomendado cuando tenemos entre manos variables con distribuciones normales, más aún debieramos tener una matriz con distribución normal multivariada. Además de esto, el investigador no tiene información de grupos *a-priori* entre las unidades de observación. Debido a que utiliza una matriz de varianzas-covarianzas, considera que entre las unidades de observación se presenta una distancia Euclídea, entonces los resultados obtenidos son equiparables a un dendrograma realizado con distancia Euclídea y un método de unión llamado de ligamiento promedio.

A partir de los resultados la pregunta que surge es: por qué razón algunos individuos se acercan a otros?, y porqué son distintos los que están en los extremos de los ejes?. En parte esas preguntas pueden ser respondidas con las mismas variables que generaron el análisis. Algunas tendrán mucho peso en el eje 1 y estarán correlacionadas con él, otras en el eje 2. De allí el investigador podrá conocer cuáles son las variables que formaron parte del análisis que marcan mayores diferencias. Pero estos grupos que se han formado,

son reflejo de algún gradientes de algo?. Podría ser un gradiente geográfico?, ambiental?. Para contestar estas pregunta surge el Análisis de Redundancia, que consta de la utilización de variables ambientales o anexas (que no fueron utilizadas previamente en el ACP) para explicar esos gradientes. Se realizan regresiones con las variables anexas oficiando de variables independientes y los componentes principales (como dependientes) con el fin de encontrar el modelo que mejor explique la ordenación de las unidades de observación obtenidas por el ACP. Dos son las formas de aplicarlo: mediante métodos de mínimos cuadrados ordinarios o mediante análisis de permutación, que son regresiones del tipo no paramétricas (Good, 2000, Mielke & Berry, 2001).

Es recomendable utilizar Análisis de Componentes Principales cuando se posee variables normales como en muchos casos son las medidas morfométricas. Por otra parte también se debe poseer una matriz con n>p, es decir más unidades que variables, aunque observando la bibliografia se intuye que esto muchas veces no se cumple. Dependiendo del objetivo del investigador, antes de realizar un ACP se pueden transformar las variables originales con los métodos descriptos anteriormente. El logaritmo le quitará peso a los datos extremos, con la estandarización se obtendrá que todas las variables tengan el mismo peso y la misma variabilidad, etc.

Los entomólogos que trabajan en Ecología deben tener en cuenta que un supuesto muy fuerte del ACP es asumir linealidad en la relación especie-gradiente. Entonces se debe utilizar sólo si los gradientes ambientales son "cortos". En un gradiente "largo" las especies tienen un valor óptimo donde son más abundantes, y a valores mayores o menores del gradiente ambiental son menos abundantes, y por lo tanto no debe ser utilizado en esos casos.

Después de haber explicado el ACP podemos ver claramente que se denomina Ordenación a un diagrama en que las unidades de observación se presentan como puntos en un espacio de (por ejemplo) dos dimensiones, que surgieron de una combinación de las originales (adaptado de Jongman *et al.*, 1995). También se comprende que el ACP muestra (si los datos así lo requieren) un gradiente indirecto entre las unidades de observación. Pero el Análisis de Redundancia, utilizado como complemento con variables ambientales, mostrará un gradiente directo (Jongman *et al.*, 1995).

Analizando el Ejemplo 1.

La figura 3 representa el *biplot* del ACP. Allí se observan las unidades de observación representadas por círculos y las variables, por triángulos. A su vez las variables están unidas al centro (el cero de ambos ejes). En este ejemplo se ordenaron 17 individuos tomándoles cinco variables a cada uno. El gráfico muestra que las unidades 2, 6 y 13 se hallan en un extremo del eje de mayor variabilidad, mientras que las unidades 7 y 14 se hallan en el extremo opuesto. Esto sugiere que estas unidades estan en lugares "diametralmente opuestos". No se observan "grupos" definidos, sino un gradiente de unidades de observación. El segundo eje de mayor variabilidad arroja a las unidades 1, 2, 7, 13 y 17 en un extremo y la unidad 4 en el otro.

Por su parte es muy pequeño el ángulo formado entre las variables ala y cuerpo lo que muestra su altísima correlación. Antena y ojos también se hallan correlacionadas, no así las dos primeras con las dos últimas, que forman un ángulo aproximado a 90 grados. Cuando el Componente Principal 1 aumenta, Ala y Cuerpo disminuyen. Fémur 3, por su parte está también correlacionada al Componente 1, pero no correlacionada al 2, ya que la "sombra" que produce en éste es muy pequeña. Este análisis y otras observaciones le servirán al autor para generar hipótesis sobre la disposición de las unidades y las variables. El porqué algunas son más parecidas a otras, por qué se produce ese gradiente, etc.

Además del gráfico se obtiene la información numérica. El primer Componente Principal (por ejemplo) arrojó un autovalor de 2,16, que representa un 60 % de la variabilidad total (suma de autovalores=3,6). El segundo eje posee un autovalor de 1,08, que representa un 30 % de la variabilidad. Es decir que en los dos primeros ejes se captura un 90% de la variabilidad de los datos.

Ejemplo 2: Análisis de Correspondencia (AC)

Se quiere ordenar distintos puntos de muestreo en un campo de maíz, sobre la base de la fauna de artrópodos del suelo. En este caso tenemos unidades de observación (puntos), y a cada una de ellas se le toman muchas variables (cada especie). Entonces se puede hacer el intento por buscar un autovector que maximice la dispersión entre las unidades (puntos) para conocer cuáles son más

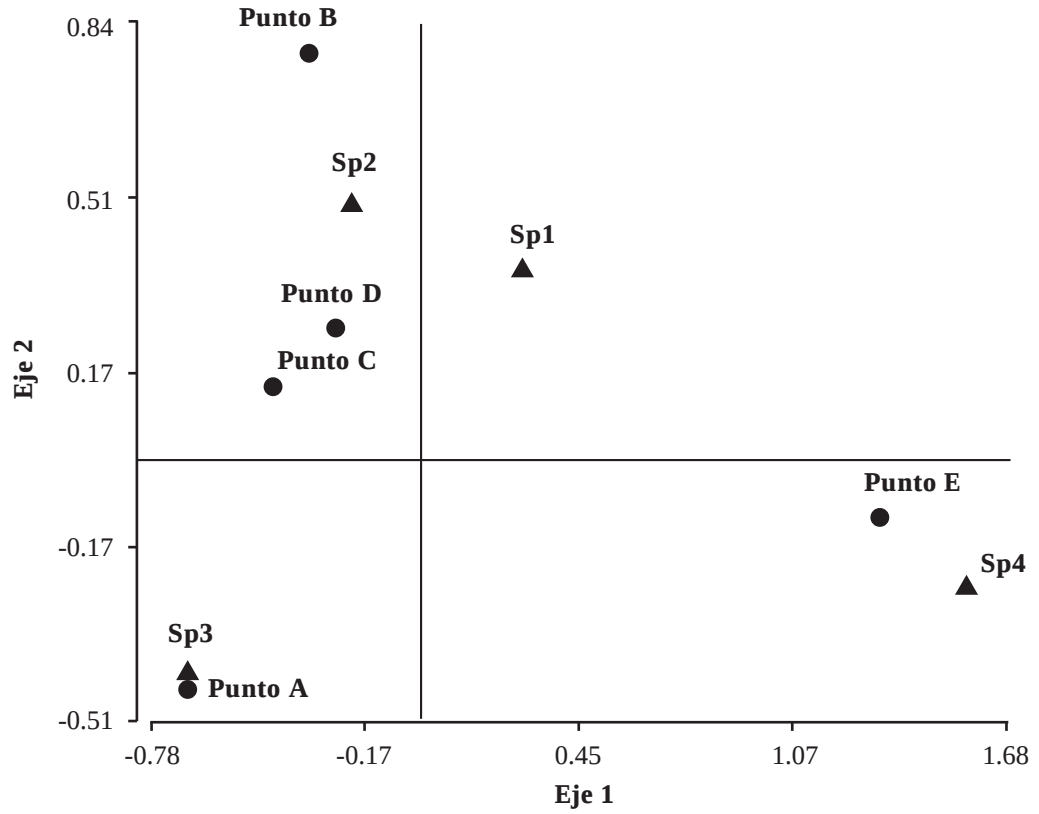

Figura 4: Gráfico (*Biplot*) de un Análisis de Correspondencia, donde figuran las unidades de observación (puntos) y las especies tomadas como variables (triángulos).

parecidos entre sí, dicho de otro modo: los puntos del borde del campo serán distintos a los del centro?. Pero el número de individuos de cada especie es un conteo, que no posee distribución normal. Para este caso se recomienda utilizar una distancia denominada chi cuadrado, que se basa en las frecuencias esperadas que se calculan en las tablas de contingencia (como en el conocido test Chi cuadrado de independencia). Después de encontrar esas distancias, el análisis traza el primer vector, luego el segundo y así sucesivamente. La interpretación de ésto es la misma que en el ACP. La diferencia consiste en que se analiza tanto el gráfico de las unidades en el espacio de las variables como el de las variables en el espacio de las unidades. Entonces se superponen los dos gráficos, obteniendo uno compuesto en este caso por puntos y especies llamado *biplot* (Figura 4).

Como la distancia ya no es euclídea ni representa la varianza entonces no se puede hablar de la varianza que explican los ejes y la variabilidad explicada entonces se denomina inercia explicada. Este análisis es conocido también como *Reciprocal Averaging* (McGarigal *et al.*, 2000).

Un análisis posterior al AC es el Análisis de Correspondencia Canónica, el equivalente al Análisis de Redundancia en este marco. Se genera un gráfico **Triplot**, que posee las unidades, las variables y las variables anexas utilizadas.

Contrariamente a las suposiciones de ACP, el Análisis de Correspondencia asume que la relación entre las variables y los gradientes ambientales es de tipo unimodal, de este modo se soluciona el problema de los gradientes "largos". Pero del mismo modo que ocurre en ACP, tiene un inconveniente: en el segundo eje se produce un "*artilugio*" matemático que deforma la ordenación, es un efecto de curvatura donde se cambia la verdadera distancia entre unidades de observación, denominado efecto herradura (Jongman *et al.*, 1995).

Analizando el Ejemplo 2:

La figura 4 representa el biplot del AC. Allí se observan las unidades de observación represen-

tadas por círculos (sitios o puntos de muestreo) y las variables, por triángulos (especies). En este ejemplo se ordenaron cinco puntos de muestreo basadas en cuatro especies. El gráfico muestra que los puntos A y E se encuentran en posiciones diametralmente opuestos con respecto al primer eje de mayor variabilidad. A su vez el punto A se encuentra "asociado" a la especie 3 y el punto E a la especie 4. El segundo eje de mayor variabilidad separa a estos dos puntos de los sitios B, C y D. A su vez también se pueden pensar en las asociaciones entre especies, las especies 1 y 2 forman un grupo que se halla relacionado a los tres sitios que están en la porción superior del gráfico.

De la información numérica obtenida se desprende que el primer autovalor es igual a 0,82, mientras que el valor del segundo es de 0,42. El primero consigue atrapar una variabilidad (inercia) del 75 %, mientras que el segundo toma un 19% de ésta. Entre ambos se consigue explicar el 94 % de la variabilidad de los datos.

Ejemplo 3: Análisis de Correspondencia Detendenciado (ACD)

Se está trabajando en el efecto de la contaminación sobre la comunidad de insectos bentónicos. Entonces se intenta ordenar distintos sitios sobre la base de las variables (especies de insectos).

El Análisis de Correspondencia Detendenciado (Hill & Gauch, 1980) es una técnica de ordenación que quita el efecto herradura explicado anteriormente. Ha sido muy utilizado en los trabajos científicos de las últimas dos décadas. El primer programa de computación que realizaba este análisis se llamó Decorana por lo que se ha generalizado el análisis con ese nombre. Utiliza igualmente que el Análisis de Correspondencia una distancia chi cuadrado, pero a los ejes los particiona en segmentos que "desarman" la herradura. La interpretación de los ejes es la misma que en Componentes Principales y Correspondencia: se busca el eje de mayor variabilidad y se intenta absorver en la menor cantidad de ejes posibles la mayor variabilidad de los datos originales, que sigue llamándose inercia explicada.

Además para entender a los ejes como gradientes ambientales, se presenta un Analisis de Correspondencia Detendenciado Canónico, que grafica un **Triplot**.

Ejemplo 4: Análisis de Coordenadas Principales. (ACoP)

Un especialista en genética posee varias variables de diferentes unidades de observación, pero él sabe que se ha llegado a un acuerdo y la mejor distancia para medir la similitud o disimilitud entre los individuos es un índice basados en distancias genéticas. Entonces le interesa estudiar cómo se ordenan sus unidades sobre la base de esas distancias, no le interesa ni la Euclidea (como ACP) ni Chi cuadrado (como AC ó ACD). Entonces se puede buscar el mayor autovector que maximice la dispersión de datos con otra distancia?. La respuesta es afirmativa, el Análisis de Coordenadas Principales lo hace. Cualquier medida de distancia propuesta, aún alguna que haya "inventado" el investigador puede ser utilizada. De esta forma el ACP pasa a ser un caso particular de AcoP para la distancia euclídea y el AC otro caso particular para la distancia Chi cuadrado. Así el ACoP amplía el análisis para aquellas distancias que esten probadas ser más representativas de las similitudes para algunas áreas de estudio. Como ejemplo, Anderson & Willis (2003) aseguran que en algunas áreas de Ecología es preferible utilizar este análisis con las medidas de Bray-Curtis o de Kulczynski.

Como hemos estado viendo en los casos anteriores, aquí se podrían pensar en análisis ulteriores: El Análisis de Coordenadas Principales Canónicas busca restringir los resultados a otros datos por ejemplo ambientales.

Un sinónimo de ACoP es Análisis de Escalamiento Multidimensional (*Multidimensional Scaling*). Con ese nombre se conocieron a técnicas que originalmente se realizaron para hacer "mapas" a partir de matrices de distancia. En ese marco se intenta priorizar las coordenadas principales de manera que se produzcan la menor distorsión con respecto a las distancias originales. La bondad de ajuste del método la calcula mediante un valor llamado *Stress* (u otro llamado *Sstress*) que da una idea de la distorsión ocurrida. Por su parte se puede tener un Escalamiento Multidimensional Métrico (*Metric Multidimensional Scaling*) que trabajan con la verdadera magnitud de las distancias o Escalamiento Multidimensional No Métrico (*Non Metric Multidimensional Scaling*). Esta es una ordenación que se puede hacer con datos que son no normales, arbitrarios, discontinuos o cuestionables. Está basado en distan-

cias, pero no utiliza la magnitud de la distancia sino sólo el número de orden de las mismas (distancias ordenadas o *rankeadas*).

OTROS ANÁLISIS

Análisis Factorial (AF)

Recordemos que en el ACP, cada componente es una combinación lineal de variables. En el Análisis Factorial se invierte esta idea: se asume que cada variable es una combinación lineal de factores desconocidos más un residuo o error no medido. De modo que:

$$X_1 = \mu_1 + l_{11} F_1 + l_{12} F_2 + ... + l_{1m} F_m + e_1$$

$$X_p = \mu_p + l_{p1} F_1 + l_{p2} F_2 + ... + l_{pm} F_m + e_p$$

donde X_i representa cada variable, μ_i es una ordenada al origen, l_i son las pendientes, F_i factores, e_i error.

Esos factores son no observables y tienen aleatoriedad.

Este análisis ha tenido muy poca utilización en las Ciencias Biológicas. Se han conocido diversos ejemplos tanto en Psicología, como en Geología (Davis, 1986). En este momento se lo está revalorizando, ya que establece las bases para otros análisis denominados Modelos de Ecuaciones Estructurales (*Structural Equation Models*). Estos MEE son herramientas que permiten contrastar hipótesis sobre relaciones causales en datos observacionales. Presentan una alternativa a los diseños experimentales sustituyendo al control experimental por un control estadístico. Iriondo *et al.*, 2003 provee una excelente explicación aplicada a datos biológicos. Para detalles teóricos ver Tabachnick & Fidell (1996).

Procrustes

Es una técnica para comparar si son similares los resultados de dos o más ordenaciones realizadas sobre las mismas unidades de observación. Este método deja fija una de las ordenaciones y corre, estira y rota la/s otras, de manera tal que la distancia entre los puntos ordenados por los distintos sistemas sea el mínimo posible (Digby & Kempton, 1991). Además esto se complementa con un análisis de permutaciones de manera tal de obtenerse una probabilidad que esa "semejanza" entre las ordenaciones pueda deberse sólo al azar o si el patrón de semejanza es significativo. Este tipo de análisis se esta utiliza con frecuencia en los análisis morfométricos conocido como "*landmark*".

Como corolario quiero expresar que este trabajo intentó presentar una sencilla explicación y puesta al día de los métodos de ordenación más comúnmente utilizados en las Ciencias Biológicas en general y en la Entomología en particular. Debiera ser tomado con ese objetivo y así servir como un nexo o introducción para que los interesados arriben a textos de mayor (y mejor) envergadura en pos de mayor claridad.

AGRADECIMIENTOS

Quiero agradecer a la Directora de la Revista de la Sociedad Entomológica Argentina, Dra. Lucía Claps y a la Editora Dra. Graciela Valladares, la invitación a participar en este foro.

BIBLIOGRAFIA CITADA

AFIFI, A. & V. CLARK. 1996. *Computer-aided multivariate analysis*. Chapman & Hall. Boca Raton.

ANDERSON, M. & T. WILLIS. 2003. Canonical Analysis of Principal Coordinates: a useful method of constrained ordination for Ecology. *Ecology*. 84:511–525.

DAVIS, J. 1986. *Statistics and data analysis in Geology*. John Wiley & Sons. New York.

DIGBY, P. & R. KEMPTON. 1991. *Multivariate analysis of ecological communities*. Chapman & Hall. London.

GOOD, P. 2000. *Permutation test. A practical guide to resampling methods for testing hypotheses.* Springer. New York.

HAIR, J., R. ANDERSON, R. TATHAM & W. BLACK. 1995. *Multivariate data analysis with readings.* Prentice-Hall. New Jersey.

HILL, M. & H. GAUCH. 1980. Detrended correspondence analysis. an improved ordination technique. *Vegetatio* 42:47-58.

IRIONDO, J., M. ALBERT & A. ESCUDERO. 2003. Structural equation modelling. an alternativa for assessing causal relationships in threatenet plant populations. *Biological conservation*. 113: 367-377.

JOHNSON, R. & D. WICHERN. 1998. *Applied multivariate statistical analysis*. Prentice-Hall. New Jersey.

JONGMAN, R., C. TER BRAAK & O. VAN TONGEREN. 1995. *Data analysis in community and landscape ecology*. Cambridge Univ. Press. Cambridge.

MCGARIGAL, K., S. CUSHMAN & S. STAFFORD. 2000. Multivariate Statistics for Wildlife and Ecology Research. Springer. New York.

MIELKE, P. & K. BERRY. 2001. *Permutation Methods. A distance function* approach. Springer. New York.

TABACHNICK, B. & L. FIDELL. 1996. *Using multivariate statistics*. HarperCollins College Publishers. New York.

Recibido: 26-VIII-2004
Aceptado: 19-XI-2004