UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS BIOLÓGICAS Y AGROPECUARIAS

ANÁLISIS MULTIVARIABLE EN ECOLOGIA MICROBIANA

TRABAJO DE TITULACIÓN EN LA MODALIDAD DE

MONOGRAFIA DE ACTUALIZACIÓN

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN BIOLOGIA

PRESENTA

ELVIS GIOVANNI EZEQUIEL GUZMÁN ORNELAS

Las Agujas, Zapopan, Jal., Agosto de 2008

Universidad de Guadalajara

Centro Universitario de Ciencias Biológicas y Agropecuarias

Coordinación de Titulación y Carrera de Licenciatura en Biología

C. Elvis Giovanni Ezequiel Guzmán Ornelas

PRESENTE

Manifestamos a ustedes que con esta fecha ha sido aprobado su tema de titulación en la modalidad de: **Investigación y Estudios de Posgrado** opción: **Trabajo Monográfico de Actualización** con el titulo: **"Análisis Multivariable en Ecología Microbiana"** para obtener la Licenciatura en Biología.

Al mismo tiempo le informamos que ha sido aceptado como Director / a de dicho trabajo el/la: **M.C. Adrián Ricardo López González.**

Sin más por el momento, le envío un afectuoso saludo.

ATENTAMENTE
"PIENSA Y TRABAJA"
Las Agujas, Zapopan., 30 de noviembre del 2007.

DR. FRANCISCO MARTÍN HUERTA MARTÍNEZ
PRESIDENTE DEL COMITÉ DE TITULACIÓN

C.A.

M en C. GLORIA PARADA BARRERA
SECRETARIO DEL COMITÉ DE TITULACIÓN

FORMA F

Dr. Fco. Martín Huerta Martínez.
Presidente del Comité de Titulación.
Licenciatura en Biología.
CUCBA.
Presente

Nos permitimos informar a usted que habiendo revisado el trabajo de titulación, modalidad **INVESTIGACIÓN Y ESTUDIOS DE POSGRADO**, opción **MONOGRAFÍA DE ACTUALIZACIÓN** con el título: "**ANÁLISIS MULTIVARIABLE EN ECOLOGÍA MICROBIANA**" que realizó el pasante **ELVIS GIOVANNI EZEQUIEL GUZMÁN ORNELAS** con número de código **300519162** consideramos que ha quedado debidamente concluido, por lo que ponemos a su consideración el escrito final para autorizar su impresión.

Sin otro particular quedamos de usted con un cordial saludo.

Atentamente

Las Agujas, Zapopan, Jalisco, 18 de agosto de 2008

V.B°

M. C. Adrián Ricardo López González
Director del trabajo

Nombre completo de los Sinodales asignados por el Comité de Titulación	Firma de aprobado	Fecha de aprobación
DR. FRANCISCO MARTIN HUERTA MARTINEZ		19/08/08
M.C. ROSALBA MIREYA HERNANDEZ HERRERA		18/08/08
M.C. PABLO TORRES MORAN		19/08/08
M.C. SALVADOR VELAZQUEZ MAGAÑA (SUPLENTE)		14/08/08

AGRADECIMIENTOS

1. A DIOS Y COLABORADORES

2. A MI MADRE Y A MI FAMILIA

3. A MIS AMIGOS Y PROFESORES:

- CYNTHIA TEMORES
- DOLORES BARRAGAN
- KARINA RODRIGUEZ
- NESTOR SANTANA
- MIREYA HERNÁNDEZ
- ADRIAN LÓPEZ
- MARGARITA BONILLA

CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

El presente trabajo presenta una visión global de los métodos de análisis multivariable aplicados en ecología microbiana, a manera de guía, que permita orientar sobre la elección de las herramientas multivariables para identificar e interpretar los patrones de diversidad microbiana.

En este trabajo se presenta una introducción al campo de estudio de la ecología microbiana y al análisis multivariable. Los métodos se dividieron en tres grupos: de ordenación, de clasificación y de inferencia. Se describen los métodos, sus ventajas, limitaciones y algoritmos; además se presentan los ejemplos sobre su aplicación e interpretación de acuerdo a los objetivos y tipo de estudio en particular.

En los diagramas de ordenación los puntos representan muestras, perfiles moleculares, objetos o especies y los vectores representan variables ambientales. Los puntos más cercanos entre sí indican una similitud en sus atributos u ocurrencias; además de mostrar su respuesta a una variable latente o ambiental. De esta forma se redujo la dimensionalidad de los datos.

En la clasificación, la interpretación depende de las características del grupo o conglomerado.

En los métodos inferenciales se comparan grupos mediante la razón entre la variabilidad entre los grupos y la variabilidad dentro de los grupos. Cuando los elementos de los grupos representan perfiles moleculares (poblaciones o comunidades microbianas) entonces una alta variación entre los grupos indica que las comunidades de cada grupo son diferentes entre grupos, mientras que una baja variación dentro de los grupos indica que las comunidades dentro de los grupos son similares.

1. INTRODUCCIÓN

1.1. GENERALIDADES SOBRE ECOLOGIA MICROBIANA

1.1.1. DEFINICIÓN DE ECOLOGIA MICROBIANA

La Ecología microbiana estudia las relaciones de los microorganismos con su ambiente biótico y abiótico. Se entiende por ambiente biótico los demás microorganismos (bacterias, protozoarios, algas y hongos), plantas y animales que le rodean; mientras que el ambiente abiótico se refiere a todos los factores físicos y químicos. (Atlas y Bartha, 2002).

1.1.2. CARACTERÍSITICAS GENERALES DE LOS MICROORGANISMOS

Los microorganismos son seres vivos que, por lo general, tienen dimensiones del orden de micras (μm) y no forman tejidos.

En la tabla 1 se muestra un resumen de las características generales de los microorganismos (Atlas y Bartha, 2002; Hurst *et al.*, 2002).

TABLA 1. Características generales de los microorganismos.

Característica	GRUPOS MICROBIANOS			
	Bacterias	Protozoarios	Algas eucariotas	Hongos
Fuente de Carbono	Autotrofos / Heterotrofos	Por lo general heterotrofos / Mixotrofos	Por lo general autotrofos / Mixotrofos	Heterotrofos
Fuente de Energía	Fototrofos / Quimiotrofos	Por lo general quimiotrofos / pocos fototrofos	Por lo general fototrofos / pocos quimiotrofos	Quimiotrofos
Aceptor de electrones	Aerobios / Anaerobios	Aerobios / Anaerobios	Aerobios	Aerobios / Anaerobios
Medio en el que viven	Acuático / Terrestre	Por lo general acuático / Terrestre	Acuático / pocas terrestres	Acuático / Terrestre
Estilo de vida	Parásitos / Vida libre / Simbiontes	Parásitos o vida libre / pocos simbiontes	Vida libre / Simbiontes	Parásitos / Saprobios

Se encuentran donde quiera que existan plantas y animales e incluso en ambientes extremos como en las regiones polares o en los manantiales de aguas termales. Su éxito se debe a las siguientes características:

- Presentan requerimientos de nutrientes moderados y una amplia tolerancia de condiciones ambientales, de esta manera crecen y se reproducen en una amplia variedad de sustratos y en una variedad de condiciones.
- Tienen las tasas de crecimiento y reproducción más rápidas.
- Tienen la capacidad de producir toxinas y al mismo tiempo tolerancia a las toxinas de otros seres vivos.

1.1.3. ESTRUCTURA DE LAS COMUNIDADES E INTERACCIONES MICROBIANAS

Los ecólogos microbianos con frecuencia desean conocer la estructura de las comunidades en los ambientes naturales y urbanos, las interrelaciones que guardan unas especies con otras, así como las variables ambientales que influyen en la dinámica y organización de esas comunidades (Gauch, 1982; Atlas y Bartha, 2002).

La estructura de la comunidad es determinada por las especies presentes en un ambiente y las proporciones relativas de esas especies. (Hurst *et al.*, 2002). De esta manera; las especies, su dinámica y los factores ambientales en un medio acuático serán diferentes de un medio edáfico.

Los factores que afectan las comunidades en el agua son la temperatura, el pH, los gases disueltos, la profundidad, la luz solar, los nutrientes, las corrientes y la conductividad.

Los factores que afectan las comunidades en el suelo son el tamaño y naturaleza de las partículas, la materia orgánica, el agua y la atmósfera del suelo, la temperatura, el pH, las enzimas extracelulares y nutrientes, además de los fenómenos de adsorción y adhesión.

Las comunidades microbianas se organizan en base a las interacciones entre especies de microorganismos y éstos con las plantas y

animales. Estas interacciones pueden ser neutrales, positivas o negativas (tabla 2).

Las relaciones neutrales son aquellas donde no existe algún beneficio o daño para ambas especies, debido a que ocupan nichos ecológicos diferentes y no necesitan uno del otro.

Las relaciones positivas son aquellas donde ambas se benefician (mutualismo: sinergismo y simbiosis) o una de ellas se beneficia y la otra le es indiferente (comensalismo).

Las relaciones negativas son aquellas en donde ambas especies se ven afectadas (competencia), una de ellas se beneficia y la otra se ve perjudicada (depredación y parasitismo) o bien una se ve afectada mientras la otra le es indiferente (amensalismo).

Por lo general, las interacciones positivas predominan cuando la densidad de la población es baja, mientras que las interacciones negativas predominan cuando la población es alta.

TABLA 2. Interacciones microbianas. Tomado de Atlas y Bartha (2002).

Tipo de interacción	Efecto de la interacción	
	Especie A	Especie B
Neutralismo	Sin efecto	Sin efecto
Comensalismo	Sin efecto	Efecto positivo
Sinergismo	Efecto positivo	Efecto positivo
Simbiosis	Efecto positivo	Efecto positivo
Competencia	Efecto negativo	Efecto negativo
Amensalismo	Sin efecto	Efecto negativo
Depredación	Efecto positivo	Efecto negativo
Parasitismo	Efecto positivo	Efecto negativo

1.1.4. APLICACIONES

Con el conocimiento de la Ecología microbiana y otras ciencias afines, se ha aprendido a aprovechar las interacciones microbianas y las actividades biosintéticas y biodegradativas de los microorganismos (Hurst *et al.*, 2002) para:

- La producción de materiales benéficos para el hombre como son antibióticos, vitaminas y biocombustibles.
- La biodegradación de materiales naturales y antropogénicos como son las aguas residuales y las compostas.
- Controlar plagas que afectan cultivos agrícolas o bien controlar microorganismos que afectan la salud pública.
- Prevenir las actividades microbianas naturales como biocontaminación, corrosión o desintegración de materiales expuestos al ambiente.
- Evaluar el ambiente mediante bioindicadores.

1.2. GENERALIDADES SOBRE ANÁLISIS MULTIVARIABLE

1.2.1. DEFINICIÓN DE ANÁLISIS MULTIVARIABLE

El análisis multivariable es un conjunto de métodos matemáticos que analizan un gran número de variables de manera simultánea. (Peña, 2000; Jonson, 2000; Gauch, 1982).

1.2.2. ANTECEDENTES

Los métodos de análisis multivariable se desarrollaron durante el siglo XX (tabla 3); comenzaron para resolver problemas de clasificación en Biología, se extendieron para encontrar variables indicadoras y factores en Psicometría, Mercadotecnia y las Ciencias sociales y han alcanzado gran aplicación en Ingeniería y Ciencias de la Computación como herramientas para resumir la información y diseñar sistemas de clasificación automática y de reconocimiento de patrones (Peña, 2000).

TABLA 3. Antecedentes del Análisis multivariable.

PERSONAJE	APORTACIÓN
K. Pearson (1900)	Obtuvo el estimador del coeficiente de correlación de Pearson.
J. P. Benzecri (1930)	Creó el análisis de correspondencia (basado en una tabla de contingencia).
P. C. Mahalanobis (1931)	Inventó la distancia de Mahalanobis.
S. M. Wilks (1932)	Construyó generalizaciones multivariables para el análisis de varianza.
H. Hotelling (1933)	Inventó el análisis de componentes principales.
R. A. Fisher (1933)	Inventó el análisis discriminante y el diseño estadístico de experimentos.
Bray y Curtis (1957)	
Shepard (1962) y Kruskal (1964)	Describieron una técnica para el Escalado multidimensional no métrico.
Rao (1964)	Inventó el análisis de redundancia.
Gower (1966)	Describió el análisis de coordenadas principales.
MacQueen (1967)	Introdujo el algoritmo de k-medias.
Mantel (1969)	Describió la prueba de Mantel.
Hill (1973)	Introdujo el análisis de correspondencia en ecología con el algoritmo de promedio ponderado
Hilll y Gauch (1980)	Desarrollaron el análisis de correspondencia sin tendencia
Ter Braak (1986)	Describió el análisis de correspondencia canónica
Clarke (1993)	Describió el análisis de similitud (ANSiM).
Legendre y Anderson (1999)	Describieron el método de análisis de redundancia basado en distancia.

Existen otros trabajos aplicados a comunidades de plantas y animales como el de Gauch (1982) donde se presenta una explicación de los métodos multivariables básicos aplicados a estas comunidades. Así mismo, se han desarrollado trabajos en ecología microbiana en donde se han aplicado los métodos de análisis multivariables, por ejemplo, en estudios de fitoplancton (Mora-Navarro *et al.*, 2004; Ariyadej *et al.*, 2004), en comunidades bacterianas (Kaneene *et al.*, 2007; Blackwood *et al.*, 2006) y en estudios de hongos (Lanoiselet *et al.*, 2005). Ramette (2007) presentó un trabajo de revisión en el cual expuso una breve explicación de algunos métodos multivariables aplicados en ecología microbiana así como algunas sugerencias sin ser exhaustivo ni profundizar.

1.2.3. OBJETIVOS DEL ANÁLISIS MULTIVARIABLE

De acuerdo a Peña (2000) y a Jonson (2000) los del análisis multivariable son:

- Resumir los datos mediante un pequeño conjunto de nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información; lo cual permite reducir la complejidad del conjunto de datos.
- Encontrar patrones en los datos, si existen.
- Clasificar nuevas observaciones en grupos definidos.
- Relacionar dos conjuntos de variables.
- Examinar numerosas variables simultáneamente.
- Obtener los mejores resultados al más bajo costo y con el menor esfuerzo.

1.2.4. CLASIFICACIÓN DE LOS MÉTODOS

En Ecología microbiana existen dos tipos de métodos multivariables: de ordenación y de clasificación.

La ordenación (sección 3) representa las muestras (u objetos) con base en la composición de especies (o con base en los atributos de los objetos) en un espacio de menos dimensiones, por lo general dos o tres dimensiones ya que esto permite una representación gráfica y extrae la mayor cantidad de variación del conjunto de datos.

La ordenación se divide en dos tipos: ordenación indirecta o sociológica (sección 3.3 a 3.8) y ordenación directa o restringida o ambiental (sección 3.9 a 3.10). La ordenación indirecta utiliza la matriz de datos de especies x muestras para encontrar las variables latentes que explican la variación en los datos y luego las relaciona con variables ambientales. La ordenación directa relaciona los datos de composición de especies y los datos de variables ambientales mediante una regresión múltiple de las variables ambientales (explicatorias) sobre los puntajes de muestras de manera que estos últimos sean combinaciones lineales de las variables ambientales.

La clasificación (sección 4) agrupa las muestras, especies u objetos en conglomerados. La clasificación puede ser jerárquica o no jerárquica (sección 4.2). La clasificación jerárquica se divide en aglomerativa (sección 4.1) y en divisiva (sección 4.3).

Además existen métodos paramétricos para comparar conjuntos de datos (sección 5.1 a 5.3) y métodos no paramétricos (sección 5.4 a 5.6).

En la tabla 4 se muestra la clasificación de los métodos multivariables.

TABLA 4. Clasificación de los métodos multivariables.

Grupo de métodos	Tipo	Método
Ordenación	Indirecta	Ordenación polar Análisis de componentes principales Análisis de coordenadas principales Escalado multidimensional no métrico Análisis de correspondencia con y sin tendencia
	Directa	Análisis de redundancia Análisis de correspondencia canónica con y sin tendencia
Clasificación	No jerárquica	K-medias
	Jerárquica	Análisis de conglomerados jerárquico TWINSPAN
Inferenciales	Paramétricos	Análisis de varianza multivariable Análisis discriminante Análisis de variables canónicas
	No paramétricos	Análisis de redundancia basado en distancia Análisis de similitud Prueba de Mantel

1.2.5 PROGRAMAS

Para hacer cálculos y mostrar los resultados tanto analíticos como gráficos, los investigadores hacen uso de la computadora y de programas estadísticos y/o matemáticos. Los gráficos varían de acuerdo al programa usado. El tipo de programa a elegir depende del método a utilizar y el tipo de información a manejar (Peña, 2000; Dixon, 2003).

Peña (2000) recomienda los siguientes programas:

- STATIGRAPHICS. Permite aplicar las herramientas multivariables básicas y tiene buenas capacidades gráficas. Tiene el inconveniente de ser poco flexible para análisis no estándar de datos.

- MINITAB. Es de fácil manejo. Es más completo que el anterior y más cómodo para la manipulación de datos y la lectura de ficheros en distintos formatos.

- SPSS. Permite mucha flexibilidad en la entrada de datos y en su manipulación, así como en la presentación de los resultados. Además, este programa tiene algoritmos bastante fiables y muy contrastados en distintas aplicaciones.

- MATLAB. Programa excelente para la manipulación matricial, por lo que es recomendable para personas que quieran escribir sus propios

programas y probar análisis nuevos. Tiene la ventaja de la flexibilidad y el inconveniente de que es menos automáticos para análisis tradicionales.

Jongman *et al.* (1995) y Ter Braak y Prentice (1988) recomiendan el programa CANOCO para análisis de datos ecológicos. Puede ejecutar diferentes métodos de ordenación.

Clarke y Gorley (2001) recomiendan el uso del programa PRIMER v5 para análisis de datos ecológicos. Puede ejecutar el análisis de componentes principales, el escalado multidimensional no métrico y el análisis de similitud (ANSIM). Tiene buenos gráficos.

Dixon (2003) recomienda el lenguaje R y el programa VEGAN para datos ecológicos. Puede ejecutar varios métodos de ordenación como el análisis de componentes principales, el análisis de correspondencia con y sin tendencia, el análisis de correspondencia canónica, el análisis de redundancia, el análisis de similitud (ANSIM) y la prueba de Mantel. Tiene la ventaja de poder descargarse gratis de la Internet.

Hill (1979) recomienda TWINSPAN como herramienta para realizar análisis de especies indicadoras por dos vías.

1.3. PLANTEAMIENTO DEL PROBLEMA

La mayoría de los obstáculos encontrados por los ecólogos microbianos cuando se trata de resumir e interpretar conjuntos de datos muy grandes se debe a la elección inadecuada de las herramientas para evaluar datos de manera estadística.

1.4. JUSTIFICACIÓN

Debido a los problemas que enfrentan los ecólogos microbianos, es importante contar con una guía básica para la elección correcta de los métodos multivariables en ecología microbiana.

1.6. OBJETIVOS

1. Describir de los métodos multivariables más usados en Ecología microbiana.
2. Mencionar las ventajas y las limitaciones de los métodos descritos.
3. Mencionar los algoritmos de métodos descritos.
4. Mencionar las aplicaciones de los métodos descritos en Ecología microbiana.
5. Explicar las posibles interpretaciones de los patrones de diversidad microbiana de acuerdo a los objetivos y al método de estudio en particular.

1.7. REFERENCIAS

Atlas, R. M. y R. Bartha (2002). *Ecologia microbiana y Microbiologia ambiental.* 4a. ed.. Pearson Addisson Wesly. España.

Blackwood, C. B., T. Marsh, S. H. Kim y E. A. Paul (2003). *Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities.* Applied and Environmental Microbiology, 69 (2): 926-932.

Campbell, R. (2001). *Ecologia microbiana.* Limusa. México.

Clarke, K. R. y R. N. Gorley (2001). *PRIMER v5: User manual/tutorial.* PRIMER-E Ltd. Reino Unido.

Dixon, P. (2003). *VEGAN, a package of R functions for community ecology.* Computer programs review. Journal of vegetation science. 14: 927-930.

Frankovich, T. A., E. E. Gaiser, J. C. Zieman y A. H. Wachnicka (2006). *Spatial and temporal distribution of epiphytic diatoms growing on Thalassia testudinum Banks ex Konig: relationships to water quality.* Hidrobiologia, 569: 259-271.

Gauch, H. G. (1982). *Multivariate analysis in community ecology.* Cambridge University Press. EUA

Hurst, C. J., R. L. Crawford, G. R. Knusen, M.J. McInerney y L. D. Stetzenbach (2002). *Manual of Environmental Microbiology.* 2a. ed.. ASM Press. EUA.

Jongman, R. G. , C. J. F. Ter Braak y O. F. R. Van Tongeren (1995). *Data analysis in community and landscape ecology.* Cambridge University Press. Reino Unido.

Jonson, D. E. (2000). *Métodos multivariadas aplicados al análisis de datos.* Internacional Thompson Editores. México.

Kaneene, J. B., R. A. Miller, R. Sayah, Y. J. Johnson, D. Gilliland y J. C. Gardiner (2007). *Considerations when using Discriminant Function Analysis of antimicrobial resistance profiles to identify sources of fecal contamination of surface water in Michigan.* Applied and Environmental Microbiology, 73 (9): 2878-2890.

Lanoiselet, V. M., E. J. Cother, N. J. Cother, G. J. Ash y J. D. I. Harper (2005). *Comparision of two total cellular fatty acid analysis protocols to differentiate Rhizoctonia oryzae and R. oryzae sativae.* Mycologia, 97 (1): 77-83.

Mora-Navarro, M. R., J. A. Vazquez-García y Y. L. Vargas-Rodríguez (2004). *Ordenación de comunidades de fitoplancton en el lago de Chapala, Jalisco-Michoacán, México.* Hidrobiológica. 14(2): 91-103

Park, S., Y. K. Ku, M. J. Seob, D. Y. Kim, J. E. Yeon, K. M. Lee, S. C. Jeong, W. K. Yohh, C. H. Hark y H. M. Kim (2006). *Principal component analysis and discriminant analysis (PCA-DA) for discriminating profiles or terminal restriction fragment length polymorphism (T-RFLP) in soil bacterial communities.* Soil Biology Biochemistry, 38: 2344-2349.

Peña, D. (2002). *Análisis de datos multivariantes.* McGraw Hill. España.

Ramette, A. (2007). *Multivariate analyses in microbial ecology.* FEMS Microbial Ecology: 1-19

Ter Braak, C. J. F. y I. C. Prentice (1988). *A theory of gradient analysis.* Advances in ecological research. 18: 271-317

Varese, G. C., P. Gonthier y G. Nicolotti (2003). *Long-term effects on other fungi are studied in biological and Chemicals stump treatments in the fight against Heterobasidion annosum coll.* Mycologia, 95 (3): 379-387.

2. DATOS ECOLÓGICOS Y MÉTODOS DE MUESTREO

2.1. PASOS DE UN PROYECTO DE INVESTIGACIÓN

En un proyecto de investigación (o estudio) ecológico se pueden distinguir varios pasos (Jongman *et al.*, 1995):

1. **Definición del problema de investigación.** Plantear el problema de manera concreta permite tener una idea clara de lo que ocurre y de lo que se va a hacer.

2. **Hipótesis.** Expresa la posible solución del problema o la relación del fenómeno que se desea examinar. Permite darle sentido a la investigación. Se debe recordar que la hipótesis nula expresa la posibilidad de que nada ha ocurrido o que el cambio no ha sido producido por la causa de interés.

3. **Objetivos.** Expresan las acciones a realizar y son la guía para las decisiones a tomar sobre las definiciones de los tratamientos, las variables respuesta y las variables explicatorias, además de los métodos multivariables a usar. También influyen en los aspectos subsecuentes de la colección y análisis de datos.

4. **Diseño del experimento.** Es un esquema que resume el procedimiento a ser ejecutado desde la colección de los datos hasta el análisis de los mismos. Ello implica:

   - Describir los objetos de estudio.
   - El tipo de muestreo estadístico a usar.
   - El proceso de toma y análisis de las muestras y de los datos.
   - Los métodos estadísticos y/o matemáticos a usar.
   - La forma de interpretar los datos.

5. **Ejecución del experimento y análisis de los datos.**

6. **Resultados y conclusiones.**

2.2. TIPOS DE VARIABLES

La clasificación de las variables se muestra en la figura 1 (Peña, 2000).

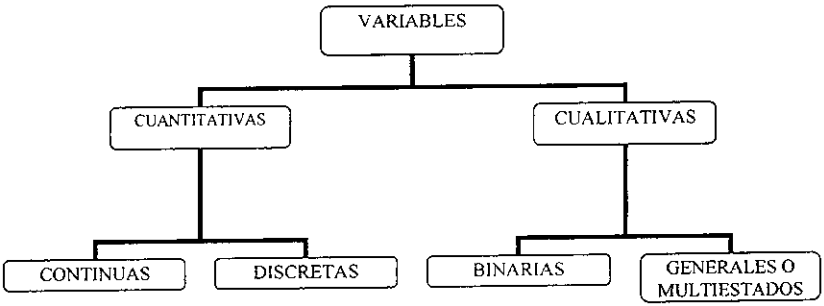

FIGURA 1. Tipos de variables.

Las variables cuantitativas expresan su valor de manera numérica; por ejemplo, la abundancia de las especies o la concentración de un ión. Las variables cuantitativas se dividen en continuas y discretas. Las variables continuas pueden tomar cualquier valor real en un intervalo (por ejemplo, rango de pH) mientras que las variables discretas sólo toman valores enteros (por ejemplo, número de individuos).

Las variables cualitativas expresan atributos o categorías; por ejemplo, tipo de tratamiento. Las variables cualitativas se dividen en binarias o generales (también conocidas como multiestados). Las variables binarias toman sólo dos valores posibles (presencia o ausencia), mientras que las variables multiestados toman varios valores (por ejemplo, zona de estudio).

Las escalas de medida indican los valores que pueden tomar las variables. Existen 4 tipos de escalas de medida (Zar, 1999):

- *Escala nominal.* Las variables bajo estudio son clasificadas por alguna cualidad; por ejemplo, nombre de la zona de estudio.

- *Escala ordinal.* Consiste en ordenar o acomodar las medidas por rangos; es decir, trata con diferencias relativas en vez de diferencias cuantitativas (por ejemplo, A es menos similar que B).
- *Escala de intervalo.* Posee un intervalo constante pero sin un valor cero real; por ejemplo, la escala de temperatura.
- *Escala de razón.* Tiene un tamaño de intervalo constante entre cada unidad sucesiva sobre la escala de medida y existe un punto cero real con significado físico; por ejemplo, volumen.

2.3. ASPECTOS DE LOS DATOS

Por lo general, los datos ecológicos son de naturaleza multivariable. Estos datos presentan las siguientes características (Gauch, 1982; Jongman *et al.*, 1995):

- Ruido. Significa que algunas muestras con condiciones ambientales similares no son idénticas en composición de especies.
- Redundancia. Significa que las muestras son más similares a otras en su composición de especies y muchas especies se parecen a otras en sus ocurrencias en las muestras.
- Datos atípicos. Son observaciones que parecen haberse generado de forma distinta al resto de los datos, bien sea por error de muestreo o bien sea a la heterogeneidad intrínseca de los elementos observados.
- Relaciones internas. Las relaciones entre muestras, entre especies y entre las muestras y las especies consideradas en conjuntos.
- Algo de información sólo es interpretable de manera indirecta.
- Gran cantidad de datos.

2.4. MATRICES DE DATOS

Una matriz es una tabla con filas y columnas cuyos elementos son los datos. Los datos se acomodan en la i-ésima fila y la j-ésima columna. Para más información sobre matrices y álgebra lineal se puede consultar el anexo 6.2.

Por lo general, en ecología se manejan 3 tipos de matrices de acuerdo al tipo de datos implicados (Ludwig y Reynolds, 1988):

- Matriz **Y** de especies x muestras. Tiene las especies en filas y las muestras en columnas. Los datos pueden ser abundancia de las especies o datos de presencia-ausencia.
- Matriz **X** de variables ambientales x muestras. Tiene las variables ambientales en filas y las muestras en columnas. Los datos son cuantitativos.
- Matriz **D** de distancia entre especies o entre muestras. Es una matriz cuadrada y simétrica que muestra la relación entre especies o entre muestras.

También existen matrices que miden la dependencia lineal entre las variables, llamadas matriz de varianza-covarianza ($\boldsymbol{\Sigma}$) y matriz de correlación (**P**) (Peña, 2000; Jonson, 2000). Para más información sobre estas matrices se puede consultar el anexo 6.3.

2.5. TRANSFORMACIÓN DE LOS DATOS

Una transformación es el cambio de los valores de una variables con alguna característica no deseada (como es no linealidad) mediante una relación matemática en otros valores con la característica deseada (Ramette, 2007).

Existen dos grupos de transformaciones: estandarización y transformación para ajustar a un modelo (por ejemplo, normalización).

La estandarización es un tipo de transformación para hacer comparables los datos de una matriz cuyos elementos se expresan en unidades diferentes o en diferentes escalas. Una consecuencia de la transformación es que las variables tienen media cero y la varianza es la unidad, de esta manera, las variables se miden en unidades de desviación estándar (Jongman *et al.*, 1995; Crisci y López-Armengol, 1983). Según Crisci y López-Armengol (1983) tiene el inconveniente de igualar la variación; es decir, una variable con escaso rango de variabilidad una vez transformada tiene la misma variabilidad que una variable de amplio rango.

En la tabla 5 se muestra un resumen de los tipos de estandarización (Jongman *et al.*, 1995; Ramette, 2007).

TABLA 5. Tipos de estandarización.

Estandarización	Descripción	Características
Transformación puntaje z	Para cada variable, se calcula la diferencia entre el valor original y la media de la variable (centrado) y luego la diferencia se divide por la desviación estándar.	▪ Los valores transformados tienen media cero y varianza unidad. ▪ Las variables son medidas en desviación estándar para hacerlas comparables.
Estandarización a la muestra total	Cada abundancia se divide por la suma de las abundancias en una muestra.	▪ Obtiene las abundancias relativas de las especies y hace una corrección al tamaño de la muestra.
Estandarización al total de especies	Se suman las abundancias de cada especie en todas las muestras y luego se divide por el total.	▪ Es recomendable sólo si la frecuencia de las especies no difieren mucho entre sí. ▪ Sobreestima las especies raras y subestima las especies comunes.

La normalización es un tipo de transformación de ajuste a un modelo para corregir la forma de la distribución de ciertas variables, las cuales se alejan de la distribución normal (Ramette, 2007).

En la tabla 6 se muestra un resumen de los tipos de transformaciones de ajuste a un modelo (Ramette, 2007; Jongman *et al.*, 1995; Zar, 1999).

TABLA 6. Transformaciones de ajuste a un modelo(c es una constante para datos con valores cercano o igual a cero y a es una constante mayor que uno).

Transformación	Fórmula	Aplicación
Transformación logarítmica	$x' = \log_{10}(x + c)$	▪ Se aplica a variables con distribuciones log-normal para ajustar a un modelo de distribución normal. ▪ Se aplica para convertir una distribución sesgada positiva en una distribución simétrica. ▪ Se aplica a variables que son multiplicativas para expresarlos en forma aditiva. ▪ Es preferible cuando los valores observados son pequeño o cercanos a cero. ▪ Dar menos peso a especies dominantes.
Transformación raíz cuadrada	$x' = \sqrt{x + c}$	▪ Se aplica cuando las varianzas de los grupos son proporcionales a las medias. ▪ Dar menos peso a las especies dominantes.
Transformación arcoseno	$x' = arcsen\ \sqrt{x + c}$	▪ Se aplica a porcentajes o proporciones.
Transformación cuadrada	$x' = x^2$	▪ Se aplica si la desviación estándar disminuye cuando la media de los grupos aumenta y/o si la distribución es sesgada a la izquierda.
Transformación exponencial	$x' = a^x$	▪ Dar más peso a las especies.
Transformación recíproca	$x' = \frac{1}{x + c}$	▪ Se aplica si las desviaciones estándar de grupos de datos son proporcionales al cuadrado de la media de los grupos.

2.6. SIMILITUD Y DISTANCIA

Los datos pueden asociarse por pares de especies o de muestras mediante una medida de similitud o de distancia y acomodarse en una matriz cuadrada y simétrica.

La similitud es una medida de qué tan semejantes son dos muestras en términos de su composición de especies o bien dos objetos con respecto a sus atributos. Presentan las siguientes características:

- Por lo general, sus valores están en el rango de cero a uno. Un valor igual o cercano a cero indica que las dos muestras son diferentes o distintos, mientras que un valor igual o cercano a uno indica que son similares.
- No presentan la propiedad métrica (no cumplen con la condición del triángulo de inequidad).
- Pueden usarse datos binarios y datos cuantitativos.

Cuando los datos son binarios, la información sobre el grado de asociación entre pares de muestras i y j se pueden mostrar en una tabla de contingencia de 2 x 2 (tabla 7).

TABLA 7. Tabla de contingencia de 2 x 2 de asociación. Tomado de Digby y Kempton (1987).

		UNIDAD j (UM j)		
		PRESENTE	AUSENTE	TOTAL
UNIDAD i (UM i)	PRESENTE	a	b	a+b
	AUSENTE	c	d	c+d
	TOTAL	a + c	b + d	

La tabla indica lo siguiente:

- a es el número de especies presentes en ambas muestras.
- b es el número de especies presentes en la muestra *i* pero ausentes en la muestra *j*.
- c es el número de especies ausentes en la muestra *i* pero presentes en la muestra *j*.
- d es el número de especies ausentes en ambas muestras.

- a + b es el total de especies presentes en la muestra *i*.
- c + d es el total de especies ausentes en la muestra *i*.
- a + c es el total de especies presentes en la muestra *j*.
- b + d es el total de especies ausentes en la muestra *j*.

Los coeficientes de similitud para datos binarios se basan en la tabla 6.

En la tabla 8 se muestra un resumen de los coeficientes de similitud para datos binarios y datos cuantitativos.

TABLA 8. Coeficientes de similitud. El símbolo yk indica la abundancia de la especie k; i y j indican las muestras.

Coeficiente De similiutd	Fórmula	Tipo de Variables	Características
Concordancia Simple	$\frac{a+d}{(a+b+c+d)}$	Binarias	▪ Incluye dobles ceros. ▪ La ocurrencia conjunta y la ausencia conjunta es divida por el total de especies en ambos sitios. ▪ La presencia y ausencia conjunta son tratadas de manera similar.
Jaccard	$\frac{a}{(a+b+c)}$	Binarias	▪ Excluye dobles ceros. ▪ Muestra la proporción de especies comunes en ambas muestras. ▪ Muy sensible a riqueza de especies. ▪ No es sensible a especies dominantes ni al tamaño de la muestra.
Sorensen y Dice	$\frac{2a}{(2a+b+c)}$	Binarias	▪ Excluye dobles ceros. ▪ Muestra la ocurrencia conjunta entre la media aritmética en ambas muestras. ▪ Sensible a riqueza de especies ▪ No es sensible al tamaño de muestra ni a especies dominantes. ▪ El Coeficiente de Dice suele ser utilizado para datos moleculares. ▪ El coeficiente de Sorensen es una medida robusta de similitud para datos ecológicos.
Ochiai	$\frac{a}{\sqrt{(a+b)(a+c)}}$	Binarias	▪ Excluye dobles ceros. ▪ La ocurrencia de la especie en ambas muestras es dividida por la media geométrica en las dos muestras.

TABLA 8. Continuación.

Razón de Similitud	$\frac{\sum_k y_{ki}y_{kj}}{\left(\sum_k y_{ki}^2+\sum_k y_{kj}^2-\sum_k y_{ki}y_{kj}\right)}$	Cuantitativa	▪ Excluye dobles ausencias. ▪ Compara las abundancias conjuntas. ▪ Muy sensible a riqueza de especies, a especies dominantes y al tamaño de la muestra.
Porcentaje de Similitud	$\frac{200*\left[\sum_k \min\left(y_{ki},y_{kj}\right)\right]}{\sum_k y_{ki}+\sum_k y_{kj}}$	Cuantitativa	▪ Excluye dobles ausencias. ▪ La abundancia mínima de la especie en ambas muestras es dividida por la suma de la abundancia de la especie en cada muestra.

Las similitudes pueden ser convertidas en distancias por medio de las siguientes relaciones matemáticas:

- $D = 1 - S$
- $D = (1 - S)^{1/2}$
- $D = \sqrt{(1 - S^2)}$

La distancia es una función matemática que mide qué tan distintos o diferentes son dos objetos con respecto a los atributos observados. Cumplen con al menos las primeras tres condiciones siguientes:

- Positividad. Sólo toman valores positivos.
- Simetría. La distancia entre el punto *i* y *j* es la misma que la distancia entre el punto *j* e *i*.
- Tomar un valor igual a cero si es medida sobre sí misma.
- Desigualdad triangular. La suma de las longitudes de dos lados del triángulo (catetos) formado por los 3 puntos debe ser siempre mayor que el tercer lado (hipótenusa).

Las distancias pueden tener las siguientes propiedades:

- Propiedad métrica. Es cuando una distancia cumple con la condición de desigualdad triangular.
- Propiedad euclídea. Es cuando las distancias pueden representarse como una línea recta entre un conjunto de puntos en un espacio real (eucídeo).

La distancia se divide en tres grupos (Legendre y Legendre, 1998):

- Métricas (cumplen con la desigualdad triangular).
- Semimétricas (no cumplen con la condición de desigualdad triangular).
- No métricas (pueden tomar valores negativos).

En la tabla 9 se muestra un resumen de las medidas de distancia (Legendre y Legendre, 1998; Jongman *et al.*, 1995; Digby y Kempton, 1987; Legendre y Gallagher, 2001; Faith *et al.*, 1987). Por lo general, se usan variables cuantitativas para distancias.

TABLA 9. Medidas de distancia. El símbolo yk indica la abundancia de la especie k; i y j indican las muestras.

Medida de Distancia	Fórmula	Características
Euclídea	$\sqrt{\sum_k \left(y_{ki} - y_{kj}\right)^2}$	▪ Es la distancia más corta entre dos puntos en un plano cartesiano. ▪ Es métrica y euclídea. ▪ Muy sensible a la riqueza de especies. ▪ Sensible a especies dominantes y al tamaño de la muestra. ▪ Depende de la escala de medida.
Ji-cuadrada X^2	$\sqrt{\sum_k \left(\frac{1}{\sum_i y_{ki}}\right)\left[\left(\frac{y_{ki}}{\sum_k y_{ki}}\right) - \left(\frac{y_{kj}}{\sum_k y_{kj}}\right)\right]^2}$	▪ Tiene una parte de la distancia euclídea calculada sobre las abundancias relativas ponderadas por la inversa de la suma de especies. ▪ Es métrica y euclídea. ▪ Sensible a especies raras.
Chord	$\sqrt{2\left(1 - \left(\frac{\left(\sum_k y_{ki} y_{kj}\right)}{\sqrt{\left(\sum_k y_{ki}^2\right)\left(\sum_k y_{kj}^2\right)}}\right)\right)}$	▪ Es la distancia euclídea calculada después de escalar los vectores a longitud 1. ▪ Es Euclídea. ▪ Es una medida poco robusta para datos ecológicos. ▪ Sensible a riqueza de especies y a especies dominantes. ▪ No es sensible al tamaño de la muestra.
Hellinger	$\sqrt{\sum_k \left(\sqrt{\frac{y_{ki}}{\sum_k y_{ki}}} - \sqrt{\frac{y_{kj}}{\sum_k y_{kj}}}\right)^2}$	▪ Es euclídea. ▪ Es útil para métodos de ordenación lineal y para datos de abundancia con ausencias (ceros). ▪ Es una medida poco robusta para datos ecológicos.

TABLA 9. Continuación.

Mahalanobis	$\sqrt{(y_{ki} - y_{kj})' \Sigma^{-1} (y_{ki} - y_{kj})}$	▪ Es una medida de distancia entre variables correlacionadas que estandariza las variables y es invariante ante los cambios de escala. ▪ Es métrica y euclídea.
Manhattan	$\sum_k \lvert y_{ki} - y_{kj} \rvert$	▪ Es la suma de las diferencias absolutas entre dos muestras. ▪ Es la distancia entre dos puntos a lo largo de un ángulo recto. ▪ Es métrica.
Canberra	$\sum_k \frac{\lvert y_{ki} - y_{kj} \rvert}{(y_{ki} + y_{kj})}$	▪ Toma en cuenta la distancia entre dos puntos y su relación con el origen. ▪ Es una medida sesgada. ▪ Sólo toma valores positivos.
Bray-Curtis	$\frac{\sum_k \lvert y_{ki} - y_{kj} \rvert}{\sum_k (y_{ki} + y_{kj})}$	▪ Es semimétrica. ▪ Es una medida robusta para datos ecológicos.

Legendre y Gallagher (2001) mencionan algunas transformaciones sobre los datos de composición de especies que contienen ceros para usarlos en métodos lineales como el análisis de componentes principales (sección 3.4) o el análisis de redundancia (sección3.10) (tabla 10).

TABLA 10. Transformaciones de los datos de especies para métodos lineales. En la tabla, y representa la abundancia de la especie k en el sitio m; el símbolo + indica suma de especies o muestras; n representa el total de fila y p el total de columna. Tomado de Legendre y Gallagher (2001).

Transformación	Fórmula	Características
Transformación Chord	$y'_{km} = \frac{y_{km}}{\sqrt{\sum_k y_{km}^2}}$	La distancia euclídea entre vectores columna de datos transformados es idéntica a la distancia Chord entre vectores columna originales de abundancia de especies.
Transformación χ^2	$y'_{km} = \left(\sqrt{y_{++}}\right)\left(\frac{y_{km}}{y_{+m}\sqrt{y_{k+}}}\right)$	La distancia euclídea entre vectores columna de datos transformados es idéntico a la distancia χ^2 entre los vectores columna originales de abundancia de especies.
Transformación Hellinger	$y'_{km} = \sqrt{\frac{y_{km}}{y_{+m}}}$	La distancia euclídea entre vectores columna de datos transformados es idéntico a la distancia Hellinger entre los vectores columna originales de abundancia de especies.

2.7. MÉTODOS DE MUESTREO

El procedimiento mediante el cual obtenemos una o más muestras recibe el nombre de muestreo. Decimos que el muestreo es probabilístico (dado que en toda muestra existe un error de muestreo) cuando puede calcularse de antemano cuál es la probabilidad de obtener cada uno de los elementos de la muestra. Para esto, es necesario que la selección pueda considerarse como un experimento aleatorio o al azar (Dos Santos-Márquez y Guzmán-Arellano, 1995; Hurst *et al.*, 2002).

Los tipos de muestreo probabilístico son (Dos Santos-Márquez y Guzmán-Arellano, 1995):

- *Aleatorio simple*: todas las unidades de la población tienen la misma probabilidad de ser extraídas.
- *Aleatorio estratificado*: se divide la población en subpoblaciones o estratos dentro de los cuales se hace una selección aleatoria simple.
- *Sistemático*: consiste en tomar lo elementos poblacionales que formarán la muestra, de k en k, a partir de uno de ellos elegido de manera aleatoria entre los que ocupan el primer intervalo de muestreo que resulta de dividir el número total de elementos entre el tamaño de la muestra. Esto es, k=N/n. Luego se escoge al azar un elemento del primer intervalo; los sucesivos están determinados con base en el primero.
- *Conglomerados*: consiste en dividir la población en grupos los cuales se llaman conglomerados. Los conglomerados deben ser seleccionados de manera que: a) las diferencias entre las unidades elementales del mismo grupo lo más pequeño posible, y b) las diferencias entre los grupos sean lo más grandes posibles.

El método de obtención de muestras está determinado por las propiedades fisicoquímicas del ecosistema sometido a estudio, por la abundancia esperada de microorganismos y por los procedimientos de medición y análisis que van a ser llevados a cabo (Atlas y Bartha, 2002).

En la tabla11 se muestran los métodos de toma de muestras de agua y sedimento.

TABLA 11. Muestreo de agua y sedimento.

Tipo de Análisis	Método	Descripción
Bacteriológico	Toma directa mediante frascos de vidrio o de plástico esterilizados	Se toma la muestra de agua de manera directa de una bomba de mano o grifo. También se puede tomar muestra de un cuerpo de agua superficial.
Cualitativo de fitoplancton	Red de fitoplancton	Se arrastra la red cónica a través del agua para obtener un concentrado de fitoplancton. La red tiene una boca ancha que se mantiene abierta por un aro metálico y éste es amarrado a un acuerda por unas bridas, la parte estrecha termina en un colector de plástico o metálico. Se recomienda el uso de redes con abertura de malla de entre 20 a 64 µm.
Cuantitativo de fitoplancton y/o protozoarios	Botella Van Dor	El principio básico de la botella es bajar el cilindro de capacidad conocida a una profundidad requerida y mediante un mensajero (plomo) se cierran. El agua es contenida a presión por lo que el agua de otros niveles no puede entrar mientras el equipo sube a la superficie. Sólo recolecta muestras puntuales.
Sedimento	Bomba	La bomba de vacío tiene un tubo de succión (entrada) y uno de repulsión (salida). El tubo de entrada se introduce en la masa de agua y el de salida se coloca en una red o recipiente receptor. Como el tanque receptor posee un volumen conocido, éste método puede ser útil en estudios cuantitativos de análisis de comunidades bacterianas.

Para el muestreo de suelo pueden tomarse de 100 g a pocos kilogramos. Las muestras se pueden obtener de cada horizonte de suelo o a diferentes profundidades con barrenas, palas o desplantadores. Para suelos agrícolas se pueden muestrear los primeros 25 cm, mientras que para suelos de pastizales se muestrean los primeros 10 cm.

Las muestras de suelo de la rizósfera pueden ser obtenidas al excavar el suelo con todo y plantas con una pala o un desplantador esterilizado. Las raíces y otras partes de la planta deben permanecer intactas en lo posible. Se debe tomar el suelo que se encuentra alrededor de las raíces.

Las muestras de suelo son colocadas en bolsas de plástico delgadas para transportarse a un laboratorio. Los métodos para tomar muestras de aire se resumen en la tabla 12.

TABLA 12. Muestreo de aire.

Método	Descripción
Impactación	Separa las partículas del viento al utilizar la inercia de las partículas para forzar su deposición sobre un medio sólido o semisólido. El proceso de impactación depende del tamaño, densidad y velocidad de la partícula, además de los parámetros físicos del impactor. Es útil para análisis cuantitativos.
Impresión líquida	Es similar a la impactación en que la inercia de las partículas es la principal fuerza pero el medio de colección es líquido. La colección de las partículas del bioaerosol en un medio líquido permite la división de la muestra y la aplicación potencial de varios métodos de análisis. Es útil para análisis cuantitativos.
Filtración	Separa las partículas del aire la hacer pasar el aire por un medio poroso, por lo general una membrana de filtro. La colección de las partículas de pende de sus propiedades físicas y del tamaño del poro del filtro. Puede usarse un extractor de aire. Es útil para análisis cuantitativos.
Gravedad	Se expone un medio agar al ambiente y los microorganismos del aire son colectados por gravedad. Es útil para análisis cualitativos.

Los métodos descritos han sido tomados de Hurst *et al.* (2002), Atlas y Bartha (2002), Baustista-Zuñiga (2004) y López-González *et al.* (2006) y se recomienda consultarlos para una explicación más exhaustiva.

Para estudiar las comunidades microbianas se pueden usar métodos moleculares o métodos basado en cultivo.

Los métodos basados en cultivo toman inoculos de las muestras ambientales y los colocan en un medio de cultivo. Se espera a que las células se reproduzcan y después son analizadas. Los microorganismos pueden ser observados al microscopio, ser identificados y en lo posible cuantificados (en el caso de protozoarios y algas). Sin embargo, muchos microorganismos no pueden ser cultivados (Campbell, 2001; Hurst *et al.*,2002; Atlas y Bartha, 2002).

Los métodos moleculares analizan los ácidos nucleicos o los ácidos grasos de las comunidades de las muestras ambientales y crean perfiles moleculares cuyos picos o bandas pueden ser tomados como unidades taxonómicas operativas (UTO). Algunos métodos que analizan los ácidos grasos de las comunidades son PLFA (*Phospholipid ester-Linked Fatty Acid*) y FAME (*Fatty Acid Methyl Ester*). Algunos métodos que analizan los ácidos nucleicos son LH-PCR (*Length Heterogeneity Polymerase Chain Reaction*), T-RFLP (*Terminal Restriction Fragment Length Polymorphism*) y DGGE (*Denaturing Gradient Gel Electrophoresis*), entre otros (Atlas y Bartha, 2002; Hurst *et al.*, 2002).

2.8. REFERENCIAS

Bautista-Zuñiga, F. (2004). *Técnicas de muestreo para manejadores de recursos naturales.* UNAM. México.

Digby, P. G. N. y R. A. Kempton (1987). *Multivariate analysis of ecological communities.* Chapman and Hall. EUA.

Dos Santos-Márquez, M. M. J. y L. M. Guzmán-Arellano (1995). *Elementos de muestreo.* UNAM FES Zaragoza. México.

Faith, D. P., P. R. Minchin y L. Belbin (1987). *Compositional dissimilarity as a robust measurement of ecological distance.* Vegetation. 69: 57-68

Gauch, H. G. (1982). *Multivariate analysis in community ecology.* Cambridge University Press. EUA

Hurst, C. J., R. L. Crawford, G. R. Knudsen, M.J. McInerney y L. D. Stetzenbach (2002). *Manual of Environmental Microbiology.* 2a. ed. ASM Press. EUA.

Johnson D. E. (2000). *Metodos multivariadas aplicados al análisis de datos.* Internacional Thompson Editores. México.

Jongman, R. G., C. J. F. Ter Braak y O. F. R. Van Tongeren (1995). *Data analysis in community and landscape ecology.* Cambridge University Press. Reino Unido.

Legendre, P. y E. D. Gallagher (2001). *Ecologically meaningful transformations for ordination of species data.* Oecologia. 129: 271-280

López-González, A. R., M. R. Mora-Navarro y R. M. Hernández-Herrera (2006). *Manual de prácticas de laboratorio y campo de ficología.* Centro Universitario de Ciencias Biológicas y Agropecuarias. Universidad de Guadalajara. México.

Ludwig, J. A. y J. F. Reynols (1988). *Statistical ecology: A primer on methods and computing.* Wiley-Interscience publications. EUA.

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw Hill. España.

Ramette, A. (2007). *Multivariate analyses in microbial ecology*. FEMS Microbial Ecology: 1-19.

Zar, J. H. (1999). *Biostatistical analysis*. Prentice Hall. EUA.

3. ORDENACIÓN

3.1. GENERALIDADES SOBRE ORDENACIÓN

La ordenación es el acomodo de un conjunto de muestras (o sitios) a lo largo de ejes en base a datos de composición de especies (Jongman *et al.*, 1995; Ter Braak y Prentice, 1988).

Los objetivos de la ordenación son:

- Acomodar las muestras representadas por puntos de manera que los puntos más cercanos son similares en composición de especies y los más alejados son disimilares en composición de especies (Jongman *et al.*, 1995; Ludwig y Reynolds, 1988).
- Ayudar a generar hipótesis sobre las relaciones entre la composición de especies de las muestras y los factores ambientales subyacentes (Digby y Kempton, 1987).
- Reducir la dimensionalidad de los datos (Digby y Kempton, 1987; Jongman *et al.*, 1995).

La ordenación está relacionada con el análisis de gradientes, la cual se refiere a un conjunto de métodos de análisis de datos que relacionan la composición de la comunidad en términos de la respuesta de las especies a los gradientes ambientales (Jongman *et al.*, 1995; Ter Braak y Prentice, 1988).

El análisis de gradientes se divide en dos tipos: análisis de gradientes indirecto (también llamado ordenación indirecta o no restringida) y análisis de gradientes directo (también conocido como ordenación canónica o restringida) (Jongman *et al.*, 1995; Ter Braak y Prentice, 1988; Digby y Kempton, 1987; Ludwig y Reynolds, 1988).

En el análisis de gradientes indirecto primero se registran las abundancias (u ocurrencias) de las especies de un conjunto de muestras o sitios y luego se busca los patrones de variación explicados por una o varias variables latentes (o variables ambientales hipotéticas), las cuales son los ejes de ordenación construidos sin referencia a variables ambientales observadas. Después las variables latentes o ejes de ordenación son

relacionados con los datos de variables ambientales mediante una regresión lineal (Jongman *et al.*, 1995; Ter Braak y Prentice, 1988; Ludwig y Reynolds, 1988). Se aplican estos métodos cuando se desea conocer los patrones de variación de las especies y no se tienen disponibles, al menos por el momento, los datos de variables ambientales o bien se desea reducir la dimensionalidad de los datos (Digby y Kempton, 1987). Un problema que enfrenta el análisis de gradientes indirecto es que se tiene que inferir acerca de las relaciones con la variable latente sólo a partir de los datos de especies (Jongman *et al.*, 1995). Algunos métodos de ordenación indirecta son la ordenación polar (sección 3.3.), el análisis de componentes principales (sección 3.4.), el análisis de coordenadas principales (sección 3.5.), el escalado multidimensional no métrico (sección 3.6.), el análisis de correspondencias con tendencia (sección 3.7.) y sin tendencia o *detrended* (sección 3.8.).

En el análisis de gradientes directo se relaciona en un solo paso los datos de las especies con los datos de variables ambientales de manera que los ejes de ordenación resultantes son una combinación lineal de las variables ambientales. Se aplican estos métodos sólo cuando se tienen los datos de variables ambientales (Jongman *et al.*, 1995; Ter Braak y Prentice, 1988). Algunos métodos de ordenación canónica son el análisis de redundancia (sección 3.9.), el análisis de correspondencia canónica con y sin tendencia (sección 3.10.).

Cuando se quiere resumir la variación de la comunidad que permanece después de que los efectos de las variables ambientales han sido removidos se puede utilizar la ordenación parcial, la cual es aplicación de un método de ordenación canónica (por ejemplo, análisis de redundancia) seguido de un método de ordenación indirecta (por ejemplo, análisis de componentes principales) (Jongman *et al.*, 1995; Ter Braak y Prentice, 1988).

Ahora, cuando se tiene un conjunto de variables ambientales (explicatorias) que pueden ser divididas en dos subconjuntos, uno de variables cuyos efectos son de interés y otro de covariables (variables que no son el objetivo primario del estudio), se puede usar la ordenación

canónica parcial para relacionar la variación residual a otras variables ambientales. La ordenación canónica parcial consiste en la aplicación de una regresión seguida de un método de ordenación canónica (Jongman *et al.*, 1995; Ter Braak y Prentice, 1988).

En los métodos de ordenación indirecta la interpretación dependerá si se enfoca bien sea en las relaciones de similitud o distancia entre muestras (modo Q) o bien en la correlación entre especies (modo R) (Ramette, 2007). El modo Q se calcula multiplicando la transpuesta de la matriz de datos de especies (**Y'**) por la matriz de datos de especies (**Y**). El modo R se calcula multiplicando la matriz de especies (**Y**) por su transpuesta (**Y'**). (Ludwig y Reynolds, 1988).

Un aspecto importante en los métodos de ordenación es el tipo de modelo de respuesta que tienen las especies con respecto a las variables ambientales. Algunos métodos como el análisis de componentes principales (ACP), el análisis de coordenadas principales (ACoP) y el análisis de redundancia (ADR) presentan un modelo de respuesta lineal, en donde la abundancia de cualquier especie aumenta o disminuye con el valor de cada variable ambiental o latente; mientras que otros métodos como el análisis de correspondencias (AC) y el análisis de correspondencias canónica (ACC) presentan un modelo de respuesta unimodal, en donde los valores esperados de las especies (variables respuesta) aumentan con la variable ambiental, alcanzan un máximo (óptimo) y luego disminuyen. (Jongman *et al.*, 1995). Algunos métodos como la ordenación polar (OP) y el escalado multidimensional no métrico (EMNM) no presentan esas características.

Los investigadores que simulan y estudian los datos multivariables ecológicos miden la longitud de los gradientes como la diferencia en la composición de especies en las muestras, diversidad beta, en unidades de HC (*Half change*). Un HC es definido como la separación en la cual las muestras son 50% similares en composición (Gauch y Wenworth, 1976). Los gradientes ambientales también pueden ser medidos en unidades de desviación estándar mediante un método como el análisis de correspondencia sin tendencia. (Jongman *et al.*, 1995; Palmer, 2006). Ramette (2007) recomienda utilizar métodos lineales cuando la longitud del

gradiente es menor a 3 desviaciones estándar, usar métodos unimodales cuando la longitud del gradiente es mayor a 4 desviaciones estándar y usar cualquier método para gradientes intermedios.

3.2. DIAGRAMAS DE ORDENACIÓN

Por lo general, los resultados de la ordenación son representados en diagramas. Estos diagramas pueden ser de dispersión, diagrama conjunto (*joint plot*) o *biplot*.

Un diagrama de dispersión representa las muestras, sitios, especies o UTO (unidad taxonómica operativa) como puntos en un espacio bidimensional (y en ocasiones tridimensional) de manera que los puntos más cercanos son similares en sus atributos con respecto a una variable latente o ambiental.

El *biplot* , cuyo prefijo "bi" se refiere a la representación conjunta de las muestras y las especies, se interpreta de la siguiente manera (Jongman *et al.*, 1995; Ter Braak y Prentice, 1988):

1. Los sitios son representados por puntos y las especies son representadas por vectores.
2. Los vectores especies apuntan en la dirección de máxima variación en la abundancia de las especies y su longitud es proporcional a la máxima razón de cambio.
3. La abundancia de una especie en las muestras se puede interpretar mediante la proyección perpendicular de las muestras sobre el vector especie de manera que si la proyección se encuentra entre el origen (del diagrama) y la cabeza del vector entonces hay una abundancia alta, mientras que si el origen se encuentra entre la cabeza del vector especie y la proyección de la muestra entonces hay una abundancia baja en esa muestra.
4. El ángulo entre vectores de cada par de especies indica una aproximación de su correlación; es decir, una mediada de su dependencia lineal.

5. Las especies en el borde del diagrama (alejados del origen) indican diferencias en los sitios.
6. Las variables ambientales son representadas por vectores cuya dirección indica la razón de cambio y su longitud indica su importancia relativa.
7. El ángulo entre vectores de variables ambientales, y entre vector de especie y vector variable ambiental se interpretan de manera similar al ángulo entre vectores de especies.
8. El ángulo y la longitud de las flechas indican la dirección y la fuerza de la relación de las variables fisicoquímicas.

El diagrama conjunto (*joint plot*) re refiere a la representación conjunta de especies y muestras como puntos. Se interpreta de la siguiente manera:

1. Cada sitio es localizado en el centro de gravedad (**centroide**) de las especies que ahí ocurren.
2. Los puntos especies son el óptimo de las mismas.
3. La abundancia o probabilidad de ocurrencia de una especie tiende a disminuir con la distancia de su localización en el diagrama.
4. Las variables ambientales, si están presentes, se representan como vectores.
5. Los puntos en el borde del diagrama con frecuencia son especies raras.
6. Los vectores indican la dirección y razón de cambio a través del espacio dimensional.

3.3. ORDENACIÓN POLAR (OP)

3.3.1. DESCRIPCIÓN

También es llamado ordenación de Bray-Curtis (Gauch, 1982; Ludwig y Reynolds, 1988). Es un método basado en distancia que usa un algoritmo geométrico el cuál toma dos muestras como polos para indicar algún gradiente (Gauch, 1982; Ludwig y Reynolds, 1988; Palmer, 2006; Poole, 1974). Por lo general, usa el coeficiente de Sorensen aunque puede usar cualquier coeficiente de distancia o de similitud.

El procedimiento implica la selección de muestras como puntos-extremo (polos) sobre un eje seguido de un simple posicionamiento geométrico de las muestras restantes con respecto a las muestras puntos-extremos, de manera que las distancias entre muestras reflejen su similitud y su relación a gradientes ambientales (Poole, 1974; Gauch, 1982).

Existen 4 métodos para la selección de puntos-extremo: original, regresión de varianza, menor desviación y subjetivo. En el método original: se seleccionan los objetos más diferentes, el primer punto-extremo es el que tenga la mayor suma de distancias, el segundo punto-extremo es el que tenga la mayor distancia al primer punto-extremo, pero tiende a seleccionar datos atípicos. El método de regresión de varianza: el primer punto-extremo es el de mayor varianza de distancia a otros puntos, el segundo punto-extremo es el que minimiza el coeficiente de regresión en la regresión de las distancias entre el punto y los demás puntos y un punto prueba y los demás puntos, además de evitar datos atípicos. Cuando se usa ordenación polar se sugiere lo siguiente:

1. Analizar las razones por las que se va a utilizar este método.
2. Las muestras que sirvieron como puntos-extremo.
3. La técnica utilizada para determinar los puntos-extremo.
4. El tipo de distancia utilizada.
5. El tipo de datos usados (cuantitativos o cualitativos).
6. El tipo de transformación aplicada sobre los datos.

Los resultados pueden representarse en un diagrama de ordenación. Su interpretación es similar al diagrama conjunto (sección 3.2).

3.3.2. VENTAJAS

- Puede usar datos cuantitativos y cualitativos.
- Puede usar cualquier medida de distancia o de similitud.
- Es eficaz para la detección de gradientes ecológicos de alta heterogeneidad (Mora-Navarro et al., 2004) si se utiliza la técnica de regresión de varianza.
- Ludwig y Reynolds (1988) mencionan que la ordenación polar evita curvilinealidades causados por relaciones polinomiales del segundo eje con respecto al primer eje. Además, dado que los investigadores tienen un conocimiento a priori acerca de la existencia de gradientes ambientales, la selección de puntos-extremo se puede hacer de manera subjetiva siempre y cuando no se elijan datos atípicos.

3.3.3. LIMITACIONES

- El segundo eje puede ser oblicuo con respecto al primer eje; es decir, los ejes pueden no ser ortogonales (independientes) entre sí.
- Si se seleccionan datos atípicos como puntos-extremo puede provocar que las muestras se amontonen en el eje, lo que llevaría a una mala interpretación.

3.3.4. ALGORITMO

1) Calcular la matriz de distancias D, usando un coeficiente de distancia.
2) Calcular la suma de cuadrados de diferencias de las distancias SS_T:

$$SS_T = \sum\sum D_{ij}^2$$

3) Seleccionar los puntos extremo mediante alguno de los 4 métodos para la selección de puntos extremo (original, regresión de varianza, menor desviación y subjetivo). Calcular las posiciones de los demás objetos mediante $y_{ij} = \frac{\left(D_{AB}^2 + D_{AI}^2 - D_{BI}^2\right)}{2D_{BA}}$

4) Calcular la matriz de distancias residuales D_R mediante $D_R = \sqrt{D_{ij}^2 - \sum (y_{fi} - y_{fh})^2}$. Calcular la suma de cuadrados residual mediante $SS_R = \sum\sum D_{Rij}^2$.

5) Calcular la varianza representada por cada eje mediante $\%Variación = 100 * \left(\frac{1 - SS_R}{SS_T} \right)$.

6) Sustituir la matriz D_R por la matriz D para calcular ejes sucesivos.

7) Repetir los pasos 3-6.

8) Para relacionar los ejes con variables ambientales, se realiza una regresión lineal entre los puntajes de los ejes y las variables ambientales. Además, se obtienen los coeficientes de correlación.

3.3.5. EJEMPLO

Ejemplo A. Mora-Navarro *et al.* (2004) ordenaron las comunidades de fitoplancton del Lago de Chapala, Jalisco-Michoacán. Su principal objetivo fue investigar las variables fisicoquímicas que explican la composición del fitoplancton.

Realizaron arrastres con una red para fitoplancton en 16 estaciones y tomaron muestras. También midieron un total de 20 variables ambientales, mismas que fueron estandarizadas. Obtuvieron un total de 96 muestras y se identificaron 116 especies. Con estos datos crearon una matriz de datos de especies x muestras y otra matriz de variables x muestras.

Usaron la ordenación de Bray-Curtis con la técnica de regresión de varianza y emplearon el coeficiente de Sorensen para ordenar las comunidades de fitoplancton y los factores fisicoquímicos y se relacionaron mediante regresiones lineales con las variables ambientales medidas.

Los resultados de la ordenación se muestran en el diagrama de ordenación (figura 2). Los 3 ejes explicaron el 38, 13 y 9 % de la variabilidad de los datos. Para el eje 1, los extremos asignados fueron la muestra 92 y la muestra 12. La variación en este eje fue explicada por la concentración de sulfatos y de manera inversa por la alcalinidad total.

Para el eje 2, los extremos asignados fueron la muestra 59 y la muestra 16. La variación de este eje fue explicada por la dureza total y por la dureza debida al calcio, ambas relacionadas entre sí.

Para el eje 3 (no mostrado), los extremos asignados fueron la muestra 85 y la muestra 64. La variación del eje 3 no fue explicada por alguna de las variables medidas.

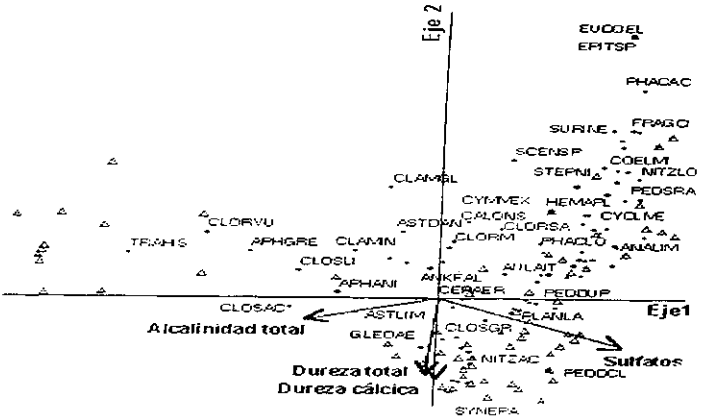

FIGURA 2. Ordenación de Bray-Curtis regresión de varianza de comunidades de fitoplancton en el lago de Chapala, con base en presencia ausencia de especies. El ánguio y la longitud de los vectores indican la dirección y la fuerza de la relación de las variables fisicoquímicas. Los triángulos muestran la ubicación de los sitios. Los puntos (círculos sólidos) muestran la posición óptima de las especies en los ejes de ordenación. Tomado de Mora-Navarro *et al.* (2004).

El diagrama se pudo interpretar de la siguiente manera:

- Los puntos representaron las especies. Los puntos más cercanos corresponden a especies que tienen una ocurrencia similar con respecto a las variables ambientales.
- Los vectores representaron variables ambientales. La dirección entre vectores correspondió a una medida de su correlación. De esta manera, los vectores que apuntaban en la misma dirección tuvieron una correlación positiva, mientras que los vectores que apuntaban en direcciones opuestas tuvieron una correlación negativa.

- El primer eje representó un gradiente que va desde valores altos de sulfatos y bajos de alcalinidad total hasta valores bajos en sulfatos y altos en alcalinidad total. De hecho, los vectores que representan las variables sulfatos y alcalinidad total apuntan en direcciones opuestas lo que indica una correlación negativa entre sí.
- El segundo eje representa un gradiente que va desde valores altos de dureza total hacia valores bajos.
- La localización de los sitios y las especies a lo largo de estos gradientes mostraron sus preferencias ambientales. Por ejemplo, las especies *Pediastrum duplex, Phacus pleuronectes, Phormidium fragile* y *Strombomonas costata* se encontraron en los niveles más altos de sulfatos y más bajos de alcalinidad total.

3.4. ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

3.4.1. DESCRIPCIÓN

Es un método exploratorio de ordenación indirecta basado en un modelo de respuesta lineal que sólo usa la distancia euclídea.

El análisis de componentes principales construye una variable teórica (variable latente) que minimiza la suma de cuadrados residual después de ajustar una línea recta a los datos de especies (Jongman et al., 1995); es decir, se minimizan las distancias entre los puntos originales y sus proyecciones perpendiculares sobre la recta que pasa cerca de todos los puntos (Figura 3) (Peña, 2000).

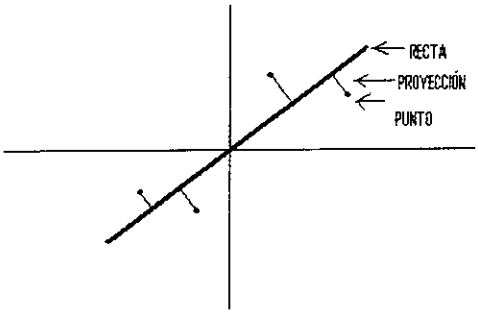

FIGURA 3. Ejemplo de recta que minimiza las distancias ortogonales de los puntos.

Los objetivos son (Jonson, 2000):

- Reducir la dimensionalida de los datos o bien descubrir la verdadera dimensionalidad de los datos.
- Identificar nuevas variables.
- Encontrar los patrones de variación de las variables.

Comprende un procedimiento matemático que usa una matriz de varianza-covarianza (una matriz de correlación cuando las variables son medidas en escalas diferentes) para calcular los valores y vectores propios (éstos son escalados) de esa matriz. Los valores propios indican la cantidad de variación de los datos y los vectores propios escalados son los respectivos componentes principales. Los valores propios son extraídos en

orden descendiente de tal forma que los correspondientes ejes (componentes principales) representan la cantidad de variación de mayor a menor. Los vectores propios proveen de los coeficientes para cada variable en la combinación lineal que define al componente principal (Peña, 2000; Jonson, 2000).

El método calcula las cargas de los componentes, las cuales son las correlaciones entre los puntajes de los componentes y las variables originales. Altas cargas indican que una variable tiene una alta correlación con un componente. Además, si se usa el ACP estandarizado o centrado entonces las cargas son provistas por los vectores escalados.

Existe otro algoritmo para calcular los componentes principales llamado suma ponderada de dos vías, el cual utiliza una matriz **Y** de especies x muestras. En este algoritmo iterativo, los puntajes de especies son sumas ponderadas de los puntajes de muestras y los puntajes de muestras son sumas ponderadas de los puntajes de especies. (Jongman *et al.*, 1995).

Los resultados del método pueden graficarse en un biplot (Peña, 2000; Jongman *et al.*, 1995) y su interpretación se explica en la sección 3.2. En el diagrama de ordenación los puntos más cercanos son similares en sus atributos. Además, los puntajes de especies positivos significa que aumenta la abundancia de la especie a lo largo del eje, mientras que los puntajes negativos indican que la abundancia de la especie disminuye (Jongman *et al.*, 1995). La posición de los puntos puede indicar la respuesta de los objetos o especies con respecto a la variable latente.

3.4.2. VENTAJAS

- Los componentes principales son combinaciones lineales de las variables originales, de esta manera se reduce la dimensionalidad.
- Los componentes pueden ser usados por otros métodos multivariables (Hurst *et al.*, 2002).

3.4.3. LIMITACIONES

- Para datos ecológicos presenta las restricciones de ser un método lineal y de usar sólo la distancia euclídea.
- Según Kessel y Whitaker (1976) aún a baja diversidad beta la ordenación de ACP es más vulnerable a la distorsión de los datos que la ordenación de Bray-Curtis.

3.4.4. ALGORITMO

Algoritmo basado en un análisis propio (Johnson, 2000; Ludwig y Reynolds, 1987):

1. Se computa una matriz de varianzas-covarianza (**Σ**) o de correlación (**P**) a partir de la matriz original de datos de especies x muestras (**Y**).
2. Se obtiene la traza de la matriz simétrica.
3. Se obtienen los valores propios de la matriz simétrica mediante $|\Sigma - \Lambda \mathbf{I}| = 0$.
4. Se obtiene el porcentaje de variación total de cada valor propio λ mediante $100\left(\frac{\lambda_i}{tr\Sigma}\right)$.
5. Se obtienen los vectores propios **u** (normalizados **u'u** =1) mediante $\Sigma \mathbf{u} = \lambda \mathbf{u}$.
6. Se escalan los vectores propios para obtener las coordenadas mediante $\mathbf{z} = \mathbf{u}\sqrt{\lambda}$.
7. Se organizan los datos en un diagrama de ordenación (biplot).

Algoritmo de Suma ponderada de dos vías (Jongman *et al*, 1995):

1. Se toma un puntaje de sitios (y_m) de manera arbitraria diferente de cero.
2. Se calculan nuevos puntajes de sitios de especies (y_k) por suma ponderada de los puntajes de sitios mediante $y_k = \sum (y_{km})(y_m)$.
3. Se calculan nuevos puntajes de sitios (y_m) por suma ponderada de los puntajes de las especies mediante $y_m = \sum (y_{km})(y_k)$.
4. Para el primer eje se va al paso 5. Para el segundo y posteriores ejes, se hacen los puntajes de sitios (y_m) incorrelacionados con el eje previo por el procedimiento de ortogonalización. Para ello
   - Se denota el puntaje de sitios del eje previo por f_m y el puntaje de prueba del presente eje por y_m.
   - Se calcula $v = \sum (y_m)(f_m)$.
   - Se calcula $y_{m\ nueva} = (y_{m\ vieja}) - (v)(f_m)$.
   - Se repiten los pasos a-c para todos los ejes previos.
5. Se estandarizan los puntajes de sitios (y_m). Para ello:
   - Se calcula la suma de cuadrados de los puntajes de sitios mediante $s^2 = \sum y_m^2$.
   - Se calcula $y_{m\ nuevo} = \frac{(y_{m\ viejo})}{s}$
6. Se detiene en la convergencia; esto es, cuando los nuevos puntajes de sitios son suficientemente cercanos a los puntajes de sitios del ciclo previo de la iteración; sino se va al paso 2.
7. Se organizan los datos en un diagrama de ordenación (biplot).

3.4.5. EJEMPLO

Ejemplo A. Ariyadej *et al.* (2004) estudiaron la diversidad de fitoplancton y su relación con los factores fisicoquímicos en la Reserva Bangland, Tailandia. Sus objetivos fueron determinar la diversidad de fitoplancton y la calidad del agua, además de estudiar los efectos de los factores fisicoquímicos sobre la densidad de fitoplancton.

Colectaron muestras de agua cada mes durante el periodo de mayo de 2000 a abril de 2001 en tres estaciones (zona de transición, zona lacustre y zona de flujo) a 3 profundidades diferentes (0, 10 y 30 m). Midieron un total de 10 parámetros fisicoquímicos, los cuales fueron estandarizados. Se identificó un total de 135 especies de algas y se estimó su abundancia como el número de células/mm^3.

Usaron el análisis de componentes principales para identificar las especies más dominantes en cada estación. Los resultados de la ordenación se muestran en el diagrama (figura 4). Los dos ejes explicaron el 67 % de la variación en la superficie en la zona de transición (ZT), el 67 % de la variación en la zona lacustre (ZL) y 54 % en la zona de flujo (ZF). El diagrama muestra las 20 especies más dominantes. Se encontró que *Cyclotella meneghiana* y *Melosira varians* fueron las especies más abundantes.

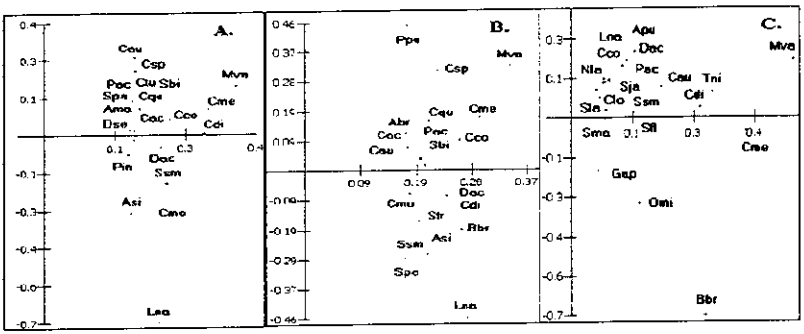

FIGURA 4. ACP de la abundancia de especies de fitoplancton en la Reserva de Bangland. Solo las primeras 20 especies dominantes fueron contadas. A es la zona de transición, B es la zona lacustre y C es la zona de flujo. Tomado de Ariyadej *et al.* (2004).

El diagrama se interpretó de la siguiente manera:

- Los puntos representaron la abundancia de las especies, de manera que las especies más cercanas son similares en abundancia.
- El eje 1 es una variable latente en donde las especies se acomodan de menor a mayor abundancia.

Ejemplo B. Park *et al.* (2006) compararon las comunidades bacterianas del suelo asociadas con la rizosfera de sandía transgénica y no transgénica mediante un análisis T-RFLP (*Terminal Restriction Fragment Length Polymorphism*) en el cual usaron las enzimas de restricción *Hae III* y *Hha I* para evaluar el impacto de un cultivo transgénico sobre el suelo. Sus objetivos fueron establecer un procedimiento en el cual el análisis de componentes principales (ACP) y el análisis discriminante (AD) fuese usado par discriminar perfiles T-RFLP obtenidos de diferentes comunidades bacterianas, además de aplicar el procedimiento para comparar comunidades bacterianas del suelo asociadas con la rizosfera de sandía.

Plantaron sandía transgénica y no transgénica (tratamientos) en dos parcelas cada uno. Después de un mes se tomaron muestras de suelo de cada parcela y se extrajo el ADN del suelo, el cual fue sometido a un análisis T-RFLP. Durante el análisis se usaron las enzimas de restricción *Hae III* y *Hha I*. De los perfiles T-RFLP se identificaron picos T-RF, los cuales fueron alineados y compilados para producir matrices de datos con 112 observaciones y 113 variables picos para cada enzima de restricción.

Usaron el ACP sobre una matriz de covarianzas con datos de los perfiles para cada enzima de restricción con el fin de reducir la dimensionalidad de los datos. Los dos componentes principales para el conjunto de datos *Hae III* explicaron el 52.6 % de la variabilidad de los datos, mientras para el conjunto de datos y *Hha I* los dos componentes principales explicaron el 59.8% de la variabilidad. Ambos diagramas mostraron que no hubo una separación significativa entre los puntajes ACP de perfiles T-RFLP asociados con sandía transgénica y no transgénica (figura 5).

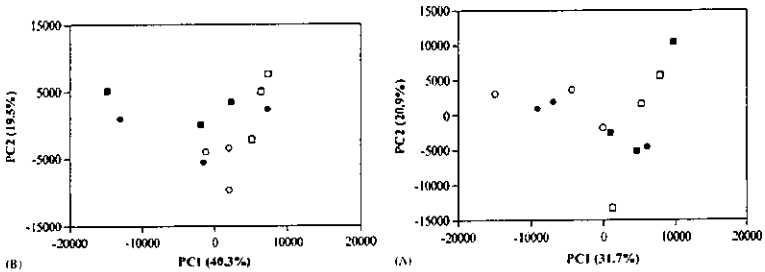

FIGURA 5. Diagrama ACP de perfiles T-RFLP de comunidades bacterianas de la rizósfera con el la enzima HaeIII (A) y con la enzima HhaI (B). Solo el primer componente principal (PC1) y el segundo componente principal (PC2) fueron mostrados. Los círculos cerrados indican perfiles T-RFLP de sandía no transgénica del cuadrante 1 mientras los cuadrados cerrados indican sandía no transgénica del cuadrante 2. Los círculos abiertos indican perfiles T-RFLP de sandía transgénica del cuadrante 1 mientras que los cuadros abiertos indican sandía transgénica del cuadrante 2. Tomado de Park *et al.* (2006).

El diagrama se pudo interpretar de la siguiente manera:

- Los puntos más cercanos entre sí indicaron perfiles T-RFLP más similares en la composición de las comunidades bacterianas del suelo.
- Los ejes son variables latentes que explicaron la estructura de los datos, de esa manera se reduce la dimensionalidad y pudo ser usado por otros métodos como el análisis discriminante.

3.5. ANÁLISIS DE COORDENADAS PRINCIPALES (ACoP)

3.5.1. DESCRIPCIÓN

También se conoce como escalado multidimensional no métrico.

Es un método exploratorio de ordenación indirecta basado en un modelo de respuesta lineal que puede usar cualquier medida de distancia o de similitud (Jongman et al., 1995; Quin y Keough, 2002).

Los objetivos son:

- Reducir la dimensionalidad de los datos.
- Encontrar los patrones de variación de los datos.

El análisis de coordenadas principales (ACoP) transforma una matriz de distancia en una matriz de similitud. Luego, se calculan los valores y vectores propios (escalados) para obtener las coordenadas principales (ejes) (Legendre y Anderson, 1999; Digby y Kemtpon, 1987).

Las coordenadas principales son funciones de las variables originales de acuerdo a la medida de distancia seleccionada, en vez de ser combinaciones lineales de las variables originales como en el análisis de componentes principales (Ramette, 2007).

En el ACoP, cuando se usan distancias métricas los ejes pueden ser representados en un espacio euclídeo y los valores propios son positivos. Sin embargo, cuando se usan medidas de distancia semimétricas o no métricas surgen valores propios negativos. Los valores propios negativos corresponden a la variación en la matriz de distancia que no puede ser representada en el espacio eucídeo. Para solucionar este problema, la matriz de distancia original puede ser ajustada para corregir la porción no euclídea. Existen dos métodos para lograr ello (Legendre y Anderson, 1999; Legendre y Legendre, 1998):

1. Una constante, c_1, puede ser usada par corregir las distancias cuadradas, dando nuevas distancias $d_{ij}' = d_{ij}^2 + 2c_1$ para i ≠

j. el valor c_1 es igual al valor absoluto del valor propio negativos más alto de la matriz Δ.

2. Una constante, c_2, puede ser agregado a todos los términos d_{ij} de la matriz D dando nuevas distancias $d_{ij}' = d_{ij} + c_2$ para i ≠ j. el valor de c_2 es igual a el valor propio más alto de la matriz asimétrica. La constante c_1 es el valor más pequeño que producirá coordenadas euclídeas; cualquier valor más alto que c_1 también eliminará todos los valores propios negativos y hará el sistema completamente euclídeo. La matriz asimétrica es $\begin{bmatrix} 0 & 2\Delta \\ -1 & -4\Delta \end{bmatrix}$. Donde Δ_2 es definido por Δ pero con elementos -1/2 d_{ij} en lugar de -1/2 d^2_{ij}. La constante c_2 es el valor más pequeño que producirá coordenadas euclídeas; cualquier valor más grande que c_2 eliminará también todos los valores propios negativos y hará el sistema completamente euclídeo.

Según Quin y Keough (2002), si se usa la distancia eucídea, las posiciones relativas de los objetos en ACoP serán similares a aquellos del diagrama de ordenación de ACP centrado sobre una matriz de covarianzas. En cambio, si los datos originales sufrieron una doble transformación por filas o columnas totales de manera que la distancia ji-cuadrada fuese usada para crear la matriz de distancias entonces las posiciones relativas de los objetos en ACoP serán similar a la ordenación de análisis de correspondencia.

3.5.2. VENTAJAS

- Se permite usar cualquier medida de distancia.
- También puede usarse las coordenadas principales para otros métodos multivariables.
- Reduce la dimensionalidad de los datos.

3.5.3. LIMITACIONES

- Pueden existir valores propios negativos.
- Sólo es útil para gradientes cortos, dado que es un método lineal.

3.5.4. ALGORITMO

1. La matriz inicial debe ser una matriz de distancia D. Es posible llevar a cabo los cálculos sobre una matriz de similitud S. Es mejor, de cualquier manera, primero transformar la matriz S en una matriz D.
2. La matriz D es transformada en una nueva matriz A definida por $a_{ij} = -\left(1_2\right)D_{ij}^2$.
3. La matriz A es centrada para dar una matriz $\Delta = (\delta_{ij})$ mediante $\delta_{ij} = a_{ij} - a_{i+} - a_{+j} + a_{++}$. Donde a_{i+} y a_{+j} son las medias de las filas y las columnas correspondientes a los elementos de a_{ij} de la matriz A, respectivamente, mientras a_{++} es la media de todos los a_{++} en la matriz. Este centrado tiene efecto de posicionar el origen del nuevo sistema de ejes en el centroide de la dispersión de los objetos, sin alterar la distancia entre los objetos. Desde que la suma de las filas y las columnas de Δ es nulo, Δ tiene al menos un valor propio nulo.
4. Se calculan los valores propios y vectores propios. Luego, los vectores propios son escalados a longitudes iguales a la raíz cuadrada de los respectivos valores propios mediante $\overline{u_k' u_k} = \lambda_k$.
5. Después del escalado, si los vectores propios son representados por columnas, las filas de la tabla resultante son las coordenadas de los objetos en el espacio de coordenadas principales, sin cualquier otra transformación.
6. La matriz de distancias puede ser ajustada para corregir la porción no euclídea para evitar los valores propios negativos, de manera que los datos puedan ser representados en un espacio euclídeo mediante dos métodos de corrección señalados en la descripción del método.

7. Se ordenan los datos en un diagrama de ordenación.

3.5.5. EJEMPLO

Ejemplo A. Branco *et al.* (2001) investigaron la distribución ecológica de las cianobacterias en ecosistemas lóticos del Estado de Sao Paulo. Sus objetivos fueron examinar los patrones de distribución de las cianobacterias en diferentes regiones e investigar los factores ambientales que influyen en la abundancia de las mismas. Tomaron muestras de las comunidades fitobentónicas de 172 ríos de 6 regiones diferentes localizados en el Estado de Sao Paulo, Brasil; además midieron 6 factores ambientales (saturación de oxígeno, turbidez, pH, velocidad, conductancia y porcentaje de cobertura), mismas que fueron estandarizadas. Se registraron 34 especies de cianobacterias y se midió el porcentaje de cobertura como una medida de la abundancia, la cual fue transformada por medio de la transformación arcoseno.

Usaron el ACoP para agrupar especies con preferencias ambientales similares para lo cual usaron el coeficiente de correlación. Los resultados de la ordenación fueron representadas en un diagrama. El dos ejes explicaron el 59.2 y el 23.8% de la variabilidad de los datos, respectivamente. La variación en el primer eje fue explicada por la conductancia y la saturación de oxígeno y de manera inversa por la turbidez. El segundo eje fue explicado por la velocidad.

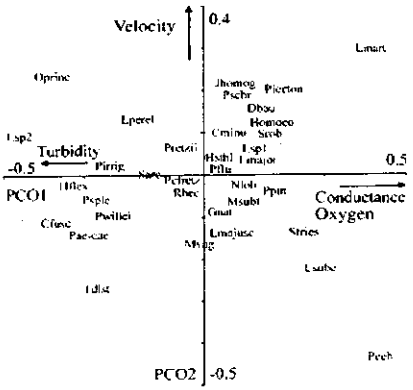

FIGURA 6. Análisis de coordenadas principales para especies agrupadas en base a los parámetros cuantitativos. Tomado de Branco *et al.* (2001).

Se interpretó de la siguiente manera:

- Los puntos representaron la ocurrencia de especies y por consiguiente las especies más cercanas son similares en su ocurrencia además de tener preferencias ambientales similares.
- El primer eje representó un gradiente ambiental que va desde baja conductividad y saturación de oxígeno con una alta turbidez hasta una alta conductividad y saturación de oxígeno con una baja turbidez.
- El segundo eje representó un gradiente ambiental que va desde una baja velocidad hasta una alta velocidad.
- Por ejemplo, *T. distorta* (Tdist) prefiere condiciones de alta turbidez, baja saturación de oxígeno y conductividad y baja velocidad.

Ejemplo B. McCaigh *et al.* (2001) evaluaron la diversidad bacteriana de tres tipos de pastizales mediante el los métodos DGGE (*Denaturing Gradient Gel Electrophoresis*) y clonación. Los objetivos de su trabajo fueron explorar el potencial del análisis DGGE para evaluar las diferencias entre las comunidades bacterianas de tres tipos de suelos y comparar el método de clonación con el DGGE.

Tomaron muestras por triplicado de tres tipos de suelos de pastizales (designados como mejorados, semimejorados y no mejorados) localizados en la estación de Investigación Sourhope en Escocia. Después se extrajo el ADN para ser analizado el método de clonación y de DGGE. El DGGE comparó las comunidades bacterianas de los tres tipos de suelos, mientras que la clonación comparó los suelos mejorados y no mejorados. Luego se determinó la intensidad y posición relativa para obtener una matriz de datos cuantitativa (ponderada) y también se obtuvo una matriz con datos de presencia-ausencia de bandas (datos no ponderados). Cada banda fue tratada como una unidad taxonómica operativa (UTO) y el número de bandas fue usado como indicador de riqueza. Se usó el coeficiente de similitud de concordancia ($S_{AB} = M_{AB}/N$, en donde M_{AB} es el número de pares de bandas para cada posible posición de bandas y N es el número de posiciones de bandas) sobre datos ponderados y no ponderados para crear dos matrices de similitud para ambos conjuntos de datos representados por WM y UM, respectivamente.

Usaron el análisis de coordenadas principales sobre las matrices WM1, WM2, UM1 Y UM2 para examinar las similitudes entre las comunidades bacterianas de los tres tipos de suelos. Los resultados de la ordenación son representados en los diagramas (figura 7). El eje 1 (PC1) de los diagramas representaron el 18.5, 18.2, 16.7 y 17.3% de la variación de los diagramas a-d. El eje 2 (PC2) representa de los diagramas representaron el 16, 16.5, 15.2 y 15.1% de la variación de los diagramas a-d. El ACoP no observó un patrón de agrupamiento y las posiciones de los puntos (muestras) entre sí se mantuvieron consistentes sin considerar el tipo de análisis; sin embargo, se encontró un alto grado de similitud entre dos tipos de muestras de suelos semimejorados.

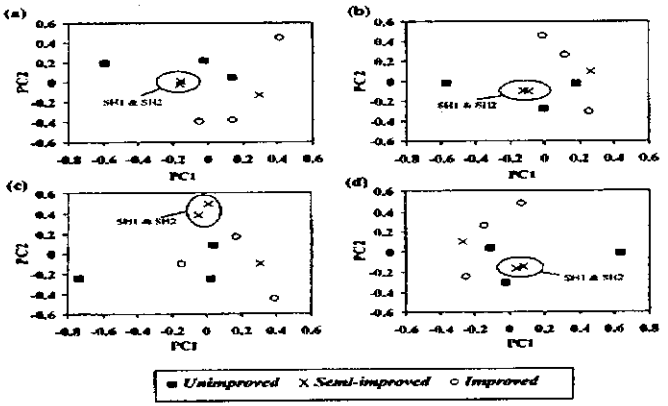

FIGURA 7. ACoP de las matrices de similitud UM1 y UM2 (a y c) para datos DGGE no ponderados y matrices WM1 y WM2 (b y d) para datos DGGE ponderados. Los símbolos representan los 3 tipos de suelos. Tomado de McCaigh *et al.* (2001).

Se interpretó de la siguiente manera:

- Los símbolos representaron perfiles moleculares (comunidades microbianas). Los símbolos más cercanos entre sí indicaron una mayor similitud en su composición de especies, mientras que los más alejados indicaron que son comunidades microbianas diferentes.
- Se observo que las comunidades microbianas fueron diferentes dentro y entre los tipos de suelos. Por lo tanto, las muestras de los tres tipos de suelos representaron comunidades bacterianas diferentes.
- Solo dos muestras de suelos mejorados tuvieron comunidades bacterianas similares.

3.6. ESCALADO MULTIDIMENSIONAL NO MÉTRICO (EMNM)

3.6.1. DESCRIPCIÓN

Es un método exploratorio de ordenación indirecta basado en el orden de rango de las disimilitudes entre objetos y puede usar cualquier medida de distancia o de similitud (Fasham, 1977; Jongman *et al.*, 1995; Ramette, 2007; Quin y Keough, 2002; Legendre y Legendre, 1998; Field *et al.*, 1982). Se puede aplicar cuando existen estructuras no lineales en los datos (Field *et al.*, 1982).

Los objetivos son (Fasham, 1977; Quin y Keough):

- Ordenar los objetos en un espacio de menor dimensión.
- Encontrar los patrones de variación de los datos.
- Medir el grado de monotonicidad entre los valores de disimilitud y la distancia en el diagrama.

En el escalado multidimensional no métrico se calcula una matriz de disimilitud y se desea representar los objetos de la matriz en un espacio de ordenación de menor dimensión. Para ello, se usa el orden de rango de las disimilitudes entre objetos y no su magnitud. Se busca una configuración inicial de los objetos en el espacio de ordenación cuyas dimensiones han sido especificadas de manera previa. La distancia entre los objetos i y j se representa por d_{ij}, mientras que la disimilitud entre esos objetos se representa por δ_{ij}. Los objetos son ordenados de manera que sus disimilitudes están en orden ascendente. Para ello, se calcula una función de optimización (*stress*) que aumenta la relación monotónica entre la disimilitud entre objetos y la distancia entre objetos en el diagrama de Shepard. La función *stress* es $Stress = \sqrt{\frac{\sum (d_{ij} - \bar{d}_{ij})^2}{\sum d_{ij}^2}}$. En donde $\bar{d}_{ij}$ representa la media de las disimilitudes. La función es una medida de monotonicidad (Kenkel y Orloci, 1986). Entre más pequeño es el valor del *stress*, mayor es la relación monotónica.

Quin y Keough (2002) mencionan que un valor de *stress* mayor a 0.3 indica que la configuración no es mejor que la arbitraria y se sugiere no tratar

de interpretar configuraciones a menos que los valores sean menor a 0.2 y de preferencia menor a 0.1.

Se ha demostrado que el valor del *stress* mejora conforme aumenta el número de dimensiones; sin embargo, muchas dimensiones hacen menos intepretable los resultados (Fasham, 1977; Field *et al.*, 1982; Quin y Keough, 2002).

Los resultados del EMNM pueden representarse en un diagrama de dispersión (sección 3.2). Los objetos más cercanos son similares en sus atributos y tienen el orden de rango de las disimilitudes similar.

3.6.2. VENTAJAS

- Se aplica cuando existen estructuras no lineales entre los datos.
- Se puede usar cualquier medida de distancia o similitud.
- Se pueden usar datos cualitativos y cuantitativos.
- Es menos susceptible a la distorsión a altas diversidades beta que el análisis de componentes principales y el análisis de correspondencia (Fasham, 1979).

3.6.3. LIMITACIONES

- Debe especificarse el número de dimensiones.
- El valor del stress cambian de acuerdo al número de dimensiones utilizado.
- Es un método que usa un algoritmo iterativo que demanda mucho tiempo en computar.

3.6.4. ALGORITMO

Según Ludwig y Reynolds (1988):

1. Se calcula una medida de distancia o de similitud para todos los pares de objetos (i, j) y éstos se acomodan en rangos de menor a mayor disimilitud (δ_{ij}).
2. Se determina el número de dimensiones a usar.
3. Se obtiene una configuración inicial. Para ello:
   - Se usan las coordenadas de los ejes de un método de ordenación (por ejemplo, análisis de coordenadas principales) con la medida de distancia usada en el paso anterior para obtener una configuración inicial. Una alternativa es usar una configuración aleatoria.
   - Se ordenan los rangos de las distancias (d_{ij}) de menor a mayor distancia.
4. Se optimiza el orden de rangos. Para ello:
   - Se comparan las disimilitudes y las distancias en un diagrama de dispersión llamado diagrama de Shepard y se calcula una regresión no paramétrica de Kruskal. Para cada par de objetos (i, j) la regresión estima un valor ($\hat{d}_{ij}$) para cada valor de disimilitud.
   - Se calcula la función de optimización *stress* mediante $Stress = \sqrt{\frac{\sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2}}$.
5. Se repiten los pasos 3-5 con una configuración diferente cada vez hasta obtener el valor más bajo de *stress*.
6. Se ordenan los datos en un diagrama de ordenación.

3.6.5. EJEMPLO

Ejemplo A. Schnitler *et al.* (2006) estudiaron las comunidades de myxomycetos del bosque deciduo localizado en Leipzig, Alemania. Sus objetivos fueron crear una lista de especies de myxomycetos que ocurren sobre madera en descomposición y analizar su ecología.

Colectaron ramas muertas localizadas en las coronas de 8 especies de árboles y también registraron varios parámetros ambientales (pH, especies de árboles, exposición a la luz solar, peso sobre el suelo, estratos verticales y características del substrato como capacidad de retención del agua, tipo y estado de descomposición). Las especies fueron cultivadas y después se procedió al conteo de los esporocarpos para tener una medida de la abundancia. Se identificaron 27 especies en 127 muestras.

Usaron el escalado multidimensional no métrico con la distancia de Sorensen para ordenar las especies. Los resultados se muestran en el diagrama conjunto (figura 8). Los valores propios de los dos ejes fueron 0.664 y 0.137, respectivamente. Se encontró que sólo 3 factores ambientales fueron significativos: el pH, la capacidad de retención de agua (*water*) y el estado de descomposición (*decay*). El diagrama mostró que la capacidad de retención de agua y el estado de descomposición estuvieron correlacionados de manera negativa con el pH.

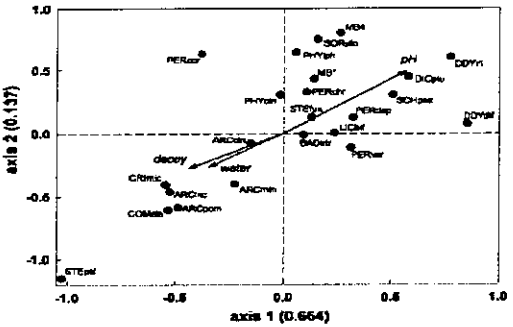

FIGURA 8. Ordenación EMNM de especies en un espacio de muestras con variables ambientales mostradas como vectores. Los puntos definen las posiciones de las especies más comunes. Las líneas radiando del centroide indican la fuerza relativa y dirección de la correlación de los parámetros ambientales más influyentes (pH, estado de descomposición y capacidad de retención de agua) con la ordenación. Tomado de Schnitler *et al.* (2006).

El diagrama se interpretó de la siguiente manera:

- Los puntos representaron la ocurrencia de las especies, de manera que las especies cercanas son similares en su ocurrencia y tienen preferencias ambientales similares.
- Los ejes representaron un gradiente ambiental que va desde una mayor capacidad de retención de agua y descomposición del substrato con un pH bajo hacia una menor capacidad de retención de agua y descomposición del substrato con un pH alto.
- Algunas especies como *Arcyria* (ARC), *Comatricha elegans* (COMele) y *Stemonis pallida* (STEpal) prefieren substratos con alta capacidad de retención de agua, alta descomposición y pH bajo.

Ejemplo B. Frankovich *et al.* (2006) analizaron las distribuciones espaciales y temporales de las diatomeas epifíticas sobre la planta *Thalassia testidinum* dentro del estuario de la Bahía de Florida. Sus objetivos fueron caracterizar la variación espacial y temporal de las diatomeas epifíticas en 8 sitios del estuario y analizar su relación con las características fisicoquímicas del agua.

Colectaron muestras de la planta *T. testidinum* de cada uno de los sitios durante 36 eventos de muestreo e identificaron un total de 92 especies de diatomeas epifíticas. También se midieron 7 características fisicoquímicas del agua. Se construyó una matriz de datos de especies x eventos de muestreo y otra matriz de datos de calidad del agua x eventos de muestreo.

Usaron el escalado multidimensional no métrico y el coeficiente de Bray- Curtis para mostrar las diferencias espaciales y temporales de las diatomeas. Los resultados se muestran en el diagrama de ordenación (figura 9). Los 3 ejes obtenidos explicaron el 40, 28 y 13.8 % de la variación de los datos de especies, respectivamente. La ordenación espacial mostró que los grupos A (sitio 8) y B (sitio 1) fueron diferentes de los grupos de sitios D (sitios 5, 6 y 7) del interior este de la Bahía. El conjunto de diatomeas de los

sitios del interior oeste de la Bahía (sitios 2, 3, y 4) exhibieron algún solapamiento con los sitios de los grupos A, B y D.

La ordenación temporal del EMNM mostró que los conjuntos de diatomeas de Febrero y Marzo fueron diferentes de los conjuntos de diatomeas de Junio y Agosto.

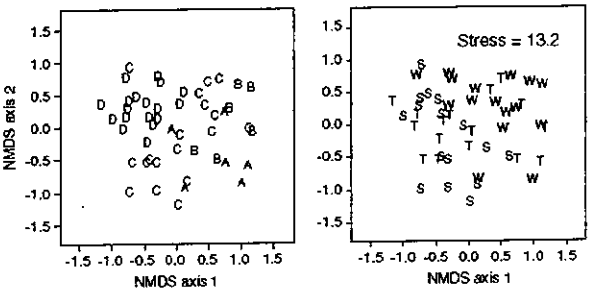

FIGURA 9. Ordenación EMNM de los puntajes de sitios. Los sitios son codificados por grupos de comunidades y grupos de periodos de muestreo. Los grupos son designados como: A (sitio 8), B (sitio 1), C (sitios 2, 3 y 4), D (sitios 5, 6 y 7). Las designaciones para grupos de periodos de muestreo son: S (verano: Junio y Agosto de 2000), W (invierno: Marzo de 2000 y febrero de 2001), T (periodos de transición: Octubre y Diciembre de 2000). Tomado de Frankovich *et al.* (2006).

Se interpretó de la siguiente manera:

- Las comunidades de diatomeas de los sitios de los grupos D fueron diferentes de las comunidades del grupo A y B.
- En el diagrama de variación espacial, se formaron 2 conglomerados: grupos D y grupos C. Esto significó que las comunidades dentro de cada conglomerado fueron más similares entre sí. En cambio, los grupos A y B no mostraron un patrón claro de similitud en las comunidades. Esto sugiere una variación espacial.
- La comunidades de diatomeas en invierno fueron diferentes de las comunidades de verano, lo que sugiere una variación estacional.

3.7. ANÁLISIS DE CORRESPONDENCIA (AC)

3.7.1. DESCRIPCIÓN

Es un método exploratorio de ordenación indirecta basado en un modelo de respuesta unimodal que usa de manera implícita sólo la distancia ji-cuadrada.

Los objetivos son:

- Ordenar las muestras (objetos) en un espacio de menor dimensión.
- Encontrar los patrones de variación en los datos.

El análisis de correspondencia construye una variable teórica (variable latente) cuyos valores son los puntajes de muestras que maximizan la dispersión de los puntajes de especies (Jongman *et al.,* 1995). Una característica importante de este método es que las correspondientes ordenaciones de muestras y especies se obtienen de manera simultánea, lo que permite examinar las interrelaciones ecológicas entre muestras y especies en un sólo análisis (Ludwig y Reynolds, 1987).

Existen 3 formas de calcular el análisis de correspondencia: mediante un algoritmo de promedio ponderado de dos vías (Jongman *et al.,* 1995), mediante un análisis propio (Ludwig y Reynolds, 1987) o bien mediante una tabla de contingencia que contiene las frecuencias de dos o más variables cualitativas (Peña, 2000).

En el algoritmo de promedio ponderado de dos vías, se calcula de manera iterativa los puntajes de especies como promedios ponderados de los puntajes de muestras y los puntajes de muestras como promedios ponderados de los puntajes de especies. Los puntajes de muestras son estandarizados para tener media cero y varianza unidad. Se obtiene el primer eje de ordenación que maximiza la dispersión de los puntajes de especies. El segundo ejes también maximiza la dispersión de los puntajes de especies pero sujeto a la restricción de no estar correlacionado con el primer eje. Los ejes de ordenación se miden en unidades de desviación estándar cuando los puntajes de muestras son estandarizados (Jongman *et al.,* 1995).

Los resultados del análisis de correspondencia pueden ser representados en el diagrama conjunto (*joint* plot) y su interpretación se explica en la sección 3.2. Los puntajes de las especies representan su óptimo en el diagrama de ordenación.

Cuando el método se ejecuta mediante un análisis propio que usa la matriz **Y** de especies x muestras o bien mediante una tabla de contingencia (también es una matriz), se transforma la matriz original mediante una doble transformación que divide las abundancias de las especies entre la raíz cuadrada del total de filas y la raíz cuadrada del total de columnas. A partir de la matriz transformada se calcula una matriz de especies x especies (modo R) y una matriz de muestras x muestras (modo Q). Luego, se calculan los valores y vectores propios de ambas matrices. Los valores propios entre ambas matrices son idénticos. Los vectores propios son escalados para obtener las coordenadas de los ejes. Los resultados pueden representarse de manera conjunta en un diagrama conjunto (*joint plot*) (Peña, 2000; Ludwig y Reynolds, 1988).

Los ejes de ordenación de AC son llamados vectores propios. Cada vector propio tiene asociado un valor propio. El valor propio es igual a la dispersión (máxima) de los puntajes de especies sobre el eje de ordenación. El primer eje de ordenación tiene el valor propio más grande, el segundo eje tiene el segundo valor propio más grande y así sucesivamente. Los valores propios de AC tienen un valor entre 0 y 1. Un valor de 0.5 frecuentemente denota una buena separación de las especies a lo largo del eje (Jongman *et al.*, 1995).

El método se puede ver afectado por el efecto de arco, el cual es un artefacto matemático que no representa una estructura real en los datos (Jongman *et al.*, 1995; Ter Braak, 1986). El efecto de arco es una representación gráfica de una distorsión en la representación de los datos. Se muestra como una compresión de los extremos de los ejes con respecto a la mitad del eje y el segundo eje muestra una relación cuadrática con el primer eje. En la comprensión de los ejes los puntajes de muestras están más cercanos en los extremos que en el medio.

3.7.2. VENTAJAS

- Es útil cuando los gradientes son largos.
- Se puede usar variables cuantitativas y variables cualitativas.
- Permite una representación conjunta de las especies y muestras.

3.7.3. LIMITACIONES

- Es sensible a especies raras.
- Sólo usa la distancia ji-cuadrada.
- No puede ser aplicado a datos que contienen valores negativos.
- Se puede ver afectado por el efecto de arco debido a las relaciones polinomiales del segundo eje con respecto al primero.

3.7.4. ALGORITMO

Algoritmo de promedio ponderada de dos vías (Jongman *et al.*, 1995):

1. Se toma un puntaje de sitio (y_m) de forma arbitraria, pero diferente de cero.
2. Se calculan nuevos puntajes de especies (y_k) por promedio ponderada de los puntajes de sitios mediante $y_k = \frac{\sum (y_{km})(y_m)}{\sum y_{km}}$.
3. Se calculan nuevos puntajes de sitios (x_i) por promedio ponderada de los puntajes de especies mediante $y_m = \frac{\sum (y_{km})(y_k)}{\sum y_{km}}$.
4. Para el eje 1 ve al paso 5. Para el segundo eje y posteriores, hacer los puntajes de sitios (y_m) incorrelacionados con el previo eje por el procedimiento de ortogonalización. Para ello:

- Se denotan los puntajes de sitios del previo eje por f_m y los puntajes de prueba del eje por y_m.

- Se calcula $v = \frac{\sum (y_{+m})(y_m)(f_m)}{y_{++}}$. En donde $y_{+m} = \sum y_{km}$; $y_{++} = \sum (y_{+m})$.
- Se calcula $y_{m\ nuevo} = (y_{m\ viejo}) - (v)(f_m)$.
- Se Repiten los 3 puntos anteriores para los ejes previos.

5. Se estandarizan los puntajes de sitio (y_m). Para ello:

- Se calcula el centroide (z) de los puntajes de sitio (y_m) mediante $z = \left(\frac{\sum (y_{+m})(y_m)}{y_{++}} \right)$.
- Se calcula la dispersión de los puntajes de sitio mediante $s^2 = \frac{\sum (y_{+m})(y_m - z)^2}{y_{++}}$.
- Se calcula $y_{m\ nueva} = \frac{(y_{m\ vieja} - z)}{s}$

6. Se detiene en la convergencia; esto es, cuando los nuevos puntajes de sitios son suficientemente cercanos a los puntajes de sitios del ciclo previo de la iteración; sino se va al paso 2.
7. Se ordenan los datos en un diagrama de ordenación (*joint plot*).

El algoritmo basado en análisis propio es (Ludwig y Reynols, 1988):

1. Se obtiene el total de filas (especies), r_i.
2. Se obtiene el total de columnas (muestras), c_j.
3. Se calcula el gran total mediante $G_t = \sum r_i = \sum c_j$.
4. Se transforman los datos en la matriz **A** mediante $a_{ij} = \frac{(y_{ij})}{(r_i)(c_j)}$.
5. Se calculan las matrices de resemblanza entre especies (modo R) y entre muestras (modo Q).

6. Se calculan los valores propios y vectores propios de las matrices de resemblanza entre especies (modo R) y entre muestras (modo Q).
7. Se calculan las coordenadas de especies escalando los vectores propios $\mathbf{u}_i$ mediante $z_i = (\mathbf{u}_i)\left(\sqrt{\frac{G_i}{r_i}}\right)$.
8. Se calculan las coordenadas de muestras al escalar los vectores propios $\mathbf{v}_j$ mediante $\mathbf{z}_j = (\mathbf{v}_i)\left(\sqrt{\frac{G_i}{c_j}}\right)$.
9. Se ordenan los datos en un diagrama de ordenación.

3.7.5. EJEMPLO

Ejemplo A. Varese *et al.* (2003) estudiaron los efectos de los tratamientos biológicos y químicos en contra del hongo *Heterobasidium annosum* sobre las comunidades fúngicas de los tocones de Norway, Italia. Sus objetivos fueron describir los efectos después de dos años de aplicación de 6 tratamientos biológicos y uno químico sobre las comunidades fúngicas, y luego comparar los efectos de cada tratamiento un año después de la aplicación con los efectos de dos años después de la aplicación.

Se aplicaron 6 tratamientos biológicos sobre tocones infectados por *H. anossum*, los cuales son: HF (*Hypholoma fasciculare*), PV (*Phanerochaete velutina*), TH (*Trichoderma harzianum*), VB (*Verticillium bulbillosum*); FVB (cultivo filtrado de *Verticillium bulbillosum*) y VC (*Vuilleminia comedens*). Además de un tratamiento químico TI (propiconazole) y los controles C1 (control con disco de madera) y C2 (control sin disco de madera). Después de uno y dos años de la aplicación se tomaron 27 astillas de madera de cada uno de los 130 tocones y procedieron al cultivo e identificación taxonómica. Se obtuvieron un total de 49 especies de hongos, los cuales fueron agrupados en 14 grupos.

Se calculó la frecuencia y densidad de colonización para cada especie.

Se usó el análisis de correspondencias para ordenar los grupos de especies y los tratamientos (después de dos años), lo que permitió analizar el efecto de cada tratamiento al nivel de la comunidad. Los resultados del AC se muestran en el diagrama. Los tres ejes explicaron el 45.51, 19.33 y 14.92 % de la variabilidad en los datos, respectivamente. Los tratamientos que seleccionaron la micocenosis también fueron agrupados. El tratamiento biológico *Trichoderma harzianum* (TH) fue separado de otros tratamientos a lo largo del primer eje y estuvo relacionado con el grupo de especies 14. Se reconocieron dos grupos a lo largo del segundo eje. El primero incluye a los tratamientos *Hypholoma fasciculare* (HF), *Phanaerochaete velutina* (PV), *Vuilleminia comedens* (VC), *Verticilum bulbillosum* (VB), cultivo filtrado de *Verticilum bulbillosum* (FVC) y tratamiento control con disco de madera (C1), los cuales estuvieron correlacionados con los grupos de especies 3-13. El segundo grupo estuvo formado por el control sin disco de madera (C2) y el tratamiento químico propiconazol (TI), correlacionados con los grupos de especies 1 y 2. Esto se muestra en el siguiente diagrama (figura 10).

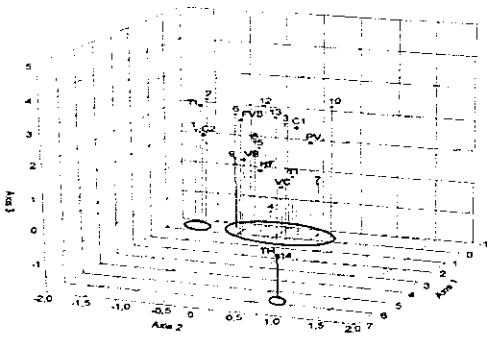

FIGURA 10. Diagrama AC de la matriz de bloques (grupos de especies y tratamientos, dos años después de las aplicaciones). Se muestran los primeros 3 ejes. Los símbolos son explicados en el texto. Tomado de Varese *et al.* (2003).

Se interpretó de la siguiente manera:

- Los grupos de especies representaron especies con densidades de colonización similar.

- Los tratamientos cercanos entre sí indicaron efectos similares sobre los grupos de especies.
- La cercanía entre un grupo de especies y un tratamiento indicó que las especies tuvieron altas densidades de colonización en los tocones sometidos a ese tratamiento.

Ejemplo B. Braker *et al.* (2001) estudiaron la estructura de las comunidades denitrificadoras, del grupo *Bacteria* y del grupo *Archeae* a lo largo de un gradiente redox mediante un análisis T-RFLP. Su principal objetivo fue investigar las comunidades de bacterias denitrificadoras, *Bacteria* y *Archaea* que responden a gradiente redox.

Tomaron muestras de sedimento de 4 estaciones de muestreo (C, 301, 304 y 306) localizados en el margen de Washington y Pugent Sound. Se extrajo el ADN y se sometió a un análisis T-RFLP. Las comunidades fueron caracterizadas por el número de picos y la altura de los picos. La abundancia relativa de T-RFs dentro las estaciones fue determinada mediante el cálculo de la proporción entre la altura del pico de cada pico y la altura total de picos para todos los picos dentro de una muestra. Las proporciones fueron convertidas en porcentajes.

Usaron el análisis de correspondencia para comparar los perfiles T-RFLP de bacterias denitrificadoras, del grupo *Bacteria* y del grupo *Archaea* en las 4 estaciones y comparar las diferencias entre y dentro de las estaciones. Los resultados del AC se muestran en el diagrama de dispersión (figura 11). Para el AC de las bacterias denitrificadoras (S) los dos ejes explicaron el 20 y 19 % de la variabilidad de los datos. Para el AC del grupo *Bacteria* los dos ejes explicaron el 38 y 20% de la variabilidad de los datos. Para el AC del grupo *Archaea* los dos ejes explicaron el 31 y 27% de la variabilidad de los datos.

Las comunidades de diferentes estaciones de muestreo fueron separadas y se mostraron patrones de conglomerado dentro de las estaciones, excepto para la muestra C-2 de las bacterias denitrificadoras y la muestra 306-2 par el grupo *Bacteria.* Este patrón de conglomerado no se

encontró en el grupo *Archeae*. Se encontró un conglomerado para las muestras de las estaciones 301, 304 y 306 excepto para las muestras 304-2 y 304-3. Las comunidades de la estación C en el grupo *Archeae* fueron separadas del otro grupo.

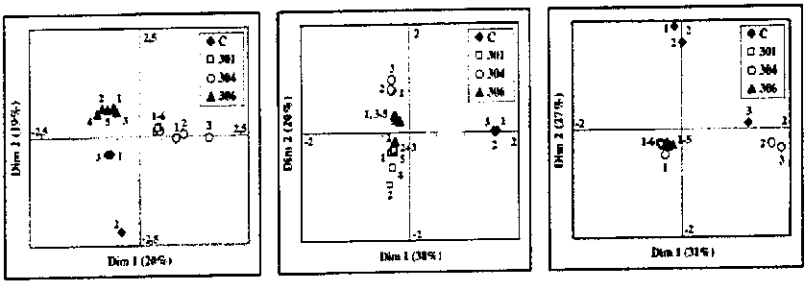

FIGURA 11. Diagramas de dispersión AC de perfiles T-RFLP. Los símbolos representan las diferentes estaciones. El primer diagrama corresponde a las bacterias denitrificadoras (S), el segundo diagrama corresponde al grupo Bacteria, y el tercer diagrama corresponde al grupo *Archeae*. Tomado de Braker *et al.* (2001).

Se interpretó de la siguiente manera:

- Los puntos (símbolos) representaron comunidades bacterianas de las diferentes estaciones. Los puntos más cercanos entre sí son similares en composición de especies y tienen una respuesta similar con respecto a la variable latente (eje).
- En los diagramas para bacterias denitrificadoras y par el grupo *Bacteria*, las comunidades de cada elemento fueron diferentes entre estaciones. Además, las comunidades dentro de cada estación fueron similares.
- En el diagrama para el grupo *Archeae*, las comunidades de la estación 304 fueron similares entre sí.

3.8. ANÁLISIS DE CORRESPONDENCIA SIN TENDENCIA

3.8.1. DESCRIPCIÓN

Es una modificación del análisis de correspondencia para corregir el efecto de arco.

La corrección se logra mediante una técnica conocida como *detrended* (sin tendencia). Este procedimiento divide el primer eje en un número de segmentos y dentro de cada segmento los puntajes de sitios en el segundo eje son ajustados al sustraer su media.

Para remover el efecto de arco, se requiere que el segundo eje no solo sea incorrelacionado (esto es, que sea independiente y ortogonal) con respecto al primer eje sino también incorrelacionado con su relación cuadrática (polinomial).

Este método reescala los ejes con el supuesto de que todas las especies muestran una curva de respuesta unimodal con varianzas homogeneas a lo largo de cada gradiente. Este reescalado tiende a expandir el espacio interpuntos entre muestras o especies localizadas en los extremos de los ejes.

El proceso de quitar tendencia mediante polinomios puede ser incorporado en el algoritmo de promedio ponderado de dos vías al extender el paso 4 de manera que los puntajes de ensayo no sólo están correlacionados con el eje previo, sino también con los polinomios del eje previo.

Esta información fue tomada de Jongman *et al.* (1995), Wartenberg *et al.* (1987), Jackson y Somers (1991) y Ter Braak y Prentice (1988) y se recomiendan para una explicación más amplia.

3.8.2. VENTAJAS

- Permite eliminar el efecto de arco que provoca una mala interpretación de los datos.
- Permite medir los ejes como unidades de desviación estándar.
- Puede usarse para determinar si usar un método lineal o un método unimodal.

3.8.3. LIMITACIONES

- Solo usa la distancia ji-cuadrada.
- Es sensible a especies raras.
- La interpretación depende del número de segmentos utilizado (Jackson y Somers, 1991).

3.8.4. EJEMPLO

Ejemplo A. Anastasi *et al.* (2005) estudiaron las características cualitativas y cuantitativas de las comunidades fúngicas en composta y vermicomposta. Sus objetivos fueron determinar la composición y abundancia de la micobiota de la composta y la vermicomposta, además de evaluar las diferencias cualitativas y cuantitativas en la composición de las dos compostas.

Tomaron 10 muestras de cada composta. Después procedieron a cultivar los hongos para determinar el número de unidades formadoras de colonias por gramo de suelo seco (UFC/g) tanto de la micobiota total como de cada especie o morfotipo (entidad fúngica). Se identificó un total de 194 entidades fúngicas.

Usaron el análisis de correspondencia sin tendencia (*detrended*) para evaluar las diferencias cualitativas y cuantitativas en la composición de la micobiota de las dos compostas y entre 10 muestras de cada composta. Los resultados del de AC sin tendencia se mostraron en el diagrama. Los valores propios de los dos ejes fueron 0.576 y 0.297, respectivamente. Hubo tres

zonas a lo largo del primer eje. En la zona I se encontraron las muestras de la composta 4-10 y las entidades fúngicas que se encontraron solo o en su mayoría a la composta. La zona III contenía las muestras de la vermicomposta 1, 2 y 4-7, además de las entidades fúngicas que se encontraron solo o en su mayoría en la vermicomposta. La zona II contenía las muestras de composta 1 y 3 y las muestras de la vermicomposta 3 y 8-10, además de las entidades fúngicas distribuidas entre la composta y la vermicomposta pero en pequeñas cantidades u ocurrencia.

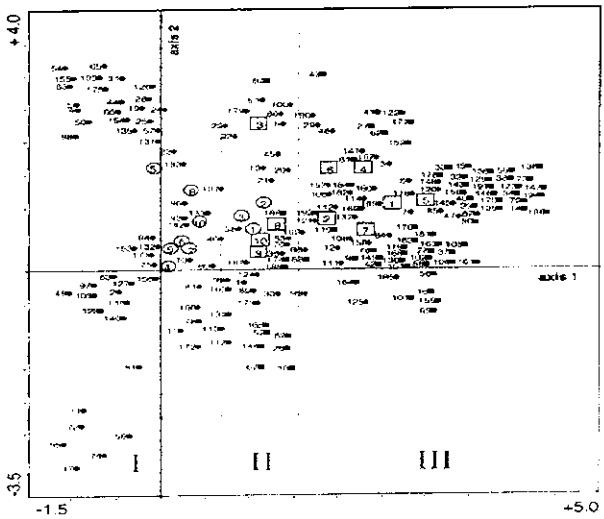

FIGURA 12. Diagrama de dispersión AC sin tendencia de 10 muestras de composta (círculo), 10 muestras de vermicomposta (cuadrado) y 194 entidades fúngicas. Los círculos representan especies exclusivas de la composta; los cuadrados representan especies exclusivas de la vermicomposta; los romboides representan especies comunes de ambas compostas. Tomado de Anastasi *et al.* (2005).

Se interpretó de la siguiente manera:

- Las muestras más cercanas entre sí son similares en la composición cualitativa de entidades fúngicas.
- Las entidades fúngicas más cercanas entre sí son similares en su abundancia y ocurrencia en al composta, en la vermicomposta o en ambas.

- Las entidades fúngicas exclusivas de la composta se encontraron en la zona I que contiene las muestras de la composta. Por consiguiente, la zona I representó la composta.
- Las entidades fúngicas exclusivas de la vermicomposta se encontraron en la zona III que contiene las muestras de la vermicomposta. Por consiguiente, la zona III representó la vermicomposta.
- La zona II representó muestras de la composta y vermicomposta que son similares en composición de entidades fúngicas, además de entidades fúngicas con una baja ocurrencia en esas muestras.

3.9. ANÁLISIS DE REDUNDANCIA (ADR)

3.9.1. DESCRIPCIÓN

Es un método de ordenación canónica basada en un modelo de respuesta lineal que utiliza sólo la distancia euclídea (Jongman *et al.*, 1995; Legendre y Anderson, 1999).

Su principal objetivo es detectar los patrones de variación en la comunidad que pueden ser explicados por las variables ambientales.

Este método es la forma canónica del análisis de componentes principales y se puede describir como una serie de regresiones múltiples seguido de un ACP mediante el algoritmo de suma ponderada de dos vías (sección 3.4.4). Los puntajes de muestras son restringidos a combinaciones lineales de las variables ambientales. Se obtienen los coeficientes canónicos que son los coeficientes de las regresiones múltiples que definen los ejes de ordenación como combinaciones lineales de las variables ambientales; además, se obtiene la correlación especies-ambiente, la cual es una correlación entre los puntajes de muestras que son sumas ponderadas de los puntajes de especies y los puntajes de muestras que son una combinación lineal de las variables ambientales. Por último se obtienen las correlaciones intraconjuntos, las cuales son los coeficientes de correlación entre las variables ambientales y estos ejes de ordenación (Jogman *et al.*, 1995; Ter Braak y Prentice, 1988).

El segundo eje también es una combinación lineal de las variables ambientales pero sujeto a la restricción de ser ortogonal con respecto al eje previo.

Previo al análisis, las variables ambientales son estandarizadas a media cero y varianza unidad.

El diagrama de ordenación puede ser interpretado como un *biplot* (sección 3.2). En el diagrama, el coseno del ángulo entre vectores de una especie y una variable ambiental es una aproximación del coeficiente de correlación entre las especies y las variables ambientales. Una medida de bondad de ajuste del *biplot* de especies y variables ambientales es

$(\lambda_1 + \lambda_2)/(\sum \lambda_i)$, la cual expresa la fracción de varianza de todas las covarianzas entre especies y variables ambientales (Jogman *et al.,* 1995).

3.9.2. VENTAJAS

- Se explican los patrones de variación por las variables ambientales, lo que da un significado ecológico.
- Es útil cuando los gradientes son cortos; es decir, las especies presentan una curva de respuesta lineal con respecto a las variables ambientales.

3.9.3. LIMITACIONES

- La cantidad de variación explicada por los ejes canónicos puede ser menor a la explicada por las variables latentes.
- Es sensible a estructuras no lineales en los datos de especies.
- Sólo usa la distancia euclídea.

3.9.4. ALGORITMO

Se pueden obtener los ejes de ADR mediante la extensión del algoritmo suma ponderada de dos vías (sección 3.4.4) de ACP, de la siguiente manera: en cada ciclo iterativo, los puntajes de sitios calculados en el paso 3 son regresados sobre las variables ambientales con la ecuación $x = B_o + \sum B_i z_i$, en donde B representa los coeficientes canónicos y z_i representa la variable ambiental. Los valores ajustados de la regresión son tomados como los nuevos puntajes de sitios para continuar con el paso 4.

En ADR, la correlación equivale a la correlación entre los puntajes de sitios que son sumas ponderadas de los puntajes de especies y los puntajes de sitios que son una combinación lineal de las variables ambientales (Jongman *et al.*, 1995).

3.9.5. EJEMPLO

Ejemplo A. González *et al.* (2003) estudiaron los patrones de respuesta química de 3 especies de líquenes a la contaminación del aire en Argentina. Sus objetivos fueron comparar los patrones de respuesta química de 3 especies de líquenes transplantados a sitios urbanos con condiciones ambientales similares, y aplicar el análisis de redundancia para analizar la influencia de las variables ambientales sobre la respuesta química de las especies de líquenes.

El estudio consistió en transplantar de manera simultánea 3 especies de líquenes (*Ramalina celastri, Punctelia micosticta* y *Canomaculina pilosa*) en 23 sitios de monitoreo. Después de 3 meses se tomaron muestras de líquenes para analizar sus compuestos químicos (parámetros químicos), los cuales fueron: concentración de sulfuro (S), malondialdehido (MDA), clorofila a (CA), clorofila b (CB), feofitina a (PA), proporción de feofitina / clorofila a (PACA), proteínas solubles (PR), peróxido (HPCD) y el índice de contaminación (PI). Las variables ambientales fueron: nivel del tráfico (Tr), densidad industrial (In), industrias en el sitio (Ti), distancia de las industrias (Di), distancia de la planta de poder (Dp), distancia del río (Dr), altura de los edificios, posición de la cuadra (Si), nivel topográfico (Tl) y cobertura el árbol (Tc).

Se aplicó el ADR por cada especie para analizar las relaciones simultáneas entre los parámetros químicos y las variables ambientales. Los resultados del ADR se muestran en el diagrama (figura 13). Los dos ejes explicaron el 29.6 y 17.7 % de la variabilidad de los datos, respectivamente. En el eje 1 las variables ambientales más importantes fueron el nivel de tráfico y distancia de la planta de poder, lo que indicó que estas variables ambientales podrían definir la respuesta de *R. celastri*. La correlación líquenes-ambiente fue de 0.895, lo que indicó que las variables ambientales explican la respuesta de *R. celastri*.

En el eje 2 las variables ambientales más importantes fueron la distancia de las industrias, distancia de la planta de poder, densidad industrial, industrias en el sitio, nivel topográfico y el nivel del tráfico.

El *biplot* mostró que las variables nivel de tráfico y distancia de la planta de poder tuvieron una correlación positiva con el índice de contaminación.

El contenido de azufre estuvo relacionado con las variables nivel del tráfico y cobertura del árbol.

El nivel de peróxido mostró una correlación negativa con la distancia del río.

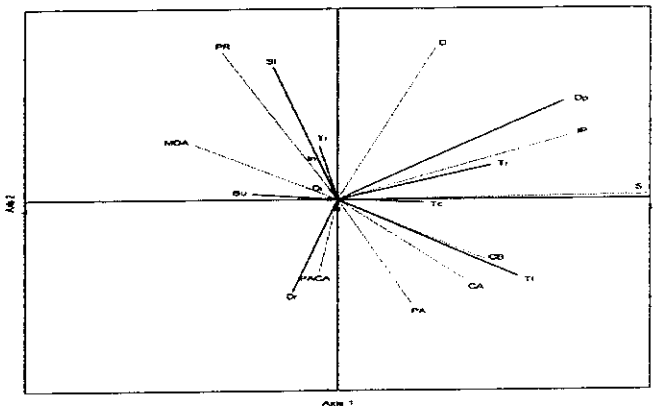

FIGURA 13. Biplot de ADR sobre las relaciones entre las variables químicas (vectores con líneas punteadas) de R. celastri y las variables ambientales (vectores con líneas definidas). Los símbolos se explican en el texto. Tomado de Gonzáles *et al.* (2003).

Se interpretó de la siguiente manera:

- La relación entre el índice de contaminación y las variables nivel de tráfico y distancia de la planta de poder indicó que los contaminantes emitidos de esas fuentes provocaron daños fisiológicos sobre el liquen.
- La relación entre el contenido de azufre y las variables nivel de tráfico y cobertura del árbol indicó que existen condiciones de poca ventilación en los sitios afectados por la acumulación.

- La relación entre el peróxido y la distancia del río indicó que los efectos de los contaminantes sobre el bioindicador se incrementó en los sitios cercanos al río donde los niveles de humedad fueron altas.
- El coeficiente de correlación liquen-ambiente fue una medida la fuerza de la relación entre las variables ambientales sobre la respuesta del liquen.
- El segundo eje representó una combinación lineal de las variables ambientales que caracterizaron las condiciones industriales.
- En el diagrama de ordenación, tanto las variables ambientales como los parámetros químicos fueron representados por vectores. Por consiguiente, es posible determinar la correlación entre variables ambientales, entre parámetros químicos y entre variables ambiéntales y parámetros químicos en base a la dirección entre los vectores. La correlación será positiva si apuntan en la misma dirección, mientras que será negativa si apuntan en direcciones opuestas.

3.10. ANÁLISIS DE CORRESPONDENCIA CANÓNICA (ACC) CON Y SIN TENDENCIA

3.10.1. DESCRIPCIÓN

Es un método de ordenación canónica basado en un modelo de respuesta unimodal que usa sólo la distancia ji-cuadrada.

El principal objetivo es detectar los patrones de variación en la composición de especies que puede ser explicada por las variables ambientales.

Este método es la forma canónica del análisis de correspondencia y se puede describir como una serie de regresiones múltiples seguido de un análisis de correspondencia mediante el algoritmo de promedio ponderado de dos vías (sección 3.7.4). Los puntajes de muestra son restringidos a combinaciones lineales de las variables ambientales que maximizan la dispersión de los puntajes de especies. Se obtienen los coeficientes canónicos que son los coeficientes de las regresiones múltiples que definen los ejes de ordenación como combinaciones lineales de las variables ambientales; además, se obtiene la correlación especies-ambiente, la cual es una correlación entre los puntajes de muestras que son promedios ponderados de los puntajes de especies y los puntajes de muestras que son una combinación lineal de las variables ambientales. Por último se obtienen las correlaciones intraconjuntos, las cuales son los coeficientes de correlación entre las variables ambientales y estos ejes de ordenación.

El segundo eje también es una combinación lineal de las variables ambientales que maximizan la dispersión de los puntajes de especies pero sujeto a la restricción de estar incorrelacionado con el eje previo.

Previo al análisis, las variables ambientales son estandarizadas a media cero y varianza unidad.

Este método, al igual que el análisis de correspondencia, se ve afectado por el efecto de arco y también existe una alternativa que elimina el efecto llamada análisis de correspondencia canónica sin tendencia. El método para eliminar ese efecto se llama *detrended* (sin tendencia) por polinomios (sección 3.8) y puede ser ejecutado dentro del algoritmo de

promedio ponderado de dos vías al extender el paso 4 de manera que los puntajes de ensayo no sólo estén incorrelacionados con el eje previo sino también con los polinomios de ejes previos.

Los resultados de la ordenación se pueden representar en un diagrama conjunto (sección 3.2) con vectores ambientales. Este diagrama es un biplot que aproxima los promedios ponderados de cada especie con respecto a cada variable ambiental. Los puntos especies son sus óptimos en el diagrama.

Una medida de bondad de ajuste del diagrama es $(\lambda_1 + \lambda_2)/(\sum \lambda_i)$, la cual es expresa la fracción de variación de los promedios ponderados explicados por el diagrama.

La información ha sido tomada de Jongman *et al.* (1995), Ter Braak (1986) y Ter Braak y Prentice (1988) y se recomienda consultarlos para una explicación más exhaustiva del método.

3.10.2. VENTAJAS

- Detecta los patrones de variación en la comunidad que puede ser explicado por las variables ambientales.
- El análisis de correspondencia sin tendencia elimina el efecto de arco para hacer los datos más interpretables.
- Es útil cuando los gradientes son largos.
- Puede usar variables cualitativas y cuantitativas (datos de especies).

3.10.3. LIMITACIONES

- Es sensible a especies raras.
- La cantidad de variación explicada por los ejes canónicos puede ser menor la explicada por las variables latentes.
- El análisis de correspondencia canónica con tendencia se ve afectado por el efecto de arco.

- La interpretación del ACC sin tendencia, al igual que el análisis de correspondencia sin tendencia, depende del número de segmentos usado.

3.10.4. ALGORITMO

Se pueden obtener los ejes del análisis de correspondencia canónica mediante la extensión del algoritmo promedio ponderada de dos vías (sección 3.7.4) de AC, de la siguiente manera: en cada ciclo iterativo, los puntajes de sitios calculados en el paso 3 son regresados sobre las variables ambientales con la ecuación $x = B_o + \sum B_i \mathbf{z}_i$, en donde B representa los coeficientes canónicos y z_i representa la variable ambiental. Los valores ajustados de la regresión son por definición una combinación lineal de las variables ambientales y son los nuevos puntajes de sitios para continuar con el paso 4.

3.10.5. EJEMPLO DE ANÁLISIS DE CORRESPONDENCIA CANÓNICA.

Ejemplo A. Griffith *et al.* (2002) analizaron las relaciones entre los conjuntos de perifiton y las características fisicoquímicas del arroyo que se localiza en la ecoregión de Rocas del Sur (Colorado, EUA). Sus objetivos fueron investigar el uso de las métricas de la comunidad y las abundancias de las especies para diagnosticar los factores de estrés ambiental sobre los ecosistemas, además de determinar cómo estos enfoques pueden usarse para diferenciar los efectos de los factores de estrés ambiental con respecto a otros gradientes.

Tomaron muestras de agua y sedimento del arroyo. Midieron 45 variables ambientales, de las cuales solo se tomaron 14 para el análisis multivariable. Identificaron 282 especies, de las cuales solo 142 fueron usadas para el análisis de las abundancias.

Usaron el ACC para relacionar los datos ambientales y las abundancias de las especies. Los resultados de la ordenación se muestran en el diagrama (figura 14). Los valores propios de los dos ejes fueron 0.213

y 0.154, respectivamente. Las correlaciones especies-ambiente para los dos ejes fue de 0.865 y 0.915.

La variable ambiental correlacionada de manera positiva con el primer eje fue la profundidad media, mientras que las variables aguas lentas y tipos de hábitat estuvieron correlacionadas de manera negativa. El eje describió conjuntos de algas con una riqueza alta en ambientes lóticos.

Las variables correlacionadas de manera positiva con el segundo eje fueron los iones, la temperatura, el tipo de substrato y pastoreo.

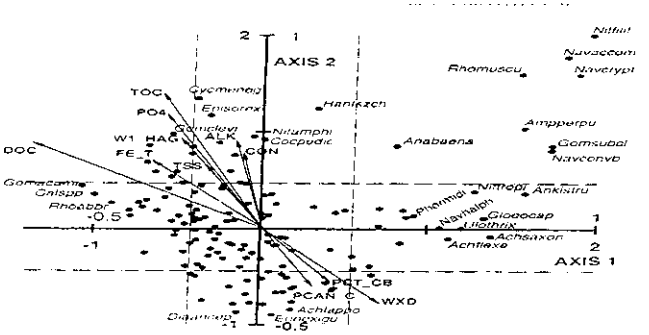

FIGURA 14. Biplot de ACC de la abundancia de especies de perifiton en ríos de la ecoregión Rocas del Sur, Colorado. La escala en unidades S es -0.5 a 1.0 para variables ambientales y -1.0 a 1.0 para puntajes de especies. Tomado de Griffith *et al.* (2002).

Se interpretó de la siguiente manera:

- El primer eje describió un gradiente de ambientes lóticos en el extremo positivo a arroyos con una gran cantidad de carbono orgánico disuelto en el extremo negativo.
- El segundo eje describió un gradiente de arroyos largos con substratos gruesos, concentración alta de nitratos en el extremo negativo a arroyos con una gran perturbación antropocéntrica y concentración de carbono orgánico disuelto, fósforo total y alcalinidad en el extremo positivo.

- Los vectores (variables ambientales) que apuntaron en la misma dirección indicaron correlación positiva, por ejemplo, el fósforo total, carbono orgánico disuelto y alcalinidad.
- Las especies (puntos) que estuvieron cercanas entre sí comparten condiciones ambientales similares y su posición en el gradiente indicó su preferencia ambiental.

Ejemplo B. Stephenson *et al.* (2004) estudiaron los conjuntos de myxomycetos asociados con el bosque nublado de Maquipucuna en el oeste de los Andes de América del Sur. Sus objetivos fueron caracterizar los conjuntos de especies de myxomycetos y registrar su abundancia; además, obtener datos sobre la distribución de los objetos a lo largo de un gradiente de altitud.

Tomaron muestras de substratos de diferentes microhabitats y de cuerpos fructíferos de myxomycetos en 3 sitios de estudio. Se registraron los parámetros del microhábitat fueron los tipos de sustrato como corteza de árboles vivos (bark), hojas esparcidas (litter), madera en descomposición (wood), inflorescencias (in), hepáticas epifílicas (epi); suelo (soil), diámetro del substrato (diam), humedad del aire (air), altura de muestreo (height), exposición al viento (wind) y exposición al sol (sun). Todas las muestras fueron colocadas en cámaras húmedas de cultivo. Se determinó el número de esporocarpos. Para determinar las abundancias ponderadas se dividió el número de esporocarpos por el valor medio de todos los registros para una especie en particular. La suma de todas las abundancias para una especie en particular es igual al número de registros.

Se usó el análisis de correspondencia canónica para determinar la respuesta de los conjuntos de myxomycetos a los factores ambientales. Para el biplot, se escalaron los puntajes de especies y de variables ambientales a media 1 y desviación estándar 1.

Los resultados se muestran en la el biplot (figura 15). Los valores propios para el primer y segundo eje fueron 0.71 y 0.5. El ACC mostró 3 conjuntos diferentes de especies.

El primer conjunto contiene especies que habitan en hojas dispersas, el segundo conjunto está asociado con madera de árboles y el tercer conjunto está asociado con inflorescencias.

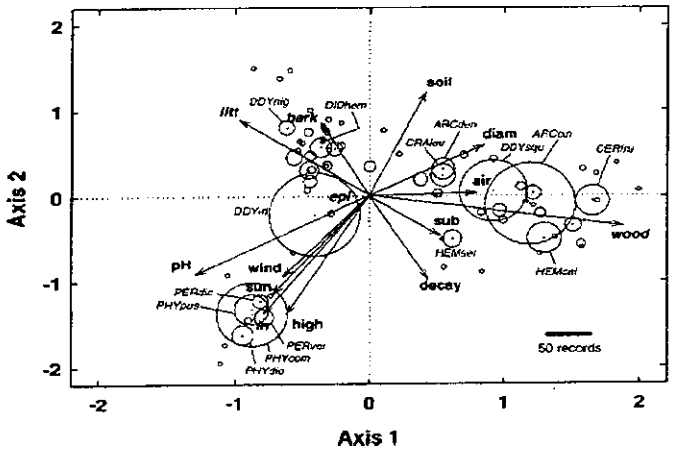

FIGURA 15. Biplot de ACC sobre las especies de myxomycetos y 10 variables ambientales. Los símbolos de las variables ambientales se muestran en el texto. Tomado de Stephenson *et al.* (2004).

Se interpretó de la siguiente manera:

- Los vectores representaron los parámetros ambientales. El diámetro del círculo alrededor de cada especie indicó su abundancia relativa.
- El conjunto asociado con hojas dispersas tuvo la riqueza de especies más alta con la mayoría de especies adaptadas a substratos con pH cercano a la neutralidad.
- El conjunto asociado con la madera de los árboles contuvo especies adaptadas a un pH ácido.
- El conjunto asociado con las inflorescencias contuvo 4 de las especies más abundantes adaptadas a un pH alto.
- Las variables substrato húmedo y aire húmedo tendieron a ser más altos sobre la madera en descomposición; es decir, estos parámetros ambientales favorecieron la descomposición de la madera.

- Los parámetros de hojas dispersas y madera de árboles estuvieron más cercanos entre sí debido a que comparten algunas especies.
- Las especies más abundantes en el diagrama fueron *Arcyria cinerea* (ARCcin), *Didimium squamulosum* (DDYsqu), *D. iridis* (DDYiri), *Physarum compressum* (PHYcom).

3.10.6. EJEMPLO DE ANALISIS DE CORRESPONDENCIA CANÓNICA SIN TENDENCIA.

Ejemplo A. Schonfelder *et al.* (2002) estudiaron las relaciones entre los conjuntos de diatomeas y las variables fisicoquímicas en lagos y ríos del noroeste de Alemania. Sus objetivos fueron determinar cuáles factores ambientales están correlacionados con los cambios en la composición de especies, y desarrollar funciones *transfer* basado en diatomeas.

Tomaron muestras de agua en 69 lagos y 15 ríos. Midieron 24 parámetros ambientales (variables ambientales), de las cuales sólo 12 variables fueron significativas: sodio, carbono orgánico disuelto, amonio, nitrógeno total, hierro disuelto, fósforo total, sulfato, silicio reactivo soluble, carbono inorgánico disuelto, pH, saturación de oxígeno y profundidad del lago). Se identificó un total de 304 especies de diatomeas.

Usaron el ACC sin tendencia para relacionar los conjuntos de diatomeas con las variables ambientales. Aplicaron el reescalado no lineal y la técnica para quitar tendencia por segmentos. Los resultados del ACC sin tendencia se muestran en el diagrama. Los valores propios para los dos ejes fueron 0.544 y 0.382, respectivamente. Las variables ambientales que estuvieron correlacionadas con el primer eje fueron carbono inorgánico disuelto y pH. Las variables ambientales que estuvieron correlacionadas con el segundo eje fueron sodio, fósforo total, nitrógeno total y amonio. La mayoría de las diatomeas oligotróficas como *Acanthes minutissima, A. flexella, A. rosenstockii, Caloneis alpestris, Cymbella amphicephala, C. lacustris, Gomphonema dichotomum, Mastogloia baltica, Navicula diluvia* y *N. gottlandica* pertenecen a lagos alcalinos y se muestran en la parte inferior del diagrama.

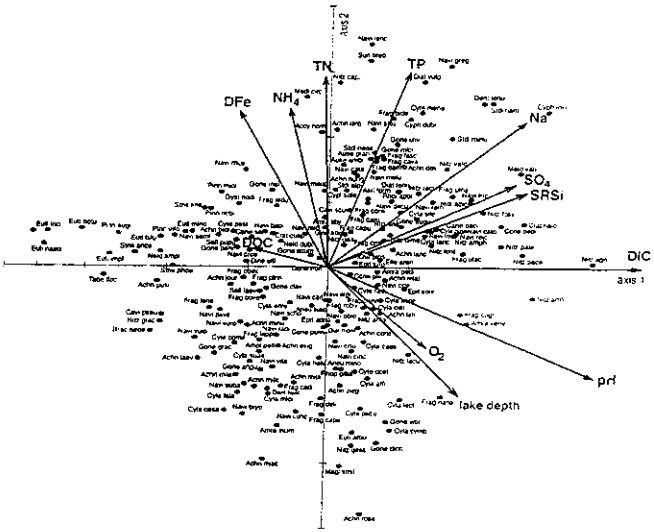

FIGURA 16. Biplot de ACCD de las especies. Sólo se mostraron los taxa más comunes en el diagrama de ordenación. Los puntajes de especies son promedios ponderados de las variables ambientales. Las variables ambientales se muestran como vectores, la longitud de las cuales indica la importancia relativa y la dirección indica la correlación de las variables a los ejes. Tomado de Schonfelder *et al.* (2002).

Se interpretó de la siguiente manera:

- El primer eje representó un gradiente de menor concentración de carbono inorgánico disuelto y pH alto a una mayor concentración de carbono inorgánico disuelto y pH bajo. El segundo eje representó un gradiente de estado trófico.
- Las especies más cercanas entre sí indicaron que son similares en su ocurrencia y comparten preferencias ambientales. Por ejemplo, las especies que se encontraron en la parte inferior del diagrama prefieren condiciones alcalinas.
- La dirección entre los vectores (variables ambientales) indicaron su correlación. Por ejemplo, el pH tuvo una correlación negativa con el amonio y hierro disuelto.

3.11. CONCLUSIONES

1. De acuerdo a los objetivos del estudio se determinó el tipo de datos a usar, qué método multivariables de ordenación aplicar y la interpretación de los resultados de los diagramas.

2. En algunos estudios aplicaron el Análisis de correspondencias sin tendencia para conocer la longitud del gradiente. Esto permitió determinar si usar métodos lineales (como ACP y ADR) que tienen gradientes cortos o bien métodos unimodales (como AC y ACC) que tienen gradientes largos.

3. El análisis de componentes principales y el análisis de redundancia usaron de manera implícita la distancia euclídea. El análisis de correspondencia y su forma canónica usaron sólo la distancia ji-cuadrada. Sin embargo, Legendre y Gallagher (2001) mencionaron que mediante transformaciones previas a los datos de especies se puede utilizar otras medidas de distancia métricas como la distancia de Hellinger.

4. La ordenación polar, el escalado multidimensional no métrico y el análisis de coordenadas principales son métodos que han permitido usar cualquier tipo de distancia.

5. Los resultados de la ordenación por lo general fueron representados de manera gráfica mediante diagramas en donde los puntos representaron muestras, especies u objetos. Los puntos más cercanos entre sí indicaron que fueron similares en sus atributos y mostraron una respuesta similar a alguna variable latente o ambiental. Las variables ambientales fueron representadas como vectores y la dirección entre variables fue una medida de su correlación lineal.

6. En los trabajos de Park *et al.* (2006), Frankovich *et al.* (2006) y Varese *et al.* (2003) usaron otros métodos multivariables a parte del método de ordenación. Esto indicó que en un estudio es posible utilizar varios

métodos multivariables tanto de ordenación como de clasificación o de inferencia.

7. Los métodos de ordenación fueron usados para reducir la dimensionalidad de los datos (Park *et al.*, 2006), mostrar las relaciones entre los datos de especies y las variables ambientales (Schonfelder *et al.*, 2002; Mora-Navarro *et al.*, 2004; Griffith *et al.*,2002; Gonzáles *et al.*, 2003; Branco *et al.*, 2001; Schnitler *et al.*,2006), mostrar la variación espacial y/o temporal de la composición de especies (McCaig *et al.*, 2001) y para encontrar grupos (Varese *et al.*, 2003).

8. En estudios donde se utilizaron métodos moleculares como T-RFLP (Park *et al.*, 2006) o DGGE (McCaig *et al.*, 2001) se crearon perfiles moleculares que representaron comunidades microbianas. En la interpretación, los perfiles más cercanos indicaron comunidades microbianas similares.

3.12. REFERENCIAS

Anastasi, A., G. C. Varese y V. F. Marchisio (2005). *Isolation and identification of fungal communities in compost and vermicompost.* Mycologia, 91(1): 33-44.

Ariyadej, C., R. Tansakul, P. Tansakul y S. Angsupanich (2004). *Phytoplancton diversity and its relationship to the physico-chemical environment in the Banglang Reservoir, Yala Province.* Songklanakarin Journal of Science Technology, 26 (5): 595-607.

Braker, G., H. L. Ayala, A. H. Devol, A. Fesefeldt y J. M. Tiedje (2001). *Communitie structure of Denitrifiers, Bacteria and Archaea along Redox gradients in Pacific Northwest Sediments by Terminal Restriction Fragment Length Polymorphism Analysis of amplified nitrite reductase (nirS) and rRNA genes.* Applied and Environmental Microbiology, 67(4): 1893-1901.

Branco L. H. Z., O. N. Junior y C. C. Z. Branco (2001). *Ecological distribution of Cyanophyceae in lotic ecosystems of Sao Paulo State.* Revista Brasileña de Botánica, 24: 99-108.

Digby, P. G. N. y R. A. Kempton (1987). *Multivariate analysis of ecological communities.* Chapman and Hall. EUA.

Fasham, M. J. R. (1977). *A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines and coenoplanes.* Ecology. 58: 551-561

Field, J. G., K. R. Clarke y R. M. Warwick (1982). *A practical strategy for analysing multispecies distribution patterns.* Marine Ecology Progress Series, 8: 37-52.

Frankovich, T. A., E. E. Gaiser, J. C. Zieman y A. H. Wachnicka (2006). *Spatial and temporal distribution of epiphytic diatoms growing on Thalassia testudinum Banks ex Konig: relationships to water quality.* Hidrobiologia, 569: 259-271.

Gauch, H. G. y T. R. Wentworth (1976). *Canonical Correlation Analysis as an ordination technique.* Vegetation. 33(1): 17-22.

Gauch, H. G. (1982). *Multivariate analysis in community ecology.* Cambridge University Press. EUA

Gauch, H. G. y W. M. Scruggs (1979). *Variants of polar ordination.* Vegetation. 40(3): 147-153.

González, C. M., M. L. Pignata y L. Orellana (2003). *Applications of redundancy analysis for the detection of chemical response patterns to air pollution in lichen.* Elsevier, 312: 245-253.

Griffith, M. B., B. H. Hill, A. T. Herlihy y P. R. Kaufmann (2002). *Multivariate analysis of periphyton assemblages in relation to environmental gradients in Colorado Rocky Mountain Streams.* Journal of Phycology, 38: 83-95.

Jackson, D. A. y K. M. Somers (1991). *Putting things in order: The ups and downs of Detrended Correspondence Analysis.* The American naturalist. 137: 704-712.

Jongman, R. G. , C. J. F. Ter Braak y O. F. R. Van Tongeren (1995). *Data analysis in community and landscape ecology.* Cambridge University Press. Reino Unido.

Jonson, D. E. (2000). *Metodos multivariadas aplicados al análisis de datos.* Internacional Thompson Editores. México.

Kenkel, N. C. y L. Orlóci (1986). *Applying metric and nonmetric multidimensional scaling to ecological studies: some new results.* Ecology, 67 (4): 919-928.

Kessell S. R. y R. H. Whittaker (1976). *Comparisons of three ordination techniques.* Vegetation. 23(1): 21-29.

Legendre P. y M. J. Anderson (1999). *Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments.* Ecological monographs. 69(1): 1-24.

Legendre, P. y E. D. Gallagher (2001). *Ecologically meaningful transformations for ordination of species data.* Oecologia. 129: 271-280

Legendre, P. y L. Legendre (1998). *Numerical ecology.* Elsevier Science. Amsterdam. 853pp

Ludwig, J. A. y J. F. Reynols (1988). *Statistical ecology: A primer on methods and computing.* Wiley-Interscience publications. EUA.

McCaig, A. E., L. A. Glover y J. I. Prosser (2001). *Numerical analysis of grassland bacterial community structure under different land management regimens by using 16S Ribosomal DNA sequence data and Denaturing Gradient Gel Electrophoresis Banding Patterns.* Applied and Environmental Microbiology, 67 (10): 45554-4559.

Mora-Navarro, M. R., J. A. Vázquez-García y Y. L. Vargas-Rodríguez (2004). *Ordenación de comunidades de fitoplancton en el lago de Chapala, Jalisco-Michoacán, México.* Hidrobiológica. 14(2): 91-103.

Palmer, M. W. (2006). *Ordenation methods for ecologists.* http://ordination.okstate.edu/.

Park, S., Y. K. Ku, M. J. Seob, D. Y. Kim, J. E. Yeon, K. M. Lee, S. C. Jeong, W. K. Yohh, C. H. Hark y H. M. Kim (2006). *Principal component analysis and discriminant analysis (PCA-DA) for discriminating profiles or terminal restriction fragment length polymorphism (T-RFLP) in soil bacterial communities.* Soil Biology Biochemistry, 38: 2344-2349.

Peña, D. (2002). *Análisis de datos multivariantes.* McGraw Hill. España.

Poole, R. W. (1974). *An introduction to quantitative ecology.* McGraw Hill Series in population biology. EUA.

Quin G. P. y M. J. Keough (2002). *Experimental design and data analysis of biologist.* Cambridge. EUA.

Ramette, A. (2007). *Multivariate analyses in microbial ecology.* FEMS Microbial Ecology: 1-19.

Schonfelder, I., J. Gelbrecht, J. Schonfelder y C. E. W. Steinberg (2002). *Relationships between littoral diatoms and their chemical environment in Northeastern German Lakes and Rivers.* Journal of Phycology, 38: 66-82.

Stephenson, S. L., M. Schnitler y C. Lado (2004). *Ecological characterization of tropical myxomycete assemblage in Maquipucuna cloud Forest Reserve, Ecuador.* Mycologia, 96 (3): 488-497.

Ter Braak, C. J. F. (1986). *Canonical Correspondence Analysis: A new eigenvector technique for multivariate direct gradient analysis.* Ecology. 65(5): 1167-1179.

Ter Braak, C. J. F. y I. C. Collin (1988). *A theory of gradient analysis.* Advances in ecological research. 18: 271-317.

Varese, G. C., P. Gonthier y G. Nicolotti (2003). *Long-term effects on other fungi are studied in biological and Chemicals stump treatments in the fight against Heterobasidion annosum coll.* Mycologia, 95 (3): 379-387.

Wartenberg D. , S. Ferson y F. J. Rohlf (1987). *Putting things in order: A critique of Detrended Correspondence Analysis.* The American naturalist. 129(3): 434-448.

4. CLASIFICACIÓN

4.1. INTRODUCCIÓN

La clasificación es la asignación de objetos o UTO (Unidad Taxonómica Operativa) a grupos o conglomerados en base a alguna medida de similitud o de distancia (Gauch, 1982; Crisci y López-Armengol, 1983; Digby y Kempton, 1987; Ludwig y Reynolds, 1988; Jongman *et al.*, 1995).

Los objetivos de la clasificación son (Jongman *et al.*, 1995; Crisci y López-Armengol, 1983):

- Dar información sobre la ocurrencia de las especies (ocurrencia).
- Establecer los tipos de comunidades para estudios descriptivos.
- Detectar las relaciones entre comunidades y el ambiente.
- Formar grupos de muestras, objetos, especies o UTO.

En la tabla 13 se muestran los tipos de clasificación (Gauch, 1982; Crisci y López-Armengol, 1983).

TABLA 13. Tipos de clasificación.

Tipo	Descripción
Exclusivas	Se originan grupos donde los objetos son exclusivos del grupo del cual forman parte y no pueden pertenecer a otro grupo que se halle en el mismo rango o nivel.
No exclusivas	Se originan grupos donde los objetos pueden pertenecer a más de un grupo en un mismo nivel o rango.
Jerárquica	Se originan conjuntos que presentan rangos, en los cuales los objetos o grupos de objetos subsidiarios forman parte de un grupo mayor o inclusivo.
No jerárquica	Se originan conjuntos que no exhiben rangos.
Aglomerativas	Se parte de *m* objetos separados, se agrupan en sucesivos conjuntos (siempre en un número mayor que *m* para llegar a un solo conjunto que contiene los *m* objetos.
Divisivas	Se parte de un conjunto que contiene los *m* objetos y luego se divide en subconjuntos.

Los resultados de una clasificación se pueden representar por un diagrama de árbol llamado dendrograma. El término dendrograma incluye dos tipos: fenogramas y cladogramas. Los fenogramas representan las relaciones fenéticas, mientras que los cladogramas representan relaciones filogenéticos.

Los valores de similitud o de distancia se expresan en una escala que suele encontrarse en el extremo superior del dendrograma. Los objetos o las UTO se colocan en el extremo derecho y dan origen cada una a un eje horizontal. Los ejes se unirán mediante ejes verticales que expresan, en relación con la escala, el valor de la similitud o distancia entre los objetos o conjuntos de ellos. En el dendrograma se tendrán núcleos (conjuntos formados por dos objetos) y grupos (conjuntos formados por más de dos objetos).

La interpretación dependerá si se desean establecer relaciones de similitud o de distancia entre objetos (modo Q) o entre UTO (modo R) (Crisci y López-Armengol, 1983; Ramette, 2007).

Por lo general, para describir la diversidad microbiana se usan matrices de distancia basadas en las secuencias de ácidos nucléicos o aminoácidos. Estas matrices pueden ser usadas para conocer las relaciones entre las especies o entre los objetos (Ramette, 2007).

Según Hair *et al.* (1999) el proceso de interpretación implica la comprensión de las características de los objetos de cada conglomerado.

4.2. ANÁLISIS DE CONGLOMERADOS

4.2.1. DESCRIPCIÓN

El análisis de conglomerados jerárquico es un método exploratorio de clasificación aglomerativa que puede usar cualquier matriz de similitud o de distancia (Crisci y López-Armengol, 1983; Digby y Kempton, 1987; Jongman *et al.*, 1995).

Los objetivos del análisis de conglomerados son (Hair *et al.*,1999; Crisci y López-Armengol, 1983):

- Obtener un conjunto de objetos en dos o más grupos en base a la similitud de sus atributos.
- Identificar las relaciones entre objetos, muestras, especies o UTO.
- Reducir el conjunto de datos.

Las principales diferencias entre los métodos aglomerativos son la forma en que se calculan las similitudes entre los conglomerados, y entre conglomerados y objetos (Quin y Keough, 2002; Crisci y López-Armengol, 1983).

Existen 5 técnicas de agrupamiento, las cuales se muestran en la tabla 14 (Quin y Keough, 2002; Crisci y López-Armengol, 1983; Jongman *et al.*, 1995).

TABLA 14. Técnicas de agrupamiento.

Técnica	Características
Ligamiento simple (vecino más cercano)	Se basa en la distancia mínima que pude ser medida entre dos miembros del conglomerado. Encuentra los dos objetos separados por la distancia más corta y los coloca en el primer conglomerado. Luego, se encuentra la distancia más corta entre el primer conglomerado y un tercer objeto para formar un nuevo conglomerado, o bien se buscan otros dos objetos para formar un nuevo conglomerado. Puede ser usado para detectar discontinuidades en los datos. Tiene la tendencia a formar cadenas entre grupos.
Ligamiento completo (vecino más	Se basa en la distancia máxima entre cualquier par de miembros (una en cada conglomerado) de ambos conglomerados. Tiende a sobreestimar las diferencias entre conglomerados.

lejano)	

TABLA 14. Continuación.

Ligamiento promedio	El criterio de aglomeración es la distancia media de todos los objetos de un conglomerado con todos los objetos de otro conglomerado. La partición se basa en todos los objetos de los conglomerados en lugar de un par único de objetos-extremo. Tiende a estar sesgado hacia la producción de conglomerados con una varianza similar. Existen dos técnicas muy usadas: UPGMA (*unweighted pair-group method using arithmetic average*) que usa la media aritmética no ponderada, y WPGMA (*weighted pair-group method using arithmetic average*) que usa la media aritmética ponderada. Este último da menos peso a las similitudes originales de los grupos más largos.
Centroide	La distancia entre los dos conglomerados es la distancia entre los **centroides** de los conglomerados. Los centroides son los valores medios de cada elemento de todos los objetos en un conglomerado. En cada paso se calcula un nuevo centroide, de manera que los centroides cambian conforme se fusionan los conglomerados.
Método de Ward	La distancia entre conglomerados se calcula como una distancia euclídea cuadrada entre todos los pares de objetos en un conglomerado ponderado por el tamaño del conglomerado, o como un aumento en las distancias hacia el centroide del conglomerado son fusionados. En cada paso del procedimiento de aglomeración, se minimiza la suma de cuadrados dentro del conglomerado para todas las particiones obtenidas mediante la combinación de los conglomerados en un paso previo. Está sesgado hacia la producción de conglomerados con similar número de elementos. Se aplica cuando se desea homogeneidad dentro del conglomerado.

La primera etapa del proceso de agrupamiento consiste en reconocer en la matriz de similitud el par de objetos con el valor de mayor similitud. La primera matriz derivada diferirá de la matriz de similitud original en que el núcleo formado será considerado como unidad con respecto a los restantes objetos. Si comparamos la matriz de similitud de la cual se parte con la matriz derivada obtenida, ésta disminuye en tamaño en una columna y en una fila. Es en la matriz derivada donde debe buscarse el próximo valor de mayor similitud.

Se vuelve a obtener una nueva matriz derivada, diferente a la anterior, en la cual los nuevos núcleos o grupos formados se consideran como unidades con respecto a los objetos no integrantes de núcleos o grupos. Se repite así el procedimiento hasta lograr que todos los núcleos y grupos constituyan un grupo que contenga a todos los objetos.

Para medir la distorsión de la técnica de agrupamiento con respecto a la matriz de (dis)similitud original se usa el coeficiente de correlación

cofenética, el cual consiste en construir una nueva matriz de similitud a partir de los valores del fenograma y se le denomina matriz cofenética. Se dispone de la matriz de similitud que dio origen al fenograma y de la matriz cofenética que representa el fenograma. Se computa el coeficiente de correlación del momento-producto entre ambas matrices. Una alta correlación entre ellas (0.6-0.95) es señal de escasa distorsión.

El procedimiento antes descrito fue tomado de Crisci y López-Armengol (1983) y se recomienda consultarlos para una explicación más exhaustiva.

4.2.2. VENTAJAS

- No es necesario especificar el número de grupos.
- Se puede usar datos cuantitativos y cualitativos.
- Se puede usar cualquier coeficiente de similitud o de distancia.
- Se muestran las relaciones entre objetos o entre especies.
- Es posible medir la distorsión de las relaciones representativas por el dendrograma.
- Se puede usar para encontrar grupos, los cuales pueden ser utilizados por otros métodos multivariables como el análisis discriminante (sección 5.2).

4.2.3. LIMITACIONES

- Se recomienda para pequeños conjuntos de datos (Gauch, 1982).
- Es sensible a datos atípicos.
- Las limitaciones dependen de la técnica de agrupación elegida.

4.2.4. ALGORITMO

Según Crisci y Armengol (1983) y Quin y Keough (2002) el algoritmo del análisis de conglomerados jeráquico es:

1. Se calcula la matriz de disimilitudes entre todos los pares de objetos.
2. Se forma el primer agrupamiento entre los dos objetos con la disimilitud más baja.
3. Las disimilitudes entre este agrupamiento y los demás agrupamientos son luego recalculados.
4. Se calcula un segundo agrupamiento entre el agrupamiento 1 y el objeto más similar al agrupamiento 1.
5. El procedimiento continúa hasta que todos los objetos son ligados en grupos.

4.2.5. EJEMPLO

Ejemplo A. van Gremberghe *et al.* (2007) estudiaron la relación entre la composición de la comunidad cianobacteriana y de la comunidad zooplanctónica en el lago Blaarmeersen (Bélgica). Su principal objetivo fue evaluar el impacto de los factores bióticos y abióticos sobre los cambios estacionales en la comunidad cianobacteriana.

Se tomaron muestras de agua del lago Blaarmeersen a dos profundidades diferentes (0.5 y 7.5 m) durante dos años (2003 y 2004). Los periodos de muestreo se dividieron dos estaciones: estación de crecimiento (final de la primavera y a principio de verano y otoño) y estación de invierno (final de otoño, invierno y a principio de primavera). Se midió la temperatura, el pH, la transparencia del agua, la conductividad, la concentración de oxígeno y la concentración de nutrientes (nitrato, amonio y ortofosfato). Se identificaron las especies de cianobacterias, fitoplancton y zooplancton; además se midió su biomasa. Se extrajo el ADN de las muestras y se sometió a un análisis DGGE (*Denaturing Gradient Gel Electrophoresis*). Se creó una matriz de intensidad de bandas. La intensidad de bandas fue convertida a intensidad relativa (esto es, la contribución relativa de cada banda al total de bandas en cada columna del gel). Los datos fueron transformados mediante log (x+1). Se usó la medida de Bray-Curtis para crear una matriz de similitud.

Se usó el análisis de conglomerados con la técnica ligamiento promedio ponderado (WPGMA) sobre la matriz de similitud de Bray-Curtis para encontrar las relaciones estacionales entre las muestras.

Los resultados del método se muestran en el dendrograma (figura 17). El dendrograma mostró dos conglomerados: uno que agrupó muestras de la estación de crecimiento y otro que agrupó muestras de la estación de invierno.

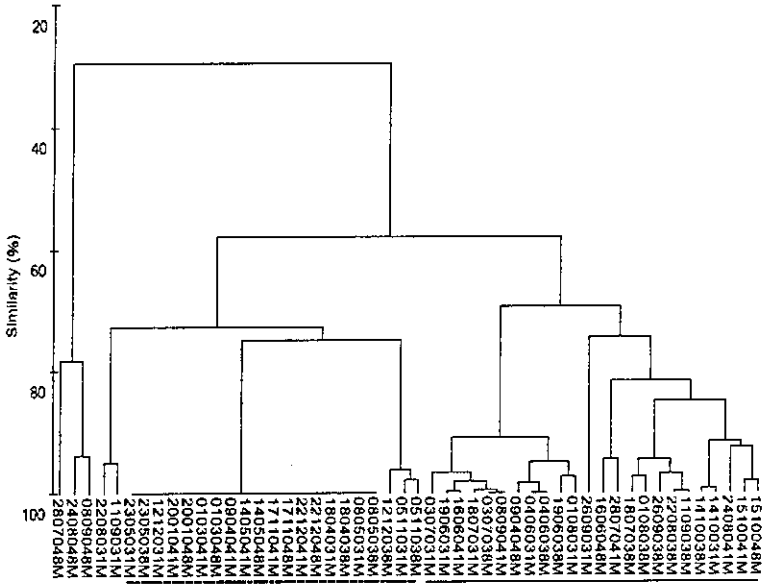

FIGURA 17. Resultados del análisis de conglomerados sobre la composición de especies de las muestras. Las líneas llenas representan las muestras tomadas en la estación de crecimiento. Las líneas punteadas representan las muestras tomadas en la estación de invierno. 1M = 0.5 m. 8M = 7.5 m. Tomado de Gremberghe *et al.* (2007).

Se interpretó de la siguiente manera:

- Las muestras dentro de cada conglomerado fueron similares en composición de especies.
- La composición de especies en la estación de invierno fue muy diferente de la composición biótica en la estación de crecimiento.

- En la estación de invierno hubo menos variación en la composición de especies. En cambio, en la estación de crecimiento se mostró una gran variación en la composición biótica.
- Las muestras de diferentes profundidades no fueron muy diferentes entre sí.

Ejemplo B. Lanoiselet *et al.* (2005) investigaron la utilidad de dos métodos de análisis de ácidos grasos (MIDI y MIDI modificado) para caracterizar y diferenciar aislados de *Rhizoctnia oryzae* y *R. oryzae-sativae* de cuatro países (Australia, Japón, Uruguay y EUA). Su principal objetivo fue evaluar y comparar los métodos MIDI y MIDI modificado para discriminar entre *R. oryzae* y *R. oryzae-sativae*.

Se obtuvo un total de 30 aislados de *R. oryzae* y *R. oryzae-sativum* de cuatro países y después procedieron a cultivarlas. Se extrajeron los ácidos grasos y los sometieron a los métodos de análisis MIDI y MIDI modificado. Obtuvieron un total de 10 ácidos grasos por el método MIDI y un total de 11 ácidos grasos por el método MIDI modificado.

Usaron el análisis de conglomerados (con la técnica de vecino más cercano y distancia euclídea) para analizar las relaciones entre los aislados en base a la composición de ácidos grasos mediante el método MIDI modificado.

Los resultados del método se muestran en el dendrograma (figura 18). Se encontró que los aislados de *R. oryzae* y *R. oryzae-sativum* formaron dos conglomerados diferentes, cada conglomerado con una distancia euclídea de 2.66 y 1.72, respectivamente. Los aislados de *R. oryzae* y *R. oryzae-sativum* fueron agrupados con una distancia de 4.26.

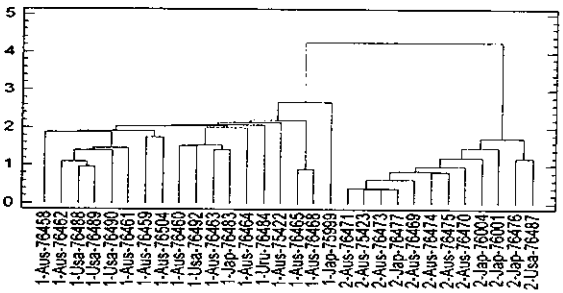

FIGURA 18. Dendrograma de *R. oryzae* (1) y *R. oryzae-sativum* (2) por el método MIDI modificado. Tomado de Lanoiselet *et al.* (2005).

Se interpretó de la siguiente manera:

- Las dos especies fueron diferentes entre sí con respecto a su composición de ácidos grasos. En el diagrama se pudo observar que todos los aislados de *R. oryzae* se encontraron en un solo conglomerado sin importar su lugar de origen. También se pudo observar que todos los aislados de *R. oryzae-sativae* se encuentran en un solo conglomerado.
- Los aislados de *R. oryzae* fueron más similares entre sí que los aislados de *R. oryzae-sativum*.

4.3. K-MEDIAS

4.3.1. DESCRIPCIÓN

Es un método de clasificación no jerárquica aglomerativa que asigna objetos a g grupos definidos *a priori* en base a la distancia eucídea más cercana a la media de los conglomerados (Ramette, 2007; Hair *et al.*, 1999).

La partición de los objetos en g grupos se determina de manera que los objetos dentro de cada conglomerado son más similares que los objetos otro conglomerado (Ramette, 2007; Legendre y Legendre, 1998).

Para el agrupamiento se minimiza la suma de cuadrados de las distancias entre los datos y los correspondientes centroides (medias) de los conglomerados.

En el proceso de agrupamiento, se calcula el centroide del conglomerado inicial y todos los objetos dentro de una distancia umbral especificada *a priori* se incluyen dentro del conglomerado resultante. Luego, se selecciona otro centroide y la asignación continua hasta que todos los objetos han sido agrupados (Ramette, 2007; Hair *et al.*, 1999).

Existen 3 técnicas para asignar objetos a cada conglomerado, las cuales se describen en la tabla 15 (Hair *et al.*, 1999).

TABLA 15. Técnicas de K-medias.

Técnica	Descripción
Umbral Secuencial	Se empieza con la selección de un centroide de un conglomerado e incluye todos los objetos que caen dentro de una distancia especificada de manera previa. Cuando todos los objetos están incluidos, se selecciona un segundo centroide y se incluyen todos los objetos dentro de las distancias especificadas. El proceso continua hasta que todos los objetos han sido agrupados. Los conglomerados son excluyentes.
Umbral Paralelo	Se seleccionan varios centroides de conglomerados de manera simultánea al principio y se asigna objetos dentro de la distancia umbral hasta el centroide más cercano. A medida que se avanza, se puede ajustar las distancias umbral para incluir más o menos objetos en los conglomerados.
Optimización	Es similar a las técnicas umbral excepto que permite la reubicación de los objetos. Si, en el curso de la asignación de los objetos, un objeto se acerca más a otro conglomerado que no es el que tiene asignado en ese momento, un procedimiento de optimización cambia el objeto al conglomerado más cercano.

4.3.2. VENTAJAS

- Es adecuado para grandes conjuntos de datos (Gauch, 1982).
- Se puede usar para confirmar la existencia de grupos (Langenheder *et al.*, 1995).

4.3.3. LIMITACIONES

- No muestra la relación entre conglomerados (Gauch, 1982).
- Se debe especificar de manera previa el número de grupos (Ramette, 2007; Hair *et al.*, 1999; Legendre y Legendre, 1998).

4.3.4. ALGORITMO

Según Hair *et al.* (1999) el algoritmo de k-medias es:

1. Se selecciona el número de grupos (g).
2. Se calcula el centroide del conglomerado inicial.
3. Se utiliza una distancia umbral para asignar los objetos a un conglomerado.
4. Se repite el paso 3 para otros conglomerados hasta que todos los objetos han sido agrupados.

4.3.5. EJEMPLO

Ejemplo A. Langenheder *et al.* (2005) realizaron un experimento en el cual el agua de 4 lagos diferentes sirvió como medio para el crecimiento de las comunidades bacterianas que pertenecían a esos lagos en todas las combinaciones posibles. Su principal objetivo fue examinar los efectos de fuente de medio u origen del inóculo sobre varias propiedades funcionales y la composición de la comunidad.

Se tomaron muestras de agua de 4 lagos. Después se preparó un medio y un inóculo de cada agua del lago. Luego, cada medio fue inoculado con cada uno de los 4 inóculos. Los tratamientos fueron: 4 medios, 4 inóculos y la interacción medios x inóculos. Los parámetros funcionales (variables) fueron: abundancia, producción de biomasa bacteriana, tasa de crecimiento máximo, respiración bacteriana, eficiencia del crecimiento bacteriano y porcentaje de carbono orgánico disuelto total. Se extrajo el ADN de las muestras de agua y se sometió a un análisis T-RFLP para analizar la composición de la comunidad.

Se usó el análisis de conglomerados k-medias para confirmar la existencia de un número de agrupamientos definidos por el análisis de componentes principales. El conglomerado k medias con un número predeterminado de 3 conglomerados confirmó la estructura de 3 conglomerados o agrupamientos del ACP (tabla 16).

TABLA 16. Resultados del análisis de conglomerados k medias con 3 grupos predefinidos. El conglomerado fue hecho para las variables funcionales. Tomado de Langenheder *et al.* (2005).

Conglomerado	Parámetro funcional
1	1.1, 1.2, 1.3
2	2.1, 2.2, 2.3, 3.1, 3.2, 3.3
3	4.1, 4.2, 4.3, 4.4

Se interpretó de la siguiente manera:

- Cada conglomerado contuvo como objetos los tratamientos que tuvieron valores similares en los parámetros funcionales.
- El conglomerado 1 mostró tratamientos con bajos valores en abundancia, producción de biomasa y respiración bacteriana.
- En el conglomerado 2 se observó tratamientos con altos valores de eficiencia de crecimiento bacteriano.
- El conglomerado 3 tuvo tratamientos con altos valores de abundancia, producción de biomasa y respiración bacteriana.

4.4. TWINSPAN

4.4.1. DESCRIPCIÓN

TWINSPAN (Two Way Indicador Species Análisis) es un método de clasificación jerárquica divisiva que caracteriza cada grupo de sitios por un grupo de especies diferenciales, las cuales dominan de uno u otro lado de la dicotomía.

Un término esencial en este método es pseudo-especie. Las pseudo-especie es un equivalente cualitativo de la abundancia de especies. La abundancia de cada es reemplazada por la presencia de una o más pseudo-especies. Cada pseudo-especie es definida por una abundancia mínima de la correspondiente especie llamado nivel de corte.

El método usa el primer eje de ordenación del análisis de correspondencia y lo divide en su centro de gravedad (centroide), el cual se divide en lado derecho (positivo) y el lado izquierdo (negativo).

Se usan las frecuencias de especies sobre el lado positivo y el lado negativo para crear una nueva dicotomía. Las especies diferenciales (para uno u otro lado de la dicotomía) son identificadas mediante el cálculo del puntaje de preferencia. Los puntajes positivos son asignados a las especies con preferencia para el lado positivo de la dicotomía, mientras que los puntajes negativos son asignados a aquellas especies con preferencia al lado negativo.

Un puntaje de preferencia absoluta de 1 se asigna a cada pseudo-especie que es al menos tres veces más frecuente de un lado de la dicotomía que del otro lado. Las pseudo-especies raras y pseudo-especies menos preferenciales (comunes) son subponderadas.

Se calcula la suma ponderada de los puntajes de preferencia para ordenar los sitios (primera ordenación) y se estandariza esa suma ponderada de manera que el valor absoluto máximo es 1.

Se calcula el promedio ponderado de los puntajes de preferencia para cada especie sin subponderar las especies raras (segunda ordenación).

Los puntajes en ambas ordenaciones son sumadas entre sí para crear una ordenación refinada, la cual es dividida en un punto cercano a su centro. Esta ordenación refinada determina la dicotomía. El procedimiento se repite un número de veces especificado por el investigador.

4.4.2. VENTAJAS

- Al usar métodos de ordenación en su algoritmo permite relacionar la composición de especies de cada grupo de sitios con las variables ambientales.
- Muestra las especies diferenciales de cada grupo.
- Muestra los sitios de cada grupo.

4.4.3. LIMITACIONES

- Según Jongman *et al.* (1995) el método es sensible a datos atípicos y a especies raras; además, la detección de los niveles de corte son arbitrarios.
- Según Groenewoud (1992) la utilidad del método depende del significado ecológico de los ejes extraídos y de qué tan bien so divididos los ejes.
- Tiene las mismas limitaciones del análisis de correspondencias.

4.4.4. ALGORITMO

Según Jongman *et al.* (1995) el algoritmo de TWINSPAN es:

1. Se determina el nivel de corte de cada pseudo-especie y el tamaño del grupo.
2. Se realiza un análisis de correspondencia con el algoritmo de promedio ponderado y se extrae el primer eje.
3. Se divide el primer eje en su centroide.

4. Se designa el lado derecho como positivo y el lado izquierdo como negativo.
5. Se calculan los puntajes de preferencia para las especies diferenciales.
6. Se calcula la suma ponderada de los puntajes de preferencia para ordenar los sitios (primera ordenación) y se estandariza esa suma de manera que el valor absoluto máximo sea 1.
7. Se calcula el promedio ponderado de los puntajes de preferencia para cada especie si subponderar las especies raras (segunda ordenación).
8. Los puntajes en ambas ordenaciones son sumadas entre sí para crear una ordenación refinada.
9. Se repiten los pasos 5-8 hasta que se llega al nivel de corte.
10. Los resultados se muestran en un diagrama de árbol.

4.4.5. EJEMPLO

Ejemplo A. Soininen *et al.* (2004) analizaron las comunidades de diatomeas bénticas en relación a gradientes ambientales o espaciales. Sus objetivos fueron clasificar sitios de muestreo en base a la composición de diatomeas, y examinar la variación en la estructura de la comunidad explicada por la variación ambiental y espacial.

Se tomaron muestras de diatomeas que estaban adheridas a piedras y al fondo de arroyos y ríos en 141 estaciones localizadas en 5 ecoregiones de Finlandia. Se midieron 7 variables ambientales (color, pH, fósforo total, conductividad, porcentaje de sombra, anchura y velocidad), mismas que fueron estandarizadas. Se identificó un total de 212 especies de diatomeas. Se aplicó la transformación ln (x +1) a datos de abundancia.

Se usó TWINSPAN para definir los tipos de conjuntos de diatomeas. Además, se definieron 5 pseudoespecies como niveles de corte y 5 como el

tamaño mínimo del grupo. Los resultados de TWINSPAN se muestran en la figura 19. Se produjo 13 grupos de sitios (A-M).

La primera división de TWINSPAN separó los arroyos oligotróficos con baja conductividad localizados en el norte y centro de Finlandia (grupos A-H, indicados por *Frustulia romboides* y *Tabellaria pusilla*) de los arroyos eutróficos con alta conductividad localizados en el sur de Finlandia (grupos I-M, indicados por *Nitzchia pallea* y *Surirella brevisonii*). Una segunda división (nivel 2) en la rama izquierda separó ríos húmicos y ácidos (grupos A-E, indicados por *Eunotia meisteri*) de arroyos oligotróficos de agua clara y con pH cercano a la neutralidad (grupos F-H, indicado por *Achnathes pusilla*). La segunda división de la rama derecha separó ríos mesotróficos (grupos L-M, indicados por *Diatoma tenuis*) de ríos eutróficos contaminados (grupos I-K, indicado por *Nittzchia pallea*).

El grupo A consistió de arroyos ácidos y polihúmicos representados por las especies del género *Eunotia*.

El grupo B consistió de arroyos oligotróficos con un pH alto y bajo contenido húmico y fue representado por *Fragillaria construens* y *Gomphonema exilissimum*.

El grupo C (representado por *Gomphonema gragile*) y el grupo D (representado por *Aulacoseira ambigua*) consistieron de arroyos oligotróficos con un pH alto y bajo contenido húmico.

El grupo E consistió de arroyos con condiciones mesotróficas, húmicas y pH cercano a la neutralidad y fue caracterizado por *Achnanthes bioretii*.

Los grupos F y H consistieron de arroyos con aguas claras y pH cercano a la neutralidad.

El grupo G consistió de arroyos con aguas claras y pH cercano a la neutralidad. Fue caracterizado por *Diatoma tenuis*.

El grupo I consistió de arroyos eutróficos.

El grupo J consistió de arroyos eutróficos con alta conductividad.

El grupo K consistió de arroyos impactados por aguas residuales.

El grupo L consistió de arroyos localizados en el sur de Finlandia.

El grupo M consistió de arroyos impactados por la agricultura.

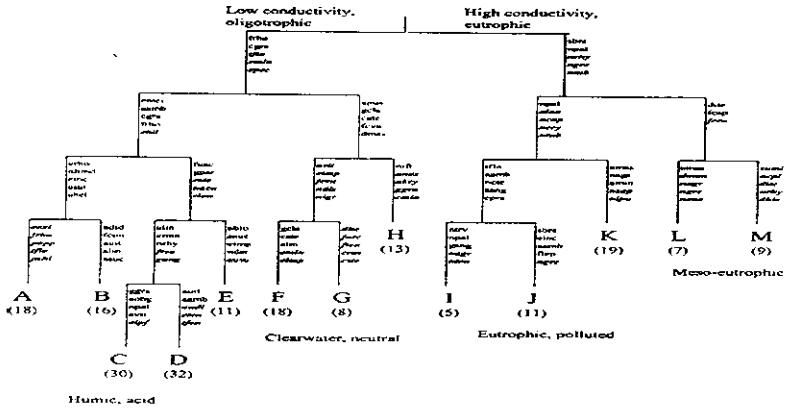

FIGURA 19. Clasificación TWINSPAN del estudio de ríos. La figura se refiere al número de sitios en cada grupo TWINSPAN (A_/M). Los taxa en itálica fueron identificados como indicadores solo por el método del valor indicador (IndVal), mientras que todos los otros fueron identificados sólo por el método del valor indicador (IndVal), mientras que todos los otros fueron identificados por TWINSPAN e IndVal. Tomado de Soininen *et al.* (2004).

Se interpretó de la siguiente manera:

- Los grupos fueron separados por variables químicas y físicas. La mayoría de los sitios en cada grupo fueron localizados dentro de áreas restringidas. Por lo tanto, cada grupo presentó un conjunto de especies con condiciones ambientales similares y ubicación geográfica específica.
- Las especies en cada nivel fueron indicadoras de sitios con condiciones ambientales particulares.

4.5. CONCLUSIONES

1. Todos métodos de clasificación usaron alguna medida de distancia. El análisis de conglomerados jerárquico pudo usar cualquier medida de distancia, mientras que k-medias y TWINSPAN solo usaron la distancia euclídea y ji-cuadrada, respectivamente.
2. En TWINSPAN se definieron grupos de sitios con especies diferenciales relacionados con variables fisicoquímicas (Soininen *et al.*, 2004).
3. El método de k-medias fue usado para confirmar la existencia de grupos (Lengenheder *et al.*, 2005).
4. Para la intepretación de los grupos o conglomerados se tomó en cuenta las características de los elementos del conglomerado.

4.6. REFERENCIAS

Crisci, J. V. y M. F. López-Armengol (1983). *Introducción a la teoría y práctica de la taxonomía numérica.* Secretaria General de los Estados Americanos. Argentina.

Digby, P. G. N. y R. A. Kempton (1987). *Multivariate analysis of ecological communities.* Chapman and Hall. EUA.

Gauch, H. G. (1982). *Multivariate analysis in community ecology.* Cambridge University Press. EUA

Hair, J. F. J., R. E. Anderson, R. L. Tatham y W. C. Black (1999). *Análisis multivariante.* 5a. edición, Prentice Hall, España.

Hurst, C. J., R. L. Crawford, G. R. Knusen, M.J. McInerney y L. D. Stetzenbach (2002). *Manual of Environmental Microbiology.* 2a. ed.. ASM Press. EUA.

Jongman, R. G. , C. J. F. Ter Braak y O. F. R. Van Tongeren (1995). *Data analysis in community and landscape ecology.* Cambridge University Press. Reino Unido.

Jonson, D. E. (2000). *Metodos multivariadas aplicados al análisis de datos.* Internacional Thompson Editores. México.

Langenheder, S., E. S. Lindstrom y L. J. Tranvik (2005). *Weak coupling between community composition and functioning of aquatic bacteria.* Limnological Oceanograghic, 50(3): 957-967.

Lanoiselet, V. M., E. J. Cother, N. J. Cother, G. J. Ash y J. D. I. Harper (2005). *Comparision of two total cellular fatty acid analysis protocols to differentiate Rhizoctonia oryzae and R. oryzae sativae.* Mycologia, 97 (1): 77-83.

Legendre, P. y L. Legendre (1998). *Numerical ecology.* Elsevier Science. Amsterdam. 853pp

Ludwig, J. A. y J. F. Reynols (1988). *Statistical ecology: A primer on methods and computing.* Wiley-Interscience publications. EUA.

Quin G. P. y M. J. Keough (2002). *Experimental design and data analysis of biologist.* Cambridge. EUA.

Ramette, A. (2007). *Multivariate analyses in microbial ecology.* FEMS Microbial Ecology: 1-19

Soininen J., R. Paavola y T. Muotka (2004). *Benthic diatom communities in boreal streams: Communitie structure in relation to environmental and spatial gradients.* Ecography, 27: 330-342.

Van Gremberghe, I., J. Van Wichelen, K. Van der Gucht, P. Vanormelingen, S. D'hondt, C. Buhote, A. Wilmotte y W. Vyverman (2007). *Covariation between zooplancton communitu composition and cyanobacterial communitu dynamics in Lake Blaarmeersen (Belgium).* FEMS Microbial Ecology, 63: 222-237.

Van Groenewoud, H. (1992). *The robustness of Correspondence, Detrended Correspondence and TWISPAN Analysis.* Journal of vegetation science. 5: 239-246

5. MÉTODOS INFERENCIALES

5.1. ANÁLISIS DE VARIANZA MULTIVARIABLE (ANVAM)

5.1.1. DESCRIPCIÓN

El análisis de varianza multivariable es la generalización del análisis de varianza (ANVA) univariable para comparar más de dos grupos (vectores de medias de los grupos) y varias variables medidas por cada unidad experimental de manera simultánea. Existen formas multivariables para cada ANVA ((Jonson, 2000; Hair *et al.*, 1999, Quin y Keough, 2002).

El modelo matemático es $x_{pig} = \mu_{gp} + \varepsilon_{pig}$. En donde x_{pig} es el valor observado en la p-ésima variable respuesta, en la i-ésima unidad experimental del g-ésimo grupo.

La hipótesis nula es que los vectores de medias de los grupos son iguales (Jonson, 2000; Hair *et al.*, 1999).

Los supuestos del ANVAM son:

- Las observaciones deben ser independientes.
- Las matrices de varianzas-covarianzas deben ser iguales para todos los grupos pero desconocidas.
- Las variables deben de tener una distribución normal multivariable.
- Linealidad entre las variables.
- Datos cuantitativos.

Su principal objetivo es determinar si existen diferencias entre los grupos. Además, requiere tamaños de muestra más grandes que ANVA univariable. Como mínimo, se sugiere que el tamaño debe de ser más grande que el número de variables incluidas. Un mínimo de 20 observaciones por grupo es recomendable (Hair *et al.*, 1999).

En el ANVAM se tiene una matriz de sumas de cuadrados y productos cruzados total (**T**) el cual es descompuesto en una matriz de suma de

cuadrados y productos cruzados dentro de grupos (o del error, denotado por **W**) y una matriz de suma de cuadrados y productos cruzados entre grupos (parte sistemática del modelo, denotado por **B**). En la diagonal principal se tienen las sumas de cuadrados de las variables y fuera de ella se tienen los productos cruzados entre pares de variables. Luego, se busca una combinación lineal de las variables (función discriminante) que maximiza la razón de variación entre grupos sobre la variación dentro de grupos; es decir, se busca una función discriminante que maximiza el estadístico F. Las funciones discriminantes son ortogonales entre sí (Jonson, 2000; Hair *et al.*, 1999, Quin y Keough, 2002).

Se tienen opciones de prueba estadística, las cuales se resumen en la tabla 17 (Hair *et al.*, 1999).

TABLA 17. Pruebas estadísticas multivariables.

Prueba	Fórmula	Características
Contraste de Roy	Valor propio más grande de $\mathbf{W}^{-1}\mathbf{B}$	Es el más adecuado cuando las variables están correlacionadas en una sola dirección. Está basado en el primer valor propio más grande.
Lambda de Wilks	$\Lambda = \frac{\lvert\mathbf{B}\rvert}{\lvert\mathbf{B}+\mathbf{W}\rvert}$	Considera todos los valores propios; es decir, compara si los grupos son de algún modo diferentes sin estar afectados por el hecho de que los grupos difieren en al menos una combinación lineal de las variables. Cuanto mayor es el dispersión entre los grupos, más pequeñño es el valor de Lambda de Wilks y mayor la significancia.
Criterio de Pillai	$V = tr\left(\mathbf{B}(\mathbf{B}+\mathbf{E})^{-1}\right)$	Es más robusto y debe ser empleado si los tamaños de muestra disminuyen o si se incumple la homogeneidad de varianzas-covarianzas. Considera todos los valores propios y puede ser aproximado por un estadístico F.
La traza de Hotelling	$T = tr\left(\mathbf{W}^{-1}\mathbf{E}\right)$	Considera todos los valores propios y puede ser aproximado por un estadístico F.

5.1.2. VENTAJAS

- Se puede tener cierto control del porcentaje del error tipo I (Hair *et al.*, 1999).
- Permite comparar varios grupos y varias variables de manera simultánea (Jonson, 2000; Hair *et al.*, 1999).

5.1.3. LIMITACIONES

- Se ve afectado por la multicolinealidad, lo cual indica que hay variables redundantes y disminuye la eficiencia estadística (Hair *et al.*, 1999).
- Se ve afectado por datos atípicos y por la heteroscedasticidad.
- No muestra qué grupos son diferentes (Jonson, 2000).

5.1.4. ALGORITMO

El siguiente algoritmo fue tomado de Peña (2000) y de Jonson (2000):

1. Se fija la prueba de hipótesis.
2. Se fija el nivel de significancia.
3. Se fija el tipo de prueba estadística a utilizar.
4. Se ordenan los datos en una matriz con elementos x_{ij}.
5. Se calcula el tamaño de cada grupo n_g.
6. Se calcula la media de cada grupo $\bar{x}_g$.
7. Se calcula la media total $\bar{x}_T$.
8. Se calcula la matriz de suma de cuadrados y productos cruzados total (**T**) mediante: $\mathbf{T} = \sum_{i=1} (x_i - x_T)(x_i - x_T)'$.
9. Se calcula la matriz de suma de cuadrados y productos cruzado entre grupos (**B**) mediante: $\mathbf{B} = \sum_{g} n_g (x_g - x_T)(x_g - x_T)'$.
10. Se calcula la matriz de suma de cuadrados y productos cruzados dentro de grupos (**W**) mediante $\mathbf{W} = \sum_{i=1}^{n_g} \sum_{g}^{G} (x_{ig} - x_g)(x_{ig} - x_g)'$.
11. Se multiplica la matriz $\mathbf{W}^{-1}$ por la matriz **B** para obtener la matriz $\mathbf{W}^{-1}\mathbf{B}$.
12. Se calculan los valores λ_i y vectores propios v_i más grandes de la matriz $\mathbf{W}^{-1}\mathbf{B}$. Los valores propios indican la variabilidad en los datos.

13. Se calcula la función discriminante z mediante los vectores propios y las variables mediante $z = c + \sum \mathbf{v}_i x_i$. En donde c representa una constante.

14. Se aplica una prueba estadística para hacer el contraste de que los vectores de medias de los grupos son iguales.

5.1.5. EJEMPLO

Ejemplo A. Langenheder *et al.* (2005) realizaron un experimento en el cual el agua de 4 lagos diferentes sirvió como medio para el crecimiento de las comunidades bacterianas que pertenecían a esos lagos en todas las combinaciones posibles. Su principal objetivo fue examinar los efectos de fuente de medio u origen del inóculo sobre varias propiedades funcionales y la composición de la comunidad.

Se tomaron muestras de agua de 4 lagos. Después se preparó un medio y un inóculo de cada agua del lago. Luego, cada medio fue inoculado con cada uno de los 4 inóculos. Los tratamientos fueron: 4 medios, 4 inóculos y la interacción medios x inóculos. Los parámetros funcionales (variables) fueron: abundancia, producción de biomasa bacteriana, tasa de crecimiento máximo, respiración bacteriana, eficiencia del crecimiento bacteriano y porcentaje de carbono orgánico disuelto total. Se extrajo el ADN de las muestras de agua y se sometió a un análisis T-RFLP para analizar la composición de la comunidad.

Se usó el análisis de varianza multivariable de dos vías (con la prueba de la traza de Pillai) sobre los puntajes de las 3 dimensiones del EMNM para probar si el origen del medio o del inóculo (o de ambos) tuvo efecto sobre la composición de las comunidades bacterianas.

Se encontró que tanto el origen del medio como del inóculo tuvieron efectos significativos sobre la estructura de la comunidad (tabla 18).

TABLA 18. Resultados de ANVAM. Tomado de Langenheder *et al.* (2005).

Tratamiento	Valor F	Valor P
Medio	$F_{9,96}$ =15.5	p < 0.0001
Inóculo	$F_{6,62}$ = 36.3	p < 0.0001
Medio x inóculo	$F_{18,96}$ = 3.1	p < 0.0001

Se interpretó de la siguiente manera:

- Tanto el origen del medio como del inóculo determinaron la estructura genética de las comunidades bacterianas.

5.2. ANÁLISIS DISCRIMINANTE (AD)

5.2.1. DESCRIPCIÓN

Es un método multivariable que busca una regla o un esquema de clasificación para predecir el grupo del que es más probable que tenga que venir una observación (Johnson, 2000).

La hipótesis nula es que las medias de los grupos son iguales.

Los supuestos son (Hair *et al.*, 1999; Hurst *et al.*, 2002):

- Normalidad multivariable de las variables.
- Linealidad de las relaciones.
- Ausencia de multicolinealidad entre las variables.
- Igualdad de dispersión entre las matrices.
- Se conoce por anticipado el número de grupos y la identidad de pertenencia de grupo para un subconjunto de muestras.

Los objetivos son (Hair *et al.*, 1999):

- Determinar si existen diferencias significativas de manera estadística entre los perfiles de las puntuaciones medias sobre un conjunto de variables de dos o más grupos definidos *a priori*.
- Producir una regla o esquema de clasificación que permita predecir la población de la que es más probable que tenga que venir una observación.
- Determinar una combinación lineal que maximice la variación entre grupos.

El análisis discriminante implica obtener una combinación lineal de dos o más variables (funciones discriminantes) que discriminen mejor entre los grupos definidos *a priori*. La discriminación se lleva a cabo al establecer las ponderaciones (coeficientes) de la función discriminante para cada variable de tal forma que se maximice la variación entre grupos y se minimice la variación dentro de grupos. Si la variación entre grupos es

grande con relación a la variación dentro de los grupos, entonces la función discriminante separa bien los grupos. La función discriminante se deriva de la ecuación que adopta la forma $z = c + \sum v_i x_i$. En donde z es la puntuación z discriminante, c es una constante, v_i es la ponderación discriminante para cada variable i, y x_i es la variable i.

Después se obtiene un promedio de las puntuaciones z para todas las observaciones dentro de un grupo particular llamado centroide. Estos indican la situación más común de cualquier observación de un determinado grupo.

Johnson (2000) menciona tres reglas para predecir de cuál de las dos poblaciones (o grupos) es más probable que venga una observación x:

- Regla de verosimilitud: Se elije la población 1 si la función de verosimilitud para el grupo 1 es mayor que la función de verosimilitud para la población 2; de lo contrario, se elije la población 2.
- Regla de función discriminante lineal: Se elije la población 1 si $x(\Sigma^{-1}(\mu_1 - \mu_2)) - (0.5)(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) > 0$; de lo contrario se elije la población 2.
- Regla de distancia de Mahalanobis: Se elige la población 1 si el cuadrado de la distancia de Mahalanobis entre x y μ_1 , $(x - \mu_1)' \Sigma^{-1}(x - \mu_1)$, es menor que el cuadrado de la distancia de Mahalanobis para la población 2; de lo contrario se elije la población 2. Esta regla clasifica una observación en la población cuya media está más próxima.

También es necesario determinar o estimar las probabilidades de las clasificaciones correctas de las mismas observaciones. Un método para lograr ello es la estimación de validación cruzada (Johnson, 2000; Kaneene *et al.*, XXXX).

Los pasos de la estimación de validación cruzada son:

1. Eliminar el primer vector de observaciones de los datos, formular una regla discriminante basada entonos los datos restantes, usar la regla

para clasificar la primera observación y observar si ésta se clasifica en forma correcta o no.

2. Reemplazar la primera observación al conjunto de datos y eliminar la segunda.
3. Formular una regla discriminante basada en todos los datos restantes y usarla para clasificar la segunda observación y determinar si se clasifica de manera correcta o no.
4. Se continúa con el mismo proceso a través del conjunto completo de datos.
5. Se crea una matriz resumen par las estimaciones validadas en forma cruzada.

El grado de confianza de clasificar una observación depende de la probabilidad de acertar; es decir, la probabilidad de que una observación pertenezca al primer grupo se puede calcular de manera previa (probabilidad *a prior*) o después de haber hecho el análisis para confirmar (probabilidad *a posteriori*) (Peña, 2000; Jonson, 2000).

En el análisis discriminante canónico se aplica cuando se desea discriminar entre varios grupos y conocer la pertenencia de una observación a alguno de ellos. En este método se crean nuevas variables al tomar las combinaciones lineales (funciones discriminantes) de las variables originales, de modo que contengan toda la información para discriminar entre grupos.

Se calcula una matriz de cuadrados y productos cruzados entre grupos (**B**) y dentro de grupos (**W**) como en el análisis de varianza multivariable. Luego, se buscan los vectores propios ($\mathbf{v}_i$) la matriz $\mathbf{W}^{-1}\mathbf{B}$ que maximizan la relación $\frac{\mathbf{v}_i' \mathbf{B}\mathbf{v}_i}{\mathbf{v}_i' \mathbf{W}\mathbf{v}_i}$. Con los vectores propios se calcula la función discriminante.

Por último, los vectores de medias pueden ser representados en un diagrama de dispersión (sección 3.2). Las distancias usadas en el diagrama corresponden a la distancia de Mahalanobis.

5.2.2. VENTAJAS

- Es posible determinar el grado de clasificación correcta de las observaciones a los grupos.
- Permite comparar dos o más grupos.
- Maximiza la separación entre grupos y disminuye la variación dentro de grupos.

5.2.3. LIMITACIONES

- Se debe conocer el número de grupos de manera previa.
- Deben de cumplir los supuestos paramétricos.
- Sólo trabaja con relaciones lineales.
- Es sensible a la razón entre el tamaño de la muestra y el número de variables.

5.2.4. ALGORITMO

Según Peña (2000) el siguiente algoritmo se aplica cuando se desea clasificar una observación en la población 1 o 2:

1. Se calcula el vector $\mathbf{v_1}$ mediante $\mathbf{v} = \Sigma^{-1}(\mu_1 - \mu_2)$.
2. Se construye la función discriminante z mediante $z = \mathbf{v}'\mathbf{x}$; en donde $\mathbf{x}$ es el vector de variables.
3. Se usa la regla de función discriminante lineal para clasificar la observación en la población más cercana.
4. Se calcula la probabilidad de clasificación correcta.
5. Se calcula la probabilidad de clasificación incorrecta.

Según Jonson (2000) el siguiente algoritmo se aplica cuando se desea conocer la pertenencia de una observación a más de dos poblaciones o grupos:

1. Se ordenan los datos en una matriz con elementos x_i.
2. Se calcula el tamaño de cada grupo n_g.
3. Se calcula la media de cada grupo $\bar{x}_g$.
4. Se calcula la media total $\bar{x}_T$.
5. Se calcula la matriz de suma de cuadrados y productos cruzados total (**T**) mediante $\mathbf{T} = \sum_{i=1} (x_i - x_T)(x_i - x_T)'$.
6. Se calcula la matriz de suma de cuadrados y productos cruzado entre grupos (**B**) mediante $\mathbf{B} = \sum_{g} n_g (\bar{x}_g - x_T)(\bar{x}_g - x_T)'$.
7. Se calcula la matriz de suma de cuadrados y productos cruzados dentro de grupos (**W**) mediante $\mathbf{W} = \sum_{i=1}^{n_g} \sum_{g}^{G} (x_{jg} - x_g)(x_{jg} - x_g)'$.
8. Se multiplica la matriz $\mathbf{W}^{-1}$ por la matriz **B** para obtener la matriz $\mathbf{W}^{-1}\mathbf{B}$.
9. Se calculan los valores λ_i y vectores propios $\boldsymbol{v}_i$ más grandes de la matriz $W^{-1}B$. Los valores propios indican la variabilidad en los datos.
10. Se calcula la función discriminante *z* mediante los vectores propios y las variables mediante $z = c + \sum \mathbf{v}_i x_i$. En donde c representa una constante.
11. Se pueden calcular dos o más funciones discriminantes con la restricción de no estar correlacionadas entre sí mediante la relación $\mathbf{v}_1' \mathbf{W} \mathbf{v}_2 = 0$.

12. Las coordenadas de los grupos se forman por parejas $(\mathbf{v}_1' \boldsymbol{\mu}_i, \mathbf{v}_2' \boldsymbol{\mu}_i)$. La proyección de una nueva observación x sobre el plano tiene las coordenadas $(\mathbf{v}_1' \mathbf{x}, \mathbf{v}_2' \mathbf{x})$.

13. Los datos se proyectan en un diagrama de dispersión.

5.2.5. EJEMPLO

Ejemplo A. Park *et al.* (2006) aplicaron el AD para probar si hubo diferencias en las comunidades bacterianas asociadas con la rizosfera de sandía transgénica y no transgénica diferían de manera estadística.

El análisis discriminante sobre los 4 grupos (transgénica 1, parcela transgénica 2, parcela no transgénica 1 y parcela no transgénica 2) usó los puntajes de los componentes principales (no mostrado) para encontrar diferencias entre grupos.

El AD sobre los puntajes de ACP de cada parcela no mostró diferencias significativas de manera estadística entre los puntajes de diferentes parcelas. También mostró que hubo diferencias considerables, aunque no significativas al 95%, entre puntajes ACP de perfiles T-RFLP de diferentes parcelas de sandía transgénica (tabla 19).

TABLA 19. Resultados del AD sobre los puntajes del ACP de los perfiles T-RFLP con las enzimas Hae III y Hha I para cada parcela (LM1, LM2, WT1 y WT2). LM indica sandía transgénica; WT indica sandía no transgénica; CP indica componentes principales; T2 para Hotelling. Tomado de Park *et al.* (2006).

No. CP	*Hae III*					*Hha I*				
	Lambda Wilks	P	Grupo	T^2	P	Lambda Wilks	P	Grupo	T^2	P
2	0.469	0.422	LM1-LM2	2.497	0.152	0.280	0.122	LM1-LM2	4.121	0.066
			LM1-WT1	0.261	0.778			LM1-WT1	1.137	0.374
			LM1-WT2	2.333	0.167			LM1-WT2	3.422	0.092
			LM2-WT1	1.197	0.357			LM2-WT1	1.941	0.214
			LM2-WT2	0.117	0.891			LM2-WT2	1.1758	0.241
			WT1-WT2	1.206	0.354			WT1-WT2	0.615	0.568

TABLA 19. Continuación.

3	0.179	0.179	LM1-LM2	4.529	0.055					
			LM1-WT1	0.153	0.924					
			LM1-WT2	1.375	0.338					
			LM2-WT1	3.511	0.086					
			LM2-WT2	2.492	0.157					
			WT1-WT2	0.710	0.581					

Se interpretó de la siguiente manera:

- No existen diferencias entre las comunidades de las 4 parcelas.
- Las comunidades bacterianas del suelo de sandía transgénica son iguales a las comunidades de suelo de sandía no transgénica.
- La sandía transgénica no tiene un impacto sobre las comunidades bacterianas del suelo.

Ejemplo B. Kaneene *et al.* (2007) estudiaron el uso del análisis discriminante basado sobre los perfiles de resistencia antimicrobiana de aislados bacterianos de *E. coli* provenientes de muestras fecales de animales domésticos, silvestres y humanos para identificar y clasificar los aislados en la fuente más probable de especies. Sus objetivos fueron: identificar los problema específicos que afectan la eficiencia de las funciones discriminantes; desarrollar reglas de decisión para AD a partir de fuente conocidas de aislados bacterianos de animales domésticos y silvestres; probar la precisión del método; identificar la fuente más probable de contaminación fecal por *E. coli* en la superficie de agua en Michigan. Se tomaron muestras fecales de animales (domésticos y silvestres) y de humanos, además de muestras de agua de varias granjas localizadas en la Cuenca Cedro Rojo, Michigan. Después se procedió al aislamiento e identificación de *E. coli*. Luego se usó el método de difusión de discos de Kirby-Bauer para desarrollar perfiles de susceptibilidad antimicrobiana para 12 agentes antimicrobianos.

Se usaron 3 modelos de análisis discriminante (lineal, cuadrático y Epanechnikov) con la distancia de Mahalanobis para clasificar los aislados en los grupos de especies. Se usó el método de validación cruzada para determinar la proporción promedio de clasificación correcta (PPCC).

Cuando se analizaron los PPCC sobre los 3 métodos, se encontró que el modelo lineal y el modelo Epanechnikov tuvieron los mejores resultados (tabla 20). Sin embargo, cuando se analizaron los resultados de los modelos sobre los grupos de especies (% de aislados clasificados de manera correcta) el modelo cuadrático fue el más eficiente para identificar aislados del grupo de animales silvestres sobre todos los grupos de especies. El modelo lineal fue capaz de identificar suinos y grupos combinados de especies que contiene suinos al menos en un 50%, mientras que el modelo Epanechnikov identificó de manera correcta los aislados bacterianos de humanos.

TABLA 20. Porcentajes promedios de clasificación correcta (PPCC) para diferentes reducciones en el número de antibióticos y clasificación de especies para los 3 modelos: resultados del método de validación cruzada. El valor entre paréntesis indica el porcentaje de aislados clasificados de manera correcta en el grupo. Tomado de Kaneene *et al.* (2007).

Método (probabilidad debido al cambio aleatorio)	% de aislados clasificados de manera correcta		
	Lineal	Cuadrático	Epanechnikov
12 antibióticos; 8 especies (12.5%)	28.7; suino (49)	25.6; silvestres (75)	27.5; humanos (67)
Antibióticos reducidos; 8 especies (12.5%)	30.8; suino (50)	27.7; silvestres (75)	37.3; silvestres (89)
12 antibióticos; 6 especies (16.7%)	35; suino (59)	33.1; silvestres (80)	33.8; humanos (67)
Antibióticos reducidos; 6 especies (16.7%)	37.1; suino (57)	35; silvestres (78)	54.5; humanos (100)
12 antibióticos; 5 especies (20%)	33.6; ganado (67)	32.7; silvestres (83)	42.3; humanos (100)
Antibióticos reducidos; 5 especies (20%)	36.3; ganado (67)	35.4; silvestres (80)	51.5; silvestres (89)
12 antibióticos; 4 especies (25%)	32.6; ganado (82)	36.8; silvestres (84)	48.7; humanos (100)
Antibióticos reducidos: 4 especies (25%)	39; ganado (80)	40.7; silvestres (81)	67.9; humanos (100)

Se interpretó de la siguiente manera:

- Las cantidades que aparecen en la tabla sobre los 3 modelos indicó la proporción promedio de clasificación correcta de los aislados a los grupos de especies.
- Las cantidades entre paréntesis indicaron el porcentaje de aislados clasificados de manera correcta en el grupo identificado. De esta manera, el modelo lineal clasificó los aislados bacterianos que pertenecían al grupo de suinos; el modelo cuadrático identificó los aislados bacterianos que pertenecían a los animales silvestres; el modelo Epanechnikov clasificó los aislados bacterianos que pertenecían al grupo de humanos.
- Dado que los humanos y animales toman diferentes antibióticos en diferentes dosis, las bacterias se vuelven selectivas. Este criterio fue utilizado para clasificar los aislados bacterianos.

5.3. ANÁLISIS DE VARIABLES CANÓNICAS (AVC)

5.3.1. DESCRIPCIÓN

También es conocido como análisis discriminante lineal (Digby y Kempton, 1987; Jongman *et al.*, 1995).

El análisis de variables canónicas es un método que permite comparar las medias de los grupos y busca una combinación lineal que maximice la variación entre grupos (Digby y Kempton, 1987).

La hipótesis nula indica que las medias de los grupos son iguales.

Los supuestos son:

- Normalidad multivariable.
- Matrices de varianzas-covarianzas iguales.
- Hay varios grupos conocidos *a priori*.

Su principal objetivo es encontrar las variables canónicas que maximizan la variación entre grupos.

Según Jongman *et al.* (1995) el análisis de variables canónicas busca la suma ponderada de las abundancias de las especies que maximizan la razón de la variación entre grupos y la variación dentro de grupos a lo largo del primer eje.

El método divide una matriz **X** de n elementos y p variables g grupos determinados de manera previa. Se obtiene una matriz de cuadrados y productos cruzados entre grupos (**B**) y una matriz de suma de cuadrados y productos cruzados dentro de los grupos (**W**). Después, se buscan los vectores propios de $\mathbf{W^{-1}B}$ que definen las variables canónicas. Las variables canónicas son ortogonales (independientes) entre sí (Digby y Kempton, 1987; Johnson, 2000).

Los resultados pueden acomodarse en un diagrama de dispersión (sección 3.2) en los que los ejes son las variables canónicas.

5.3.2. VENTAJAS

- Permite determinar donde ocurren las diferencias entre las medias de las poblaciones cuando se comparan las poblaciones sobre varias variables diferentes mediante el uso simultáneo de todas las variables medidas (Johnson, 2000).
- Permite conocer cómo la composición de especies difieren entre muestras de diferentes grupos.

5.3.3. LIMITACIONES

- El número de observaciones u objetos debe ser mayor que número de variables (Digby y kempton, 1987, Jongman *et al.*, 1995).

5.3.4. ALGORITMO

Según Digby y Kempton (1987) el algoritmo del análisis de variables canónicas es:

1. Se ordenan los datos en una matriz con elementos x_i.
2. Se calcula el tamaño de cada grupo n_g.
3. Se calcula la media de cada grupo $\bar{x}_g$.
4. Se calcula la media total $\bar{x}_T$.
5. Se calcula la matriz de suma de cuadrados y productos cruzados total (**T**) mediante $\mathbf{T} = \sum_{i=1} (x_i - x_T)(x_i - \bar{x}_T)'$.
6. Se calcula la matriz de suma de cuadrados y productos cruzado entre grupos (**B**) mediante $\mathbf{B} = \sum_{g} n_g (\bar{x}_g - \bar{x}_T)(\bar{x}_g - \bar{x}_T)'$.
7. Se calcula la matriz de suma de cuadrados y productos cruzados dentro de grupos (**W**) mediante $\mathbf{W} = \sum_{i=1}^{n_g} \sum_{g}^{G} (x_{ig} - \bar{x}_g)(x_{ig} - \bar{x}_g)'$.
8. Se multiplica la matriz $\mathbf{W}^{-1}$ por la matriz **B** para obtener la matriz $\mathbf{W}^{-1}\mathbf{B}$.

9. Se calculan los valores λ_i y vectores propios $\mathbf{v}_i$ más grandes de la matriz $W^{-1}B$. Los valores propios indican la variabilidad en los datos.

10. Se calcula la primera variable canónica definida por la combinación lineal del vector propio por el vector variable mediante $z = \sum \mathbf{v}_1 \mathbf{x}$.

11. Se pueden calcular dos o más funciones discriminantes con la restricción de no estar correlacionadas entre sí mediante la relación $\mathbf{v}_1' \mathbf{W} \mathbf{v}_2 = 0$.

12. Las coordenadas de los grupos se forman por parejas $(\mathbf{v}_1' \boldsymbol{\mu}_i, \mathbf{v}_{2i}' \boldsymbol{\mu}_i)$. La proyección de una nueva observación x sobre el plano tiene las coordenadas $(\mathbf{v}_1' \mathbf{x}, \mathbf{v}_2' \mathbf{x})$.

13. Los datos se proyectan en un diagrama de dispersión.

5.3.5. EJEMPLO

Ejemplo A. Lanoiselet *et al.* (2005) investigaron la utilidad de dos métodos de análisis de ácidos grasos (MIDI y MIDI modificado) para caracterizar y diferenciar aislados de *Rhizoctnia oryzae* y *R. oryzae-sativae* de cuatro países (Australia, Japón, Uruguay y EUA). Su principal objetivo fue evaluar y comparar los métodos MIDI y MIDI modificado para discriminar entre *R. oryzae* y *R. oryzae-sativae*.

Obtuvieron un total de 30 aislados de *R. oryzae* y *R. oryzae-sativum* de cuatro países y después procedieron a cultivarlas. Extrajeron los ácidos grasos y los sometieron a los métodos de análisis MIDI y MIDI modificado. Obtuvieron un total de 10 ácidos grasos por el método MIDI y un total de 11 ácidos grasos por el método MIDI modificado.

Usaron el análisis de variables canónicas por cada método para determinar las diferencias en la composición celular de ácidos grasos de *R. oryzae, R. oryzae-sativae* y cuatro aislados de *Rhizoctnia* sp de Uruguay. El AVC con el método MIDI reveló que, aunque *R. oryzae, R. oryzae-sativae* y cuatro aislados de *Rhizoctnia* sp de Uruguay formaron 3 grupos distintos, la

composición celular de ácidos grasos difirió entre sus respectivos aislados (figura 20-A).

El AVC con el método MIDI modificado reveló que la composición celular de ácidos grasos difirió poco entre aislados de las mismas especies pero tuvieron una gran diferencia entre especies (figura 20-B).

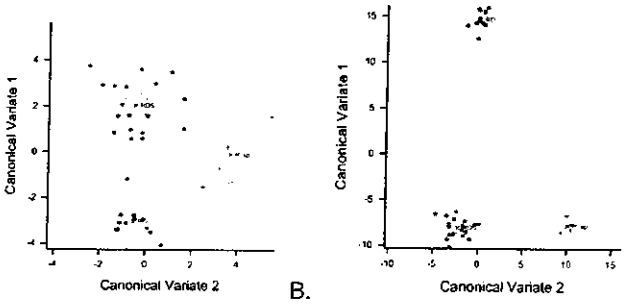

FIGURA 20. AVC por el método MIDI (A) y por el método de MIDI modificado (B). Tomado de Lanoiselet *et al.* (2005).

Se interpretó de la siguiente manera:

- Los dos métodos mostraron una gran diferencia entre grupos (especies) debido a que son especies diferentes y tienen diferencias cualitativas y cuantitativas en su composición de ácidos grasos.
- El método MIDI mostró una gran diferencia dentro de grupos, mientras que el método MIDI modificado mostró poca diferencia dentro de grupos. La diferencia fue debida al tipo de método usado.

5.4. ANÁLISIS DE REDUNDANCIA BASADO EN DISTANCIA (ADR bd)

5.4.1. DESCRIPCIÓN

Es un método multivariable no paramétrico para probar la significancia de los términos individuales en un modelo de análisis de varianza (ANVA) multifactorial y puede usar cualquier medida de distancia.

Este método usa una matriz de distancia (por ejemplo, distancia de Bray-Curtis) entre réplicas y lo somete a un análisis de coordenadas principales. Se crea una matriz de variables ficticias que corresponden al diseño del experimento (esto es, los términos individuales en un modelo lineal). Luego, se usa el análisis de redundancia para analizar la relación entre las coordenadas principales (datos de especies) y las variables ficticias (modelo, variables independientes). Por último implementa una prueba de permutación para estadísticos particulares que corresponden a los términos particulares en el modelo.

Esta información fue tomada de Legendre y Anderson (1999) y se recomienda consultarlos para una explicación más exhaustiva del método.

5.4.2. VENTAJAS

- Puede usar cualquier medida de distancia.
- Puede ser usado para probar los términos de interacción o cualquier término de un modelo lineal.
- Usa métodos de permutación no paramétricos y por consiguiente evade el supuesto de normalidad multivariable.

5.4.3. LIMITACIONES

- Según MacArdle y Anderson (2001) aumenta la suma de cuadrados total en el análisis. Además, la corrección de los valores propios negativos afecta los valores P (del proceso de permutación) en diseños multifactoriales.

5.4.4. ALGORITMO

Según Legendre y Anderson (1999) el algoritmo es:

1. Se fija la prueba de hipótesis.
2. Se fija el nivel de significancia.
3. Se calcula la matriz de distancia entre réplicas.
4. Se realiza un análisis de coordenadas principales.
5. Se calcula una matriz **Y** de variables respuesta cuyas columnas son las coordenadas principales.
6. Se calcula una matriz **X** de variables independientes que contiene las variables ficticias del modelo de ANVA.
7. Se realiza un análisis de redundancia sobre las matrices **Y** y **X** de la siguiente manera:
   - Cada vector de la matriz **Y** es regresado sobre la matriz **X**.
   - Se calcula la matriz **C** de los estimados de los coeficientes de regresión mediante $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{'}\mathbf{X}'\mathbf{Y}$.
   - Se calcula la matriz $\hat{\mathbf{Y}}$ de valores ajustados de la regresión mediante $\hat{\mathbf{Y}} = \mathbf{XB}$.
   - Se calcula la matriz de covarianzas $\Sigma_{\hat{y}}$ a partir de la matriz $\hat{\mathbf{Y}}$.
   - Se calculan los valores propios canónicos y los vectores propios mediante $(\Sigma_{\hat{S}} - \lambda\mathbf{I})\mathbf{v} = 0$.
8. Se calcula la suma de todos los valores propios de **Y** ($\sum \lambda_{\hat{Y}}$).

9. Se calcula la traza, la cual es igual a la suma de valores propios canónicos de **Y** sobre **X** (equivale a la suma de cuadrados de los tratamientos).

10. Se calcula h, el cual es igual al número de términos en el modelo.

11. Se calcula los grados de libertad para cada tratamiento (df_i), los cuales son el número de columnas en la matriz **X**.

12. Se calculan los grados de libertad del error (df_e) mediante $df_e = N - \left(\sum_{i=1}^{h} df_i\right) - 1$.

13. Se calcula la suma de cuadrados del error (SC_e) mediante $SC_e = \left(\sum \lambda_Y\right) - \sum_{i=1}^{h} traza_i$.

14. Se calcula el cuadrado medio de los tratamientos (CM_i) mediante $CM_i = \frac{traza_i}{df_i}$.

15. Se calcula el cuadrado medio del error (CM_e) mediante $CM_e = \frac{SC_e}{df_e}$.

16. Se calcula el estadístico F^* mediante $F^* = \frac{CM_i}{CM_e}$.

17. Se emplea una prueba de permutaciones.

5.4.5. EJEMPLO

Ejemplo A. Blackwood *et al.* (2003) buscaron un procedimiento óptimo para comparar perfiles T-RFLP de muestras ambientales. Sus objetivos fueron examinar las reglas para incluir picos en el análisis, determinar la diferencia (o distancia) entre perfiles, analizar las relaciones entre perfiles.

Se tomaron muestras de suelo de un cultivo de alfalfa. Se extrajo el ADN de las muestras para ser analizados por T-RFLP y se crearon perfiles moleculares. Se usó la distancia euclídea, la distancia de Hellinger

(mediante la transformación de Hellinger) y la distancia de Jaccard sobre los datos de perfiles moleculares.

Se aplicó el ADR basado en distancia para examinar si la variabilidad que puede ser atribuida a las diferencias entre perfiles es significativa. El ADR basado en distancia detectó diferencias significativas entre perfiles de comunidades con el análisis de variables transformadas (Hellinger) y el análisis de coordenadas principales con la distancia de Jaccard pero no con el análisis de las variables peso relativo de los picos (distancia euclídea) (tabla 21).

TABLA 21. Resultados del Análisis de redundancia basado en distancia para el conjunto de muestras de suelo de alfalfa. Valor de P< 0.01. Tomado de Blackwood *et al.* (2003).

Variable	Muestra	
	Proporción explicada (%)	Valor P
Peso relativo	30	0.12
Transformación Hellinger	31	0.0093
Distancia de Jaccard	32	0.0001

Se interpretó de la siguiente manera:

- La variabilidad significativa atribuida a las diferencias entre perfiles indicó que las comunidades microbianas de las muestras del suelo del cultivo de alfalfa fueron diferentes.

5.5. ANÁLISIS DE SIMILITUD (ANSIM)

5.5.1. DESCRIPCIÓN

Es un método no paramétrico que prueba las diferencias entre dos o más grupos. Está basado en cualquier medida de distancia (Ramette, 2007; Clarke y Gorley, 2001).

La hipótesis nula es que no hay diferencias entre los conjuntos de especies entre grupos de muestras (Clarke y Gorley, 2001).

Compara los rangos de distancia entre grupos con los rangos de distancia dentro de grupos. Por lo general se usa una matriz de disimilitud de Sorensen para calcular el estadístico R, el cual mide si existe una separación entre grupos (Ramette, 2007; Clarke y Gorley, 2001).

El estadístico R es $R = \frac{(r_b - r_w)}{(0.25(n(n-1)))}$. En donde r_b es la medida del rango de las disimilitudes para muestras dentro del mismo grupos, mientras que r_w es la medida de rangos de disimilitudes para muestras entre diferentes grupos.

Los valores de R son escalados en el rango de -1 a+1 (Ramette, 2007). Los valores cercanos a 1 indican una completa separación de los grupos de muestras, mientras que valores cercanos a cero indican que no existe separación entre grupos (Quin y Keough, 2002; Clarke y Gorley, 2001; Yannarell y Triplett, 2004; Ramette, 2007).

Una vez determinado el estadístico R, ANSIM asigna muestras de manera aleatoria a diferentes grupos para generar una distribución nula R y examinar si las muestras dentro de los grupos son más cercanos entre sí de lo esperado (Yannarell y Triplett, 2004).

Los valores mayores a 0.75 indican una excelente separación, los valores R mayor a 0.5 indican que los grupos están separados pero un valor menor a 0.25 indica solapamiento (Ramette, 2007; Clarke y Gorley, 2001).

Un requisito es que la dispersión dentro de los grupos de todos los grupos tiene que ser comparable (Ramette, 2007).

El programa PRIMER ejecuta el Análisis de similitud. Se puede consultar a Clarke (1993) para mayor información sobre el método.

5.5.2. VENTAJAS

- Permite comparar grupos sin cumplir con supuestos paramétricos.
- Se usa en conjunto con el escalado multidimensional no métrico (Clarke y Gorley, 2001).

5.5.3. LIMITACIONES

- No es útil para modelos con términos de interacción debido a que la hipótesis de interacción no pueden ser expresados en términos de reasignación aleatoria (Quin y Keough, 2002).
- Sólo usa los rangos de las disimilitudes.
- Sólo compara matrices de distancia o de disimilitud.

5.5.4. EJEMPLO

Ejemplo A. Yannarell y Triplett (2004) describieron los cambios en la composición de la comunidad bacteriana a varias escalas espaciales dentro de los lagos y entre lagos. Sus objetivos fueron determinar la variación en la composición de la comunidad bacteriana a diferentes escalas espaciales dentro y entre lagos, además de determinar qué condiciones ambientales provocan comunidades bacterianas similares a gran escala.

Se distribuyeron 90 estaciones a través de todos los lagos para investigar la variabilidad en las 4 escalas diferentes (niveles de estación, base, lago y regional). Luego, se tomaron muestras de agua de 13 lagos localizados en el norte y sur de Winsconsin y además se registraron los parámetros fisicoquímicos del agua. Se extrajo el ADN de las muestras de agua y después se analizaron por el método ARISA (*Automated Ribosomal Intergenic Spacer Analysis*) para construir perfiles moleculares de las comunidades bacterianas. Con los resultados se creó una matriz con datos

de presencia-ausencia de perfiles ARISA y después utilizaron el coeficiente de Sorensen para analizar la similitud de los perfiles ARISA de diferentes comunidades.

Se usó el análisis de similitud para probar la hipótesis de que la similitud en los perfiles ARISA dentro de grupos es más grande que la similitud entre grupos. Los resultados del ANSIM revelaron que no hubo diferencias significativas en la composición de las comunidades bacterianas en las escalas base y lago, pero no detectó diferencias en la composición bacteriana en las escalas estación o región (tabla 22).

TABLA 22. Resultados del análisis de similitud sobre las 4 escalas. Tomado de Yannarell y Triplett (2004).

Escala	Estadístico R	No. Permutaciones de Monte Carlo con puntajes ≥ R	Valor P
Estaciones (anidado en bases)	-0.13	999	1.000
Bases (anidados en lagos)	0.231	0	0.001
Lagos (con bases anidadas)	1.0	0	.0001
Regiones (con lagos anidados)	-0.066	693	0.694

Los resultados se pueden interpretar de la siguiente manera:

- Los valores R de las escalas estaciones y regiones tuvieron valores negativos, lo que indica valores cercanos a cero y por consiguiente indica que no hubo separación entre grupos. En cambio, el valor R de la base es cercano a uno y el valor de la escala lago es igual a 1 lo que indica que hubo una separación completa de los grupos.
- Las escalas de estaciones y regiones mostraron una gran variabilidad dentro de los grupos (un valor R negativo; es decir, las comunidades microbianas dentro de los grupos fueron muy diferentes entre sí.

5.6. PRUEBA DE MANTEL

5.6.1. DESCRIPCIÓN

Compara dos matrices basadas sobre dos conjuntos de datos independientes, los cuales contienen los mismos objetos o muestras (Ramette, 2007; Legendre y Legendre, 1998; Ritchie *et al.*,2000).

La hipótesis nula es que no existe correlación entre la matriz de distancia **A** y la matriz de distancia **B** (Ritchie *et al.*, 2000).

Se calcula el coeficiente de correlación entre las posiciones correspondientes en las dos matrices y luego se realiza la prueba de hipótesis por un procedimiento de aleatorización en el cual el valor original del estadístico es comparado con la distribución encontrada al reasignar de manera aleatoria el orden de los elementos en una de las matrices (Ramette, 2007; van Gremberghe *et al.*, 2007).

5.6.2. VENTAJAS

- La mayoría de las matrices ecológicas tienen los mismos objetos, lo que permite determinar el grado dependencia lineal entre ambas matrices.

5.6.3. LIMITACIONES

- No es útil con términos de interacción debido a que la hipótesis de interacción no puede ser expresada en términos de una reasignación aleatoria (Quin y Keough, 2002).
- Sólo compra matrices que contienen las mismas muestras u objetos.

5.6.4. EJEMPLO

Ejemplo A. Ritchie *et al.* (2000) estudiaron la composición de comunidades bacterianas del suelo mediante el uso de los métodos FAME (*Fatty Acid Methyl Ester*) y LH- PCR (*Length Heterogeneity Polymerase*

Chain Reaction). Sus objetivos fueron determinar la conveniencia y reproducibilidad del método LH-PCR para medir la composición de la comunidad microbiana del suelo y comprar el método LH- PCR con el método FAME.

Colectaron muestras de suelo de 4 tipos de campos en el Valle Oregon con 4 cuadrantes como réplicas cada uno. Extrajeron el ADN y los ácidos grasos del suelo para ser analizado por LH- PCR y FAME, respectivamente. Se obtuvieron picos relativos de la longitud del fragmento con respecto a la intensidad de la fluorescencia y se creó una matriz de perfiles LH-PCR. También se obtuvieron picos relativos del método FAME (*Fatty Acid Methyl Esther*) y se creó una matriz de perfiles FAME.

Se usó la prueba de Mantel con el coeficiente de Sorensen para examinar la significancia de la correlación entre la estructura de la comunidad basado en FAME y la estructura de la comunidad basado en LH-PCR (tabla no mostrada). Además usaron el procedimiento de aleatorización (Monte Carlo) con 1000 corridas aleatorias.

Encontraron que varias correlaciones significativas entre 19 de los fragmentos LH- PCR y 22 de los ácidos grasos extraídos. La mayoría (66%) de esas correlaciones fueron positivas. En general, se encontró una gran proporción de correlaciones positivas (74%) entre los fragmentos LH-PCR y ácidos grasos asociados con bacterias.

Se interpretó de la siguiente manera:

- La existencia de correlación indicó que existe una dependencia lineal entre la estructura de la comunidad determinada por FAME y la estructura determinada por LH-PCR. Por consiguiente, ambos métodos determinaron la misma estructura de la comunidad microbiana a cierto nivel.

Ejemplo B. van Gremberghe *et al.* (2007) estudiaron la relación entre la composición de la comunidad cianobacteriana y de la comunidad zooplánctico en el lago Blaarmeersen (Bélgica). Su principal objetivo fue evaluar el impacto de los factores bióticos y abióticos sobre los cambios estacionales en la comunidad cianobacteriana.

Se tomaron muestras de agua del lago Blaarmeersen a dos profundidades diferentes (0.5 y 7.5 m) durante dos años (2003 y 2004). Los periodos de muestreo se dividieron dos estaciones: estación de crecimiento (final de la primavera y a principio de verano y otoño) y estación de invierno (final de otoño, invierno y a principio de primavera). Se midió la temperatura, el pH, la transparencia del agua, la conductividad, la concentración de oxígeno y la concentración de nutrientes (nitrato, amonio y ortofosfato). Se identificaron las especies de cianobacterias, fitoplancton y zooplancton; además se midió su biomasa. Se extrajo el ADN de las muestras y se sometió a un análisis DGGE (*Denaturing Gradient Gel Electrophoresis*). Se creó una matriz de intensidad de bandas. La intensidad de bandas fue convertida a intensidad relativa (esto es, la contribución relativa de cada banda al total de bandas en cada columna del gel). Los datos fueron transformados mediante log (x+1). Se usó la medida de Bray-Curtis para crear una matriz de similitud.

Se usó la prueba de Mantel para investigar la relación entre la composición de la comunidad cianobacteriana y las variables bióticas (composición de fitoplancton y de zooplancton) y abióticas.

Los resultados de la prueba de Mantel se muestran en la tabla 23. El método reveló que la composición de la comunidad cianobacteriana tuvo una correlación significativa con la composición de fitoplancton y de zooplancton; además de una correlación negativa con las variables abióticas (concentración de nutrientes y variables físicas). También se mostró una correlación significativa entre las variables bióticas y abióticas.

TABLA 23. Resultados de la prueba de Mantel que relaciona la comunidad cianobacteriana con las variables bióticas y abióticas. Se muestran los coeficientes de correlación de Pearson. Los valores entre paréntesis indican los valores P de significancia. Tomado de van Gremberghe *et al.* (2007).

	Comunidad cianobacteriana	Comunidad de zooplancton	Comunidad de fitoplancton	Nutrientes	Variables físicas
Comunidad cianobacteriana	--------	0.306 (P < 0.0001)	0.218 (P < 0.001)	-0.335 (P < 0.0001)	-0.280 (P < 0.0001)
Comunidad de zooplancton		------	0.341 (P < 0.0001)	-0.272 (P < 0.0001)	-0.135 (P < 0.001)
Comunidad de fitoplancton			------	-0.275 (P < 0.0001)	-0.220 (P < 0.0001)
Nutrientes				------	0.302 (P < 0.0001)

Se interpretó de la siguiente manera:

- La correlación positiva entre la comunidad cianobacteriana y las variables bióticas indicó que un aumento en la comunidad cianobacteriana estimuló el crecimiento de las especies de fitoplancton y de zooplancton.
- La comunidad cianobacteriana, de fitoplancton y de zooplancton tuvieron una correlación negativa con las variables abióticas. Esto indicó que un aumento en las comunidades disminuyó la concentración de nutrientes y afectaron los parámetros físicos.

5.6 CONCLUSIONES

1. En todos los métodos se hicieron contrastes de hipótesis.
2. El análisis de varianza multivariable, el análisis discriminante y el análisis de variables canónicas son métodos paramétricos que crearon combinaciones lineales, las cuales contenían la información en sus coeficientes (vectores propios) para maximizar la variación entre grupos.
3. El análisis de similitud (ANSIM) es un método no paramétrico que usó los rangos de las disimilitudes para encontrar la separación entre los grupos.
4. El análisis de redundancia basado en distancia tiene la ventaja de poder usar cualquier medida de distancia y puede probar la significancia de los términos de interacción de un modelo lineal.
5. Se usaron perfiles moleculares para conocer si las comunidades microbianas de los grupos difirieron entre grupos y dentro de grupos (Park *et al.*, 2006; Lanoiselet *et al.*, 2005; Yannarell y Triplett, 2004).
6. Se usó el análisis de variables canónicas (Yannarell y Triplett, 2004) y la Prueba de Mantel (Ritchie *et al.*, 2000) para comparar la eficiencia de los métodos moleculares en la determinación de la estructura de la comunidad.
7. Una alta variación entre grupos indicó que los elementos de un grupo fueron diferentes con respecto a los de otro grupo. En cambio, una baja variación dentro de grupos indicó que los elementos dentro de cada grupo fueron más similares en sus atributos.

5.7 REFERENCIAS

Blackwood, C. B., T. Marsh, S. H. Kim y E. A. Paul (2003). *Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities.* Applied and Environmental Microbiology, 69 (2): 926-932.

Clarke, K. R. y R. N. Gorley (2001). *PRIMER v5: User manual/tutorial.* PRIMER-E Ltd. Reino Unido.

Digby, P. G. N. y R. A. Kempton (1987). *Multivariate analysis of ecological communities.* Chapman and Hall. EUA.

Hair, J. F. J., R. E. Anderson, R. L. Tatham y W. C. Black (1999). *Análisis multivariante.* 5a. edición, Prentice Hall, España.

Jongman, R. G., C. J. F. Ter Braak y O. F. R. Van Tongeren (1995). *Data analysis in community and landscape ecology.* Cambridge University Press. Reino Unido.

Jonson, D. E. (2000). *Metodos multivariadas aplicados al análisis de datos.* Internacional Thompson Editores. México.

Kaneene, J. B., R. A. Miller, R. Sayah, Y. J. Johnson, D. Gilliland y J. C. Gardiner (2007). *Considerations when using Discriminant Function Analysis of antimicrobial resistance profiles to identify sources of fecal contamination of surface water in Michigan.* Applied and Environmental Microbiology, 73 (9): 2878-2890.

Langenheder, S., E. S. Lindstrom y L. J. Tranvik (2005). *Weak coupling between community composition and functioning of aquatic bacteria.* Limnological Oceanograghic, 50(3): 957-967.

Lanoiselet, V. M., E. J. Cother, N. J. Cother, G. J. Ash y J. D. I. Harper (2005). *Comparision of two total cellular fatty acid analysis protocols to differentiate Rhizoctonia oryzae and R. oryzae sativae.* Mycologia, 97 (1): 77-83.

Legendre, P. y L. Legendre (1998). *Numerical ecology.* Elsevier Science. Amsterdam. 853pp

Legendre, P. y M. J. Anderson (1999). *Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments.* Ecological monographs. 69(1): 1-24

Mantel, N. (1967). *The detection of disease clustering and a generalized regression approach.* Cancer research. 27(1): 209-220.

McArdle, B. H. y M. J. Anderson (2001). *Fitting multivariate models to community data: A comment on distance-based redundance analysis.* Ecology. 82(1): 290-297

Park S., Y. K. Ku, M. J. Seob, D. Y. Kim, J. E. Yeon, K. M. Lee, S. C. Jeong, W. K. Yohh, C. H. Hark y H. M. Kim (2006). *Principal component analysis and discriminant analysis (PCA-DA) for discriminating profiles or terminal restriction fragment length polymorphism (T-RFLP) in soil bacterial communities.* Soil Biology Biochemistry, 38: 2344-2349.

Peña, D. (2002). *Análisis de datos multivariantes.* McGraw Hill. España.

Quin G. P. y M. J. Keough (2002). *Experimental design and data analysis of biologist.* Cambridge. EUA.

Ramette, A. (2007). *Multivariate analyses in microbial ecology.* FEMS Microbial Ecology: 1-19.

Ritchie, N. J., M. E. Schutter, R. P. Dick y D. D. Myrold (2000). *Use of Length Heterogeneity PCR and Fatty Acid Methyl Ester Profiles to characterize microbial communities in soil.* Applied and Environmental microbiology, 66(4): 1668-1675.

Van Gremberghe, I., J. Van Wichelen, K. Van der Gucht, P. Vanormelingen, S. D'hondt, C. Buhote, A. Wilmotte y W. Vyverman (2007). *Covariation between zooplancton communitu composition and cyanobacterial communitu dynamics in Lake Blaarmeersen (Belgium).* FEMS Microbial Ecology, 63: 222-237.

Yannarell, A. C. y E.W. Triplett (2004). *Within and between lake variability in the composition of bacterioplankton communities: Investigations using multiple spatial scales.* Applied and Environmental microbiology, 7 (1): 214-233.

6. ANEXOS

6.1. SIMBOLOGIA

- Las matrices se representan por letras mayúsculas en negrita **X**.
- Los vectores se representan por letras minúsculas en negrita **x**.
- Los elementos de las matrices se representan por letras minúsculas en cursiva *x*.
- **n** se refiere a elementos
- **p** se refiere a variables
- **i** se refiere a la fila
- **j** se refiere a la columna.
- **g** se refiere a grupos
- **k** se refiere a especies
- **m** se refiere a muestra
- **d** se refiere a distancia
- **δ** se refiere a disimilitud.
- **SCPC** se refiere a la suma de cuadrados y productos cruzados
- **W** es la matriz de SCPC dentro de los grupos
- **B** es la matriz de SCPC entre grupos.
- **T** es la matriz de SCPC total
- **Y** es la matriz de especies x muestras.
- **X** es la matriz de variables x muestras o bien una matriz general.
- **A** es la matriz de datos transformados
- **D** es una matriz de distancias o en algunos casos denotará una matriz diagonal.
- **Σ** es una matriz de varianzas covarianzas.
- **R** es una matriz de correlación.
- **Λ** es una matriz de valores propios en la diagonal.

- **I** es una matriz identidad.
- **U** es la matriz de vectores propios de una matriz
- **V** es la matriz de vectores propios de la segunda matriz
- **Z** es la matriz de puntajes de los ejes.

6.2. MATRICES

- Escalar: es un número real (sin indicar sentido ni dirección).
- Vector: un conjunto de n datos de una variable; puede representarse geométricamente asociando cada valor de la variable a una dimensión del espacio n dimensional, obteniendo un punto en ese espacio y también la línea que une el origen con dicho punto. Es un segmento orientado que une el origen de coordenadas con el punto x.
- Matriz: conjunto de números dispuestos en filas y columnas y puede verse como un conjunto de vectores columna o un conjunto de vectores fila.
- R^n: es el espacio de todos los vectores de n coordenadas o componentes.

Normalización y ortogonalización de vectores

- La norma de un vector **x** (la longitud del vector) se define por $\|\mathbf{x}\| = \sqrt{(\mathbf{x}'\mathbf{x})}$. Ésta mide la distancia de **x** al origen.
- El ángulo entre **x** e **y** se expresa por $\cos\theta = \dfrac{\mathbf{x}'\mathbf{y}}{(\|\mathbf{x}\|)(\|\mathbf{y}\|)}$.
- La distancia entre dos vectores **x** e **y** es $\|\mathbf{x}-\mathbf{y}\|$.
- La proyección de **x** sobre **y** es $\left(\dfrac{\mathbf{x}'\mathbf{y}}{\|\mathbf{y}\|}\right)(\mathbf{y}) = (\|\mathbf{x}\|\cos\theta)(\mathbf{y})$.
- La ortogonalización (esto es, dos vectores son perpendiculares entre sí, además de ser independientes) de un vector . : =. .

Tipos de matrices

- Matriz singular: son aquellas que su determinante es nulo y no tienen inversa; es decir, son matrices que tienen una fila o columna de ceros.
- Matriz transpuesta **X'**: dada una matriz **X** de n x p, su transpuesta es la que resulta de cambiar filas por columnas y viceversa **X'** de p x n.
- Una matriz simétrica **X**: si **X** =**X'**.
- Una matriz antisimétrica: - **X'**= **X**.
- Una matriz es cuadrada si todos tiene el mismo número de filas y de columnas **X** de n x n.
- Una matriz diagonal D es aquella que presenta números en su diagonal y ceros fuera de ella; un ejemplo de matriz de este tipo es la matriz identidad, la cual presenta unos en su diagonal.
- Se dice que una matriz simétrica **X** es: a) positiva definida si $\mathbf{x'Xx} > 0$, para toda x≠0; b) positiva semidefinida si $\mathbf{x'Xx} \geq 0$. para toda $x \neq 0$; c) no negativa si es positiva definida o semidefinida.
- Una matriz **X** es positiva definida si y sólo si existe una matriz no singular **Q** tal que **X=QQ'**.
- Matrices ortogonales (perpendiculares entre sí): se dice que una matriz cuadrada **X** es ortogonal si **XX'=I = X' X**. Luego, **|X|** es igual a 1 o a -1.

Álgebra de matrices

- Suma: Se define sólo cuando ambas matrices tienen las mismas dimensiones. Cada elemento de la matriz suma se obtiene sumando los elementos correspondientes a los sumandos (**A + B = C**). La suma es conmutativa.
- Multiplicación: Para multiplicar dos matrices, el número de columnas en la primera deber ser igual al número de filas en la segunda. Cuando todos los productos de matrices están definidos, se cumple la ley asociativa para la multiplicación de las mismas (**A (BC) = (AB) C**).

La multiplicación de matrices no siempre se conmutativa ($\mathbf{AB} \neq \mathbf{BA}$), pero si es distributiva. El elemento de la primera fila de la matriz **A** se multiplica por el elemento de la primera columna de **B** y luego se suma al producto del primer elemento de **A** por el segundo elemento de **B** y así sucesivamente para obtener el primer elemento del producto de la matriz **AB**.

- Inversa de una matriz ($\mathbf{X}^{-1}$): en álgebra lineal no existe el concepto de división, en lugar de ello, uno multiplica una matriz por el inverso de otra (por ejemplo, $\mathbf{C}=\mathbf{A}\mathbf{B}^{-1}$). El cálculo de la inversa de una matriz es un proceso iterativo que esta fuera del objetivo de este trabajo (Para saber sobre su forma de calcularse puede consultar libros de álgebra lineal).

Funciones escalares

- La traza de una matriz es la suma de los elementos de la diagonal de matriz; se representa por tr (**X**). La traza es una medida de la variabilidad del conjunto de datos.
- El determinante de una matriz **X** es el escalar resultante de multiplicar todos los términos diagonales de la matriz. Se representa como $|\mathbf{X}|$. El determinante es una medida de la dependencia lineal del conjunto de variables.
- El rango de una matriz se define como el número máximo de filas (o columnas) en **X** que son linealmente independientes. De modo equivalente, el rango de **X** es la dimensión del subespacio vectorial generado por las filas (o columnas) de la matriz **X**.
- Se dice que un conjunto de vectores ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$) son linealmente dependientes si existen las constantes ($\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_p$) las cuales no son todas iguales a cero, tales que $\sum c_i x_i = 0$. De lo contrario, se dice que los vectores son linealmente independientes. Un conjunto de vectores es linealmente dependiente si y sólo si por lo menos uno de los

vectores se puede escribir como una combinación lineal de los restantes.

Forma cuadrática

Una forma cuadrática en p variables, $x_1, x_2, \ldots, x_p$, es una función de la forma . ! para alguna matriz simétrica **X**; **x** es un vector.

Valores y vectores propios

Los valores y vectores propios de una matriz son funciones especiales de los elementos de esta última que desempeñan un papel en extremo importante en muchas técnicas del análisis multivariable.

En cada uno de los incisos siguientes sea **X** una matriz simétrica de p x p.

Los valores propios (también llamados raíces características o raíces latentes) de **X** son las raíces de la ecuación polinomial mediante ¦. $-\lambda$¦ =. ; nótese que esta ecuación en términos de determinantes es una ecuación polinomial en λ, de p-ésimo grado.

A cada valor propio de **X** le corresponde un vector diferente de cero llamado vector propio (también conocido como vector característico o latente) que satisface $\mathbf{X}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, para i = 1, 2, ..., p.

Si **X** es una matriz simétrica de números reales, entonces sus valores propios y vectores propios también consistirán en números reales.

Los valores propios de **X** se denotan por $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Los vectores propios no son únicos, de modo que a menudo se normalizan de tal forma que $\mathbf{v}_i' \mathbf{v}_i = 1$.

Cuando dos valores propios no son iguales, sus vectores propios correspondientes serán ortogonales. Cuando dos o más valores propios de

X son iguales, se pueden elegir los vectores propios correspondientes de modo que sean ortogonales entre sí, y siempre se hará así.

La traza de una matriz simétrica es igual a la suma de sus valores propios, es decir, $tr(\mathbf{X}) = \sum \lambda_i$.

El determinante de una matriz simétrica siempre es igual al producto de sus valores propios, es decir, $|\mathbf{X}| = \prod \lambda_i$.

Una matriz simétrica **X** es positiva definida si y sólo si $\lambda_i > 0$, para cada i.

Una matriz simétrica **X** es positiva semidefinida si y sólo si $\lambda_i \geq 0$, para cada i.

Si **X** es una matriz no negativa de rango m, entonces habrá exactamente m valores propios diferentes de cero.

Si **X** es una matriz simétrica, existe una matriz ortogonal, V, tal que $\mathbf{V}'\mathbf{X}\mathbf{V} = \mathbf{\Lambda}$, en donde **Λ** es un matriz diagonal. Además, los elementos en la diagonal de **Λ** son los valores propios de **X** y las columnas de **V** son sus vectores propios correspondientes. Por tanto,

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \quad \text{y} \quad \mathbf{V} = \begin{bmatrix} v_1 & v_2 & \cdots & v_p \end{bmatrix}.$$

El resultado anterior implica que $\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' = \sum \lambda_i v_i v_i'$; esto se conoce como descomposición espectral de **X**. Donde **V** es una matriz ortogonal de orden n (sus columnas son los vectores propios de **X**) y **Λ** es la matriz diagonal con los elementos diagonales λ_i (valores propios de **X**).

Descomposición del valor singular $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$; donde **U** tiene el mismo tamaño que **X** (n x p), y **D** y **V** son matrices cuadradas de orden p. Las matrices tienen las siguientes propiedades: las columnas de **U** son ortonormales, la matriz **V** es ortogonal, la matriz **D** es diagonal y sus valores diagonales son no-negativos y acomodados en orden descendiente de magnitud (sus valores son llamados valores singulares de **X**).

6.3. MATRIZ DE VARIANZA-COVARIANZA Y MATRIZ DE CORRELACIÓN

Matriz de varianza-covarianza

Para una variable multivariable se define la matriz de varianzas y covarianzas mediante la siguiente fórmula:

$$\Sigma = (1/n)\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})'$$

Es una matriz cuadrada y simétrica que contiene en la diagonal las varianzas y fuera de la diagonal las covarianzas entre las variables. La covarianza mide la relación lineal entre dos variables. La varianza mide la variabilidad de los datos respecto a la media. Este tipo de matriz se utiliza cuando los datos se miden en unidades comparables.

Matriz de correlación

La dependencia lineal entre dos variables se estudia mediante el coeficiente de correlación lineal o simple (rij), el cual se calcula de la siguiente manera:

$$r_{ij} = \frac{s_{ij}}{s_i s_j}$$

La dependencia por pares entre las variables se mide por la matriz de correlación (R). Llamaremos matriz de correlación (R) a la matriz cuadrada y simétrica que tiene unos en la diagonal y fuera de ella los coeficientes de correlación lineal entre pares de variables. Esta matriz es positiva semidefinida. Este tipo de matrices se utiliza cuando los datos se miden en diferentes unidades. La fórmula para calcularse es:

$$\mathbf{R} = \mathbf{D}^{-1/2}\Sigma\mathbf{D}^{-1/2}$$

En este caso D-1/2 es una matriz diagonal que contiene las desviaciones típicas y la matriz R está relacionada con la matriz de varianzas y covarianzas Σ.

7. GLOSARIO

Aglomerativo: empieza con cada objeto en un grupo y luego fusiona los grupos en base a una medida de similitud o de distancia hasta que todos los objetos están en un grupo.

Algoritmo: es un conjunto de reglas o instrucciones para efectuar algún cálculo.

Análisis de gradientes: se refiere a métodos de análisis de datos que relacionan la composición de la comunidad en términos de la respuesta de las especies a las variables ambientales.

Análisis multivariable: es un conjunto de métodos matemáticos que analizan varias variables de manera simultánea.

Análisis propio: es una operación matemática sobre una matriz cuadrada y simétrica que calcula una serie de valores propios y vectores propios.

Biplot: es un diagrama de ordenación, cuyo prefijo "bi" se refiere a la representación conjunta de sitios (o muestras) y especies, representa los sitios como puntos y las especies como vectores (flechas). Las variables ambientales son representadas como vectores.

Cargas de los componentes: son correlaciones entre los puntajes de los componentes y las variables originales.

Centroide: es la media (o centro) de un grupo de objetos.

Coeficiente de correlación cofenética: es una medida de la distorsión del agrupamiento jerárquico y se calcula como la correlación entre la matriz de similitud original y la matriz cofenética.

Coeficientes canónicos: coeficientes que definen los ejes de ordenación como combinaciones lineales de las variables ambientales.

Correlación canónica: correlación entre la combinación lineal de la matriz **Y** y la combinación lineal de la matriz **X**.

Correlación especies-ambiente: correlación entre los puntajes de sitios que son sumas (o promedios) ponderados de los puntajes de especies y los puntajes de sitios que son una combinación lineal de las variables ambientales.

Correlación interconjuntos: coeficientes de correlación entre las variables ambientales y los puntajes de sitios que son derivados de los datos de especies.

Correlación intraconjuntos: coeficientes de correlación entre las variables ambientales y los ejes de ordenación.

Datos atípicos: datos de la muestra que no parecen ser coherentes con la mayoría de los datos.

Desigualdad triangular: la suma de las longitudes de dos lados cualesquiera del triángulo formado por los tres puntos debe ser siempre mayor que el tercer lado.

Diagrama conjunto (*jointplot*): es un diagrama de ordenación en donde tanto las especies como los sitios son representados por puntos. Las variables ambientales son representadas por vectores (flechas).

Diagrama de Shepard: es un diagrama de dispersión en donde el eje horizontal representa la disimilitud y el eje vertical representa la distancia.

Dimensiones: son las características de un objeto.

Distancia de Mahalanobis: es una medida de distancia entre variables correlacionadas que estandariza las variables y es invariante ante cambios de escala.

Distancia de Manhattan: es la suma de las diferencias de sus correspondientes componentes y es una distancia entre dos puntos medida a lo largo de ejes en ángulo recto.

Distancia euclídea: es la distancia más corta entre dos puntos y es la ecuación del teorema de Pitágoras.

Distancia X^2: esta medida tiene una parte de la distancia euclídea computada sobre abundancias relativas ponderado por la inversa de la suma de especies.

Distancia: contrario a similitud, es una función matemática que mide qué tan diferentes son dos objetos con respecto a los atributos observados y cumple con al menos las primeras tres condiciones siguientes: positividad, simetría, tomar valor igual a cero si es medida sobre sí misma y desigualdad triangular.

Distribución normal multivariable: es la generalización de la distribución normal univariable para el caso de p-variables.

Distribución normal: es una distribución de probabilidad continua teórica en el cual el eje horizontal representa todos los posibles valores de una variable y el eje vertical representa la probabilidad de que ocurran esos valores. Los valores de la variable están agrupados en torno a la media de forma simétrica y unimodal.

Divisivo: empieza con todos los objetos en un solo grupo y después es dividido en dos grupos más pequeños, lo cual es repetido de manera iterativa hasta que el último objeto está en un grupo.

Efecto de arco: el diagrama de ordenación toma la forma de un arco debido a que el segundo eje (y posteriores ejes) son funciones polinomiales del primer eje. Esto provoca que los sitos en los extremos del primer eje estén más cercanos entre sí que en el medio del eje.

Estandarización: es un proceso de transformación para hacer comparables los datos de una matriz cuyos elementos se expresan en unidades diferentes (o de diferente escala).

Estrés: es una medida de la desviación de monotonicidad.

Fenético: cualquier tipo de caracter utilizable en la clasificación incluyendo las morfológicas, fisiológicas, ecológicas, moleculares, anatómicas, citológicas y otros.

Filogenia: es la historia evolutiva de los seres vivos.

Función discriminante lineal: es una función lineal de los elementos en **x** que resume toda la información contenida en este vector, de la que se dispone para realizar una discriminación efectiva entre dos (o más) poblaciones.

Funciones de resemblanza: son relaciones cuantitativas del grado de similitud o disimilitud entre dos objetos (muestras o sitios) en base a las observaciones sobre un conjunto de descriptores (especies o variables).

Gradiente compuesto: es una combinación lineal de variables ambientales medidas o una variable teórica.

Gradiente: es el cambio gradual en los valores de alguna variable ambiental.

Heterocedasticidad: cuando el término de error tiene una varianza en aumento u ondulante.

Homocedasticidad: descripción de los datos en los que la variación del término error aparece constante sobre un rango de variables independientes.

Ji-cuadrado: es un método de estandarización de datos en una tabla de contingencia en el que se compara la frecuencia de la celda observada con la frecuencia de la celda esperada.

Linealidad: es utilizada para expresar el concepto de que el modelo posee las propiedades de aditividad y homogeneidad; es decir, presenta relaciones lineales entre las variables.

Métrica de Canberra: es una sumatoria de series de proporciones entre los correspondientes valores, y toma en cuenta la distancia entre dos puntos y su relación con el origen.

Mínimos cuadrados: procedimiento que estima los coeficiente de regresión para disminuir la suma de cuadrados de residuos total.

Monotético: divide un conjunto de muestras de acuerdo a la presencia-ausencia de una sola especie.

Monotónico: es una curva de respuesta en donde los valores de la variable respuesta aumentan (o disminuyen) cuando la variable explicatoria aumenta.

Normalización: es una transformación de los datos para ajustarlos a una distribución normal.

Ordenación: es un término colectivo para designar a métodos multivariables que acomodan sitios (o muestras) a lo largo de ejes en base a datos de composición de especies.

Ortogonalidad: son vectores o ejes que son perpendiculares entre sí y significa que son independientes y no representan una relación lineal entre sí.

Politético: considera toda la composición de especies en las muestras durante el proceso de clasificación.

Ponderación: es dar más o menos peso (o importancia) a las variables (o a las especies).

Positividad: la distancia entre dos puntos es un número no negativo.

Propiedad euclídea: es cuando las distancias pueden originarse como una línea recta entre un conjunto de puntos en un espacio euclídeo.

Propiedad métrica: es cuando una distancia cumple con la condición de desigualdad triangular, además de las condiciones de positividad, simetría y tomar valor de cero si es medido sobre sí mismo.

Redundancia: las muestras son más parecidas a otras en su composición de especies y muchas especies se parecen a otras en sus ocurrencias en las muestras.

Residual: es la diferencia entre los valores observados y los valores esperados o predichos y representa el error del modelo.

Robusto: es la flexibilidad de algunas técnicas o métodos para funcionar aún bajo violaciones a algunos de sus supuestos o bien es poco restrictivo en sus supuestos.

Ruido: significa que muestras con condiciones ambientales similares no son idénticas en composición de especies.

Simetría: significa que la distancia entre la i-ésima y la j-ésima unidad es la misma respecto a la j-ésima y la i-ésima unidad.

Similitud: es una medida de qué tan similares (o parecidos) son dos muestras en términos de composición de especies o bien dos objetos en término de sus atributos.

Singularidad: es un caso en el que se puede predecir una variable independiente por una o más variables independientes.

Transformación: consiste en el cambio de los valores de una variable con alguna característica no deseada (tal como no normalidad) mediante una relación matemática en otros valores con la característica deseada.

Unidad taxonómica operativa (UTO): es una unidad a clasificar como puede ser especie o género.

Unimodal: es una curva de respuesta en donde los valores de la variable respuesta incrementan con la variable explicatoria, alcanza un máximo y luego decrece. Tiene forma de campana.

Valor P: es la probabilidad de observar los datos u observar los efectos cuando la hipótesis nula es verdadera, la hipótesis menciona que no hay influencia sistemática. Su rango es de cero a uno. Un valor cercano a 1 indica que el efecto no es significativo. En cambio, un valor cercano a cero indica que un efecto significativo es debido a algún tipo de influencia sistemática.

Valores propios: son medidas básicas del tamaño de una matriz que no se ven alterados si se hace un cambio de coordenadas que equivale a una rotación de los ejes y se interpretan como una medida de la cantidad de variación a lo largo de un eje. En un sentido matemático, son las raíces características de la ecuación polinomial definida por $|\Sigma = \lambda \mathbf{I}| = 0$.

Variables cualitativas: variables cuyo valor es un atributo o categoría.

Variables cuantitativas: variables cuyo valor se expresa de manera numérica.

Variables ficticias (*dummy*): variables dicotómicas que representan la categoría de una variable independiente cualitativa.

Vectores propios: son vectores que representan las direcciones características de la matriz y no modifican su posición en el espacio, solo pueden cambiar de norma. En el análisis multivariable son los puntajes de las variables, muestras o especies. En un sentido matemático, es un vector no cero que satisface la ecuación matricial $\Sigma a = \lambda a$.

ANÁLISIS MULTIVARIABLE EN ECOLOGÍA MICROBIANA

PRESENTA: ELVIS GIOVANNI EZEQUIEL GUZMÁN ORNELAS

FE DE ERRATAS

PÁGINA (PÁRRAFO)	ERRATA	CORREGIDO
Varias	Jonson	Johnson
6 (2)	los ___ del análisis multivariable son	los objetivos del análisis multivariable son
10 (3)	1.6 Objetivos	1.5 Objetivos
11 (1)	1.7 Referencias	1.6 Referencias
24 (1)	análisis de redundancia (sección 3.10)	análisis de redundancia (sección 3.9)
112 (tabla 17)	$V=tr\left(B(B+E)^{-1}\right)$	$V=tr\left(\mathrm{B}(\mathrm{B}+\mathrm{W})^{-1}\right)$
112 (tabla 17)	$T=tr\left(W^{-1}E\right)$	$T=tr\left(\mathrm{W}^{-1}\mathrm{B}\right)$
117 (7)	Kaneene *et al.*, XXXX	Kaneene *et al.*, 2007
143	**R** es una matriz de correlación	**P** es una matriz de correlación
148 (2)	es una función de la forma ____ para alguna matriz simétrica **X**;	es una función de la forma $x'Xx$ para alguna matriz simétrica **X**;
148 (5)	ecuación polinomial mediante ____;	ecuación polinomial mediante $\lvert\mathbf{X}-\mathrm{I}\lambda\rvert=0$;
148 (6)	que satisface ____, para	que satisface $\mathbf{X}\mathbf{v}_i=\lambda_i\mathbf{v}_i$, para
148 (9)	de tal forma que ____.	de tal forma que $\mathbf{v}_i'\mathbf{v}_i=1$.

149 (4)	si y sólo si ___, para	si y sólo si $\lambda_i > 0$, para
149 (5)	si y sólo si ___, para	si y sólo si $\lambda_i \geq 0$, para
149 (7)	tal que ____, en donde	tal que $V'XV = \Lambda$, en donde
149 (8)	implica que ___; esto se	implica que $X = V\Lambda V' = \sum \lambda_i v_i v_i'$; esto se
149 (9)	Descomposición del valor singular ____ ; donde	Descomposición del valor singular $\mathbf{X = UDV'}$; donde