2ej
16

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES
" Z A R A G O Z A "

EL ANALISIS DE DISTANCIA COMO UNA HERRAMIENTA PARA LA INVESTIGACION BIOLOGICA

T E S I S

QUE PARA OBTENER EL TITULO DE

B I O L O G O

P R E S E N T A :

PATRICIA RIVERA GARCIA

DIRECTOR DE TESIS: M. en C. ARMANDO CERVANTES SANDOVAL

MEXICO, D. F. 1991

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

CAPITULO 1

INTRODUCCION

Dentro de la diversidad que lo rodea, el hombre ha tratado de agrupar organismos u objetos tomando en cuenta los atributos o características compartidos por ellos y que los hacen semejantes a otros. Por eso recurre a la clasificación como un medio para evitar la confusión e instintiva o conscientemente, clasifica a su mundo circundante. Por ejemplo, el lenguaje mismo no sería posible sin una clasificación implícita, ya que cada objeto individual sería considerado como tal, sin llegar a una generalización. Por decir algo, cada mesa debería tener un nombre propio, ya que el término mesa implica un concepto colectivo inaplicable sin una clasificación previa (Everitt, 1981).

Debido a que muchos de los estudios para analizar diversos fenómenos son de naturaleza observacional, deben procesarse grandes volumenes de datos para explicar su comportamiento. En Biología es muy común que los datos se agrupen de acuerdo a las semejanzas que presentan y que caracterizan a cada observación. Por ejemplo, en Ecología es importante reconocer clases ecológicas que corresponden a comunidades; del mismo modo, en Taxonomía Numérica se busca identificar grupos de organismos relacionados por su máxima semejanza, que pueden corresponder con alguna categoría taxonómica como especie, género, familia.

Durante las tres últimas décadas han proliferado los métodos de clasificación, como puede apreciarse en la cantidad de literatura en la que se hace mención de diversos métodos de agrupamiento para la clasificación de datos de tipo ecológico. La disponibilidad de programas y equipo de cómputo que indudablemente facilitan el manejo de grandes cantidades de datos es una de las razones que ha propiciado este desarrollo.

Entre los múltiples elementos disponibles para apoyar los estudios de clasificación se encuentran las medidas de similaridad y distancia, que son la base de diferentes métodos estadísticos debido a que mediante el valor obtenido es posible conocer la semejanza existente en el conjunto de datos, es posible agrupar y describir el fenómeno en cuestión. Estas herramientas son poco conocidas y utilizadas en Biología, ya que se han generado en diferentes áreas del conocimiento.

Dado su gran potencial de aplicación en el estudio y descripción de los fenómenos naturales, este trabajo recopila y describe algunas medidas de distancia utilizadas en diferentes ámbitos de la investigación. De esta forma se busca promover su uso como una herramienta eficaz para apoyar los trabajos de investigación biológica. Con este fin se proponen los siguientes objetivos:

OBJETIVO GENERAL

Analizar y conjuntar las diferentes medidas de similaridad y distancia, con el fin de mostrar su utilidad y potencial de aplicación como herramienta para apoyar la descripción y estudio de diversos fenómenos biológicos.

OBJETIVOS PARTICULARES

- Adquirir una visión global de las medidas de similaridad y distancia mediante la recopilación bibliográfica del material referente al tema.

- Comprender y manejar las diferentes medidas de similaridad y distancia, a través del análisis de cada una de ellas.

- Efectuar la selección y aplicación de las medidas trabajadas de acuerdo al tipo de estudio y de los datos disponibles mediante la aplicación de las medidas seleccionadas a un estudio de caso.

- Proponer una medida de similitud y distancia que cubra la mayoria de los casos de la investigación biológica.

Para cumplir estos objetivos se presentan siete capítulos además del presente, que muestran diferentes aspectos de las medidas de similaridad y distancia, cuyo contenido está estructurado de la siguiente manera:

En el capítulo 2 se pretende dar una visión global de la importancia de la clasificación dentro de la investigación en Biología y la importancia de conocer las medidas de similitud y distancia.

En el capítulo 3 se presentan los diferentes tipos de datos y su arreglo para el análisis de similitud y distancia, así como las diferentes matrices generadas por estas medidas.

El capítulo 4 está dedicado a la revisión y análisis de las medidas de similaridad.

En el capítulo 5 se describen y analizan las diferentes medidas de distancia.

El capítulo 6 muestra un estudio de caso, donde se analiza la aplicación de las medidas de similaridad y distancia seleccionadas, en un estudio de tipo limnológico acerca de la distribución fitoplanctónica a diferentes niveles de profundidad.

En el capítulo 7 se discute sobre las ventajas de utilizar medidas de distancia en la Biología.

Finalmente, en el capítulo 8 se presentan las conclusiones del trabajo.

CAPITULO II

Clasificación, una breve revisión de su desarrollo histórico.

El origen de la clasificación se remonta hasta la antigua Grecia y Roma, en la que los médicos griegos desarrollaron algunas tipologías psicológicas. Lo más importante de esta tipología fué generado por Galeno (199-129 A.C.), quién definió 9 tipos de temperamentos que se presentan en relación a la susceptibilidad de personas con enfermedades referentes al comportamiento.

También en la antigua Grecia Aristóteles elaboró un sistema de clasificación de las especies del Reino Animal e hizo los dividión en dos grandes grupos: Aquellos que tienen sangre roja y que pertenecen a la clase de los vertebrados; y los que no contienen sangre roja y que clasificó como invertebrados. De la misma forma, subdividió a esos grupos de acuerdo a su desarrollo y a la forma de vida, los separó en aquellos que crecen en huevo, en pupas, o como larvas; creando con ello, la primera estructura clasificatoria.

A la muerte de Aristóteles, Theophrastus siguió trabajando sobre la misma línea logró la primera explicación fundamental de la estructura y clasificación de las plantas. El resultado de este trabajo es una completa y profunda documentación que proporciona el cimiento en la investigación biológica a través de varios siglos.

Todos estos trabajos fueron superados hasta los siglos XVII y XVIII cuando los grandes científicos europeos empezaron a realizar investigaciones sobre los restos del mundo antiguo con lo que dió la oportunidad de iniciar la búsqueda de nuevos esquemas clasificatorios y así conformar grupos semejantes, a través de colecciones de ellos (Sneath & Sokal, 1973).

Entre los trabajos más notables se encuentra el del sueco naturalista Linneo, quién publicó en 1737 su obra *"Genero Plantarum"* en la que se menciona lo siguiente:

> "Todo el conocimiento que poseemos sobre los objetos depende de la forma en que nosotros podamos ser capaces de distinguir lo similar de lo disimilar. El gran número de distinciones naturales cambia la concepción de los objetos, dando una mejor comprensión de las cosas".

Con base en esta idea, Linneo logró efectuar la separación de organismos en categorías definidas, tales como *Genero y especie* creando de esta manera un sistama de clasificación taxonómica que ha sido empleado a través de los años (Sneath & Sokal, 1973).

Posteriormente, el sistema de clasificación taxonómico linneano fué reemplazado por el concepto de Lindley que se basa en la afinidad natural, descrito en el "Sistema Natural de Botánica", publicado en 1836. Este se convirtió en el sistema de clasificación biológica que se utilizó hasta 1859, cuando Darwin propone la teoría clasificatoria de la evolución basada en la selección natural.

A pesar de que la mayor parte de los trabajos clasificatorios se han realizado en Biología, la clasificación se presenta como una herramienta importante en disciplinas tales como la Química, Astronomía, Física, Evolución, Sociología y otras, entre las que destaca la Taxonomía, que efectua la ordenación de objetos semejantes para agruparlos en categorías.

Partiendo de esta idea se puede definir a la clasificación como el proceso de ubicar objetos dentro de un conjunto de categorías, donde los atributos inherentes de cada categoría pueden ser conocidos con cierta incertidumbre al efectuar la asignación de grupos, ya sea porque se superpongan o porque sus

características no esten bien definidas. Independientemente de los problemas encontrados al aplicarla, la clasificación juega un papel fundamental en la ciencia.

Específicamente dentro del ámbito biológico, es común describir los fenómenos que ocurren en la naturaleza mediante la agrupación de entidades semejantes, conociendo a este procedimiento como similaridad.

La disimilaridad es también una manera indirecta de conocer la similaridad, ya que permite agrupar entidades parecidas y separar las no parecidas. Su gran relevancia dentro de la Biología se debe a que partiendo de un conjunto de observaciones se pueden generar grupos que ayuden a describir algún fenómeno biológico en estudio.

Es de gran interés conocer cuantitativamente el grado de separación o semejanza establecido entre entidades, haciendo uso de diversos coeficientes, ya que de esta forma es posible establecer que tan distante se encuentra una entidad de otra. Esta forma de medir las diferencias se conoce como distancia y expresa el grado de similaridad o disimilaridad existente dentro de las observaciones, que de alguna manera describe la estructura de un conjunto de datos.

Los coeficientes que miden la semejanza se conocen comúnmente como coeficientes de similaridad, de distancia o de asociación, que tienen un frecuente uso en diversas áreas de la investigación, sobre todo en Biología donde interesa la agrupación del conjunto de observaciones para describir el comportamiento del fenómeno biológico en cuestión.

El uso de los coeficientes de similaridad y distancia son mencionados ampliamente en diversos trabajos, como los efectuados por Sokal & Sneath (1973) cuyas contribuciones fueron desarrolladas dentro de la Taxonomía, donde describen algunas

medidas de similaridad y distancia. De manera análoga, Anderberg (1973) y Orlocí (1979) han mencionado trabajos similares sobre este tema.

Diferentes autores hacen una revisión de medidas tanto de similitud como de distancia, Pielou (1984) hace hincapié en aquellas medidas que usan datos de presencia-ausencia dentro de la comunidad; Gower (1971), plantea un coeficiente de similitud en el cuál abarca una generalidad de datos cualitativos como cuantitativos, siendo una buena medida de similaridad.

Clifford y Stephenson (1975) mencionan algunos índices de similaridad como: El índice de Hamman, el de Canberra, la medida euclidiana y algunas derivadas de ella. Estos también son reportados por Legendre y Legendre (1983), y Reynolds (1988), quienes realizan una amplia descripción de varias medidas.

Mueller (1974), efectúa una descripción de algunas medidas de similitud aplicando estos índices en trabajos que se encuentran enfocados al análisis de vegetación. Orlocí (1979) cita algunas medidas de disimilaridad como la euclidiana y el coeficiente de correlación producto-momento y su aplicación en algunos datos ecológicos.

Además de estos trabajos existe una gran cantidad de bibliografía en la que se hace referencia a la utilización de coeficiente de similitud y distancia. En la mayoría de ellos se refleja la necesidad de adaptar los datos para poder aplicar las diferentes medidas, ya que cada tipo de datos requiere de una forma diferente de aplicar los coeficientes, lo cual también depende de los objetivo del estudio.

Algunas medidas están restringidas a datos binarios, otras permiten considerar las desigualdades establecidas en la frecuencia de los atributos y minimizan la influencia de pequeños y grandes valores, mientras que otras se basan en ideas que pueden

aproximarse a una distribución estadística inherente a las medidas. El tipo de datos en algunas ocasiones puede influenciar la selección de la medida a trabajar.

Al utilizar la clasificación en Biología es importante conocer y manejar el tipo de datos presentes, así como las medidas de similaridad y distancia a trabajar; ya que los valores obtenidos conforman las matrices de similaridad y distancia, que son la base de las diferentes técnicas clasificatorias y descriptivas multivariadas o univariadas existentes.

CAPITULO III

TIPOS DE DATOS Y SU ARREGLO PARA EL ANALISIS DE SIMILARIDAD Y DISTANCIA

Las bases de una metodología para analizar datos se puede derivar de la interacción entre las condiciones en las que se efectúan las observaciones y los resultados obtenidos. Como un apoyo, tres ciencias proveen a los investigadores de herramientas para analizar la complejidad de sus datos. La física, que auxiliada por la matemática, proporciona el análisis dimensional, dando una elegante y simple solución a los problemas. La comunicación que contribuye con la información teórica para efectuar el procesamiento general de los distintos tipos de datos. Finalmente, la estadística permite disponer, entre otras cosas, de las técnicas multivariadas para trabajar datos biológicos (Reynolds, 1988).

Al considerar esas tres ciencias como herramientas que contribuyen de una manera fundamental en el manejo de los datos, se deben identificar las variables más importantes o las que presentan un mayor peso en el estudio a realizar, siendo importante tomar en cuenta el costo y el nivel de precisión requerido para describir y analizar el fenómeno en cuestión.

Por otra parte, al seleccionar las variables que serán utilizadas se debe tener una idea definida y clara del objetivo del trabajo. Para este fin es de gran importancia el juicio y la experiencia del investigador en la selección de las variables y el desarrollo del trabajo.

Las variables utilizadas en la descripción y clasificación de entidades y/o objetos; se definen con base en características físicas, químicas, ecológicas o biológicas presentes en los objetos de interés para el investigador. Esto suele ocurrir con

bastante frecuencia en las especies biológicas, que pueden ser definidas como las variables de estudio (Reynolds,1988).

Una vez seleccionadas las variables a medir y su escala de medición, es importante conocer el tipo de datos obtenidos, ya que con base en su clasificación se determina el tratamiento numérico que se puede aplicar. En este contexto las variables se dividen en: Binarias, Cualitativas y Cuantitativas (Anderberg, 1973).

1) Variables binarias

Presentan dos estados contrastantes, tales como presencia-ausencia, (0,1 o +,-). Dentro de la Biología este tipo de datos son utilizados frecuentemente, sobre todo en Ecología y Taxonomía.

Las variables binarias más utilizadas en Ecología son del tipo presencia-ausencia, que sirven para definir asociaciones biológicas, nichos de especies, o bien, para efectuar reagrupamiento de muestras en ausencia de datos cuantitativos confiables.

2) Variables cualitativas

Este tipo de varibles tiende a representar características que no son medibles cuantitativamente, abarcando ciertos rangos presentes en las entidades, tales como las variaciones en color, tipo de alas, o definición de sexo. Cabe mencionar que las variables binarias son consideradas también como datos cualitativos. No obstante se presentan algunas discrepancias entre ellos, ya que en los primeros solo es posible encontrar dos estados, mientras que en los segundos se pueden encontrar más de dos estados comparables que pueden ser jerarquizados.

Dentro de los datos cualitativos, mencionados por algunos autores como datos multiestados (Clifford, 1975) pueden encontrarse dos formas:

- **Variables cualitativas desordenadas.**- Presentan más de tres formas contrastantes, donde cada valor tiene el mismo peso. Por ejemplo, el color de las flores que puede ser rojo, azul o blanco.

- **Variables cualitativas ordenadas.**- Poseen una jerarquía de formas contrastantes que abarcan al total de variaciones que pueden encontrarse dentro de los valores que pueden tomar las entidades sujetas a estudio. Como por ejemplo, la abundancia de organismos en cuadrantes puede ser agrupadas en términos de series como: raras, comunes o abundantes. De una manera semejante, las plantas pueden ser agrupadas de acuerdo a su desarrollo (corto o largo). Estas variables también son conocidas como ordinales o de grado.

3) Variables cuantitativas.

Son el tipo más usual, en ellas se incluyen las variables de abundancia, así como otras medidas que puedan ubicarse en un espacio geométrico.

- **Variables cuantitativas contínuas.** Estas variables expresan cualidades, cuya variabilidad se distribuye dentro de una escala contínua. Pueden ser expresadas como números enteros o fraccionarios. En estudios taxonómicos, las medidas de longitud o de proporción de longitudes produce un tipo común de datos contínuos.

- **Variables cuantitativas discontínuas.** Son conocidas como variables de ennumeración, en las cuales se contabilizan a los objetos comparados asignándoles un número entero.

ARREGLO DE LOS DATOS

La manera más frecuente de encontrar una matriz de datos, principalmente en ecología, es mediante la ubicación de entidades en las hileras y las variables, en las columnas.

Una matriz de entidades por variables y/o atributos, es un tipo de matriz que toma la siguiente forma:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \cdots\cdots x_{1p} \\ x_{21} & x_{22} & x_{23} \cdots\cdots x_{2p} \\ x_{31} & x_{32} & x_{33} \cdots\cdots x_{3p} \\ \vdots & \vdots & \vdots \quad \vdots \\ x_{n1} & x_{n2} & x_{n3} \cdots\cdots x_{np} \end{bmatrix}$$

Esta matriz constituye un arreglo de datos para una muestra o estudio determinado. En ella, la x_{ij} es una medida de la i-ésima entidad para la j-ésima variable. Las medidas involucradas para formar esta matriz pueden conformarse de datos cualitativos, cuantitativos o binarios, o de una mezcla de tipos de variables. Conviene aclarar que una matriz contiene n hileras correspondientes a las entidades y p columnas que corresponden a las variables, siendo por lo tanto, de tamaño n×p.

Dichas matrices se utilizan frecuentemente para la organización y presentación de datos, teniendo la facilidad de efectuar manipulaciones matemáticas muy diversas. Es aquí donde radica la importancia del uso de las matrices, ya que presentan la facilidad de condensar un conjunto de símbolos que contiene una gran riqueza de manipulaciones matemáticas.

Como resultado, las matrices auxilian de una manera eficaz en el análisis de los datos, a los cuales se enfrenta el biólogo a través del constante aumento de los métodos cuantitativos en su trabajo. Dependiendo de la forma en que se presenten los datos y

del objetivo de estudio, se pueden construir diferentes tipos de matrices para datos ecológicos como son (Crisci, 1983):

1) Matrices de similitud.

Este tipo de matrices son el resultado de calcular el grado de asociación entre variables aplicando los coeficientes de similitud. La similaridad es el grado de semejanza que existe entre dos individuos y toma valores entre 0 y 1, donde 0 indica una nula similaridad y 1 una similitud total. La representación de este tipo de matrices está dado por:

$$S = \begin{bmatrix} 1 & s_{12} & s_{13} & \cdots & s_{1n} \\ s_{21} & 1 & s_{23} & \cdots & s_{2n} \\ s_{31} & s_{32} & 1 & \cdots & s_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ s_{n1} & s_{n2} & s_{n3} & \cdots & 1 \end{bmatrix}$$

2) Matrices de disimilitud

La disimilaridad es lo contrario de la similaridad, ya que toma en cuenta las diferencias para separar un conjunto de datos en grupos.

Para construir una matriz de disimilaridad se usan los coeficientes de distancia conociéndo de esta manera el grado de separación entre dos puntos. La representación de este tipo de matrices está dada como sigue:

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots\cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots\cdots & d_{2n} \\ d_{31} & d_{32} & 0 & \cdots\cdots & d_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots\cdots & 0 \end{bmatrix}$$

3) Matrices de asociación.

Este tipo de matrices proviene de una matriz básica de datos y se genera al analizar entidades o variables, a las que se les aplican medidas de similaridad y distancia.

En estas matrices los elementos a_{ii} representan la asociación de las hileras o columnas. En el caso de las columnas, la medida de asociación de una entidad consigo misma puede tomar un valor de 0 o 1, dependiendo de que si se trabaja una medida de similaridad o de distancia.

$$A = \begin{bmatrix} 1 & a_{12} & a_{13} & \cdots\cdots & a_{1n} \\ a_{21} & 1 & a_{23} & \cdots\cdots & a_{2n} \\ a_{31} & a_{32} & 1 & \cdots\cdots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots\cdots & 1 \end{bmatrix}$$

Con $a_{ii}=1$

3) Matrices de correlación.

La correlación entre dos vectores de variables puede considerarse una medida de similaridad entre individuos, tomando un rango de valores $-1 \leq r_{ij} \leq 1$.

La matriz de correlación puede construirse a partir de valores obtenidos por el coeficiente de correlación producto momento r.

Como puede verse en este capítulo se muestran las formas y características de los datos que deben ser consideradas al utilizar las diferentes medidas de similaridad y distancia, cuidando su aplicación, forma de trabajo y su representación al momento de realizar su cálculo, tópicos que ayudan a entender las medidas de similaridad y distancia que se describen en los dos siguientes capítulos.

CAPITULO IV
MEDIDAS DE SIMILITUD

A través del tiempo, investigadores de diferentes disciplinas han tratado de clasificar diversos fenómenos por la agrupación de entidades o variables de acuerdo a la semejanza presente en ellas. Para cuantificar la similitud se ha hecho uso de una extensa gama de medidas de similaridad que de alguna forma describen al fenómeno en cuestión.

La aplicación de las diferentes medidas de similaridad a un conjunto de datos permite tres tipos de comparaciones:

- Entre entidades.
- Entre una entidad y un grupo de entidades.
- Entre grupos de entidades.

Para medir la similaridad, generalmente se hace la comparación entre entidades dependiendo del objetivo planteado en el estudio, aunque no se descartan las otras dos opciones.

Si la similaridad se expresa como una semejanza entre entidades, puede definirse como:

Una función real no negativa, denominada $S(x,y)$, de pares de puntos que pertenecen a un conjunto de entidades E. Entonces $S(x,y)$ es considerado como una medida de similaridad para ese conjunto E si:

1) $0 \leq S(x,y) \leq 1$ para toda $x \neq y$
2) $S(x,x)=1$
3) $S(x,y)=S(y,x)$

Estas tres condiciones marcan la pauta para comparar dos o más entidades. El primer axioma establece la similaridad como una

medida de semejanza cuyo intervalo de valores oscila entre 0 y 1. El manejar al 0 como valor mínimo se interpreta que existe una discordancia total entre entidades, y si vale 1 hay una similitud total.

El segundo axioma define que la similaridad de una entidad consigo misma toma el valor de 1. El tercer axioma marca la simetría de las medidas de similaridad, ya que $S(x,y)=S(y,x)$.

Para cuantificar el grado de similitud existe una amplia gama de medidas cuyo objetivo es mostrar numéricamente la semejanza entre dos entidades. Las medidas de similaridad se aplican generalmente a datos de tipo cualitativos o binarios que se representan en un cuadro de doble entrada donde el grado de asociación de las entidades i y j es expresado de la siguiente manera:

		entidad j		
		+ (1)	- (0)	total
entidad i	+ (1)	a (1,1)	b (1,0)	a+b
	- (0)	c (0,1)	d (0,0)	c+d
	total	a+c	b+d	p

Las letras a,b,c,d se refieren a la suma de atributos, donde a representa la presencia conjunta de atributos; b y c, la presencia de uno de ellos y la ausencia del otro y d la ausencia de atributos. La suma a + b + c + d se representa por p.

Con este cuadro se puede seleccionar un coeficiente de similaridad de los más utilizados en Biología, los cuales se describen a continuación.

1) Coeficientes de similitud

Son los coeficientes más citados en la literatura, su objetivo principal es medir la similitud entre entidades. Contrastando con la mayoría de las medidas de distancia, las medidas de similitud nunca presentan la propiedad de ser métricas y su rango de valores oscila entre 0 y 1 (Crisci, 1983).

2) Coeficientes de asociación (Clifford, 1975).

En este tipo de coeficientes se estima el grado de asociación entre variables y/o entidades. Pueden ser trabajados datos binarios, categóricos o cuantitativos pero se preferentemente con datos de tipo continuo. Esto refleja la manera en que se presentan las características en n entidades que pueden estar correlacionadas. Su intervalo de valores oscila de +1 (indicando que los cambios en las dos variables son diréctamente proporcionales), hasta -1 (en donde hay asociación pero los cambios se consideran inversamente proporcionales, es decir cuando los valores de una entidad aumentan los de la otra disminuyen).

Los coeficientes de similaridad más utilizados en diferentes disciplinas se listan a continuación, dejando claro que la utilización de cada uno dependerá del criterio del investigador y los objetivo del estudio planteado.

Estos coeficientes han sido concebidos y aplicados en diferentes disciplinas, como: Ecología, Taxonomía, Botánica, Zoología, etc,. Para ejemplificar su cálculo, a continuación se presenta una matriz de datos, en la cuál se comparan a tres especies (entidades) y diez atributos (variables), teniéndo lo siguiente:

Especies	Atributos									
	1	2	3	4	5	6	7	8	9	10
A	1	1	0	1	1	0	0	0	0	0
B	0	1	1	1	1	1	1	1	1	0
C	0	0	0	0	0	1	0	0	1	1

Datos de presencia ausencia para 3 especies en diez sitios

A continuación se hace la comparación de las especies A, B y C mediante las cuales se obtienen valores para a, b, c, y d, que se muestran en el siguiente cuadro:

		entidad j		
		+ (1)	- (0)	total
entidad i	+ (1)	a (1,1)	b (1,0)	a+b
	- (0)	c (0,1)	d (0,0)	c+d
	total	a+c	b+d	p

Para efectuar el cálculo de a, b, c y d se procede como sigue:

Tabla de dos entradas para efectuar la comparación entre dos especies.

Para las especie A y B

A \ B	+	-
+	3	1
-	5	1

Para las especies A y C:

A \ C	+	-
+	0	4
-	3	3

Para las especies B y C:

		C: +	C: -
B	+	2	6
	-	1	1

donde + = presencia y - = ausencia.

El siguiente paso consiste en aplicar algunos de los coeficientes, para analizar el grado de similitud por entidades a través de diferentes sitios. A continuación se listan una serie de coeficientes, acompañados de una breve descripción de cada uno de ellos y se efectua el cálculo para las especies A y B, y solo se indica el valor obtenido para las comparaciones entre A-C y B-C.

1) Indice de Jaccard.

El índice de Jaccard expresa la similaridad entre entidades y se basa en las relaciones de presencia de variables o atributos comunes. Considera al total de atributos que existen en las entidades, es decir, expresa la proporción de características o atributos comunes para las entidades presentes en la muestra. Esta idea se expresa como sigue:

$$IS_J = \frac{\text{Atributos comunes}}{\text{Todos los atributos}}$$

Originalmente este índice se utilizó en la comparación de sitios de gran extensión, y se aplicó a datos de vegetación, aunque también se utiliza para comparar sitios menos extensos (Mueller, 1974).

Este índice puede trabajarse con base en listados de entidades con características comunes por la suma o cantidad de cada característica presente. Su expresión es:

$$S_{ij} = \frac{a}{a + b + c}$$

El valor obtenido para esta medida es interpretado como una probabilidad condicional que toma en cuenta el número de veces en las que dos entidades coinciden en la muestra. En relación al total de variables presentes, es decir, partiendo del total de veces en las que las variables aparecen, se contabiliza el número de coincidencias que ocurren en la comparación. Se recomienda el uso del índice de Jaccard cuando los datos son estrictamente de presencia-ausencia.

A partir del índice de Jaccard se han derivado una serie de índices que toman en cuenta las veces que se presenta cierta característica en las entidades que es una forma indirecta de medir la similitud entre ellas. Es importante tomar en cuenta que esta medida no sólo es función de las características comunes o únicas, sino que también de la suma de cada atributo presente.

El concepto de similaridad se puede expandir a la inclusión de formas de vida, composición, y abarcar otros criterios que pueden utilizarse dependiendo del objetivo de estudio (Mueller, 1974).

El desglose numérico para esta medida se dá como sigue:

$$S_{A,B} = \frac{3}{3 + 1 + 5} = 0.333, \quad S_{A,C} = 0, \quad S_{B,C} = 0.444$$

2) Indice de similaridad de Sorensen

El índice de Sorensen se considera una modificación del índice de Jaccard, siendo expresando en su forma original de la siguiente manera (Mueller D., 1974):

$$IS_{ij} = \frac{2c}{a + b} \quad o \quad IS_{ij} = \frac{c}{\frac{1}{2}(a + b)}$$

Este índice es mencionado con frecuencia en la literatura de la siguiente forma (Gauch, 1982; Matteuci, 1982):

$$S_{ij} = \frac{2a}{2a + b + c}$$

Donde c = Número de atributos comunes localizados en la comparación.
a = Número de atributos que se contabilizan en una de las entidades comparadas.
b = Número de atributos contabilizados para la otra entidad.

Este índice es también conocido como el índice de coincidencia de Dice (Crisci, 1983), y se utiliza frecuentemente en Taxonomía numérica.

En teoría este índice considera que cada variable tiene igual oportunidad de presentarse en dos entidades, implicando que cualquier variable puede ser localizada en las dos entidades o solamente en una. Por ello, la expresión 1/2(a+b) representa la suma de coincidencias teóricamente realizables, donde el numerador a expresa la coincidencia conjunta de entidades (Mueller, 1974).

El índice de Jaccard toma en cuenta solo la presencia de las variables en las entidades e ignora completamente su ausencia, asignando un mayor peso a la coincidencia de atributos si y sólo si estos se encuentran presentes. Por ello, este índice pierde la posibilidad de ser interpretado probabilísticamente. Su cálculo es como sigue:

$$S_{A,B} = \frac{2(3)}{2(3)+1+5} = 0.65 \qquad S_{A,C} = 0 \qquad S_{B,C} = 0.3636$$

3) Coeficiente de Hamann

Este coeficiente se formula de la siguiente manera (Crisci, 1983):

$$H = \frac{(a+d) - (b-c)}{a + b + c + d}$$

En él se toman en cuenta las diferencias generadas por las relaciones de presencia-ausencia de variables en las entidades. De manera implícita se encuentra el concepto de probabilidad, dado que el denominador toma en cuenta la totalidad de atributos presentes en las entidades, mientras que en el numerador la importancia que tienen las ausencias como las presencias es la misma, independientemente del lugar en el cual aparezca la variable que es comparada. La desventaja que presenta es que el número de características o atributos puede ser mayor al número de ausencias y presencias conjuntas, dando como resultado un valor mayor que 1 o incluso negativo, saliendo en esta situación del rango de similaridad. Su cálculo numérico está dado como sigue:

$$S_{A,B} = \frac{(3+1) - (1-5)}{3 + 1 + 5 + 1} = 0.80 \qquad S_{A,C} = 0.2 \qquad S_{B,C} = -0.2$$

4) Coeficiente de Roger y Tanimoto

El coeficiente de Roger y Tanimoto se formula de la siguiente manera (Crisci, 1983; Legendre, 1983):

$$S_{ij} = \frac{a + d}{a + 2(b + c) + d}$$

Esta expresión toma en cuenta las relaciones de presencia-ausencia de atributos comunes en las entidades, donde, en el numerador estas relaciones tienen igual peso, mientras que en el denominador se realiza una consideración semejante, pero se enfatiza la presencia de atributos sin importar en cual entidad se

encuentre, interesando solo que aparezca en la comparación. Esto se representa por el doble valor que se asigna a las letras b y c 2(b + c). Por esta razón, este coeficiente pierde su interpretación probabilística.

El coeficiente de Roger y Tanimoto es una extensión del coeficiente de Sorensen el cual da un cierto peso a la coincidencia de atributos (Mueller, 1974). Su desglose numérico está dado como sigue:

$$S_{A,B} = \frac{3 + 1}{3 + 2(1+5) + 1} = 0.25 \qquad S_{A,C} = 0.1764 \qquad S_{B,C} = 0.1764$$

5) Coeficiente de Sokal y Sneath

Matemáticamente este coeficiente se expresa de la siguiente forma:

$$S_{ij} = \frac{2(a + d)}{2(a + d) + b + c}$$

Esta fórmula pondera las relaciones de presencia-ausencia conjunta de atributos o variables, asignándole un mayor peso, sin importar si los atributos se presentan o no. Esta situación se plantea tanto en el numerador como en el denominador, pero en el segundo se considera a la aparición de los atributos en cualquier entidad. Por esta razón, este coeficiente pierde interpretación probabilística. Su cálculo se dá como sigue:

$$S_{A,B} = \frac{2(3 + 1)}{2(3 + 1) + 1 + 5} = 0.5714 \quad S_{A,C} = 0.4615 \quad S_{B,C} = 0.46153$$

6) Coeficiente de pares simples (Simple Matching Coefficient)

El coeficiente de pares simples es ampliamente trabajado dentro de la Ecología y se define de la siguienta manera (Matteuci, 1982):

$$S_{ij} = \frac{a + d}{a + b + c + d} = \frac{a + d}{p}$$

En el numerador se toma en cuenta la presencia y ausencia conjunta de atributos y/o variables, mientras que en el denominador se considera el total de apariciones de los atributos en las entidades comparadas. Cada coincidencia se considera de la totalidad de atributos presentes en el estudio. Por esta razón se menciona que este coeficiente tiene una interpretación netamente probabilística. A continuación se muestra el cálculo numérico para este índice.

$$S_{A,B} = \frac{3 + 1}{3 + 1 + 5 + 1} = 0.40 \qquad S_{A,C} = 0.30 \qquad S_{B,C} = 0.30$$

7) Coeficiente de Ochiai

Ochiai (1957) utilizó este coeficiente como una medida de similaridad considerándose como una variante del índice de Jaccard, cuya expresión es (Gauch, 1982):

$$ISo = \frac{\text{Coocurrencias}}{\text{Media geométrica de ocurrencia en las dos entidades}}$$

y su formulación se dá como sigue:

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

Este coeficiente ignora completamente la ausencia de atributos y/o variables, importando solo la presencia conjunta. Esto se representa en el numerador, mientras que el denominador se toma en cuenta a los atributos que se localizan en la comparación y en función de su número se dá la coincidencia; es decir, las variables son comunes para dos entidades, si y solo si se

presentan en un cierto número. El coeficiente de Ochiai maneja probabilidades condicionales de aparición de atributos en el denominador, considerando que las coincidencias estan en función de cuantas entidades se presenten. Su cálculo es:

$$ISo_{A,B} = \frac{3}{\sqrt{(3+1)(3+5)}} = 0.53033 \quad ISo_{A,C} = 0 \quad ISo_{B,C} = 0.4082$$

8) Coeficiente de Mozley (1936), Coeficiente de Margaleff (1958)

Inicialmente este coeficiente se planteó de la siguiente manera (Gauch, 1982):

$$M = \frac{\text{Coocurrencias}}{\text{Coocurrencias expresadas al distribuirse independientemente}}$$

Esta relación se expresa como sigue:

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

Este coeficiente, presenta como punto importante la presencia de atributos y/o variables comunes a través del número total de variables presentes, donde dependiendo de su cantidad, se obtiene la probabilidad de aparición de las k variables en el estudio.

En el numerador, se toma en cuenta la coincidencia conjunta tomada del total de entidades presentes, mientras que en el denominador se considera a la probabilidad de coincidencia de dos atributos para ambas entidades, sin importar en cual se presenta, dando este hecho una interpretación de tipo probabilistico. Sus rangos caen entre 0 y p/a, donde p expresa el número de variables que se utilizan para su clasificación. Su valor crece cuando se dá la coincidencia de atributos, sobre todo cuando las entidades y/o atributos son "raros" o poco comunes. Su cálculo numérico está dado por:

$$S_{M_{A,B}} = \frac{3(3 + 1 + 5 + 1)}{(3 + 1)\ (3 + 5)} = 0.9375 \quad S_{M_{A,C}} = 0 \quad S_{M_{B,C}} = 0.5555$$

9) Coeficiente de Mountford (1962)

Este coeficiente toma en cuenta solamente las relaciones de presencia conjunta de entidades, dandole a esta relación un doble valor. Mountford derivó este coeficiente de la siguiente forma (Gauch, 1983):

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

En el numerador, se da un doble valor a los atributos comunes para dos entidades, mientras que en el denominador las coincidencias son contabilizadas como importantes a partir del número total de veces en la que los atributos se presentan en cualquier entidad que es comparada. Por ello, la presencia de los atributos puede estar en función del tamaño de la muestra variando, por tanto, el valor calculado para este coeficiente.

Mountford utilizó este coeficiente para comparar muestras de diferentes tamaños mencionando que el número de entidades sigue una distribución logarítmica. El cálculo para este coeficiente es:

$$S_{A,B} = \frac{2(3)}{3(1 + 5) + 2((1)(5))} = 0.21428 \quad S_{A,C} = 0, \ S_{B,C} = 0.1538$$

10) Coeficiente de Russell y Rao

Este coeficiente se caracteriza por tomar en cuenta la coincidencia de atributos a lo largo de la comparación. Se expresa como sigue (Anderberg, 1973):

$$S_{ij} = \frac{a}{a + b + c + d} = \frac{a}{p}$$

En esta expresión se plantea que la coincidencia de atributos se dá a partir del número total de variables presentes, teniendo todos ellos la misma oportunidad de coincidir con otro, lo que da a este coeficiente una interpretación netamente probabilística. Su cálculo es el siguiente:

$$S_{A,B} = \frac{3}{3 + 1 + 5 + 1} = 0.3 \qquad S_{A,C} = 0 \qquad S_{B,C} = 0.2$$

11) Coeficiente de Kulczynski

Kulczynski (1928), sugirió un coeficiente de similaridad basado en datos de presencia-ausencia que se representa de la siguiente manera:

$$S_{ij} = \frac{a}{b + c}$$

En la expresión anterior se plantea la relación de presencia sin contemplar las ausencias, resaltando que las coincidencias van a estar dadas en las entidades con una igual probabilidad, independientemente de donde se encuentren registradas, condicionando las coincidencias al número de variables registradas. Su cálculo numérico está dado como sigue:

$$S_{A,B} = \frac{3}{1 + 5} = 0.5 \qquad S_{A,C} = 0 \qquad S_{B,C} = 0.28571$$

Clitford y Stephenson (1975), reportaron el coeficiente de Kulczynski como una expresión alternativa, de la siguiente manera:

$$IS_K = \frac{a}{2}\left[\frac{1}{(a + b)} + \frac{1}{(a + c)}\right]$$

A continuación, se hace una breve descripción de una serie de índices, que dentro de la literatura no se encuentran con un

nombre definido, o no son recomendados para un uso generalizado, pero que pueden ser tomados en cuenta para medir la similitud entre entidades (Cliffort, 1975; Crisci, 1983).

12) $$S_{ij} = \frac{a + d}{a + b + c}$$

En esta fórmula se considera la relación presencia-ausencia conjunta de atributos en las entidades. En el numerador se considera tanto a la presencia como la ausencia con el mismo peso, mientras que en el denominador se hace referencia al número total de veces en las que aparecen las variables dentro de las entidades, sin importar la ausencia de ellas en la comparación. El valor calculado para este índice está dado a continuación:

$$S_{A,B} = \frac{3 + 1}{3 + 1 + 5} = 0.4444 \qquad S_{A,C} = 0.42857 \qquad S_{B,C} = 0.333$$

13) $$S_{ij} = \frac{a + d}{a + 2(b + c)}$$

Este índice puede interpretarse de manera semejante al índice anterior pero sufre cierta modificación al hacer una ponderación de la presencia de los caracteres en cualquiera de las entidades trabajadas, asignandole un doble peso, perdiendo por esta razón una interpretación probabilística. Su cálculo es el siguiente:

$$S_{A,B} = \frac{3 + 1}{3 + 2(1 + 5)} = 0.2666 \qquad S_{A,C} = 0.21428 \qquad S_{A,C} = 0.1875$$

14)

$$S_{ij} = \frac{2(a + d)}{2a + b + c}$$

En este índice, el numerador considera tanto la presencia como ausencia de atributos comunes en las entidades dando a esta relación un doble peso, mientras que en el denominador se pondera a la coincidencia, considerando la posibilidad de encontrar a los atributos en la comparación. Su cálculo numérico está dado por:

$$S_{A,B} = \frac{2(3+1)}{2(3)+1+5} = 0.6666 \qquad S_{A,C} = 0.8571 \qquad S_{B,C} = 0.5454$$

Los índices 12, 13 y 14 tratan de considerar a los pares ausentes dentro del numerador como importantes, pero los excluye del denominador. En términos de presencia estos índices pueden ser útiles pero si se toman en cuanta las ausencias, sesgan de alguna manera el cálculo y la interpretación, ya que al efectuar comparaciones en base a las ausencias no son válidas y su resultado no indica realmente similitudes entre entidades.

15) Coeficiente de Gower.

Para calcular similaridades mediante este índice, debe tomarse en cuenta que:

- La semejanza entre dos individuos, i y j, al medirse con respecto a un caracter k, está definida por s_{ijk}.

- La posibilidad de comparar a las entidades se representa por a_{ijk}, cuyos valores oscilan entre 1 cuando el caracter k puede ser comparado y 0 cuando no es posible la comparación. Si $a_{ijk} = 0$, el valor de S_{ijk} se desconoce, pero por convención se le asigna el valor de 0.

La similaridad de los individuos i y j es definida como un valor promedio, tomando en cuenta todas las posibles comparaciones, expresándose como sigue:

$$S_{ij} = \frac{\sum_{k=1}^{v} s_{ijk}}{\sum_{k=1}^{v} \delta_{ijk}}$$

Donde:

S_{ij} = Es la medida de similaridad entre el individuo i y j

s_{ijk} = Es el valor calculado de la comparación del individuo i y j con respecto a un caracter k.

δ_{ijk}= Representa la posibilidad de realizar la comparación entre el individuo i y j.

Cuando la variable $\delta_{ijk} = 0$, para todos los k-caracteres comparados, el valor de la S_{ij} es indefinido. Cuando es posible comparar entidades, $\sum \delta_{ijk} = v$, en donde v es el total de atributos.

Alternativamente se dá una expresión equivalente a la ecuación 1 de la siguiente forma:

$$S_{ij} = \frac{\sum_{k=1}^{v} (s_{ijk})(\delta_{ijk})}{\sum_{k=1}^{v} \delta_{ijk}}$$

En donde se pondera a las entidades comparadas y el valor de δ_{ijk} es válido solo cuando es posible realizar comparaciones entre las entidades. El valor de s_{ijk} puede asignarse como:

1) Para un caracter dicotómico, la presencia del caracter es denotado como positivo (+) y la ausencia como negativo (-). Esto puede ser representado de la siguiente manera:

Valor de k

	1	2	3	4
i	+	+	-	-
j	+	-	+	-
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

2) Para un caracter cualitativo, los valores de las s_{ijk} se dan como sigue:

s_{ijk} = 1 Si los individuos i y j coinciden en el k-ésimo caracter.

s_{ijk} = 0 Si los individuos i y j presentan discrepancias en el k-ésimo caracter.

Para representar datos cualitativos, se presenta la siguiente tabla, donde se comparan algunos colores a través de algunas entidades como sigue:

Simbología:

n = negro a = azul b = blanco

Valor de k

	1	2	3	4	5	6
i	n	a	b	n	n	a
j	a	a	n	b	n	a
Sijk	0	1	0	0	1	1
δijk	1	1	1	1	1	1

3) Para un caracter cuantitativo con valores X_1, X_2,.....X_n, de los k caracteres para la muestra total, denominada como n individuos, el valor de *sijk* se da como a continuación se muestra:

$$s_{ijk} = 1 \left| \frac{X_{ik} - X_{jk}}{R_k} \right|$$

donde R_k = Rango de valores sobre el caracter k que es dado como el rango total en la población o el rango en la muestra.

Cuando $X_{ik} = X_{jk}$, entonces, $S_{ijk} = 1$ si $X_{ik} \neq X_{jk}$, el valor de $s_{ijk}=0$. Con valores intermedios, la s_{ijk} es una fracción positiva que oscila entre 0 y 1.

Análogamente, Gower realiza una ponderación en las variables, quedando la expresión de la siguiente forma:

$$S_{ij} = \frac{\sum_{k=1}^{v} (s_{ijk})(W_k)}{\sum_{k=1}^{v} (\delta_{ijk})(W_k)}$$

Donde S_{ij} y s_{ijk} tienen el mismo significado que en la fórmula anterior, y W_k es una constante de ponderación para cada caracter.

Las ponderaciones son susceptibles de ser registradas como función del valor que asuma cada caracter comparado. Así, la diferencia en un caracter puede ser considerado como más importantes que las coincidencias conjuntas, de forma que el coeficiente de similaridad toma la siguiente fórmula:

$$S_{ij} = \frac{\sum_{k=1}^{v} (s_{ijk})(W_k(X_{ik}, X_{jk}))}{\sum_{k=1}^{v} (\delta_{ijk})(W_k(X_{ik}, X_{jk}))}$$

Donde $W_k(X_{ik}, X_{jk})$ es una expresión que pondera al caracter k, que está en función de los valores que asume X_{ik} y X_{jk}, teniendo

este coeficiente una mayor potencialidad de uso, al permitir de una manera accesible tener un mayor conocimiento de la similitud o la diferencia en los atributos comparados (llamese a estos sitio, variables, características, etc).

Este coeficiente presenta una gran flexibilidad para abarcar las diferentes formas en las que una variable puede ser trabajada, ya sea como valores cualitativos o cuantitativos, a diferencia de los coeficientes de similitud que se utilizan comúnmente, ya que este coeficiente no requiere para su uso de alguna recodificación para los caracteres cualitativos o cuantitativos. Para concluir, es importante mencionar algunas de las razones por las que se recomienda el uso de este coeficiente:

1) Que al igual que con las matrices de correlación, permite aplicar los métodos numéricos que operan solamente con matrices positivas con confianza, una vez que no haya valores faltantes.

2) Ayuda a la interpretación de aquellos métodos de Cluster y a las técnicas de ordenación que se basan en métricas euclidianas.

COMENTARIOS

En los coeficientes de similitud que se han analizado puede mencionarse lo siguiente:

La mayoría de los coeficientes que miden la semejanza entre entidades, hacen uso de variables del tipo presencia-ausencia, donde se manejan las letras a, b, c y d referidos como los valores que son representados dentro de una tabla de doble entrada, en la cual se hace referencia a un total definido como p.

En los coeficientes dados se busca contar el número total de pares semejantes entre dos individuos sobre las p variables. Su número es grande debido a las diversas formas en las que puede ser interpretada la frase 'pares semejantes'. Esencialmente, las diferentes interpretaciones surgen de dos factores:

1) El primero y más importante es la incertidumbre de cómo

incorporar el número de pares negativos (o ausentes, en términos de d) en las medidas.

2) La cuestión de que si deben ponderarse de igual forma los pares semejantes y los ausentes.

Algunas medidas excluyen totalmente a los pares ausentes (pares negativos) como sucede en las medidas 1, 2, 7, 9, 13, mientras que otros dan una mayor ponderación a los pares semejantes (referidos como coincidencias conjuntas), como es el caso de las medidas 2 y 5.

Otras medidas pueden dar una interpretación probabilística, como lo es la medida núm. 9 (Coeficiente de pares simples), donde se expresa simplemente la probabilidad de que un par de individuos logren el mismo valor sobre una variable seleccionada aleatoriamente. En un caso semejante está el coeficiente num. 10, pero a diferencia del núm. 6, el primero sólo considera la coincidencia conjunta de variables, mientras que el segundo, da igual valor a las relaciones de presencia-ausencia.

Similarmente, la medida núm. 1 expresa la probabilidad condicional de que en un par de individuos, tenga la manera de seleccionar aleatoriamente a una variable que se encuentre presente, dado que los pares ausentes son descartados primero. Una cuestión similar ocurre en los coeficientes num.7 y 13, donde, se expresa la probabilidad de seleccionar una variable en dos individuos, indistintamente de cual sea ella, importando el hecho de que coincida. Las medidas que dan un mayor peso a los pares presentes o ausentes, generalmente no tienen tal interpretación probabilística.

Al analizar los coeficientes de similaridad, se observa que, en su mayoría, no toman en cuenta las ausencias de variables en la comparación, siendo este un punto que es sujeto a discusión, ya que algunos autores consideran que las ausencias (o pares no

presentes), no implican que esa característica no sea registrada, porque su proporción sea baja en las entidades, argumentando que por esta razón deben ser tomadas en cuenta. Un punto que debe ser tomado en cuenta al calcular los coeficientes de similaridad es que no es válido efectuar comparaciones en base a las ausencias porque no indican realmente la medición de similitudes. Algunas veces, la similaridad se liga al concepto de asociación, ya que al conocer la forma en que se asocian ciertas características en las entidades, se pueden generar grupos semejantes. Por lo anteriormente expuesto se puede decir que:

Los coeficientes de similaridad presentan un gran uso dentro de la Biología, ya que mediante ellos se puede llegar a generar agrupaciones que ayuden a la descripción del fenómeno de estudio. Su desventaja es que para su cálculo son utilizados solamente datos cualitativos, principalmente aquellos del tipo presencia-ausencia, aunque algunas medidas de asociación pueden parcialmente solventar este problema, ya que algunos de ellos pueden trabajar datos cuantitativos como lo es el caso del coeficiente de correlación de Pearson.

Para seleccionar el coeficiente a trabajar, es muy importante no perder de vista el objetivo de estudio, el juicio y la experiencia del investigador, así como el tipo de datos que presenta el conjunto de datos.

Los coeficientes de similaridad que se recomiendan con ventajas sobre otros son:

1) Indice de Jaccard
2) Indice de comunidad de Dice.
3) Indice de Ochiai.
4) Indice de Mozley.
5) Indice de Mountford.
6) Coeficiente de Gower.

En ellos se encuentran ciertas ventajas, entre las que pueden mencionarse:

- Presentan una escala de 0 a 1, cayendo estos valores dentro del rango establecido por la similaridad.
- Son aplicables para datos tanto de abundancia como de presencia-ausencia.

A excepción del Coeficiente de Gower, los cuatro primeros índices toman en cuenta la coincidencia de especies básicamente y se recomienda su uso para el cálculo de la similitud, datos del tipo presencia-ausencia.

El coeficiente de Mozley se recomienda porque considera el número de sitios utilizados en la comparación efectuada, siendo este un buen coeficiente cuando se trabajan entidades con atributos comunes e interesa el número de sitios trabajados. El coeficiente de Mountford se recomienda cuando se trabajan tamaños de muestra diferentes en una misma comparación, siguiendo una distribución logarítmica.

Mountford mostró empíricamente que cuando los coeficientes de similaridad tienen una distribución logarítmica son más robustos a las diferencias en los tamaños de muestra. Por esta razón, se recomienda la aplicación de este coeficiente de similaridad

Debido a que los coeficientes 1-5 consideran datos de presencia-ausencia, se hace necesario considerar un coeficiente que abarque tanto datos de tipo cualitativo como cuantitativo, siendo muy adecuado para esto el Coeficiente de Gower, en el cuál, se abarcan casi todas las maneras en que un caracter puede ser codificado a diferencia de los otros coeficientes utilizados.

Por ello, se recomienda este coeficiente como la mejor medida de similaridad asimismo, la ventaja es que no se presentan problemas de dimensionalidad al momento de efectuar el análisis, siendo por ello, preferido a los anteriormente utilizados.

CAPITULO V
MEDIDAS DE DISTANCIA.

Una medida de distancia tiene como objetivo cuantificar diferencias entre entidades. Para ello, se han generado una serie de medidas que pueden ser clasificadas en:

1) Métrica
2) No métricas

Las métricas son utilizadas más frecuentemente porque presentan ciertas propiedades geométricas que las hacen más faciles de trabajar e interpretar, por esta razón en el presente trabajo se analizan prefentemente distancias métricas.

Propiedades formales de las métricas.

Llámese E a un conjunto de medidas de distancia en un espacio geométrico, y sean X,Y,Z los tres puntos que se representan en E, la función de distancia D, puede ser considerada como una métrica si y sólo si satisface las siguientes condiciones:

1) $D(X,Y)=0$ si y solo si $X=Y$
2) $D(X,Y)\geq 0$ para todo X y Y en E
3) $D(X,Y)=D(Y,X)$ para toda X y Y en E
4) $D(X,Y) \leq D(X,Z) + D(Y,Z)$ para todo X,Y y Z en E

La primera propiedad indica que la distancia entre Y y X es igual a 0, si X e Y son una misma entidad. La segunda propiedad expresa que el valor mínimo de las medidas de distancia es 0. La tercera propiedad menciona la simetría de valor de distancia entre dos puntos, X e Y. La cuarta condición se conoce como el axioma de la desigualdad del triángulo, en la cuál, se menciona que la suma de dos lados cualesquiera de un triángulo debe ser necesariamente mayor a la del tercer lado. Este axioma se representa en un espacio geométrico de la siguiente manera.

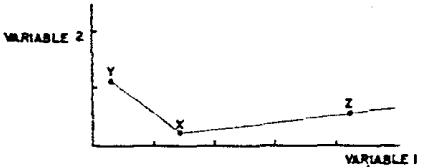


FIG. 1 Puntos X, Y y Z ubicados en un espacio geométrico. Claramente se observa $D(X,Y) \leq D(X,Z) + D(Y,Z)$

El fundamento teórico de las medidas de distancia métricas está dado por el teorema de Pitágoras. El cual permite representar por la hipotenusa de un triángulo rectángulo, la distancia que separa a dos entidades, tomando como base la siguiente expresión:

$$C = \sqrt{a^2 + b^2}$$

Esta expresión puede ser representada gráficamente ubicando cada punto dentro de un espacio geométrico de la siguiente forma:

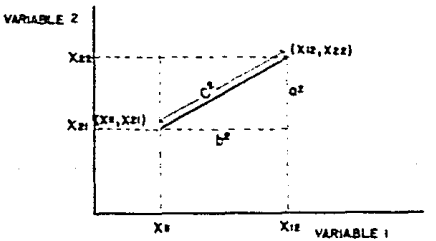
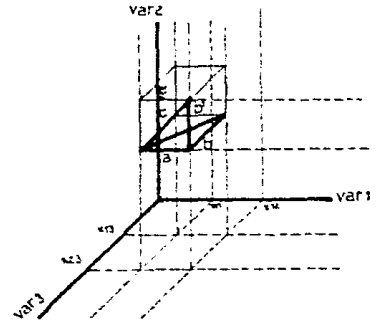


FIG. 2 Presentación geométrica del teorema de Pitágoras aplicado a las distancias métricas.

En este gráfico bidimensional la distancia entre dos puntos X_1 y X_2, puede expresarse como:

$$d_{(1,2)} = \left[(X_{21} - X_{11})^2 + (X_{22} - X_{12})^2 \right]^{1/2}$$

Donde solo se comparan dos entidades con dos atributos. En un espacio tri-dimensional, se hace una extensión del Teorema de Pitágoras el cual se aplica dos veces, ya que la distancia entre los puntos P_1 y P_2 es desconocido al igual que los catetos. Para visualizarlo considere la representación geométrica de tres puntos en un espacio tridimensional:



Para obtener el valor del primer cateto se calcula el segmento | P_1A|, de forma que:

$|P_1P_2|$= Hipotenusa

$|P_1\ A|= (X_{21}-X_{11})$ = Cateto a

$|A\ P_2|= (X_{21}-X_{11}) + (X_{22}-X_{12}) + (X_{23}-X_{13})$= Cateto b

Por lo tanto:

$$|P_1P_2| = \sqrt{(X_1-X_{11})^2 + (X_{21}-X_{11})^2 + (X_{22}-X_{12})^2 + (X_{23}-X_{13})^2}$$

Esta fórmula conduce a la representación generalizada de la distancia aplicada a n dimensiones, dada por:

$$d_{(i,j)} = \left[\sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \right]^{1/2}$$

Esta medida también se conoce como distancia euclidiana o pitagórica, que presenta las propiedades de las métricas, las cuales definen las relaciones entre los puntos dentro de un espacio geométrico, pudiendo de esta manera, representar adecuadamente a las entidades en un espacio n-dimensional.

MEDIDAS DE DISTANCIA METRICAS

Las medidas que a continuación se describen se han desarrollado preferéntemente para datos de tipo cuantitativo. Con el fin de ilustrar su uso se presenta una matriz de datos, con 3 entidades y 5 variables, a partir de estos datos, se hace el cálculo de las distancias entre las entidades 2 y 3.

	X_1	X_2	X_3	X_4	X_5
1	10	4	5	1	0
2	10	4	55	1	0
3	5	8	10	5	2

Fig 4 Matriz de datos originales

A continuación se mencionan una serie de medidas de distancia, dando una breve descripción y un desglose numérico de cada una de ellas.

1) Distancia euclidiana.

La euclidiana es la medida métrica más utilizada y familiar en las diferentes áreas de la Biología. Su fundamento teórico está sustentado por el Teorema de Pitágoras que define las relaciones que se establecen entre puntos en un espacio euclidiano (n dimensional), representado de la siguiente manera:

$$D_{ij} = \left(\sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \right)^{1/2}$$

Si se consideran solo dos entidades y dos variables, esta expresión mide la hipotenusa del triángulo rectángulo que se forma por los valores correspondientes de entidades y variables (Legendre, 1983). Gráficamente se expresa como:

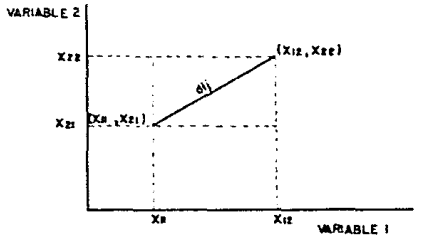


FIG. 4 La distancia euclidiana entre los individuos i y j para dos variables.

El rango de valores para esta expresión es de 0 a cualquier valor positivo. A continuación se muestra el desglose numérico para esta medida:

$$DE_{2,3} = \left((10-5)^2 + (4-8)^2 + (55-10)^2 + (1-5)^2 + (0-2)^2 \right)^{1/2}$$

$$DE_{2,3} = \left((5)^2 + (4)^2 + (45)^2 + (4)^2 + (2)^2 \right)^{1/2}$$

$$DE_{2,3} = 45.727$$

Dado que la distancia euclidiana es afectada por cambios de escala en las variables, se pueden trabajar con datos estandarizados o bien se puede incluir la expresión el denominador r_k como una normalización que frecuentemente se incluye cuando las variables son medidas en diferentes escalas.

Gauch (1982) reporta una expresión alternativa de la distancia euclidiana que incluye la expresión r_k de la siguiente manera:

$$DE_{i,j} = \frac{1}{p} \sum_{k=1}^{p} \left[\frac{(X_{ik} - X_{jk})^2}{r_k^2} \right]$$

donde p es el número de variables.

r_k = Rango de valores de la k-ésima variable.

El valor 1/p minimiza las diferencias que se presentan al incrementarse el número de variables, también hace que el valor obtenido en la expresión se encuentre del rango de 0 y 1. El cálculo numérico para esta forma alternativa se dá a continuación:

$$DE_{2,3} = \frac{1}{5} \left[\frac{(10-5)^2}{5^2} + \frac{(4-8)^2}{4^2} + \frac{(55-10)}{50^2} + \frac{(1-5)^2}{4^2} + \frac{(0-2)^2}{2^2} \right]$$

$$DE_{2,3} = \frac{1}{5} \left[\frac{25}{25} + \frac{16}{16} + \frac{2025}{2500} + 1 + 1 \right]$$

$$DE_{2,3} = 0.2 \ (1 + 1 + 0.81 + 1 + 1)$$

$$DE_{2,3} = 0.962$$

La distancia euclidiana se conoce en Taxonomía numérica como Taxonomic distance (Sokal, 1961) y se utiliza con propósitos taxonómicos.

2) Distancia cuadrada (SED).

En técnicas de agrupamiento como Cluster, esta medida se expresa como el cuadrado de la distancia euclidiana y se escribe de la siguiente forma:

$$SED^2_{i,j} = \sum_{k=1}^{p} (X_{ik} - X_{jk})^2$$

En ella se formula el cuadrado de las diferencias existentes al comparar dos entidades con respecto a las k variables. Esta medida se considera como una modificación de la distancia euclidiana, que elevada al cuadrado ($D^2_{i,j}$) se transforma en una medida semimétrica o no métrica, porque no sigue el axioma de la desigualdad del triángulo (Legendre, 1983). Su cálculo es:

$$DE_{2,3} = \left((10-5)^2 + (4-8)^2 + (55-10)^2 + (1-5)^2 + (0-2)^2 \right)$$

$$DE_{2,3} = \left((5)^2 + (4)^2 + (45)^2 + (4)^2 + (2)^2 \right)$$

$$DE_{2,3} = 2086$$

3) Distancia media o distancia promedio. (MED).

Esta modificación a la distancia euclidiana considera al número de variables (k) presentes en las entidades ya que influyen en la estimación de la distancia (Gauch, 1982). Su expresión matemática es:

$$MED_{i,j} = \frac{1}{p} \left(\sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \right)^{1/2}$$

o

$$MTD_{i,j} = \frac{\sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})}}{p}$$

Matemáticamente es semejante a la distancia euclidiana pero no se incluye la fracción $1/r_k$ lo cual ocasiona que esta medida tenga problemas de dimensiones al momento de trabajar con diferentes escalas de medición. Se representa en el espacio de una manera muy semejante a la distancia euclidiana, pero la distancia promedio se refiere al media de las diferencias entre los individuos i y j, para las k-ésimas variables. Matemáticamente, se desglosa como sigue:

$$MED_{2,3} = \frac{1}{5}\left[(10-5)^2 + (4-8)^2 + (55-10)^2 + (1-5)^2 + (0-2)^2\right]^{1/2}$$

$$MED_{2,3} = \frac{1}{5}\left[(5)^2 + (4)^2 + (45)^2 + (4)^2 + (2)^2\right]^{1/2}$$

$$MDE_{2,3} = 9.13454$$

4)Métrica de Manhattan o distancia absoluta (DA).

La métrica de Manhattan, tambien conocida como "City Block", se expresa como el cálculo de la sumatoria de las diferencias absolutas de las entidades i y j para las k variables. Matemáticamente se expresa como:

$$AD_{i,j} = \sum_{k=1}^{p} |X_{ik} - X_{jk}|$$

Gráficamente puede ser vista como:

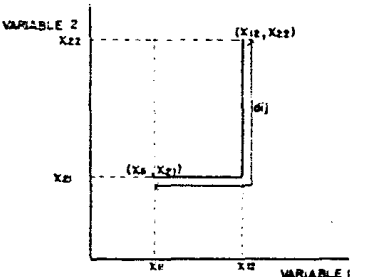


FIG. 5 Representación especial de la métrica de Manhattan para los los individuos i y j para dos variables.

Esta fórmula establece que la distancia entre dos puntos es el recorrido de los catetos. Su desglose numérico está dado como sigue:

$$MM_{2,3}=\left(|10-5| + |4-8| + |55-10|+|1-5|+|0-2|\right)$$

$$MM_{2,3}=\left(5 + 4 + 45 +4 + 2 \right)$$

$$MM_{2,3}=60$$

Al igual que la distancia euclidiana, la métrica de Manhattan se ve afectada por el número de variables y es sensible a los cambios en escala, problema que es solventando introduciendo el término r_k (rango de las k variables) en la cual Gauch (1982) da una expresión alternativa para la distancia de Manhattan de la siguiente forma:

$$MM_{ij} = \frac{1}{p} \sum_{k=1}^{p} \left(|X_{ik}-X_{jk}| / r_k \right)$$

Donde p= Número de variables cuantitativas.

r_k = Rango de valores de la k-ésima variable.

Su cálculo numérico es:

$$MM_{2,3}= \frac{1}{5} \left[\frac{|10-5|}{} + \frac{|4-8|}{} + \frac{|55-10|}{} + \frac{|1-1|}{} + \frac{|0-2|}{} \right]$$

$$MM_{2,3}= 0.2 \ (1 + 1 + 0.9 + 1 + 1)$$

$$MM_{2,3}= 0.2(4.9)$$

$$MM_{2,3}= 0.98$$

En Taxonomía Numérica, la métrica de Manhattan se conoce como la Diferencia Media de Caracteres (Mean Character Difference) y fué propuesta como una medida taxonómica por Cain y Harrison en 1958 (Crisci, 1983).

5) Distancia media absoluta

Esta medida de distancia presenta una gran semejanza con la Métrica de Manhattan, solo que la distancia media absoluta

considera las diferencias absolutas que se establecen entre la entidad i y j, sin ser afectada al incrementar el número de variables. Su expresión es:

$$MAD_{ij} = \frac{1}{p} \left(\sum_{k=1}^{p} |X_{ik} - X_{jk}| \right)$$

6) Métrica de Minkowski.

En esta medida de distancia, se incluyen como caso especial a la Euclidiana y la City Block, definiendose a la métrica de Minkowski de la siguiente manera (Reynolds, 1988):

$$M_{ij} = \left(\sum_{k=1}^{p} |(X_{ik} - X_{jk})|^{r} \right)^{1/r}$$

Cuando r=1, esta medida se convierte en la City Block, y cuando r=2, se transforma en la distancia euclidiana. Si el valor de r es mayor que dos su utilidad en la ecología es reducida, ya que al tomar un valor mayor que 2 se da una mayor importancia a las desviaciones generadas al efectuar la diferencia $|X_{ik} - X_{jk}|$. A continuación, se ejemplifica su cálculo para r=1, r=2 y r=3.

Con r=1:

$$M_{2,3} = \left((10-5)^1 + (4-8)^1 + (55-10)^1 + (1-5)^1 + (0-2)^1 \right)^1$$

$$M_{2,3} = \left(5^1 + 4^1 + 45^1 + 4^1 + 2^1 \right)^1$$

$$M_{2,3} = 5 + 4 + 45 + 4 + 2$$

$$M_{2,3} = 60$$

Con r=2

$$M_{2,3}= \left((10-5)^2+ (4-8)^2+ (55-10)^2+ (1-5)^2+ (0-2)^2 \right)^{1/2}$$

$$M_{2,3}= \left(5^2 + 4^2 + 45^2 + 4^2 + 2^2 \right)^{1/2}$$

$$M_{2,3} = 45.6727$$

Con r=3

$$M_{2,3}= \left((10-5)^3+ (4-8)^3+ (55-10)^3+ (1-5)^3+ (0-2)^3 \right)^{1/3}$$

$$M_{2,3}= \left(5^3 + 4^3 + 45^3 + 4^3 + 2^3 \right)^{1/3}$$

$$M_{2,3}= 45.042$$

7) Métrica de Canberra.

Los australianos Lance y Williams (1967), modificaron la métrica de Manhattan, originando la métrica de Canberra, cuya expresión está dada como sigue (Gauch, 1982):

$$C_{ij}= \frac{1}{p} \sum_{k=1}^{p} \frac{|X_{ik} - X_{jk}|}{(X_{ik} + X_{jk})}$$

Por esta fórmula se suma el total del valor absoluto de las diferencias generadas al comparar a los individuo i y j para las k variables presentes, dividido por el total de comparaciones. Esto muestra que las diferencias entre los individuos contribuyen en el cálculo de la distancia, así como puede hacerse con especies raras. Por esta razón, es importante que la entidad se encuentre presente en la comparación y trata de no tomar en cuenta a las ausencias en su cálculo. El rango de valores que toma es de 0 a 1.

Geométricamente se representa como una proporción que expresa a la división de las diferencias y la suma de valores para los individuos i y j, es decir, indica una proporción que representada en los catetos que no recorre el total de ellos sino solo una proporción de ellos y se genera la siguiente gráfica:

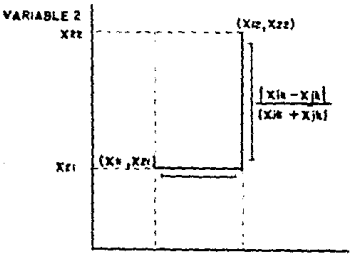


FIG. 6 Representación geométrica de la distancia de Canberra de los individuos i y j para dos variables. Esta medida recorre una proporción de catetos.

El cálculo de la métrica de Canberra está dado como sigue:

$$C_{23} = \frac{1}{5}\left[\frac{|10-5|}{(10+5)} + \frac{|4-8|}{(4+8)} + \frac{|55-10|}{(55+10)} + \frac{|1-5|}{(1+5)} + \frac{|0-2|}{(0+2)} \right]$$

$$C_{23} = \frac{1}{5}\left(\frac{5}{15} + \frac{4}{12} + \frac{45}{65} + \frac{4}{6} + \frac{2}{2} \right)$$

$$C_{23} = 0.2(0.333 + 0.333 + 0.6923 + 0.666 + 1) = 0.605113$$

Stephenson et al (972) y Moreau & Legendre, (1979) dieron una expresión alternativa de la métrica de Canberra, que es reportada por Legendre (1983) de la siguiente forma:

$$S = 1 - \left[\left[\frac{1}{p} \sum_{k=1}^{p} \frac{|X_{ik} - X_{jk}|}{(X_{ik} + X_{jk})} \right] \right]$$

Esta expresión es apropiada para ser trabajada con datos normalizados, y tiene como rango a 0 y 1.

8) Coeficiente de divergencia.

Este coeficiente es considerado como una versión normalizada de la métrica de Canberra, pero se menciona como una modificación de la distancia euclidiana. Fué utilizada por Clark (1952) con propósitos taxonómicos, tomando para este efecto el nombre del coeficiente de divergencia (Legendre 1983), expresado como:

$$D_{ij} = \left[\frac{1}{p} \sum_{k=1}^{p} \left[\frac{|X_{ik} - X_{jk}|}{(X_{ik} + X_{jk})} \right]^2 \right]^{1/2}$$

Para su cálculo, se toma en cuenta la diferencia existente entre entidades para cada variable y se divide por el total de comparaciones realizadas, donde esta relación se expresa como una fracción elevada al cuadrado que y toma en consideración el número de variables presentes. Una vez obtenida esta fracción, se representa de manera semejante a la euclidiana en la que los catetos se encuentran representando una proporción de los valores originales. Geométricamente, es representada de manera similar a la euclidiana, pero es necesario aclarar que la distancia representada por la hipotenusa no es recorrida completamente, sino solo una proporción de ella. Geométricamente se puede representar como sigue:

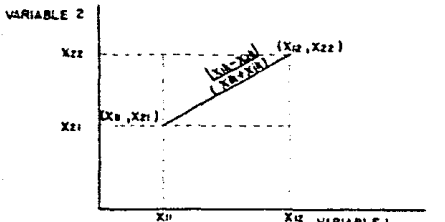


FIG.7 Representación geométrica del Coeficiente de divergencia. Esta medida solo recorre una proporción de la hipotenusa.

A continuación se muestra su cálculo de la siguiente forma:

$$CD_{ij}=\left[\frac{1}{5}\left[\frac{10-5}{10+5}\right]^2+\left[\frac{4-8}{4+8}\right]^2+\left[\frac{55-10}{55-10}\right]^2+\left[\frac{1-5}{1+5}\right]^2+\left[\frac{0-2}{0+2}\right]^2\right]^{1/2}$$

$$CD_{ij}=\left((0.2)\left((0.333)^2+(0.333)^2+(0.6923)^2-(0.666)^2+1\right)\right)^{1/2}$$

$$C_{ij}=[(\ 0.2\ (2.512399)\ ]^{1/2}$$

$$C_{ij}=0.781183$$

9) Medida de Asociación de Whittaker.

Originalmente el índice de asociación de Whittaker se utilizó para trabajar con datos de abundancia (Legendre, 1983) y se aplica también al comparar entidades para k variables. El valor obtenido indica la diferencia que existe entre el individuo i y j con respecto a cada variable, considerando para cada uno de ellos, el número total de entidades presentes en la muestra, expresado como una fracción.

Esto se expresa como la sumatoria del valor absoluto de la sustracción de la entidad i y j tomando en cuenta al número total de entidades. Su fórmula es:

$$W_{ij} = \frac{1}{2}\left[\sum_{k=1}^{p}\left|\left[\frac{X_{ik}}{\sum_{k=1}^{p} X_{ik}}\right] - \left[\frac{X_{jk}}{\sum_{k=1}^{p} X_{jk}}\right]\right|\right]$$

La diferencia de 0 para una entidad, indica la existencia de una proporción idéntica de ellas en ambas muestras. Su desglose numérico es como sigue:

$$W_{2,3} = \frac{1}{2}\left|\left[\frac{10}{70} - \frac{5}{30}\right] \cdot \left[\frac{4}{70} - \frac{8}{30}\right] \cdot \left[\frac{55}{70} - \frac{10}{30}\right] \cdot \left[\frac{1}{70} - \frac{5}{30}\right] \cdot \left[\frac{0}{70} - \frac{2}{30}\right]\right|$$

$$W_{2,3} = \frac{1}{2}\left|(0.1428-0.1666)+(0.05714-0.2666)+(0.7857-0.3333)+(0.01428-0.1666)+(0-0.0666)\right|$$

$$W_{2,3} = \frac{1}{2}(0.00046)=0.00023$$

10) Distancia relativa (RED)

Esta ecuación se considera una modificación del índice de Asociación de Whittaker (1952), en el cual se considera a las diferencias existentes entre las entidades que son divididas por la totalidad presente de ellas expresadas como una fracción y eleva cada sustracción al cuadrado. La forma que presenta esta medida es la siguiente (Reynolds, 1988):

$$DER_{ij} = \sum \left[\left[\left[\frac{X_{ik}}{\sum_{k=1}^{p} X_{ik}} \right] - \left[\frac{X_{jk}}{\sum_{k=1}^{p} X_{jk}} \right] \right]^2 \right]^{1/2}$$

Esta expresión tiene una gran semejanza tanto geométrica como matemática con la distancia euclidiana, donde cada valor referido como una fracción relativa de el. A diferencia de la euclidiana, su rango de valores cae entre 0 y $\sqrt{2}$. Su cálculo numérico está dado como sigue:

$$W_{2,3} = \frac{1}{2}\left[\left[\frac{10}{70} - \frac{5}{30} \right]^2 + \left[\frac{4}{70} - \frac{8}{30} \right]^2 + \left[\frac{56}{70} - \frac{10}{30} \right]^2 + \left[\frac{1}{70} - \frac{5}{30} \right]^2 + \left[\frac{0}{70} - \frac{2}{30} \right]^2 \right]^{1/2}$$

$$W_{2,3} = \frac{1}{2}\left\{ (0.1428-0.1666)^2 + (0.05714-0.2666)^2 + (0.7857-0.333)^2 + (0.01428-0.1666)^2 + (0-0.06663)^2 \right\}^{1/2}$$

$W_{2,3} = 0.526363$

11) Distancia relativa absoluta (DRA)

Al igual que la distancia relativa, está medida se considera como una correción hecha al índice de asociación de Whittaker, pero la distancia relativa absoluta no dá el mismo peso a las diferencias obtenidas. Se define de la siguiente manera:

$$RAD_{i\,j} = \sum_{k=1}^{p} \left| \left[\frac{X_{ik}}{\sum_{i=1}^{p} X_{ik}} \right] - \left[\frac{X_{jk}}{\sum_{j=1}^{p} X_{jk}} \right] \right|$$

A diferencia de los índices 9 y 10, el rango de valores para la distancia relativa absoluta varia de 0 a 2. El manejo numérico para esta expresión es:

$$DRA_{2,3} = \left| \left[\frac{10}{70} - \frac{5}{30} \right] \cdot \left[\frac{4}{70} - \frac{8}{30} \right] \cdot \left[\frac{5}{7} - \frac{10}{30} \right] \cdot \left[\frac{1}{70} - \frac{5}{30} \right] \cdot \left[\frac{0}{70} - \frac{2}{30} \right] \right|$$

$$DRA_{2,3} = \left| (0.1428-0.1666)+(0.05714-0.2666)+(0.7857-0.3333)+(0.01428-0.1666)+(0-0.0666) \right|$$

12) Chord Distance (Distancia cuerda).

La Chord Distance se considera como una distancia euclidiana, que puede calcularse por la normalización de las muestras, considerandolas como vectores, donde la longitud para cada vector formado es de 1, que es equivalente al valor obtenido para la distancia euclidiana, por lo tanto, la longitud que existe entre dos pares de puntos, se representa por una cuerda que toma la forma de una esfera con radio igual a 1. Esto puede ser representado de la siguiente manera:

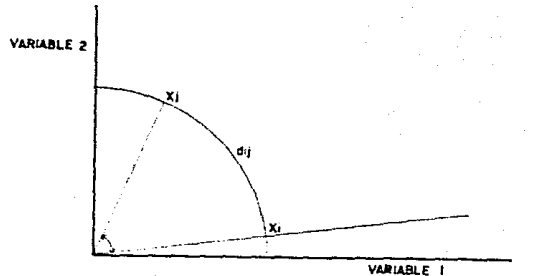


FIG. 8 Representación espacial de la Chord Distance entre dos entidades i y j.

Si se trabajan dos entidades, la normalización de los puntos se coloca dentro de la circunferencia de un círculo con radio ígual a 1. La Chord Distance puede calcularse directamente de datos no normalizados, por la siguiente expresión (Legendre, 1983):

$$CRD_{ij} = \left[2 \left[1 - \frac{\sum_{k=1}^{p} [(X_{ik})(X_{jk})]}{\sqrt{\sum_{i}^{p}(X_{ik}^2) \sum_{j}^{p}(X_{jk}^2)}} \right] \right]^{1/2}$$

Esta medida toma un valor máximo cuando las entidades presentes en la muestra son completamente diferentes, siendo la separación entre ellas de 90°, tomando como máximo a $(p)^{1/2}$. Asimismo, solventa el problema causado por la escala de medición. de una manera reducida, esta fórmula se encuentra reportada bibliográficamente como:

$$CRD_{ij} = \sqrt{[2\ (1-\cos_{ij})}$$

donde:

$$\cos_{i,j} = \frac{\sum_{k=1}^{p} [(X_{ik})(X_{jk})]}{\sqrt{\sum_{i}^{p}(X_{ik}^2) \sum_{j}^{p} (X_{jk}^2)}}$$

Su cálculo está dado como sigue:

$$CRD_{23} = \left[2 \left[1 - \frac{(10)(5)+(4)(8)+(55)(10)+(1)(5)}{((3142)(218))^{1/2}} \right] \right]^{1/2}$$

$$CRD_{23} = (0.460647466)^{1/2}$$

$$CRD_{23} = 0.678710149$$

13)Distancia Geodésica:

La distancia geodésica se considera una transformación de la medida Chord Distance, con la diferencia que la geodésica mide la longitud del arco formado por los vectores que representan a las entidades i y j a través de la superficie de la esfera de radio 1. Esto se representa como sigue:

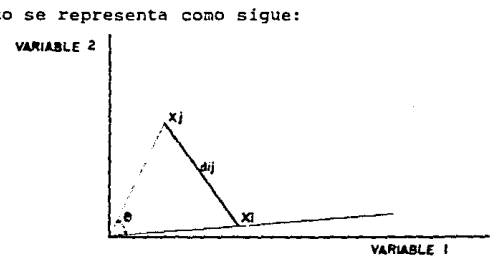


FIG. 9 Representación especial de la Distancia Geodésica entre las entidades i y j

La representación matemática de esta medida es:

$$=\arccos\left[1-\left[2\left[\frac{1-\dfrac{\sum_{k=1}^{p}\left[(X_{ik})(X_{jk})\right]}{\sqrt{\left(\sum_{i}^{p}X_{ik}^{2}\right)\left(\sum_{j}^{p}X_{jk}^{2}\right)}}}{2}\right]^{1/2}\right]^{2}\right]$$

El rango de valores que toma esta medida es de o a $\Pi/2$, (0 a 1.57).Se calcula como sigue:

$$G_{23}=\arccos\left[1-\left[2\left[\frac{1-\dfrac{(10)(5)+(4)(8)+(55)(10)+(1)(5)}{(470)^{2}(30)^{2})^{1/2}}}{2}\right]^{1/2}\right]^{2}\right]$$

-

$G_{2,3} = \arccos[1-(2(1-0.30333))]$

$G_{2,3} = \arccos 1-(1.18039)$

$G_{2,3} = \arccos(-0.1838)$

$G_{2,3} = 100.5911$

14) Métrica de Bray-Curtis

La métrica de Bray-Curtis fué sugerida por Lance y Williams (1966), se considera una extensión de la métrica de Manhattan que toma el valor absoluto de las diferencias entre los individuos i y j tomando en cuenta el total de variables. Su representación está dada por:

$$Dac = \frac{\sum_{k=1}^{p} |X_{ik} - X_{jk}|}{\sum_{k=1}^{p} |X_{ik} + X_{jk}|}$$

La métrica de Bray-Curtis presenta ciertas ventajas al trabajar con datos estandarizados, aunque también pueden ser útiles para su cálculo datos originales. Es sensitiva a los valores extremos (outliers) y su uso en Biología es restringido porque se considera no métrica, pues no cumple con el axioma de la desigualdad del triángulo.

Su desglose numérico está dado como sigue:

$$Dac = \frac{|10-5| + |4-8| + |55-10| + |1-5| + |0-2|}{|10+5| + |4+8| + |55+10| + |1+5| + |0+2|}$$

$$Dac = \frac{5 + 4 + 45 + 4 + 2}{20 + 12 + 65 + 6 + 2}$$

$$Dac = \frac{60}{105}$$

$$Dac = 0.571428$$

Otra medida no métrica puede mencionarse dentro de este contexto, citada por Duran y Odell (1974), expresandose como sigue:

$$D_{i,j} = \left[\sum \sqrt{X_{ik}} - \sqrt{X_{jk}} \right]^{1/2}$$

15) Coeficiente de correlación producto-momento.

Para llegar a la medida de distancia de correlación, es necesario considerar lo siguiente:

La media de la variable X y Y, tomada de los datos, se calcula como:

$$\bar{X} = \sum_{i=1}^{i=p} \frac{X_i}{p} \qquad \bar{Y} = \sum_{i=1}^{i=p} \frac{Y_i}{p}$$

Si la variable de la media se resta al valor original, el vector central de valores se obtiene como:

$$\hat{X}^T = [(X_1-\bar{X}), (X_2-\bar{X}) \ldots (X_n-\bar{X})],$$

$$\hat{Y}^T = [(Y_1-\bar{Y}), (Y_2-\bar{Y}) \ldots (Y_n-\bar{Y})]$$

Tanto $\hat{X}^T$ y $\hat{Y}^T$ presentan una media 0. El producto interno de dos vectores centrales (conocido como el vector de desviaciones), se toma como la dispersión de X y Y. Si la dispersión se divide por n, entonces la covarianza y la varianza se reconoce como:

$$\text{Cov}(X,Y) = \frac{\hat{X}^T \hat{Y}}{p} = \frac{1}{p} \sum_{i=1}^{i=p} (X_i-\bar{X})(Y_i-\bar{Y})$$

La varianza y la covarianza utilizan la media de cada variable, produciendo la matriz de dispersión S de las variables, que resulta de la multiplicación de la matriz de datos centralizados en la media $(X - \bar{X})$, con su transpuesta $(X - \bar{X})'$, siendo esta una medida de la dispersión conjunta entre dos variables con respecto a la media.

$$\text{Var}(X) = \frac{\hat{X}^T \hat{Y}}{p} = \frac{1}{p} \sum_{i=1}^{i=p} (X_i-\bar{X})^2$$

siendo la diferencia de la dispersión de una sola variable con respecto a la media.

La covarianza de Y y X es también conocida como el producto momento de Y y X, lo mismo que la varianza var(x) es el producto momento de X. La correlación, producto momento se define como:

$$r = \frac{Cov\ (X,Y)}{[Var\ (X)\ Var\ (Y)]^{1/2}} = \frac{\sum_{i=1}^{i=p} (X_i-\bar{X})(Y_i-\bar{Y})}{\left[\sum_{i=1}^{i=p} (X_i-\bar{X})^2 \sum_{i=1}^{i=p} (X_i-\bar{X})^2\right]^{1/2}}$$

El coeficiente de correlación de Pearson es la covarianza, estandarizada entre la desviación estandar de las dos variables y/o entidades comparadas, que produce valores entre -1 y 1. Asimismo, puede ser descrita como la contabilización de la covarianza de los datos estandarizados, que introducen a la media y a la desviación estandar de la distribución de frecuencias de los datos

De esta manera se expresa el coeficiente de correlación producto-momento como una medida de similitud, que puede interpretarse también como una medida de distancia entre entidades y que se conoce como el complemento del coeficiente de correlación, teniendo la siguiente forma:

$$Cc = 1 - \left| \frac{\sum_{i=1}^{i=p} (X_i-\bar{X})(Y_i-\bar{Y})}{\left[\sum_{i=1}^{i=p} (X_i-\bar{X})^2 \sum_{i=1}^{i=p} (X_i-\bar{X})^2\right]^{1/2}} \right|$$

Medidas de distancia y similaridad inter-grupo

Las medidas de distancia y similaridad intergrupo citadas anteriormente, están relacionadas con la medición entre pares de individuos donde no hay causa-efecto. En muchos casos, lo que interesa es saber que tanto de ciertas mediciones pueden realizarse entre dos grupos, cada uno conteniendo algunos individuos. Esto puede darse cuando los datos contiene grupos formados "a priori" y el objetivo es examinar de alguna manera, la relación que se establece entre grupos, o bien, cuando los datos han estado sujetos a alguna forma de análisis (por ejemplo Cluster) y la configuración de los grupos resultantes es sometida a investigación, utilizando algunas técnicas de ordenación, por ejemplo, el análisis de Componentes Principales (ACP), Escalamiento Multidimensional (MS), BIPLOT, etc.

La similaridad o distancia establecida entre grupos, ha sido definida de muchas formas. Un procedimiento obvio puede consistir en tomar la similaridad o distancia "promedio" entre pares de individuos, uno de cada grupo.

Esto es aceptable sólo si el concepto de un promedio es lo suficientemente adecuado para la particular medida de distancia o similaridad inter-individuo utilizada.

Un procedimiento intuitivo y razonable puede ser laa medición de la similaridad intergrupal, utilizando una medida interindividual sobre la media grupal. Por ejemplo, se puede ejemplificar a la distancia euclidiana entre las medias como una medida entre grupos. Algunas de las medidas intergrupales son dadas a continuación:

16) Coeficiente de semejanza racial (Racial Likeness).

Este coeficiente fué desarrollado por Pearson (1926), fundamentalmente con propósitos taxonómicos. Hace posible la medición de la distancia entre grupos de muestras de manera

semejante a como lo hace la distancia generalizada de Mahalanobis, que elimina en su cálculo, los efectos originados por la correlación entre variables. Su fórmula es:

$$CSR_{ij} = \left[\frac{1}{p} \sum_{k=1}^{p} \left[\frac{(x_{i,k} - x_{j,k})}{\frac{s^2_{i,k}}{P_1} \cdot \frac{s^2_{j,k}}{P_2}} \right] - \frac{2}{p} \right]^{1/2}$$

Para dos grupos de muestras, X_1 y X_2 que contienen a las muestras P_1 y P_2 respectivamente, X_i y X_j son el valor promedio de la entidad i y j en la muestra, S_i^2 y S_j^2 son las varianzas de ambas entidades.

17) Distancia generalizada para variables discretas.

Para datos categóricos, Kurczynski (1970), definio la siguiente medida de distancia generalizada entre grupos:

$$d_{(G_1G_2)} = \frac{1}{?} \log \frac{\left| \frac{1}{2} (W_{g_1} + W_{g_2}) \right| + \frac{1}{4} (\bar{X}G_1 - \bar{X}G_2)^T (\bar{X}G_1 - \bar{X}G_2)}{|WG_1|^{1/2} |WG_2|^{1/2}}$$

Donde WG_1 y WG_2 son las matrices de varianza-covarianza para los grupos G_1 y G_2, y $|WG_1|$ denota al determinante de la matriz WG. Cuando $WG_1 = WG_2$ esta expresión se reduce a la distancia de Mahalanobis.

18) Distancia de Mahalanobis (1936).

La distancia de Mahalanobis es una distancia generalizada de distancia para variables contínuas, que deja fuera los efectos de la correlación entre variables y es independiente de las escala de las variables. Esta medida dá la distancia entre dos puntos en un

espacio cuyos ejes no son ortogonales y toma en consideración la correlación entre variables.

La distancia de Mahalanobis es utilizada principalmente para la comparación entre grupos de entidades. Cuando se agrupan muestras, es importante tener en cuenta los efectos de la correlación entre variables, constituyendo esto la base de la estructura buscada. Cuando se comparan dos grupos de muestras W_1 y W_2 que tienen P_1 y P_2 muestras respectivamente, todas descritas por las n variables, el cuadrado de la distancia generalizada toma la siguiente forma matricial:

$$D^2_{(W_i,W_j)} = \bar{d}_{ik}' \; W^{-1} \; \bar{d}_{ik}$$

Donde:

$\bar{d}_{ik}$ = Vector de las diferencias entre las medias de las n variables en los dos grupos de muestras.

W = Matriz de dispersión conjunta (conocida como la matriz de varianza-covarianza) de los dos grupos de muestras estimadas de la suma de las matrices y de los productos cruzados entre variables centralizadas para cada uno de los dos grupos adicionados término a término y dividido por los grados de libertad:

(P_i+P_j-2): $S=(1/P_i+P_j-2)((P_i-1)W_i+(P_j-1)W_j)$

donde W_i y W_j son las matrices de dispersión de los dos grupos; por lo tanto, el vector $\bar{d}$ mide la diferencia entre las medias de puntos de dos grupos en un espacio n dimensional (de n variables), donde W toma en consideración la covarianza de algunas variables.

COMENTARIOS

En el análisis de las diferentes medidas de distancia se puede mencionar que:

Para seleccionar una medida de distancia adecuadamente, se requiere de un amplio conocimiento de las propiedades del conjunto de datos y de los efectos que tengan sobre cada medida.

Las medidas de distancia trabajan preferentemente con datos cuantitativos contínuos.

Es necesario considerar la escala de medición de las variables, ya que si hay diferencias en escalas en las k-variables la comparación no sería adecuada porque algunas medidas son sensibles a los cambio en escala. Una alternativa para solucionar este problema es estandarizar las variables. La estandarización se recomienda cuando las variables cambian drásticamente o bien cuando se tienen variables medidas en diferentes escalas, pues lo que se busca es disminuir las diferencias entre las variables.

Puede considerarse que las medidas de distancia son derivadas básicamente de la distancia Euclidiana y la distancia de Manhattan, porque conceptualmente presentan una forma matemática semejante, midiendo completamente la hipotenusa, los catetos, o bien una proporción de ellos. aunque su interpretación depende del objetivo y del juicio del investigador.

Existen algunas medidas que no necesariamente se derivan de la Distancia Euclidiana o la distancia de Manhattantal es el caso de la Chord Distance, la distancia Geodésica y el Coeficiente de Correlación producto momento, en los cuales se plantean vectores que representan a las entidades comparadas para las k- variables, considerando el ángulo formado entre ellos.

Las medidas que fueron seleccionadas por sus características son:

- Medida de la distancia euclidiana.
- Métrica de Manhattan.
- Distancia Geodésica.
- Chord distance.
- Métrica de Canberra.
- Coeficiente de correlación producto momento.

Al aplicar datos originales en la distancia euclidiana, se presenta la desventaja de que se ve seriamente afectada por las diferentes escalas de medición que se presentan en la matriz de datos originales. La métrica de Manhattan es preferible sobre la distancia euclidiana para este tipo de datos. La Chord Distance no presenta este tipo de problemas, no siendo necesario trabajar datos estandarizados.

La métrica de Canberra se ve afectada por la ausencia de datos en la comparación, no siendo una medida eficaz cuando en el conjunto de datos se presenten ausencias o bien no se hayan registrado datos para alguna muestra en especial.

Una medida adecuada para comparar grupos de entidades es la distancia de Mahalanobis (D^2), permite la correlación entre variables y la varianza es considerada. Cuando la correlación es 0, la distancia de Mahalanobis es equivalente a la distancia euclidiana utilizando variables estandarizadas.

CAPITULO VI

ESTUDIO DE CASO : LAS MEDIDAS DE DISTANCIA COMO UNA HERRAMIENTA PARA EL ANALISIS DE LA DISTRIBUCION DEL FITOPLANCTON

Se analizó un conjunto de datos mediante el cálculo de algunas medidas de similitud y distancia seleccionadas de acuerdo con las características de los datos.

Para mostrar la aplicación de las medidas seleccionadas y apoyar los criterios biológicos de la distribución del fitoplancton, se usaron datos originales para el cálculo de las siguientes medidas de distancia: Canberra, Manhattan, Euclidiana, Geodésica, Chord Distance y el Coeficiente de similaridad de Gower. Con los valores obtenidos se construyeron las matrices de similaridad y distancia correspondientes, analizando sólo 2 de ellas con el fin de describir de manera general la distribución fitoplanctónica, así como la elección de la medida más adecuada para este tipo de datos.

DISTRIBUCION ESPACIAL Y TEMPORAL DEL FITOPLANCTON.

Los organismos que habitan en un ecosistema de agua dulce se pueden clasificar en tres grandes grupos: descomponedores, consumidores y productores. Los productores primarios predominantemente lo constituye el fitoplancton que está formado por organismos unicelulares productores de materia orgánica a base de fotosíntesis. El fitoplancton incluye organismos que pertenecen a distintas divisiones algales como las Cianophytas, Chlorophytas, Bacillariphytas, Euglenophytas, Crisophytas y Pirrophytas entre otras.

Este estudio se realizó en la laguna "el Rodeo", Mor., que se localiza en una zona tropical y su importancia radica en que es el principal aporte de agua a la piscifactoria "Fernando Obregón", además de que sus aguas alimentan el riego de cultivos aledaños; también se lleva a cabo la pesca de tipo local y funciona como centro recreativo.

UBICACION GEOGRAFICA.

La laguna "El Rodeo" se localiza en el municipio de Miacatlán, situado en la parte centro-oeste del estado de Morelos, colindando al Norte con el Estado de México y municipio de Cuernavaca, al Este con los municipios de Temixco y Xochitepec, al Oeste con el municipio de Coatlán del Rio y al Sur con el municipio de Mazatepec y Puente de Ixtla (Detenal, 1979).

METODO

Para analizar la distribución del fitoplancton en la laguna "El Rodeo" se efectuaron 10 muestreos, uno por mes en el periódo comprendido de septiembre de 1989 a junio de 1990. Para obtener al fitoplancton se hicieron muestreos tanto cualitativos, con una red de arrastre tipo Zeppelin, como cuantitativos utilizando para ello una botella Van Dorn a diferentes profundidades. Las muestras así obtenidas se fijaron con Acetato de Lugol.

Para realizar el recuento del fitoplancton, se pusieron a sedimentar 9 ml de la muestra tomada de la botella Van Dorn en una cámara de sedimentación durante 24 hrs; posteriormente se revisaron de 10 a 20 campos elegidos al azar con el objetivo de 40X en un microscopio invertido, para determinar los organismos presentes por División y la obtención de las densidades fitoplanctónicas a través del número de organismos presentes en las muestras. La cantidad de organismos pertenecientes a cada división localizada a diferentes niveles se representó en tablas de densidades fitoplanctónicas (expresadas en unidades biológicas/litro).

Ya formadas las tablas de densidades fitoplanctónicas se procedió a analizar las características de los datos, esto con el fin de aplicar de una manera adecuada las medidas de similitud y distancia que fueron seleccionadas. Para calcular las diferentes medidas se elaboró un programa de computación en lenguaje fortran.

La aplicación de las medidas de similitud y distancia se realizó con el enfoque multivariado comparando los niveles correspondientes a cada estación muestreada con las diferentes divisiones fitoplanctónicas interpretando solo dos matrices correspondientes al mes de noviembre y febrero, esto con el fin de describir adecuadamente los datos y sobre todo, mostrar su aplicación a un conjunto de datos especifico.

RESULTADOS.

A través de muestreos cualitativos y cuantitativos se obtuvieron tablas de densidades fitoplanctónicas, para los meses muestreados (tablas 1 y 2), las cuales fueron transformadas en matrices de similaridad y distancia por el cálculo de la medida de similitud de Gower y las distancias: Euclidiana, Manhattan, Canberra, Chord distance y Geodésica, seleccionadas de acuerdo al tipo de datos del estudio, procesando solo dos matrices correspondientes a los meses de noviembre y febrero para cada medida seleccionada (matrices 3-10).

Se definió la estructura de la matriz de datos verificando si estos presentaban o no diferencias en escala y se observó que este problema no existia en las variables comparadas por lo que no se necesito estandarizar. Se recomienda estandarizar cuando hay cambios en las variables medidas o cuando se presentan en distintas escalas, pues lo que se busca es disminuir las variaciones entre los valores de las variables.

Con los datos seleccionados se hizo el cálculo de las medidas de distancia comparando dos entidades (i y j) para todas las k-variables formando vectores de comparación entre entidades, que pertenecen a cada uno de los valores de la matriz de similitud y/o distancia. Esto dá como resultado una matriz simétrica que contiene a los valores de distancia y similitud de cada comparación.

Para analizar las matrices de similaridad y distancia correspondientes a los meses de noviembre y febrero, se compararon

los niveles 0.3 y 1.0 de las 5 estaciones muestreadas y puede decirse que:

No existe mucha variación entre niveles siendo semejantes entre sí. Esto se refleja en los valores obtenidos en las matrices de similaridad con valores altos (cercanos a 1) y en los de distancias valores bajos (cercanos a 0).

Al comparar las matrices de distancia correspondientes a los meses de noviembre y febrero para las distancias Euclidiana y Chord Distance, se observa una tendencia hacia valores bajos (varían de 0 a 0.549) en las 5 estaciones muestreadas. Con respecto a la matriz generada por la distancia de Manhattan el valor más alto es de 3.33 y el menor es de 0.129.

Estos resultados indican que las diferencias entre cada nivel para las 5 estaciones son pocas y son semejantes entre sí en cuanto a cantidad de organismos que las demás estaciones, exceptuando a la estación I la cual tiene algunas variaciones en los valores obtenidos. Esta situación se corrobora con la matriz de similaridad de Gower, en la que los valores obtenidos para las 5 estaciones son altos (entre 0.84 y 0.96) indicando una gran semejanza en la cantidad de organismos en cada nivel muestreado.

Esta situación es muy similar en ambas matrices (pertenecientes al mes de noviembre y febrero) pero en el mes de febrero las matrices de distancia presentan valores más pequeños entre los niveles comparados cuyos valores oscilan entre 0.020 y 0.29) para las distancias Euclidiana, Chord Distance, Geodésica, excepto para Manhattan en la que los valores son entre 0.93 y 1.63). Esto indica una mayor semejanza entre los niveles 0.3 y 1.0 respectivamente. En la matriz de similaridad de Gower se tienen valores de similaridad altos (0.8 a 1.0) corroborando que la abundancia de fitoplancton entre estaciones y a diferentes niveles comparados son semejantes, excepto para la estación I en la cual se tiene cierta variación en los resultados, esto puede indicar que la estación I no es muy semejante a las demás estaciones.

El análisis efectuado entre cada nivel de las 5 estaciones muestreadas indica que para el mes de noviembre, en la estación I existen algunas diferencias en la abundancia fitoplanctónica, situación que se corrobora en las matrices de distancia trabajadas, mientras que en el mes de febrero, esta situación es semejante pero el valor obtenido es menor con relación al mes de noviembre.

En general se puede decir que en los meses de noviembre y febrero, existe una gran semejanza en la abundancia fitoplanctónica para los diferentes niveles de la laguna y que las estaciones II, III IV y V son muy semejantes entre sí mientras que la estación I (en términos de similaridad) tiende a ser menos semejante que las estaciones mencionadas anteriormente. Entonces se puede decir que el fitoplancton presenta una distribución casi homogénea a lo largo de las estaciones muestreadas. Esto puede deberse en parte a la morfometría de la laguna - que tiende a ser circular según Govea (1986)-, permitiendo una distribución homogénea del fitoplancton y las variaciones que pudiesen existir pueden deberse en parte al método de muestreo utilizado.

Las matrices que expresan mejor los resultados que se mencionan anteriormente son las generadas por la métrica Chord Distance y la medida de similaridad de Gower ya que con Chord Distance hay una mejor expresión de los valores porque indican que existe una gran similitud entre los niveles, concordando esto con las observaciones realizadas en campo y con algunos criterios biológicos, reflejandose con valores de distancias pequeños y de similaridad altos.

TABLA NUM 1

DENSIDADES FITOPLANCTONICAS CORRESPONDIENTES AL MES DE NOVIEMBRE

Est.	Niv	Cyanophyta	Chlorphyta	Bacillariophyta	Euglenophyta
1	0.3	3057511.00	9287911.00	192295.00	692267.00
1	1.0	87629.00	898194.00	0.00	0.00
1	3.0	285051.00	2178604.00	203608.00	0.00
1	5.7	1332652.00	4931470.00	167484.00	0.00
2	0.3	1244524.00	1944519.00	38891.00	0.00
2	1.0	1233560.00	4455548.00	257959.00	128880.00
2	3.0	814431.00	2606190.00	122164.00	0.00
3	0.3	1445392.00	2072996.00	76073.00	0.00
3	1.0	1054314.00	1730667.00	99464.00	0.00
3	3.0	228220.00	1806740.00	437421.00	0.00
4	0.3	629333.00	1140667.00	39333.00	0.00
4	1.0	546526.00	1038400.00	36435.00	0.00
4	3.0	736031.00	2964015.00	298391.00	0.00
4	4.0	1145813.00	1325617.00	368230.00	0.00
5	0.3	475548.00	1393245.00	190183.00	0.00
5	1.0	456440.00	838836.00	0.00	0.00
5	3.0	1994029.00	6508812.00	56435.00	18812.00
5	6.0	202525.00	902156.00	478695.00	0.00
5	7.0	288444.00	326904.00	480741.00	19230.00

TABLA NUM. 2

DENSIDADES FITOPLANCTONICAS CORRESPONDIENTES AL MES DE FEBRERO

Est.	Niv	Cyanophyta	Chlorphyta	Bacillariophyta	Euglenophyta
1	0.3	868545.00	1916760.00	334968.00	0.00
1	1.0	1358183.00	3555351.00	986294.00	37623.00
1	2.0	376232.00	1956406.00	846522.00	188116.00
1	3.0	1239243.00	3952757.00	1367440.00	384593.00
2	0.3	223312.00	3126366.00	74437.00	0.00
2	1.0	285051.00	1323451.00	407216.00	0.00
2	2.0	358069.00	2566161.00	556996.00	0.00
3	0.3	1730666.00	3847972.00	736454.00	92057.00
3	1.0	1486337.00	3136561.00	671906.00	20361.00
3	2.0	1018039.00	2015718.00	1268369.00	0.00
4	0.3	1334028.00	3534979.00	331404.00	92057.00
4	1.0	91088.00	1348098.00	236828.00	18218.00
4	2.0	1964912.00	3297375.00	1566709.00	72870.00
5	0.3	155566.00	2333483.00	155566.00	0.00
5	1.0	1325617.00	2062071.00	570752.00	36823.00
5	3.0	716138.00	1074207.00	377962.00	0.00
5	4.0	1026325.00	2092899.00	583597.00	0.00

MATRIZ NUM 3

MATRIZ DE DISTANCIA (EUCLIDIANA) CORRESPONDIENTE
AL MES DE NOVIEMBRE

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.000	0.758	0.835	0.997	0.937	0.937	0.937	0.900	0.931	0.716	0.849	0.867	0.939	0.984	0.839	0.704	0.813	0.789	0.794
2	0.758	0.000	0.091	0.121	0.043	0.157	0.040	0.083	0.039	0.210	0.010	0.007	0.122	0.129	0.044	0.004	0.205	0.248	0.252
3	0.835	0.091	0.000	0.054	0.058	0.054	0.016	0.058	0.038	0.260	0.098	0.096	0.017	0.059	0.003	0.091	0.145	0.087	0.084
4	0.997	0.121	0.054	0.000	0.046	0.018	0.038	0.095	0.080	0.142	0.075	0.082	0.040	0.085	0.059	0.109	0.095	0.100	0.102
5	0.937	0.043	0.058	0.046	0.000	0.060	0.014	0.003	0.005	0.101	0.015	0.016	0.085	0.110	0.041	0.013	0.081	0.149	0.145
6	0.937	0.157	0.054	0.018	0.060	0.000	0.044	0.063	0.080	0.094	0.105	0.111	0.024	0.053	0.039	0.139	0.090	0.131	0.138
7	0.937	0.040	0.016	0.038	0.014	0.044	0.000	0.014	0.005	0.110	0.019	0.018	0.084	0.076	0.018	0.090	0.002	0.157	0.151
8	0.900	0.083	0.054	0.095	0.003	0.063	0.014	0.000	0.005	0.183	0.023	0.028	0.070	0.007	0.042	0.059	0.070	0.223	0.229
9	0.931	0.039	0.039	0.080	0.005	0.080	0.005	0.000	0.000	0.148	0.010	0.013	0.080	0.078	0.019	0.029	0.008	0.178	0.180
10	0.716	0.210	0.080	0.142	0.101	0.094	0.110	0.183	0.177	0.000	0.177	0.178	0.092	0.030	0.048	0.211	0.216	0.004	0.000
11	0.848	0.010	0.098	0.075	0.015	0.105	0.019	0.023	0.010	0.177	0.000	0.000	0.083	0.125	0.035	0.005	0.147	0.214	0.210
12	0.867	0.007	0.096	0.082	0.016	0.111	0.018	0.028	0.013	0.178	0.000	0.000	0.087	0.130	0.036	0.001	0.151	0.219	0.217
13	0.939	0.122	0.017	0.040	0.085	0.024	0.084	0.070	0.080	0.092	0.083	0.087	0.000	0.018	0.023	0.118	0.147	0.056	0.068
14	0.984	0.129	0.059	0.085	0.110	0.053	0.076	0.007	0.078	0.030	0.125	0.130	0.018	0.000	0.047	0.161	0.209	0.039	0.038
15	0.839	0.044	0.003	0.059	0.041	0.039	0.018	0.042	0.019	0.048	0.035	0.036	0.023	0.047	0.000	0.040	0.180	0.093	0.090
16	0.704	0.004	0.091	0.109	0.013	0.139	0.030	0.059	0.023	0.211	0.003	0.001	0.118	0.161	0.040	0.000	0.171	0.250	0.252
17	0.813	0.205	0.145	0.095	0.081	0.090	0.002	0.070	0.009	0.216	0.143	0.151	0.147	0.209	0.180	0.171	0.000	0.182	0.186
18	0.789	0.248	0.087	0.100	0.149	0.131	0.157	0.223	0.178	0.004	0.214	0.219	0.056	0.039	0.093	0.250	0.182	0.000	0.001
19	0.794	0.252	0.084	0.102	0.145	0.138	0.151	0.229	0.180	0.000	0.210	0.217	0.068	0.038	0.090	0.252	0.186	0.001	0.000

MATRIZ NUM.4

MATRIZ DE DISTANCIA (EUCLIDIANA) CORRESPONDIENTE
AL MES DE FEBRERO

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.000	0.148	0.106	0.305	0.081	0.093	0.087	0.189	0.085	0.100	0.109	0.054	0.922	0.043	0.024	0.028	0.100
2	0.148	0.000	0.186	0.226	0.191	0.272	0.134	0.024	0.018	0.091	0.093	0.923	0.068	0.228	0.087	0.250	0 093
3	0.106	0.186	0.000	0.260	0.170	0.094	0.080	0.250	0.181	0.109	0.186	0.107	0.815	0.131	0.112	0.116	0.098
4	0.305	0.226	0.260	0.000	0.282	0.027	0.487	0.207	0.203	0.268	0.271	0.660	0.210	0.578	0.284	0.629	0.427
5	0.081	0.193	0.170	0.282	0.000	0.111	0.087	0.241	0.154	0.242	0.115	0.100	0.478	0.020	0.151	0.155	0.107
6	0.085	0.272	0.094	0.027	0.111	0.000	0.049	0.908	0.210	0.186	0.241	0.007	0.478	0.089	0.068	0.015	0.060
7	0.087	0.134	0.080	0.487	0.087	0.049	0.000	0.202	0.103	0.097	0.116	0.061	0.828	0.028	0.077	0.080	0.089
8	0.198	0.024	0.250	0.207	0.241	0.908	0.202	0.000	0.029	0.184	0.088	0.417	0.091	0.298	0.118	0.934	0.143
9	0.085	0.018	0.181	0.203	0.154	0.210	0.103	0.029	0.000	0.084	0.028	0.256	0.112	0.178	0.088	0.181	0.049
10	0.100	0.091	0.109	0.268	0.242	0.186	0.097	0.184	0.084	0.000	0.106	0.105	0.132	0.105	0.064	0.122	0.053
11	0.109	0.093	0.186	0.271	0.115	0.241	0.116	0.088	0.028	0.106	0.000	0.265	0.202	0.180	0.077	0.226	0.091
12	0.054	0.825	0.107	0.660	0.100	0.007	0.062	0.417	0.256	0.105	0.265	0.000	0.568	0.091	0.197	0.088	0.093
13	0.922	0.068	0.815	0.210	0.478	0.478	0.828	0.091	0.112	0.132	0.202	0.568	0.000	0.494	0.189	0.428	0.324
14	0.043	0.228	0.131	0.578	0.020	0.089	0.028	0.298	0.178	0.105	0.180	0.091	0.494	0.000	0.121	0.076	0.076
15	0.024	0.087	0.112	0.284	0.151	0.068	0.077	0.118	0.088	0.064	0.077	0.197	0.189	0.428	0.000	0.062	0.009
16	0.028	0.250	0.116	0.629	0.155	0.015	0.080	0.934	0.181	0.122	0.226	0.088	0.428	0.076	0.062	0.000	0.048
17	0.100	0.093	0.098	0.427	0.107	0.060	0.089	0.143	0.049	0.053	0.091	0.093	0.324	0.076	0.009	0.048	0.000

MATRIZ NUM. 5

MATRIZ DE DISTANCIA (CHORD DISTANCE) CORRESPONDIENTE

AL MES DE NOVIEMBRE

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.000	0.232	0.218	0.092	0.260	0.074	0.076	0.208	0.241	0.393	0.415	0.388	0.089	0.277	0.243	0.065	0.072	0.468	0.838
2	0.232	0.000	0.098	0.164	0.468	0.183	0.210	0.507	0.449	0.237	0.409	0.388	0.173	0.634	0.264	0.399	0.200	0.487	0.953
3	0.218	0.098	0.000	0.141	0.441	0.147	0.178	0.477	0.415	0.269	0.376	0.357	0.119	0.579	0.201	0.378	0.187	0.392	0.880
4	0.092	0.164	0.141	0.000	0.310	0.058	0.068	0.349	0.288	0.261	0.345	0.328	0.066	0.478	0.119	0.249	0.060	0.443	0.887
5	0.260	0.468	0.441	0.310	0.000	0.302	0.267	0.042	0.030	0.496	0.066	0.086	0.334	0.257	0.264	0.073	0.271	0.562	0.608
6	0.074	0.183	0.147	0.058	0.302	0.000	0.044	0.338	0.277	0.231	0.236	0.217	0.057	0.460	0.097	0.239	0.060	0.422	0.848
7	0.076	0.210	0.178	0.068	0.267	0.044	0.000	0.305	0.243	0.258	0.201	0.182	0.078	0.434	0.088	0.201	0.057	0.457	0.842
8	0.208	0.507	0.477	0.349	0.042	0.338	0.305	0.000	0.065	0.514	0.105	0.126	0.380	0.204	0.295	0.114	0.311	0.574	0.789
9	0.241	0.449	0.415	0.288	0.030	0.277	0.243	0.065	0.000	0.452	0.047	0.065	0.306	0.227	0.231	0.068	0.252	0.524	0.779
10	0.393	0.237	0.269	0.261	0.496	0.231	0.258	0.514	0.452	0.000	0.426	0.407	0.180	0.505	0.226	0.435	0.288	0.257	0.758
11	0.399	0.405	0.370	0.345	0.066	0.236	0.201	0.105	0.047	0.424	0.000	0.020	0.268	0.271	0.200	0.091	0.208	0.517	0.803
12	0.381	0.388	0.357	0.328	0.086	0.217	0.182	0.126	0.065	0.407	0.020	0.000	0.249	0.252	0.183	0.091	0.198	0.508	0.806
13	0.120	0.175	0.119	0.066	0.334	0.057	0.079	0.380	0.306	0.180	0.268	0.249	0.000	0.472	0.090	0.272	0.104	0.379	0.819
14	0.436	0.634	0.579	0.478	0.257	0.460	0.434	0.204	0.227	0.505	0.271	0.252	0.472	0.000	0.384	0.205	0.454	0.328	0.617
15	0.130	0.264	0.201	0.119	0.264	0.097	0.088	0.295	0.231	0.226	0.200	0.183	0.090	0.384	0.000	0.213	0.124	0.368	0.762
16	0.193	0.399	0.378	0.249	0.073	0.239	0.201	0.114	0.068	0.435	0.091	0.091	0.272	0.205	0.213	0.000	0.202	0.561	0.831
17	0.072	0.200	0.187	0.060	0.271	0.060	0.057	0.311	0.252	0.288	0.208	0.198	0.104	0.454	0.124	0.202	0.000	0.471	0.874
18	0.468	0.487	0.392	0.443	0.562	0.422	0.457	0.574	0.524	0.257	0.517	0.508	0.379	0.328	0.368	0.561	0.471	0.000	0.938
19	0.838	0.953	0.880	0.887	0.608	0.848	0.842	0.789	0.779	0.758	0.803	0.806	0.819	0.617	0.762	0.831	0.874	0.938	0.000

MATRIZ NUM.6

MATRIZ DE DISTANCIA (CHORD DISTANCE) CORRESPONDIENTES
AL MES DE FEBRERO

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.000	0.118	0.361	0.217	0.375	0.247	0.286	0.020	0.038	0.354	0.096	0.353	0.254	0.366	0.160	0.208	0.092
2	0.118	0.000	0.284	0.117	0.389	0.152	0.224	0.099	0.049	0.274	0.107	0.300	0.211	0.347	0.202	0.217	0.089
3	0.361	0.284	0.000	0.186	0.401	0.141	0.214	0.324	0.328	0.284	0.358	0.266	0.326	0.354	0.404	0.308	0.307
4	0.217	0.117	0.186	0.000	0.380	0.127	0.211	0.197	0.202	0.258	0.245	0.280	0.239	0.352	0.296	0.286	0.185
5	0.375	0.389	0.401	0.380	0.000	0.301	0.199	0.378	0.402	0.607	0.286	0.950	0.973	0.048	0.530	0.564	0.487
6	0.247	0.152	0.141	0.127	0.301	0.000	0.107	0.237	0.245	0.320	0.253	0.384	0.320	0.366	0.354	0.358	0.280
7	0.286	0.224	0.214	0.211	0.199	0.107	0.000	0.281	0.287	0.429	0.253	0.081	0.418	0.167	0.420	0.438	0.310
8	0.020	0.099	0.324	0.197	0.378	0.237	0.281	0.000	0.081	0.341	0.105	0.348	0.241	0.368	0.158	0.197	0.083
9	0.038	0.049	0.328	0.202	0.402	0.245	0.287	0.081	0.000	0.320	0.185	0.367	0.276	0.301	0.192	0.168	0.058
10	0.354	0.274	0.284	0.258	0.607	0.320	0.429	0.341	0.320	0.000	0.433	0.490	0.140	0.577	0.298	0.252	0.266
11	0.096	0.107	0.358	0.245	0.286	0.253	0.253	0.105	0.198	0.433	0.000	0.302	0.344	0.294	0.251	0.207	0.185
12	0.353	0.300	0.266	0.280	0.950	0.384	0.081	0.348	0.367	0.490	0.302	0.000	0.408	0.108	0.401	0.319	0.284
13	0.254	0.211	0.326	0.239	0.973	0.320	0.418	0.241	0.276	0.140	0.344	0.408	0.000	0.551	0.162	0.115	0.188
14	0.366	0.347	0.354	0.352	0.048	0.366	0.167	0.368	0.301	0.577	0.294	0.108	0.551	0.000	0.510	0.550	0.470
15	0.160	0.202	0.404	0.296	0.530	0.354	0.420	0.158	0.192	0.298	0.251	0.401	0.162	0.510	0.000	0.133	0.114
16	0.208	0.217	0.308	0.286	0.564	0.358	0.438	0.197	0.168	0.252	0.207	0.319	0.115	0.550	0.133	0.000	0.188
17	0.092	0.089	0.307	0.185	0.487	0.280	0.310	0.083	0.058	0.266	0.185	0.284	0.188	0.470	0.114	0.188	0.000

MATRIZ NUM.7

MATRIZ DE SIMILITUD (GOWER) CORRESPONDIENTE

AL MES DE NOVIEMBRE

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1.000	0.166	0.812	0.468	0.818	0.474	0.838	0.958	0.922	0.710	0.239	0.227	0.923	0.275	0.811	0.105	0.510	0.127	0.124
2	0.166	1.000	0.842	0.898	0.853	0.824	0.828	0.813	0.844	0.785	0.927	0.939	0.782	0.708	0.855	0.967	0.967	0.741	0.710
3	0.812	0.842	1.000	0.819	0.827	0.782	0.901	0.833	0.868	0.863	0.857	0.859	0.891	0.818	0.955	0.842	0.952	0.814	0.797
4	0.468	0.898	0.819	1.000	0.845	0.887	0.870	0.881	0.854	0.882	0.771	0.760	0.820	0.782	0.820	0.727	0.838	0.889	0.810
5	0.818	0.853	0.827	0.845	1.000	0.769	0.902	0.860	0.847	0.703	0.926	0.915	0.794	0.803	0.841	0.863	0.794	0.854	0.838
6	0.474	0.824	0.782	0.887	0.769	1.000	0.790	0.775	0.780	0.702	0.890	0.885	0.840	0.801	0.769	0.859	0.784	0.859	0.850
7	0.838	0.828	0.901	0.870	0.902	0.790	1.000	0.908	0.944	0.764	0.900	0.889	0.902	0.808	0.902	0.827	0.751	0.716	0.690
8	0.958	0.813	0.833	0.881	0.860	0.775	0.908	1.000	0.945	0.703	0.886	0.875	0.800	0.802	0.840	0.848	0.818	0.858	0.837
9	0.922	0.844	0.868	0.854	0.847	0.780	0.944	0.945	1.000	0.753	0.916	0.905	0.835	0.841	0.895	0.873	0.758	0.708	0.681
10	0.710	0.785	0.863	0.882	0.703	0.702	0.764	0.703	0.753	1.000	0.741	0.745	0.853	0.873	0.859	0.726	0.515	0.951	0.926
11	0.239	0.927	0.857	0.771	0.926	0.890	0.900	0.886	0.916	0.741	1.000	0.989	0.805	0.780	0.902	0.957	0.720	0.729	0.712
12	0.227	0.939	0.859	0.760	0.915	0.885	0.889	0.875	0.905	0.745	0.989	1.000	0.794	0.768	0.904	0.968	0.708	0.737	0.720
13	0.923	0.782	0.891	0.820	0.794	0.840	0.902	0.800	0.835	0.853	0.805	0.749	1.000	0.883	0.878	0.782	0.865	0.804	0.787
14	0.275	0.708	0.818	0.782	0.803	0.801	0.808	0.802	0.841	0.873	0.780	0.768	0.883	1.000	0.849	0.727	0.815	0.851	0.835
15	0.811	0.855	0.955	0.820	0.841	0.769	0.902	0.840	0.895	0.859	0.902	0.904	0.878	0.849	1.000	0.896	0.893	0.815	0.798
16	0.105	0.967	0.842	0.727	0.863	0.859	0.827	0.848	0.873	0.726	0.957	0.968	0.782	0.727	0.896	1.000	0.876	0.728	0.715
17	0.510	0.967	0.952	0.838	0.794	0.784	0.751	0.818	0.758	0.515	0.720	0.708	0.865	0.815	0.893	0.876	0.000	0.466	0.483
18	0.127	0.741	0.814	0.889	0.854	0.859	0.716	0.858	0.708	0.951	0.729	0.737	0.804	0.851	0.815	0.728	0.466	0.000	0.069
19	0.124	0.710	0.797	0.810	0.838	0.850	0.690	0.837	0.681	0.926	0.712	0.720	0.787	0.835	0.798	0.715	0.483	0.069	1.000

MATRIZ NUM.8

MATRIZ DE SIMILITUD (GOWER) CORRESPONDIENTE
AL MES DE FEBRERO

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1.000	0.659	0.728	0.954	0.765	0.859	0.838	0.590	0.742	0.815	0.737	0.819	0.480	0.819	0.863	0.899	0.922
2	0.659	1.000	0.609	0.600	0.694	0.541	0.694	0.848	0.893	0.749	0.850	0.501	0.777	0.570	0.706	0.572	0.727
3	0.728	0.609	1.000	0.496	0.626	0.787	0.774	0.574	0.611	0.716	0.586	0.607	0.476	0.700	0.720	0.677	0.725
4	0.954	0.600	0.496	1.000	0.826	0.898	0.876	0.629	0.568	0.8534	0.587	0.895	0.610	0.562	0.665	0.804	0.939
5	0.765	0.694	0.626	0.826	1.000	0.729	0.853	0.565	0.717	0.597	0.718	0.789	0.455	0.909	0.653	0.705	0.718
6	0.859	0.541	0.787	0.898	0.729	1.000	0.857	0.473	0.603	0.698	0.505	0.931	0.568	0.858	0.746	0.916	0.905
7	0.838	0.694	0.774	0.876	0.853	0.857	1.000	0.616	0.768	0.745	0.669	0.798	0.506	0.886	0.801	0.798	0.869
8	0.590	0.848	0.574	0.629	0.565	0.473	0.616	1.000	0.868	0.597	0.852	0.432	0.789	0.501	0.727	0.504	0.668
9	0.742	0.893	0.611	0.568	0.717	0.603	0.768	0.868	1.000	0.727	0.841	0.586	0.738	0.653	0.858	0.656	0.820
10	0.815	0.749	0.716	0.8534	0.597	0.698	0.745	0.597	0.727	1.000	0.600	0.684	0.605	0.671	0.814	0.729	0.877
11	0.737	0.850	0.586	0.587	0.718	0.505	0.669	0.852	0.841	0.600	1.000	0.580	0.676	0.649	0.705	0.654	0.732
12	0.819	0.501	0.607	0.895	0.789	0.931	0.798	0.432	0.586	0.684	0.580	1.000	0.932	0.880	0.705	0.857	0.761
13	0.480	0.777	0.476	0.610	0.455	0.568	0.506	0.789	0.738	0.605	0.676	0.932	1.000	0.801	0.617	0.804	0.558
14	0.880	0.570	0.700	0.562	0.909	0.858	0.886	0.501	0.653	0.671	0.649	0.880	0.801	1.000	0.727	0.779	0.791
15	0.863	0.706	0.720	0.665	0.653	0.746	0.801	0.727	0.858	0.814	0.705	0.705	0.617	0.727	1.000	0.777	0.931
16	0.899	0.572	0.677	0.804	0.705	0.916	0.798	0.504	0.656	0.729	0.656	0.857	0.804	0.779	0.777	1.000	0.858
17	0.922	0.727	0.725	0.939	0.718	0.905	0.869	0.668	0.820	0.877	0.732	0.761	0.558	0.791	0.931	0.858	1.000

MATRIZ NUM. 9

MATRIZ DE DISTANCIA (MANHATTAN) CORRESPONDIENTE

AL MES DE NOVIEMBRE

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.000	2.336	2.704	2.139	2.749	2.104	2.047	2.390	2.711	3.027	3.045	3.000	2.708	2.898	2.755	3.219	1.924	3.498	3.505
2	2.336	0.000	0.688	1.208	0.587	1.506	0.680	0.747	0.625	1.009	0.291	0.346	1.070	1.170	0.581	1.151	1.413	1.035	1.150
3	2.704	0.688	0.000	0.725	0.000	0.893	0.875	0.105	0.068	0.326	0.347	0.573	0.503	0.727	0.180	0.691	1.991	0.743	0.813
4	2.139	1.208	0.725	0.000	0.620	0.451	0.518	0.557	0.581	1.271	0.916	0.902	0.683	0.878	0.721	1.090	0.667	1.467	1.585
5	2.749	0.587	0.893	0.620	0.000	0.926	0.892	0.159	0.214	1.487	0.208	0.841	0.825	0.787	0.685	0.470	0.825	1.883	1.449
6	2.104	1.506	0.875	0.451	0.920	0.000	0.816	0.902	0.880	1.194	1.214	1.200	0.804	0.794	0.924	1.388	1.063	1.889	1.401
7	2.047	0.680	0.893	0.518	0.892	0.816	0.000	0.908	0.220	0.942	0.308	0.443	0.493	0.766	0.901	0.571	0.997	1.138	1.205
8	2.390	0.747	0.068	0.557	0.159	0.902	0.908	0.000	0.219	1.191	0.455	0.901	0.801	0.792	0.640	0.929	0.748	1.887	1.496
9	2.711	0.625	0.526	0.581	0.214	0.880	0.220	0.219	0.000	0.998	0.884	0.979	0.859	0.885	0.421	0.508	0.996	1.668	1.395
10	3.027	1.009	0.347	1.271	1.187	1.194	0.942	1.191	0.998	0.000	1.037	1.027	0.580	0.807	0.644	1.005	1.989	0.195	0.303
11	3.045	0.291	0.573	0.916	0.208	1.214	0.308	0.455	0.884	1.037	0.000	0.045	0.778	0.879	0.304	0.474	1.121	1.084	1.158
12	3.000	0.346	0.503	0.902	0.841	1.200	0.443	0.901	0.979	1.027	0.045	0.000	0.824	0.924	0.988	0.520	1.167	1.051	1.118
13	2.708	1.070	0.497	0.683	0.825	0.804	0.493	0.801	0.859	0.580	0.778	0.824	0.000	0.456	0.488	0.952	1.350	0.785	0.858
14	2.898	1.170	0.727	0.878	0.787	0.794	0.766	0.792	0.885	0.807	0.879	0.924	0.456	0.000	0.604	1.059	1.540	0.595	0.662
15	2.755	0.581	0.180	0.721	0.685	0.924	0.901	0.640	0.421	0.644	0.304	0.988	0.488	0.604	0.000	0.604	1.388	0.747	0.814
16	3.219	0.151	0.691	1.090	0.470	1.388	0.572	0.929	0.508	1.005	0.474	0.520	0.952	1.059	0.604	0.000	1.792	1.089	1.141
17	1.924	1.413	1.991	0.667	0.825	1.063	0.997	0.748	0.996	1.989	1.121	1.167	1.350	1.540	1.388	1.792	0.000	2.134	2.147
18	3.498	1.035	0.743	1.467	1.883	1.889	1.138	1.887	1.668	0.195	1.084	1.051	0.785	0.595	0.747	1.089	2.134	0.000	0.129
19	3.505	1.150	0.813	1.585	1.449	1.401	1.205	1.496	1.395	0.303	1.158	1.118	0.858	0.662	0.814	1.141	2.147	0.129	0.000

MATRIZ NUM. 10

MATRIZ DE DISTANCIA (MANHATTAN) CORRESPONDIENTE

AL MES DE FEBRERO

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.000	1.189	1.108	2.397	0.939	0.566	0.647	1.639	1.083	0.740	1.052	0.726	2.080	0.645	0.548	0.403	0.312
2	1.189	0.000	1.868	1.259	1.484	1.834	1.268	0.000	0.470	1.903	0.001	1.996	0.894	1.721	0.817	1.710	1.058
3	1.108	1.868	0.000	2.014	1.495	1.052	0.905	1.701	1.933	1.135	1.854	1.216	2.006	1.701	1.122	1.301	1.060
4	2.397	1.259	2.014	0.000	2.068	3.068	2.495	1.483	1.939	1.857	1.051	2.228	1.559	2.958	2.141	2.962	2.385
5	0.939	1.484	1.495	2.068	0.000	0.882	0.560	1.798	1.131	1.010	1.148	0.845	2.478	0.366	1.396	1.179	1.129
6	0.566	1.834	1.052	3.068	0.882	0.000	0.571	2.108	1.301	1.208	1.018	0.276	2.960	0.569	1.017	0.936	0.781
7	0.647	1.268	0.905	2.495	0.560	0.571	0.000	1.537	0.990	1.020	1.248	0.828	1.978	0.439	0.796	0.820	0.530
8	1.639	0.000	1.703	1.483	1.798	2.108	1.537	0.000	0.608	1.613	0.502	2.270	0.923	1.905	1.041	1.985	1.827
9	1.083	0.470	1.933	1.939	1.131	1.301	0.990	0.608	0.000	1.092	0.695	1.668	1.048	1.368	0.566	1.377	0.720
10	0.740	1.903	1.135	1.857	1.010	1.208	1.020	1.613	1.092	0.000	1.564	1.405	1.340	1.310	0.743	1.085	0.490
11	1.052	1.001	1.854	1.140	1.018	1.248	0.502	0.695	1.564	0.000	1.678	1.207	1.403	0.820	1.455	1.074
12	0.726	1.996	1.216	2.228	0.845	0.276	0.828	2.270	1.668	1.405	1.678	0.000	2.710	0.479	1.170	0.571	1.028
13	2.080	0.894	2.006	1.559	2.478	2.960	1.978	0.923	1.048	1.340	1.207	2.710	0.000	2.426	1.931	2.625	1.768
14	0.645	1.721	1.701	2.958	0.366	0.569	0.439	1.905	1.368	1.310	1.403	0.479	2.426	0.000	1.093	0.886	0.853
15	0.548	0.817	1.122	2.141	1.396	1.017	0.796	1.041	0.566	0.743	0.820	1.170	1.931	1.093	0.000	0.848	0.275
16	0.403	1.710	1.301	2.962	1.179	0.936	0.820	1.985	1.377	1.085	1.455	0.571	2.625	0.886	0.848	0.000	0.657
17	0.312	1.058	1.060	2.385	1.129	0.781	0.530	1.827	0.720	0.490	1.074	1.028	1.768	0.853	0.275	0.657	0.000

CAPITULO VII

DISCUSION

Debido a que la mayoría de los estudios de fenómenos biológicos son de naturaleza observacional, el hombre ha tratado de agrupar al conjunto de observaciones, organismos, u objetos que conforman el medio que lo rodea de acuerdo con las características que comparten y que los hacen semejantes. Por esta razón recurre a la clasificación como medio para evitar la confusión, e instintiva o conscientemente clasifica a su mundo circundante.

El proceso de clasificar conlleva a ubicar objetos dentro de un conjunto de categorías donde los atributos inherentes de cada una de ellas se conocen con cierta incertidumbre al efectuar la asignación de grupos, ya sea porque se superpongan o porque sus características no esten bien definidas. Independientemente de los problemas encontrados al aplicarla, la clasificación juega un papel fundamental en la ciencia.

Los fenómenos biológicos se encuentran conformados por conjuntos de observaciones que contienen una gran cantidad de información que se traduce en grandes volumenes de datos que deben procesarse para explicar su comportamiento, es común que los datos se clasifiquen en clases o grupos similares de acuerdo a sus semejanzas y así describir la estructura de los datos.

Debido a la naturaleza multivariada de los fenómenos biológicos los datos obtenidos de ellos son dependientes unos de otros, se hace necesario encontrar una estructura e interpretación de ellos. Esta búsqueda ha propiciado el desarrollo de programas y equipo de cómputo que facilita el manejo de los datos. Dentro de los múltiples elementos con los que se dispone para apoyar a los estudios de tipo clasificatorio, se encuentran como una valiosa herramienta las medidas de similaridad y distancia que son la base de diferentes métodos estadísticos, debido a que a través del valor obtenido, se puede conocer la semejanza existente entre un conjunto de datos, efectuandose entonces la agrupación y

descripción del fenómeno en cuestión. Estas herramientas son poco conocidas y utilizadas dentro de la Biología, ya que se han generado en diferentes áreas del conocimiento.

El análisis de datos multivariados se basa en matrices de asociación y similaridad en donde cada una de ellas puede conducir a resultados diferentes. En esta parte radica la importancia de las medidas de distancia, ya que a través de ellas se generan las matrices de asociación y similaridad necesarias para algunas técnicas multivariadas.

Dado que cada matriz presenta características particulares en su cálculo, es importante conocer el tipo de datos contenidos en ellas, ya que en base a su clasificación se determina el tratamiento numérico que se puede aplicar. Los datos conforman una serie de variables que son clasificadas como binarias, cualitativas y cuantitativas.

Las medidas de similaridad manejan preferentemente datos binarios o cualitativo aunque el coeficiente de similaridad de Gower es una medida que puede trabajar mezclas de datos (tanto cualitativos como cuantitativos). Por la existencia de datos cuantitativos contínuos y discretos, las medidas de distancia trabajan solamente datos de tipo cuantitativo contínuos y no se aplican a datos cuantitativos discretos. Una alternativa para trabajar a los datos cuantitativos discretos es la medida de similaridad de Gower. Aunque se puede hacer una recodificación de los datos cuantitativos a cualitativos para trabajar con medidas de similaridad, no es recomendable porque se pierde una gran cantidad de información del fenómeno en cuestión. Por ello, si los datos son cuantitativos contínuos es mejor trabajar con distancias que con similitudes.

Una vez conocidas las propiedades de los datos es necesario definir las entidades y las variables involucradas. En esta etapa es importante tomar en cuenta el objetivo del trabajo y el juicio del investigador. Una vez definidas interesa comparar a cada variable en función de las entidades siendo este enfoque el

tradicionalmente manejado. En este trabajo no se ha seguido el enfoque tradicional de comparar a cada variable para todas las entidades sino que interesa comparar que ocurre en cada una de las entidades en función de todas las variables presentes, generando vectores que permiten efectuar la comparación entre cada una de ellas dando con ella un enfoque multivariado al estudio efectuado.

Dentro del enfoque clásico las medidas de similaridad han sido utilizadas como un parámetro para medir la semejanza de una entidad con respecto a otra y en base al valor obtenido se agrupan las entidades de acuerdo con la similitud presente en ellas. Esta situación no es muy clara ya que al efectuar el análisis de las medidas de similaridad se observó que realmente no miden la semejanza sino más bien la presencia o ausencia de atributos o bien la probabilidad de aparición de los atributos en las entidades comparadas y en base a ello agrupar las entidades que presentan atributos comunes o presentes como los más semejantes y las que no las presenten como diferentes. En este contexto las medidas de similaridad pueden considerarse como probabilísticas (aquellas que manejan la probabilidad de aparición de atributos en las entidades que pueden estar o no condicionadas a la aparición de alguna característica o atributo en especial) o no probabilísticas (aquellas que solo indican la presencia de las entidades para poder compararlas).

Por esta razón se recomienda que si se desea medir correctamente la semejanza entre entidades se utilicen las medidas de distancia como un buen parámetro para cuantificar la semejanza entre entidades. Las medidas de distancia tienen algunas ventajas que las hacen una herramienta útil para medir la semejanza entre entidades siendo entre ellas las siguientes:

1.- Expresan cuantitativamente la semejanza entre entidades, dando con ello el grado de similaridad para las entidades comparadas.

2.- Pueden ser representadas dentro de un espacio geométrico, dando con ello una mayor facilidad para ubicar a las entidades

comparadas y observar gráficamente la distancia que existe entre una y otra entidad.

Las medidas de distancia tienen como principal objetivo, medir cuantitativamente las diferencias entre entidades. Las entidades cuyo valor de distancia sea bajo o cercano a 0, será muy semejante y prácticamente las diferencias entre ellas son nulas. Análogamente si el valor de distancia es mayor, la semejanza entre ellas disminuye, siendo ambas diferentes. En este contexto, las medidas de distancia son aplicadas como un buen parámetro para medir la similaridad entre entidades.

Geométricamente la distancia entre las entidades puede ser medida por la hipotenusa de un triángulo rectángulo, por el recorrido de los catetos o bien la medición de los arcos formados por los vectores correspondientes a las entidades comparadas. Al parecer las medidas de distancia presentan estas tres alternativas como base para la medición de las distancias, ya que en su mayoría son derivadas de la distancia euclidiana, la distancia de Manhattan o la Chord Distance, teniendo diversas modificaciones de acuerdo a la filosofía de cada medida, teniendo por tanto una representación semejante, claro está dependiendo de que trabaja cada una de ellas.

Es importante dejar en claro que cada medida de distancia tiene un significado geométrico diferente y que no todas son aplicables al mismo tipo de datos y/o problema. Así mismo cada técnica de agrupamiento utiliza medidas de distancia diferentes, dependiendo de los objetivos y la técnica a utilizar. Por ello es muy importante tener una idea clara del significado de cada medida de distancia, ya que cada medida genera valores diferentes, aún con los mismos datos.

El estudio de caso presentado en este trabajo tuvo la finalidad de mostrar la manera en que las medidas tanto de similaridad como de distancia son aplicadas a un conjunto de datos específico, así como la generación de las matrices de similaridad y distancia correspondientes a partir de los datos originales.

Uno de los principales problemas que se presentaron al aplicar las medidas de distancia en los datos de densidades fitoplanctónicas fueron al momento de comparar entidades que no se registraron en los niveles, generando esta situación un problema en el cálculo de las distancias, ya que al momento de obtener el rango de valores para algunas variables el valor era cero y al dividirlo con cualquier otro valor no permitia trabajar esa comparación. Esta situación se presentó al calcular la métrica de Canberra, así como en la distancia Geodésica, considerando esto una desventaja para trabajar esas medidas, recomendando por ello aplicarlas a datos que no tengan valores faltantes.

El usuario de las diferentes técnicas multivariadas, se interesa en alimentar a la computadora con los datos de sus estudios y espera tener resultados satisfactorios. Aunque esto puede hacerse, no es recomendable, pues cada una de las técnicas multivariadas son una herramienta estadística que necesitan para ser más poderosas, de una gran interrelación entre el usuario y sus datos y en la medida en que se conozca mejor el fenómeno estudiado, el investigador estará posibilitado para dar la medida de distancia más adecuada para el estudio realizado.

Aunque existen actuamente muchos paquetes de cómputo que realizan diversas técnicas multivariadas, muchas de ellas utilizan indistintamente algunas medidas de distancia siendo esta situación riesgosa, ya que al interpretar los datos resultantes no concuerdan con la realidad. Por ello, es importante conocer adecuadamente las medidas de distancias ya sea para trabajarse con los datos originales o para trabajar con datos transformados, de manera que aunque se calcule la distancia, una transformación puede resultar en otro tipo de medida, que sea en realidad la que nos interesa.

Si es un usuario de las diferentes técnicas multivariadas debe considerar lo siguiente:

- Si los datos seleccionados para el análisis no son importantes para el estudio, cualquier tipo de herramienta estadística, ya sea

univariada o multivariada, revelará resultados irrevelantes para el estudio.

- Cada una de las medidas de distancia tiene un significado estadístico diferente. Es recomendable conocerlas para elegir la más adecuada, ya que los resultdos pueden variar mucho de una medida a otra.

- Si no se conocen a fondo los datos o aún conociendolos bien no se sabe cual medida de distancia elegir; la mejor solución es indudablente probar varias y una vez que se tengan los resultados realizar la selección de la mejor medida que refleje las relaciones de las entidades con las variables. Mucho del éxito de este proceso depende del conocimiento y la habilidad del investigador.

- Si las variables están medidas en diferentes escalas, es recomendable realizar una estandarización de las variables y probar a su vez con los datos originales; si se observa que los resultados son semejantes a los originales, no es necesaria la estandarización.

Por todo lo anteriormente expuesto se puede decir que se cumplieron casi en su totalidad los objetivos planteados en este trabajo ya que se reunió un buen número de medidas tanto de similaridad como de distancia, se pudo hacer un análisis detallado de cada una de ellas para poder comprender su significado matemático y geométrico y se logró mostrar su potencial de aplicación en la Biología. Su recopilación fué difícil ya que la mayoría de las medidas se han originado en diferentes áreas de investigación, pero a pesar de esto se presentan un buen número de ellas y con un buen potencial de aplicación en la investigación biológica.

Análogamente se logró analizar las diferentes medidas y se concluyó que para la elección de cualquier medida, es muy importante considerar el tipo de datos presentes y en función de ello elegir la medida más adecuada. Esto se representó en el caso

de estudio en el cual se aplicaron diferentes medidas de similaridad y distancia construyendo las correspondientes matrices de similaridad y distancia.

CAPITULO VIII

CONCLUSIONES

Clasificar implica reconocer clases o grupos similares. Por ejemplo, en Ecología es importante reconocer clases que corresponden a comunidades; del mismo modo, en Taxonomía numérica se busca identificar grupos de organismos relacionados por su máxima semejanza, que pueden corresponder con alguna categoria taxonómica; como especie, género ó familia.

El interés por conocer cuantitativamente la separación o semejanza entre entidades, es que al expresar el grado de similaridad o disimilaridad entre observaciones, ayuda a describir de alguna manera la estructura de un conjunto de datos.

Actualmente existen una gran cantidad de elementos que apoyan los estudios de clasificación, entre los cuales se encuentran las medidas de similaridad y distancia que son la base de diferentes métodos estadísticos; ya que a partir del valor obtenido en su cálculo es posible conocer la semejanza entre un conjunto de datos efectuando de esta manera la agrupación y descripción del fenómeno en cuestión.

Las medidas de similaridad pueden ser un parámetro para registrar de una manera indirecta la semejanza entre las entidades, mientras que las medidas de distancia son una alternativa adecuada para medir la semejanza entre entidades.

De la misma forma, antes de iniciar el cálculo de cualquier medida ya sea de similaridad o de distancia es necesario considerar el tipo de datos presentes, seleccionando en base a ellos y al objetivo del trabajo la medida más adecuada a trabajar.

Los coeficientes de similaridad presentan un gran uso dentro de la Biología, ya que mediante ellos se puede llegar a generar agrupaciones que ayuden a la descripción del fenómeno de estudio. Su desventaja es que para su cálculo son utilizados solamente

datos cualitativos, principalmente aquellos del tipo presencia-ausencia, aunque algunas medidas de asociación pueden parcialmente solventar este problema, ya que algunos de ellos pueden trabajar datos cuantitativos como lo es el caso del coeficiente de correlación de Pearson.

Los coeficientes de similaridad que se recomiendan con ventajas sobre otros son:

1) Indice de Jaccard
2) Indice de comunidad de Dice.
3) Indice de Ochiai.
4) Indice de Mozley.
5) Indice de Mountford.
6) Coeficiente de Gower.

En ellos se encuentran ciertas ventajas en su uso entre las cuales pueden mencionarse:

- Presentan una escala de 0 a 1, cayendo estos valores dentro del rango establecido por la similaridad.
- Son aplicables para datos tanto de abundancia como de presencia-ausencia.

Con respecto a las diferentes medidas de distancia se puede concluir que:

Para seleccionar una medida de distancia adecuadamente, se requiere de un amplio conocimiento de las propiedades del conjunto de datos y de los efectos que tengan sobre cada medida.

Es necesario considerar la escala de medición de las variables, ya que si hay diferencias en escalas en las k-variables la comparación no sería adecuada porque algunas medidas son sensibles a los cambio en escala.

Puede considerarse que las medidas de distancia son derivadas básicamente de la distancia Euclidiana y la distancia de Manhattan, porque conceptualmente presentan una forma matemática semejante, midiendo completamente la hipotenusa, los catetos, o

bien una proporción de ellos. aunque su interpretación depende del objetivo y del juicio del investigador.

Existen algunas medidas que no necesariamente se derivan de la Distancia Euclidiana o la distancia de Manhattan, tal es el caso de la Chord Distance, la distancia Geodésica y el Coeficiente de Correlación producto momento, en los cuales se plantean vectores que representan a las entidades comparadas para las k-variables, considerando el ángulo formado entre ellos.

Las medidas que fueron seleccionadas por sus características son:

- Medida de distancia euclidiana.
- Métrica de Manhattan.
- Distancia Geodésica.
- Chord distance.
- Métrica de Canberra.
- Coeficiente de correlación producto momento.

Si desea representar las medidas de distancia geométricamente, se recomienda trabajar las medidas de distancia métrica, aunque también es conveniente mencionar que las no métricas miden distancias entre entidades, pero la desventaja que presentan es que no pueden ser ubicadas dentro de un espacio geométrico. Las medidas de distancia presentan las siguientes ventajas:

- Pueden ser representadas dentro de un espacio geométrico, dando con ello la opción de observar gráficamente el comportamiento de los datos.

- Generan las matrices de distancia correspondientes para ser utilizadas en las diferentes técnicas multivariadas existentes.

- Como las medidas de distancia son la base de las diferentes técnicas multivariadas, tienen la propiedad de poder describir en cierta forma la estructura de los datos, no llegando por ello a generar conclusiones de manera absoluta, sino que son utilizadas en conjunto con técnicas inferenciales.

- El hecho que involucra que el investigador esté tomando decisiones acerca de los datos, de la medida de distancia más adecuada para la generación de las matrices y su aplicación con alguna técnica de agrupamiento (Cluster, Componentes Principales, Escalamiento Multidimensional, etc) obliga a buscar una mayor información acerca del tipo de estudio a realizar, de las medidas de similitud y distancia más adecuada y del tratamiento estadístico posterior a los datos. Esta es una de las mayores ventajas, pero la buena o mala elección de las medidas, depende del conocimiento que tenga el usuario de su problema y la técnica más adecuada a trabajar.

Por todo lo anteriormente expuesto se puede decir que se cumplieron casi en su totalidad, los objetivos planteados en este trabajo ya que se reunió un buen número de medidas tanto de similaridad como de distancia, se realizó un análisis detallado de cada una de ellas para comprender su significado matemático y geométrico y se logró mostrar su potencial de aplicación en la Biología. Su recopilación fué difícil ya que la mayoría de las medidas se han originado en diferentes áreas del conocimiento pero a pesar de ello, se presentan un buen número de ellas.

B I B L I O G R A F I A

1) Anderberg, M.R. (1973) *Cluster analysis for aplication*, Academic Press, New York. 359 pp.

2) Arredondo, W., *et al* (1984) *Estudio de la distribución espacial y variación temporal de la comunidad fitoplanctónica en la laguna "El Rodeo", Mor. de Enero a Marzo de 1984*, ENEP Zaragoza, U.N.A.M., Mex.

3) Balakrishnan, V. and Z. D. Sanghvi (1968) *Distance between population on the basis of atribute data*, Biometrics 24:859-865

4) Begon, Harper & Townsend (1986) *Ecology (Individuals, population and communities)*, Blackwell Scient. Pub. London. 784 pp.

5) Bold, H.C., M.C. (1978) *Introduction to the algae*, Prentice Hall, New Jersey 270 pp.

6) Clifford, H.T. and Stephenson W. (1975) *An introduction to numerical classification*, Academic Press, New York. 73-95 pp.

7) Crisci, J.V. & López A.M.F. (1983) *Introducción a la teoría práctica de la taxonomía numérica*, Sria. Gral. de la O.E.A., Washington D.C 78-110 pp.

8) Crovello, T. J. (1970) *Analysis of character variation in ecology and systematics*, Ann. Rev. Ecol. Syst 1:55-98.

9) DETENAL (1979) *Síntesis geográfica de Morelos, México.*

10) Dirección General de Protección y Ordenación Ecológica (1982) *Manual de ténicas de muestreo y análisis de plancton y perifiton*, 3a. Edición, México. 130 pp.

11) Duncan, T. (1981) *Numerical phenetic: Its uses in botanical systematics*, Ecol. Syst 12:387-404.

12) Ocegueda C. S. (1991), *El análisis estadístico Cluster, métodos y aplicaciones en Biología*, ENEP Zaragoza U.N.A.M., México.

13) Everitt, B.S. (1981) *Cluster analysis* 2da Ed Halsted Press, New York 35-45 pp.

14) Fleiss, J.L. (1975) *Measuring agreement between two judges on the presence o absense of trait* Biometrics, 31:651-659.

15) Fritz, E. Z. & Jos, M.F.(1985) *A family of association coefficient for metric scales*, Psychometrika 50:1 17-24.

16) Gauch, H.G. (1982) *Multivariate analysis in community ecology*, Cambridge University Press. 298 pp.

17) Gómez Aguirre, Arenas Fuentes. (1987) *Los lagos de México. Contribuciones en hidrobiología*, U.N.A.M, México. 35-50 pp.

18) Goodman, M.M. (1967) *The races of maize: I. The use of Mahalanobis generalized distances to measure morphological similarity.* Fitotecnia Latinoamericana 4:1-22.

19) Goodman, M.M. (1972) Distance analysis in biology, Syst Zool. 21(2) 174-186.

20) Gower, J.C. (1971) *A general coefficient of similarity and some of its properties*, Biometrics, 27:857-872.

21) Green, R.H. (1979) *Sampling design and statistical methods for enviromental biologist*, John Wiley & Sons New York. 73-95 pp.

22) Hartigan, J. A. (1967), *Representation of similarity matrices by trees*, American Statistical Vol 62 Num. 320: 1140-1158.

23) Kendall, M. (1980) *Multivariate analysis*, 2a. Edic. Mac Millan Publishing C.O. New York. 235 pp.

24) Kurczynski, T.W. (1976) *Generalized distance and discrete variables*, Biometrics 27:525-534.

25) Lance & Williams,(1966) *Computer programs for herarchial polythetic clasification*, Comp.J 960-964 pp.

26) Legendre, L. & Legendre, P. (1983) *Numerical ecology*, Elsevier Scientific Publishing Company 3-300 pp.

27) Macnaughton-Smith, P. (1963) *The classification of individual by possession of attributes associated with a criterion* Biometrics 27:364-368.

28) Mahalanobis, P. C. (1936) *On the generalized distance in statistical*, Ins. Sci. India 2(1):49-55.

29) Mainly, B.F.J. (1986) *Multivariate statistical methods* Ed. Chapman and Hall, New York 42-51 pp.

30) Margaleff, R. (1983) *Limnología*, Ed. Omega, Barcelona España.

31) Marshlall, D. (1987) *Biología de las algas*, Ed. Limusa, México 210 pp.

32) Matteucci, S y Colma A. (1982) *Metodología para el estudio de la vegetación*, Dpto de asuntos Científicos y técnicos de la Sria Gral de la O.E.A., Washintong, D.C. Serie de Biología. Monografía 33 num. 23, 163 pp.

33) Mueller-Dumbois, D y H Ellenberg (1974) *Aims and methods of vegetation ecology*, Ed John Wiley & Sons, New York 400 pp.

34) Orlocí, I. C. R. & Stieler Rao and W.M (1979) *Multivariate methods in ecological works*, International Cooperative Publishing House, Fairland U.S.A.

35) Pielou, E.C. (1984) *The interpretation of ecological* data, Ed John Wiley & Sons, New York 263 pp.

36) Reynolds, F.J. and Ludwing A.J. (1988) *Statistical ecology*, John Wiley & Sons, New York 3-82 pp.

37) Secretaria de Programación y Presupuesto (S.P.P) (1981) *Síntesis geográfica de Morelos*, Coordinación General de los Servicios Nacionales de Estadística, Geografía e Informática.

38) Sneath, P.H.A. & Sokal, R.R. (1973) *Numerical taxonomy. The principles and practice of numerical classification*, Ed Freedman, Sn Fco C.O 170-210 pp.

39) Sokal, R.R. & Sneath P.H.A. (1963) *Principles of numerical taxonomy*, Freedman Sn Fco C.A. 300 pp.

40) Wetzel, R.G. (1981) *Limnología*, Ed Omega, Barcelona,España.

41) Williams, W. T. and Lambert, J. M. (1960) *Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis*, J. Ecol. 48:689-710.

42) Zeitchel, B. *Why study phytoplankton?*, En. A. Sournia, Ed. Phytoplankton Manual, UNESCO.