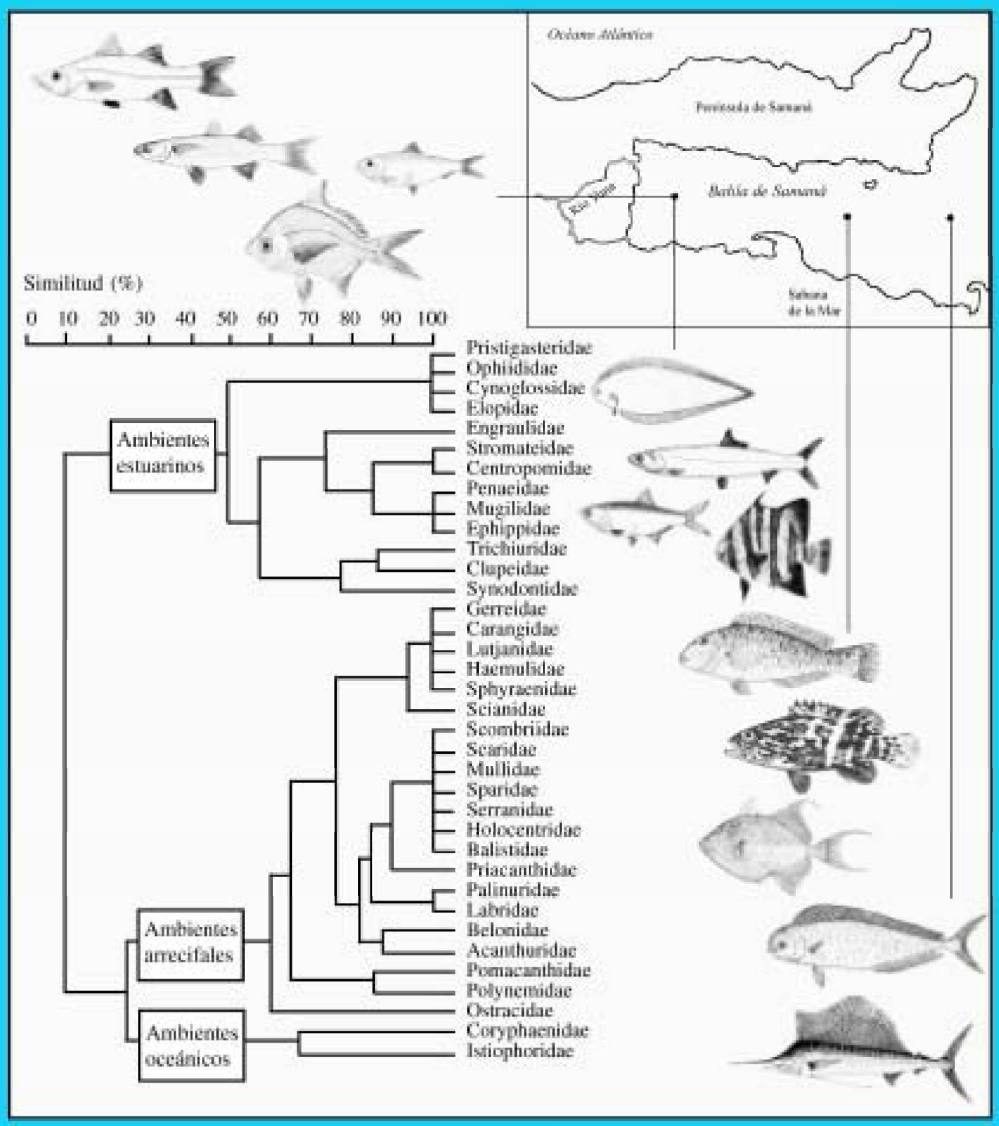
# LA CLASIFICACIÓN NUMÉRICA Y SU APLICACIÓN EN LA ECOLOGÍA

**Alejandro Herrera Moreno**




**INSTITUTO TECNOLÓGICO DE SANTO DOMINGO**
**Santo Domingo, República Dominicana**

# LA CLASIFICACIÓN NUMÉRICA Y SU APLICACIÓN EN LA ECOLOGÍA

ISBN 99934-25-12-5
Primera edición 2000

Edición y diseño
Alejandro Herrera-Moreno y Liliana Betancourt Fernández
Editora Sammenycar C. x A.
Santo Domingo, República Dominicana
Referencia: Herrera-Moreno, Alejandro (2000) La clasificación numérica y su aplicación en la ecología. Programa EcoMar/ Universidad INTEC, Editorial Sanmenycar, Santo Domingo, 121 pp.

# LA CLASIFICACIÓN NUMÉRICA Y SU APLICACIÓN EN LA ECOLOGÍA

**Alejandro Herrera Moreno**
**Programa EcoMar, Inc.**

**Publicación auspiciada por:**


**INSTITUTO TECNOLÓGICO DE SANTO DOMINGO**
**Santo Domingo, R. D.**
**2000**

*A mi adorada esposa Lily, quien "clasificó"*
*definitivamente mi vida del lado de la felicidad.*

Estimado lector:

No es pretensión de este libro -al margen de aportes personales- abordar originalmente un campo en el cual existen obras clásicas y autores consagrados, pero un propósito sí me anima: allanar a otros un camino arduo que he debido transitar. Mucho se ha escrito sobre clasificación, pero dispersos los trabajos en tiempo y espacio, escritos en su mayoría en un lenguaje esencialmente matemático, en ocasiones con un tratamiento parcial del tema, con una extensa terminología técnica no siempre bien explicada, y en idiomas diferentes del español; no es fácil para el que comienza, orientarse y avanzar. La experiencia de numerosos cursos por el mundo me ha mostrado la necesidad de una guía que recoja lo esencial de los métodos y permita al que comienza, particularmente estudiantes e investigadores jóvenes de Nuestra América, orientarse en este campo no siempre bien aplicado de la clasificación y hacerle más accesible el estudio de los textosoespecializados. Siguiendo en lo fundamental las ideas de los clásicos, particularmente a Boesch (1977), que es a mi juicio lo más didácticamente escrito sobre la materia; con aspectos nuevos de todos los trabajOs a los cuales he podido acceder y con la experiencia de mi trabajo práctico se ha confeccionado esta guía, que es además un reclamo de mis alumnos y colegas.

En el logro de este empeño no son pocos a los que debo gratitud por su ayuda desinteresada. Agradezco en primer lugar a la MSc. Rosa del Valle que me enseñó mi primer «cluster». al Dr. Jesús Sánchez del Instituto de Matemática y Cibernética Aplicada de Cuba, que como maestro me acercó por primera vez a estos métodos. Al Dr. Pedro Alcolado del Instituto de Oceanología de Cuba que revisó con su acostumbrada meticulosidad la primera versión de este libro (sin que ello le haga responsable de ninguno de sus errores) llenándolo de acertadas sugerencias.También debo agradecer a profesores de algunas Universidades que me facilitaron la presentación de mis cursos y a través de ellos a los alumnos que nutrieron esta obra de nuevas experiencias. En Brasil, agradezco a la Dra. Erika Schlenz de la Universidad de Sao Paulo y a la Dra. Maria Julia da Costa Belem de la Universidad de Río de Janeiro. En Venezuela, a la Dra. Elisabeth Méndez de Elguezábal de la Universidad de Oriente, en Cumaná. En Mëxico agradezco a los colegas del Centro de Investigaciones de Quintana Roo en Chetumal, el Instituto Tecnológico Agropecuario de Oaxaca y el Centro de Investigaciones Avanzadas de Mérida. En Cuba a los colegas del Instituto de Oceanología y del Instituto de Ecología y Sistemática, donde hallé entusiastas seguidores. En República Dominicana, tuve la oportunidad de exponer estos temas por primera vez gracias a la amabilidad del Dr. José Contreras, Decano del Area de Ciencias Básicas y Ambientales, y la Profesora Lic. Ana Mercedes Henríquez, ambos de la Universidad INTEC; la misma Universidad que hoy publica este libro bajo la dirección del Lic.Antonio Fernández, del Departamento de Investigaciones y Publicaciones Científicas, a quien extiendo un agradecimiento especial. Finalmente, deseo dar gracias al Sr. José A. Mari Mutt, Editor del Caribbean Journal of Science, por su gentileza en el envio electrónico del trabajo que sirve de base al último ejemplo de nuestro capítulo sobre la práctica de la clasificación.

Como pienso con José Martí, que «cada alumno que progresa es un maestro que nace», he intentado con este libro desbrozar de malezas un camino. Si en algunas partes éste es aún escabroso, si la hierba aquí o allá no ha sido bien cortada, no ha sido por dejadez ni apuro. Asumo la responsabilidad, sabiendo que los que vienen lo harán mejor y cumplirán su parte como he tratado yo, sencillamente- de cumplir la mía.

El Autor.

# CONTENIDO

*"La estadística es una forma de control social sobre la conducta profesional de los investigadores"*
*Richard J. Harris*

# 1. INTRODUCCIÓN

Aunque la interacción de la matemática y la estadística con la ecología no es algo nuevo, el extraordinario desarrollo de la informática en los últimos tiempos ha contribuido a popularizar el empleo de nuevos y más complejos métodos de análisis e interpretación, que han venido a dar respuesta a las necesidades de esta materia cuyo carácter multidisciplinario hace del análisis de numerosas variables bióticas y abióticas y sus interrelaciones, un objetivo esencial.

La diversidad de métodos existentes, cada uno con particularidades metodológicas y prácticas[1], y ajustados a los más diversos intereses, es algo con lo cual el ecólogo debe familiarizarse -al menos en las intenciones básicas de cada uno de ellos- a fin de orientarse debidamente tanto en la obtención como en la interpretación de sus datos. En este sentido, no son una excepción los métodos de clasificación numérica.

Definida en su forma más general la *clasificación* no es más que el ordenamiento de las entidades en grupos sobre la base de las relaciones entre sus atributos (Boesch, 1977); o en palabras más sencillas: el agrupamiento de cosas similares en clases (Pielou, 1984), con lo cual se distingue de la *identificación* que pretende encontrar la clase en la cual ha de ubicarse un nuevo individuo, entre clases ya establecidas (Orlóci, 1978). Esta subdivisión parece correcta para definir dos acciones relacionadas pero diferentes: la clasificación, en su acepción común de "ordenar o disponer por clases" y con ordenación, separación, distribución u organización como sinónimos; y la identificación en el sentido de "reconocer si una cosa es la misma que se supone o se busca" con sinónimos como filiación, identidad, reconocimiento o unificación.

Sin embargo, existe ambigüedad en el uso de estos términos en la literatura. En numerosos textos (Anderson, 1984; Morrison, 1990; Johnson y Wichern, 1992; Esbensen *et al*., 1994; Rencher, 1995) el término clasificación se emplea para denominar la ubicación de individuos en categorías pre-establecidas como parte de lo que Rencher (1995) llama el aspecto predictivo del análisis discriminante, lo cual usualmente recibe el nombre de asignación, diagnosis o como habíamos dicho, identificación, quedando el vocablo clasificación para los métodos de construir grupos (Chatfield y Collins, 1992).

Al margen de estas definiciones técnicas, la clasificación es algo inherente a la vida humana pues en este proceso aumentamos nuestro conocimiento, hacemos un uso más eficiente de la información,

[1] Existen campos bien definidos de interacción de la matemática y la estadística en la bioecología. Son algunos de ellos, además de la llamada estadística clásica (paramétrica y no paramétrica), diseño de muestreos, diseño experimental, distribuciones espaciales, índices ecológicos, métodos de clasificación, de ordenamiento y el modelaje matemático.

logramos acceder a ella más fácilmente cuando lo requerimos y además valoramos las cosas que nos rodean. El niño clasifica sus juguetes: algunos privilegiados, generalmente los más simples y desdeña otros, casi siempre los más caros. El adolescente clasifica asignaturas y amores en difíciles y fáciles. El adulto clasifica opiniones, amistades, oficios y no pocas veces en demasía los objetos del mundo material olvidando que su sentido clasificatorio infantil le acercaba más a la felicidad. Tal vez por ello Stuessy (1990), que trata en amplitud la historia de la clasificación vegetal, dice que el ser humano tiene compulsión por el orden.

En el pensamiento científico hallamos a los biólogos clasificando taxones; los geógrafos, regiones; los médicos, enfermedades; los químicos, compuestos; los historiadores, épocas; los arqueólogos, hallazgos; y los astrónomos, estrellas. En cualquier estudio, para poder interpretar la realidad circundante, el investigador clasifica y ordena los componentes de su sistema de una forma intuitiva buscando una estructura natural entre sus observaciones basada en su perfil multivariado (Hair *et al*., 1995).

Desde Aristóteles hasta Linneo, la teoría y práctica de la clasificación aplicada al estudio de los animales y plantas ha jugado un papel importante en la biología, que tuvo su momento cumbre en las teorías de Darwin. La clasificación correcta de los organismos es tan importante en la biología, que se ha desarrollado una disciplina particular: la taxonomía, para tratar la teoría y práctica de la clasificación biológica (Fielding, 1999). La clasificación de los elementos de la Tabla Periódica por Mendeleiev ha tenido un profundo impacto en las concepciones sobre la estructura atómica. La clasificación de las estrellas en "gigantes" y "enanas" sobre la base de su temperatura y luminosidad ha afectado fuertemente la teoría de la evolución estelar (Everitt, 1993).

En el campo de la ecología estructural los biólogos marinos daneses en la primera década del siglo clasificaron con acierto las comunidades bentónicas. A partir de especies comunes de moluscos y equinodermos, junto a otros taxones marinos, establecieron varias comunidades en relación con el tipo de fondo y la profundidad, cada una de las cuales se definía por la proporción de sus grupos componentes y era denominada acorde a las dos especies dominantes. Estos trabajos marcaron el punto de partida de los actuales estudios cuantitativos del bentos. Pero sin duda alguna, corresponde a los botánicos en este campo un mérito especial, pues ya desde principios de siglo la escuela fitosociológica de Braun-Blanquet (1979) estableció bases muy claras para la clasificación de las comunidades vegetales. Sus principios de ordenación espacial y cartografía, basados en inventarios florísticos y factores climáticos y edáficos, son un ejemplo de combinación de parámetros ecológicos para lograr una subdivisión coherente de la vegetación.

Estas tempranas clasificaciones, que se han dado en llamar "subjetivas", fueron la base de la actual ecología vegetal cuantitativa que nos presenta Greig-Smith (1983), donde la botánica entra de lleno en el campo de la clasificación numérica, a la cual se le atribuyen ventajas por encima de cualquier otro tipo de clasificación subjetiva, pues además de permitir el análisis de un número mayor de atributos de lo que es capaz de considerar la mente humana, tiene la propiedad de ser repetible una vez determinados los criterios clasificatorios.

Tales características están implícitas en su definición que establece que la *clasificación numérica* o el análisis de grupos, conjuntos, cúmulos o conglomerados ("cluster analysis") comprende una amplia variedad de técnicas para ordenar entidades en grupos sobre la base de ciertos criterios formales objetivamente establecidos (Boesch, 1977). Kaufman y Rousseeuw (1990) la definen como el arte de encontrar grupos en los datos; y más recientemente Krzanowski y Marriott (1966a) reiteran que con esta denominación los estadísticos distinguen las técnicas de análisis que dividen los datos en grupos. Chatfield y Collins (1992) además de considerar su objetivo de simplificación mediante la agrupación, señalan su valor para la exploración de los datos, generar hipótesis acerca de la estructura de la población o derivar predicciones a partir de las tendencias de agrupamiento.

Para Sharma (1996) el análisis de grupos es una técnica que combina observaciones en conjuntos de modo que, sobre la base de determinadas características, éstos sean homogéneos o compactos internamente y a su vez bien diferentes de otros. Este proceso brinda un resumen conveniente de los datos multivariados en los cuales se basa pues tiene el efecto de reducir la dimensionalidad de la tabla de datos (Fielding, 1999), pero generalmente rinde mucho más, pues ayuda a memorizar y entender mejor nuestros datos, facilita la comunicación entre especialistas y puede tener importantes implicaciones teóricas y prácticas (Everitt y Dunn, 1991).

Actualmente, estos métodos se aplican no solo en ecología sino en los más variados dominios (ver Statistica, 2000) como la inteligencia artificial, los patrones de reconocimiento, química, biología, economía, geociencias, estudios de mercado, medicina, ciencias políticas, sicología y otros. Aunque en esta diversidad de materias se le ha conocido con diferentes nombres: taxonomía cuantitativa o matemática, sistemática numérica, morfometría multivariada, taxometría (Sneath y Sokal, 1973), cladística numérica (Neff y Marcus, 1980), clasificación automática, análisis Q, clasificación jerárquica, botriología, análisis tipológico o taxonomía numérica (Kaufman y Rousseeuw, 1990), las denominaciones generales de clasificación numérica o análisis de grupos se han ido imponiendo.

En este contexto el término "cluster" es frecuentemente empleado -como si no fuera bien rico nuestro idioma- y empleado además ambiguamente para denominar cualquier paso del proceso clasificatorio, incluidos los diagramas arbóreos que como resultado final de éste se derivan, cuya estructura no siempre son grupos en sentido estricto. Siempre que deseemos discriminar las tendencias de relación grupal de nuestros datos podemos emplear métodos de clasificación; agruparemos, si entre las posibilidades de este método usamos determinadas técnicas; que nos darán siempre una subdivisión de nuestros datos originales pero que podrán ser grupos o no, según determinadas reglas.

Los métodos de clasificación no deben ser confundidos con los de otro campo relativamente cercano: los métodos de ordenamiento, cuyo objetivo no es establecer grupos ni delimitar clases sino expresar las relaciones entre entidades en modelos espaciales simplificados de varias dimensiones. Son ejemplo de ellos: el análisis de componentes y coordenadas principales, el factorial de correspondencias, la correlación canónica, el ordenamiento Gaussiano y el escalado multidimensional, entre otros (ver Bakus, 1990; Fielding, 1999).

La literatura recoge cierto debate acerca de cuál es el tipo de análisis multivariado más apropiado para los datos ecológicos, y distintas teorías favorecen los de clasificación o los de ordenamiento. En la ecología estructural estas discrepancias surgen al considerar la variación de las comunidades como un continuo pues con este concepto, los modelos de ordenamiento espacial se ajustarían mejor a la realidad, mientras que la partición en grupos que propone la clasificación puede parecer inapropiada.

Esta diferencia de enfoques es artificial (Digby y Kempton, 1991) pues el empleo de métodos de clasificación no implica que se desconozca la naturaleza continua de la variación ecológica y de hecho, en un gradiente de cambios muchas clasificaciones son capaces de identificar grupos precisos que identifican a zonas estructuralmente estables y grupos de "ruido" que son zonas de cambio. La clave, como en todo análisis ecológico no radica tanto en los métodos de interpretación sino en su base: el diseño del muestreo, especialmente en lo concerniente a qué puntos del gradiente se eligen y cuán representativos son los parámetros seleccionados y la calidad del muestreo.

Desde el punto de vista práctico, Boesch (1977) comenta que esta discusión no tiene sentido. Ambos tipos de métodos son herramientas interpretativas útiles, que incluso se complementan, si bien algunos pueden ser más relevantes en determinadas circunstancias. La clasificación es más útil para simplificar conjuntos complejos de datos y se le concede actualmente un alto valor exploratorio y descriptivo, aunque sus propiedades estadísticas no han sido totalmente desarrolladas (Morrison, 1990). El ordenamiento es más recomendable para el análisis de conjuntos de datos más pequeños y homogéneos donde sea de interés interpretar en detalle las relaciones entre entidades, y sus técnicas están mucho más elaboradas estadísticamente. De cualquier forma, todas estas técnicas multivariadas pueden ser utilizadas para analizar los mismos datos y ayudar en la búsqueda y comprobación de los fenómenos latentes en ellos, muchas veces con resultados muy similares (Popper y Heymann, 1996).

Hasta aquí hemos hablado de métodos, pero hay otros aspectos que sería útil comentar. Es natural que tras el muestreo se busque la mejor forma de analizar, interpretar y expresar los resultados, pero: ¿estamos seguros de la calidad de nuestros datos? No puede levantarse un edificio sobre cimientos débiles y no pocas veces se ven grandes intenciones interpretativas sobre datos pobres.

Por otra parte, si aún el muestreo y los datos que de él se derivan son óptimos, todas las variantes de análisis directo de la matriz de datos deben ser agotadas antes de pensar en métodos matemáticos o estadísticos complejos. No hay método matemático ni estadístico que supere la intuición de un investigador que conoce a fondo el material con que trabaja. Los gráficos de relación entre variables bióticas y abióticas, de variaciones estacionales o locales, la mapificación de variables, la inspección de la tendencia de variación entre estaciones y en los patrones de distribución de las especies y el empleo de métodos sencillos de correlación o pruebas de la estadística clásica -tanto paramétricas como no paramétricas- son algo insustituíble que proporcionan la clave para orientarse hacia métodos más complejos, si son realmente necesarios.

Algunas importantes revisiones sobre análisis multivariado (Digby y Kempton, 1991; Everitt, 1993; Hair *et al*., 1995) incluyen de hecho desde su inicio, resúmenes de métodos para el examen inicial de los datos. Krzanowski y Marriott (1966a) dicen que el primer paso de cualquier análisis estadístico es mirar a los datos e identificar sus peculiaridades fundamentales; simples ploteos de los datos pueden revelar fácilmente algunas facetas como las tendencias grupales de los individuos, las relaciones entre variables o la presencia de datos aberrantes, además de que pueden destacar aspectos relevantes, brindar puntos de análisis y hasta generar hipótesis para futuras investigaciones.

En relación con el campo que nos ocupa no es poco común que el interesado en clasificar sus datos aplique el primer índice de afinidad que tenga a mano; o éste u otro método de agrupamiento, sin conocer en esencia los mecanismos que le conducirán a subdividir sus datos en ciertos conjuntos. El resultado final, generalmente un árbol de clasificación, es considerado erróneamente como algo estático a lo cual debe buscarse necesariamente una explicación desconociendo que la estructura de grupos obtenida es simplemente el producto de todo un proceso matemático que de conocerse a cabalidad puede ser usado de manera objetiva en favor de una mejor interpretación ecológica.

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

*"El camino más largo comienza con el primer paso"*
*Lao Tsu*

# 2. PASOS GENERALES DE LA CLASIFICACIÓN

Antes de iniciar algún paso del proceso de clasificación debemos definir con claridad los objetivos de su empleo, que esencialmente pueden resumirse en dos: explorar los datos buscando las tendencias grupales que en ellos subyacen y/o confirmar patrones relacionados con la situación ecológica en estudio. Para cualquiera de estos fines, en un análisis de clasificación adecuado pueden seguirse varios pasos fundamentales, algunos de los cuales son obligatorios y otros opcionales dentro de las posibilidades e intereses del que lo emplee. Cada uno de ellos a su vez, involucra una serie de posibles consideraciones (Fig. 2.1).

El punto de partida es por supuesto la confección de la matriz original de datos que se obtiene en el muestreo ecológico. Generalmente, ésta resume la composición cualitativa, cuantitativa o ambas de un conjunto de parámetros físicos, químicos o biológicos (aquí nos referiremos a especies) en diferentes localidades, estaciones de muestreo o épocas (según sea el enfoque del estudio espacial o temporal). En cualquier caso los elementos que son objeto de la clasificación se denominan *entidades,* término análogo al de objetos (Legendre y Legendre, 1979; Everitt, 1993), dato unitario (Anderberg, 1973) o al de OTU («operational taxonomic units») que emplea la taxonomía numérica (Sneath y Sokal, 1973; Clifford y Stephenson, 1975). Los elementos que brindan el contenido de información de la tabla serán los *atributos;* descriptores para Legendre y Legendre (1979), variables para Anderberg (1973); individuos para Everitt (1993) o los carácteres en la taxonomía numérica (Sneath y Sokal, 1973; Clifford y Stepehenson, 1975).

Los datos obtenidos -que pueden ser de varios tipos- nos llevan a un segundo paso: el análisis de los datos, donde examinaremos si es necesario su *reducción,* qué criterios emplear, así como si procede o no su *transformación* o *estandarización.* La naturaleza y las característícas de los datos define el tercer paso: la selección de una o varias medidas de afinidad, que nos expresarán numéricamente el grado de parecido entre las entidades objeto de la clasificación, bien sea en forma de *similitud, disimilitud, correlación* o *distancia.*

En tal sentido debemos aclarar que la matriz original de datos encierra una doble información según se analice la afinidad entre filas o entre columnas. Así, si clasificamos localidades, estaciones de muestreo o intervalos temporales estaremos realizando lo que se conoce como *análisis normal* (o análisis Q); pero si clasificamos las especies (u otros parámetros) que caracterizan a cada estación o época entonces se trata del *análisis inverso* (o análisis R). Existen numerosos índices, con distintos intervalos de variación y que difieren en sus propiedades matemáticas, por lo que su selección adecuada es un paso clave en la clasificación.

1. Confección de la matriz original de datos

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 20 | 2 | 0 | 3 | 18 |
| B | 5 | 12 | 2 | 16 | 3 |
| C | 1 | 0 | 13 | 1 | 1 |
| D | 2 | 10 | 10 | 11 | 1 |
| E | 17 | 1 | 1 | 1 | 25 |

2. Análisis de los datos

3. Selección de una o varias medidas de afinidad

4. Confección de matrices de afinidad

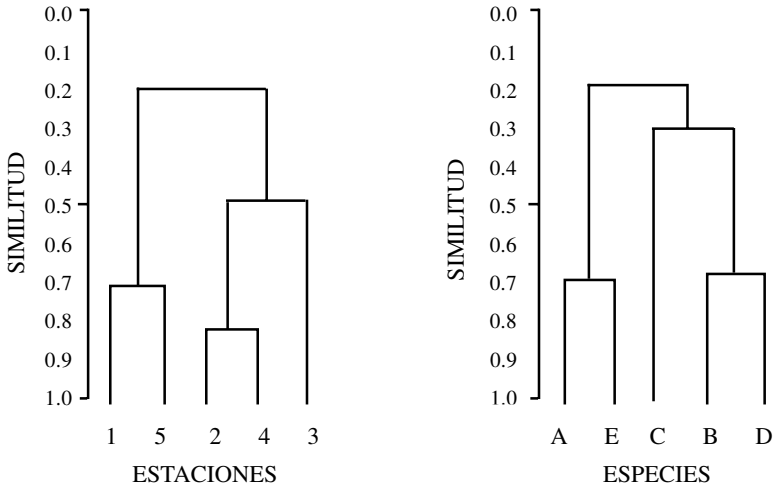
Estaciones

| 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|
| 1.00 | 0.29 | 0.17 | 0.31 | 0.86 | 1 |
| | 1.00 | 0.51 | 0.88 | 0.19 | 2 |
| | | 1.00 | 0.48 | 0.14 | 3 |
| | | | 1.00 | 0.23 | 4 |
| NORMAL | | | | 1.00 | 5 |

Especies

| A | B | C | D | E | |
|---|---|---|---|---|---|
| 1.00 | 0.32 | 0.10 | 0.21 | 0.84 | A |
| | 1.00 | 0.19 | 0.72 | 0.27 | B |
| | | 1.00 | 0.52 | 0.13 | C |
| | | | 1.00 | 0.15 | D |
| INVERSA | | | | 1.00 | E |

5. Selección de uno o varios métodos de agrupamiento




6. Interpretación de las clasificaciones

7. Relación de las clasificaciones normal e inversa: análisis nodal

Grupos de estaciones

| Grupos de especies | 1 | 5 | 2 | 4 | 3 |
|---|---|---|---|---|---|
| A | 20 | 18 | 2 | 3 | 0 |
| E | 17 | 25 | 1 | 1 | 1 |
| C | 1 | 1 | 0 | 1 | 13 |
| B | 5 | 3 | 12 | 16 | 2 |
| D | 2 | 1 | 10 | 11 | 10 |

8. Análisis global de resultados y relación con otras variables

Figura 2.1. Secuencia de pasos de la clasificación numérica.

Los resultados de la afinidad entre entidades, calculados a partir de las medidas seleccionadas, se resumen en las *matrices de afinidad* normal o inversa. La normal expresa las relaciones entre localidades (o intervalos de tiempo), mientras que la inversa la de los conjuntos de especies. Ambas matrices son simétricas, de modo que los dos lados de la diagonal que recorre las cuadrículas de autoafinidad de cada entidad son imágenes especulares. Por ello no es necesario presentar la matriz completa sino solo una de sus mitades triangulares (Fig. 2.1).

Estas matrices pueden ser empleadas por sí mismas para la interpretación en alternativas como el *diagrama de Trellis* o la *proyección de similaridad cenoclínica*, sin llegar necesariamente al quinto paso: la selección de métodos de agrupamiento. En este último se resume la esencia de la formación de grupos en el proceso clasificatorio; y como veremos, existen diferentes métodos de agrupamiento de los cuales los *jerárquicos, aglomerativos* y *combinatorios* son los más empleados. Dentro de éstos existen distintas estrategias que difieren en su forma de cálculo y en las propiedades sobre el espacio del *árbol de clasificación* o *dendrograma* que de ellos se derivan.

La interpretación de las clasificaciones -tanto normal como inversa- es un sexto paso indispensable que brindará, sobre la base de determinadas *reglas de decisión* para formar los grupos, la relación de las clasificaciones con los factores externos que determinan la creación de los conjuntos, así como la posible necesidad de *reasignación* de entidades para lograr una mayor homogeneidad y la evaluación de las diferencias entre grupos.

Como es recomendable el empleo de distintas estrategias aglomerativas a veces es conveniente, como parte de este paso interpretativo, la evaluación del grado de ajuste entre las clasificaciones y la matriz de afinidad mediante métodos sencillos de correlación, para conocer cuál dendrograma está reflejando con mayor objetividad las relaciones implícitas en la matriz.

Un séptimo paso opcional pero muy recomendable si se han realizado las clasificaciones normal e inversa es el *análisis nodal*, en el cual la información aislada de las clasificaciones se relacionan íntimamente en un gráfico nodal que permite evaluar el comportamiento simultáneo de los varios grupos, a través del cálculo de diferentes tipos de índices.

Queda finalmente el análisis global de los resultados y la relación con otras variables, paso en el cual toda la responsabilidad recae sobre el investigador, quien deberá interpretarlos considerando cuestiones metodológicas de las técnicas usadas; y ecológicas, propias de la realidad viva que desea explicar.

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

*"Los números son el principio de las cosas"*
*Pitágoras*

# 3. TIPOS DE DATOS

No es raro que aún los que están acostumbrados a manejar datos ecológicos desconozcan su correcta denominación. Sin embargo, existen distintas categorías y varias clasificaciones (véanse a Sokal y Sneath, 1963; Clifford y Stephenson, 1975; Boesch, 1977; Orlóci, 1978; Crisci y López Armengol, 1983) cuyo conocimiento permite hacer más eficiente su obtención primaria, convertir un tipo en otro, explotar al máximo su contenido de información, conocer las limitaciones que cada tipo encierra y determinar las exigencias que a ellos pueden aplicarse (Tabla 3.1).

Las clasificaciones y categorías existentes están referidas a distintas manifestaciones del quehacer científico, por lo que de ellas tomaremos solo aquellas que son de uso más común en ecología (Tabla 3.2). Para la taxonomía numérica el interesado puede consultar a Sokal y Sneath (1963) o Dunn y Everitt (1982) que tratan en detalle los tipos de datos de esta disciplina. De cualquier forma, sea cualfuere el tipo de dato, su clasificación procede a partir del número de estados que alcanza (doble o múltiple), su naturaleza (cualitativos y cuantitativos) y su escala de medición.

Tabla 3.1. Clasificación general de los tipos de datos.

| Número de estados | Naturaleza | Forma de estado |
|---|---|---|
| DOBLE ESTADO | Cualitativos | Presencia-Ausencia<br>Estados excluyentes |
| MULTIESTADO | Cualitativos | Ordenado<br>Desordenado |
| | Cuantitativos | Continuos<br>Discontinuos |
| | Semicuantitativos | Rangos<br>Escalas |

Tabla 3.2. Tipos de datos de mayor aplicación ecológica y ejemplos.

| Tipo de dato | Naturaleza | Posibles estados | Ejemplos de parámetros o situaciones ecológicas |
|---|---|---|---|
| DOBLE ESTADO | Cualitativo | Presente<br>Ausente | Composición cualitativa de especies |
| MULTIESTADO | Cualitativo ordenado | Abundante<br>Moderado<br>Escaso | Ordenamiento cualitativo de parámetros ecológicos |
| | Cuantitativo discontinuo | 0 a $\alpha$ | Número de taxones, número de individuos |
| | Cuantitativo continuo | 0 a $\alpha$ | Abundancia, biomasa, cobertura, diversidad |
| | Semicuantitativo | 1 a $\alpha$ | Asignación de rangos y escalas de abundancia |

## Datos cualitativos

*Datos de doble estado: presencia-ausencia.*- Los datos de doble estado, siempre cualitativos, son los llamados *binarios* o predicados dicotómicos (Crisci y López Armengol, 1983). El tipo de mayor aplicación ecológica es el de *presencia-ausencia*, ejemplificado cuando en nuestras estaciones, unidades muestrales o intervalos de tiempo, evaluamos el número de especies y dónde éstas están o no presentes. Poseen solamente dos estados de carácter, en este caso presencia o ausencia de las especies, de ahí su denominación de binarios.

En el muestreo ecológico el dato de presencia-ausencia identifica si la especie aparece o no en una estación determinada, por lo que podemos expresar la presencia con un signo "+" ó una cruz y la ausencia con un signo "-" ó un espacio en blanco. Sin embargo, como la clasificación exige una expresión cuantitativa del dato primario para los cálculos, el dato cualitativo de doble estado debe ser codificado (Fig. 3.1). Por ello, numéricamente se expresa como «1» la presencia y como "0" la ausencia, donde el número desempeña sólo una función de rótulo o marca de identificación para facilitar el posterior tratamiento cuantitativo (Crisci y López Armengol, 1983).

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | + | + | + | | |
| B | + | | + | + | + |
| C | + | | + | | + |
| D | + | + | + | + | + |
| E | + | + | + | + | + |

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 1 | 1 | 1 | 0 | 0 |
| B | 1 | 0 | 1 | 1 | 1 |
| C | 1 | 0 | 1 | 0 | 1 |
| D | 1 | 1 | 1 | 1 | 1 |
| E | 1 | 1 | 1 | 1 | 1 |

Figura 3.1. Matriz original de datos de presencia ausencia y de datos codificados.

A este tipo de dato se denomina *asimétrico* pues los estados que se asignan no tienen el mismo peso a la hora de agrupar (Kaufman y Rousseeuw, 1990). Ello se debe a que cuando decimos que dos especies comparten una estación implica que tienen algo en común pero no puede decirse exactamente lo mismo si no están presentes, por lo que los "ceros" pueden ser menos importantes que los "unos". Contrariamente, el otro tipo de dato cualitativo de doble estado: el de *estados excluyentes* se identifica como *simétrico* pues ambas entidades tienen el mismo peso. Un ejemplo conocido es el sexo donde si adjudicamos un código de "1" a los machos y "0" a las hembras ambos son igualmente importantes a la hora de agrupar, aunque el dato de estados excluyentes no es de mayor aplicación en la ecología estructural.

Al emplear datos binarios en el análisis normal deseamos conocer el grado de parecido entre las listas de especies en las localidades muestreadas. Como sólo se cuenta con la presencia-ausencia como atributo puede ocurrir que dos especies tengan una afinidad cualitativa muy alta por poseer similar distribución en las dos localidades, aún cuando sus dominancias o abundancias sean bien diferentes. Por ello, Pielou (1977) plantea que el dato de presencia-ausencia puede conducir a una

sobreestimación desproporcionada de las especies escasas en relación con lo que representan sus valores reales de abundancia, que cuantitativamente serían muy pequeños o despreciables. En el análisis inverso, donde nos interesa conocer cuál es el grado de concurrencia de las especies, o sea cuáles comparten determinadas localidades puede ocurrir que dos especies aparezcan siempre juntas -lo que también brindaría una afinidad cualitativa máxima entre ellas-, pero por la acción de determinados factores bióticos o abióticos que promuevan su incremento en algunos hábitats o su disminución en otros, se presenten con abundancias o dominancias diferentes.

Aunque lo anteriormente señalado son desventajas reales de los datos binarios, algunos trabajos los señalan como más ventajosos que los cuantitativos al no estar sujetos, en la misma medida, a la influencia de las variaciones estacionales o locales, o a los errores propios del muestreo cuantitativo. Independientemente de ésto, los datos binarios suelen ser la única alternativa en algunos campos como la Biogeografía, donde solo se cuenta con listas de especies por regiones. Además, si la matriz original posee un alto porcentaje de ceros, o sea, el ambiente estudiado es pobre en especies, la información que aporta la cuantificación puede ser mínima.

Por otra parte, siempre que se estudie un gradiente ecológico de cambios y la variación a estudiar tenga un reflejo en la composición cualitativa de la comunidad, es preferible esta alternativa. Tal es el caso de la zonación de las especies de moluscos en el litoral rocoso (Herrera *et al*., 1987) que muestra un patrón de variación del mar hacia la costa, como reflejo de adaptaciones particulares a los tensores propios de este ambiente (Tabla 3.3). Especies como *Drupa nodulosa*, *Astraea tuber* y *Chiton marmoratus*, por mencionar algunas, tipifican los primeros pisos litorales como componentes de la región infralitoral; *Acanthopleura granulata* y *Acmaea cubensis* la región mesolitoral; mientras que *Cenchritis muricatus*, *Littorina lineolata* y *Echininus nodulosus* caracterizan el extremo ambiental supralitoral. Nótese como la propia tendencia de aparición de los datos en la matriz, aún sin ordenar, sugiere la distribución por horizontes litorales.

Para brindar otro ejemplo de cómo la tendencia de grupos puede observarse en la propia matriz de datos cualitativos tomamos la información del estudio de la composición y diversidad de las pesquerías en ocho sitios de desembarco de la Bahía de Samaná, República Domicana (Sang *et al.,* 1997) y reordenamos sus datos con un sentido ecológico-geográfico, del estuario al océano. La Tabla 3.4 muestra cambios paulatinos en las familias representadas en los diferentes sitios de desembarco, que varían según su posición geográfica respecto a los diferentes ambientes de pesca adyacentes: estuario, manglares, pastos marinos y arrecifes coralinos.

En las áreas del NO de la Bahía (Sánchez) se registran las mayores capturas de camarones peneidos y una ictiofauna típicamente estuarina, con especies demersales pertenecientes a las familias Centropomidae y Mugilidae o pelágicas de Engraulidae. En la medida que abandonamos la influencia del estuario y comienza el desarrollo de pastos marinos y arrecifes coralinos, como ocurre en Las Pascualas, Los Cacaos (al N y centro de la bahía), Sabana la Mar y Miches (al SE casi fuera de la bahía) aumenta la representación de familias demersales neríticas arrecifales como Serranidae, Holocentridae, Balistidae o Scaridae aunque por la situación de gradiente también se registran algunos representantes estuarinos y/o típicamente pelágicos. En Las Galeras y Las Terrenas (fuera

Tabla 3.3. Distribución cualitativa de treinta y cinco especies de moluscos en veinte unidades muestrales del litoral rocoso de Puerto Escondido, (NO de Cuba) ubicadas del mar a la tierra (según Herrera *et al.,* 1987).

| | Unidades muestrales | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *Drupa nodulosa* | X | | | | | | | | | | | | | | | | | | | |
| *Astraea tuber* | | X | | | | | | | | | | | | | | | | | | |
| *Chiton marmoratus* | X | | | | X | X | | | | | | | | | | | | | | |
| *Acantochitona astriger* | | | X | X | X | | | | | | | | | | | | | | | |
| *Leucozonia ocellata* | | | | X | | | | | | | | | | | | | | | | |
| *Thais rustica* | | | | | | | X | | | | | | | | | | | | | |
| *Hemitoma octoradiata* | X | X | X | X | | | | | | | | | | | | | | | | |
| *Tricolia adamsi* | | X | X | | X | | | | | | | | | | | | | | | |
| *Fissurella barbouri* | X | X | X | | | X | | | | | | | | | | | | | | |
| *Fissurella nodosa* | X | | X | X | | | X | | | | | | | | | | | | | |
| *Pinctada radiata* | X | | | X | | | | | | | | | | | | | | | | |
| *Planaxis lineatus* | | X | | | X | | | | | | | | | | | | | | | |
| *Diodora viridula* | X | X | | | | X | | | | | | | | | | | | | | |
| *Columbella mercatoria* | X | X | X | | X | | X | | | X | | | | | | | | | | |
| *Fissurella angusta* | X | | | X | | | | | | | X | | | | | | | | | |
| *Fissurella rosea* | | X | | | | X | | | X | | | | | | | | | | | |
| *Thais deltoidea* | | | | | | | X | X | | | | | | | | | | | | |
| *Fisurella barbadensis* | | | | X | | | X | X | | | X | X | | | | | | | | |
| *Petaloconchus irregularis* | | | | | X | X | | X | X | | X | | | | | | | | | |
| *Ceratozona squalida* | | | | X | | | X | | X | X | X | | | | | | | | | |
| *Acmaea leucopleura* | | X | | X | X | | | X | X | X | | | | | | | | | | |
| *Cittarium pica* | | | | X | | | X | | | X | X | | | | | | | | | |
| *Littorina meleagris* | | | X | X | X | X | X | X | X | X | X | | | | | | | | | |
| *Brachidontes exustus* | X | | X | X | X | | X | X | X | X | X | | X | | | | | | | |
| *Isognomum alatus* | X | | | X | | X | X | X | | | | | | | | | | | | |
| *Acmaea jamaicensis* | X | X | X | X | X | X | X | X | | | X | X | | | | | | | | |
| *Purpura patula* | | | | | | | | X | | | | | | | | | | | | |
| *Acanthopleura granulata* | | | | | | | | | | X | | | | | | | | | | |
| *Littorina ziczac* | | | | | | | | | X | | | | | | | | | | | |
| *Acmaea cubensis* | | | | | | | | | | | X | | | | | | | | | |
| *Littorina angustior* | | | | | | | | | | | | X | | X | | X | | | | |
| *Nodilittorina tuberculata* | | | | | | | | | | | | | | | X | X | | X | | |
| *Cenchritis muricatus* | | | | | | | | | | | | | | X | | | X | X | X | |
| *Littorina lineolata* | | | | | | | | | | | | | | X | | X | | X | | |
| *Echininus nodulosus* | | | | | | | | | | | | | | X | X | X | | X | X | X |

de la bahía y al N de la Península), que representan el extremo del gradiente hacia el océano, algunas de las anteriores familias está ausentes, aparecen otras del complejo de pastos marinos- arrecifes coralinos (indicando su relación con estos ecosistemas) y de manera exclusiva las familias Coryphaenidae e Istiophoridae, de hábitos pelágicos, en concordancia con su ubicación en las cercanías de mar abierto.

De cualquier forma, la elección de la variante cualitativa no excluye que debamos tener en cuenta elementos cuantitativos. El número de especies que se registra en una comunidad guarda una estrecha relación con el número de individuos muestreados, de forma tal que sólo con un tamaño de muestra

Tabla 3.4. Matriz de presencia-ausencia de las principales familias de crustáceos y peces de las pesquerías de Samaná en siete sitios de desembarco, reordenados en un gradiente del estuario al océano, a partir de datos de Sang *et al*. (1997). Las letras indican los sitios de desembarco: SZ: Sánchez, M: Miches, LC: Los Cacaos, SM: Sabana de la Mar, LP: Las Pascualas, LG: Las Galeras y LT: Las Terrenas.

| Ambientes: | Estuario-Manglares-Pastos-Arrecifes-Océano | | | | | | |
|---|---|---|---|---|---|---|---|
| | Sitios de desembarco | | | | | | |
| Familias | SZ | M | LC | SM | LP | LG | LT |
| Pristigasteridae | X | | | | | | |
| Ophidiidae | X | | | | | | |
| Cynoglosidae | X | | | | | | |
| Elopidae | X | | | | | | |
| Engraulidae | X | X | | | | | |
| Stromateidae | X | X | | X | | | |
| Trichiuridae | X | X | | | X | | |
| Clupeide | X | X | X | | X | | |
| Penaidae | X | X | | X | X | | |
| Synodontidae | X | | X | | X | | |
| Mugilidae | X | X | | X | X | | |
| Centropomidae | X | X | | X | | | |
| Ephippidae | X | X | | X | X | | |
| Gerreidae | X | X | X | X | X | X | X |
| Scianidae | X | X | | X | X | X | X |
| Carangidae | X | X | X | X | X | X | X |
| Lutjanidae | X | X | X | X | X | X | X |
| Haemulidae | X | X | X | X | X | X | X |
| Sphyraenidae | X | X | X | X | X | X | X |
| Scombridae | X | X | X | X | X | X | X |
| Polynemidae | | X | | X | X | X | |
| Scaridae | | X | X | X | X | X | X |
| Mullidae | | X | X | X | X | X | X |
| Sparidae | | X | X | X | X | X | X |
| Serranidae | | X | X | X | X | X | X |
| Holocentridae | | X | X | X | X | X | X |
| Priacanthidae | | X | X | X | X | | X |
| Balistidae | | X | X | X | X | X | X |
| Belonidae | | | X | X | X | X | X |
| Palinuridae | | X | X | X | | X | X |
| Labridae | | X | X | X | | X | X |
| Acanthuridae | | | X | X | | X | X |
| Ostracidae | | | X | | X | X | |
| Pomacanthidae | | | | X | X | X | |
| Coryphaenidae | | | | | | X | X |

elevado el total de especies aparece adecuadamente representado. Quiere esto decir, por ejemplo, que no es conveniente para propósitos clasificatorios comparar dos localidades cuyas listas provengan en un caso de una muestra de 300 individuos y en la otra de 50, a causa de serias diferencias en el esfuerzo de muestreo. En el segundo caso, el requisito de tamaño de muestra que hace que la lista de especies obtenida sea representativa de la localidad estudiada, probablemente no se cumpla.

El tamaño mínimo de muestra para obtener una representación adecuada de la composición específica se estima haciendo un gráfico acumulativo del número de especies contra el número de individuos o unidades muestrales. La curva acumulativa, muy variable al inicio, va tendiendo a una estabilización al crecer el tamaño de la muestra hasta llegar a su nivelación horizontal, punto en el cual se considera que el número de individuos o unidades muestrales es suficiente para estimar adecuadamente al número real de especies.

No obstante, una comunidad puede tener una abundancia tan baja que sea prácticamente imposible obtener un tamaño de muestra grande. Por ejemplo, en el arrecife del Rincón de Guanabo, en la costa Norte de la Habana, Cuba, 30 unidades muestrales de 1 m² son suficientes para colectar 400 individuos, representados por 20 especies de gorgonáceos en los pavimentos rocosos de 10 m; mientras que en la laguna arrecifal, un esfuerzo de muestreo igual en términos de unidades muestrales, arroja solamente 21 individuos y 3 especies (Herrera *et al.,* 1997). En tales casos la representación de especies es adecuada para fines clasificatorios pues la comunidad está condicionada por un conjunto de tensores (presencia de sedimento, cobertura vegetal, turbidez y batimiento) que no permiten una mayor abundancia y no tiene sentido incrementar el esfuerzo de muestreo.

*Datos cualitativos de multiestado ordenado.-* Los datos de *multiestado ordenado***,** también denominados *de secuencia lógica,* se refieren a datos cualitativos que pueden ser ordenados en una secuencia de magnitud de la cualidad estudiada (Crisci y López Armengol, 1983), pues poseen una jerarquía de formas diferentes que abarcan la variación total del intervalo de entidades. Un ejemplo típico es cuando en una lista de especies vamos más allá de señalar su simple presencia o ausencia y tenemos criterios para subdividirlas en: abundantes, comunes, escasas y raras; dentro de la lista varias especies pueden recibir la misma categoría.

Como se trata en definitiva de un dato cualitativo de presencia-ausencia con un poco más de información, esta ganancia informativa solo puede ser aprovechada si se plasma cuantitativamente, o sea, codificando las especies mediante puntajes del 1 en adelante (Fig. 3.2). En tal caso, los números desempeñan una función ordinal ya que indican la posición de una cualidad en una secuencia de grados (Crisci y López Armengol, 1983). Este ordenamiento en categorías diferentes debe partir de un criterio de magnitud que permita la jerarquización.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | Ab | Ab | Co | Au | Ab |
| B | Co | Co | Ab | Co | Au |
| C | Es | Ra | Co | Ab | Au |
| D | Ra | Ra | Co | Co | Au |
| E | Au | Ra | Ra | Ra | Au |

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 4 | 4 | 3 | 0 | 4 |
| B | 3 | 3 | 4 | 3 | 0 |
| C | 3 | 1 | 3 | 4 | 0 |
| D | 1 | 1 | 3 | 3 | 0 |
| E | 0 | 1 | 1 | 1 | 0 |

Figura 3.2. Matriz original de datos de multiestado ordenado y de datos codificados, considerando: Ab: abundante (4), Co: común (3), Es: escasa (2) , Ra: rara (1), Au; ausente (0).

Esta codificación de los datos en clases implica que mientras mayor es el número de clase, mayor es la abundancia, pero una misma clase no siempre significa una abundancia igual. Los intervalos entre clases no tienen casi significación y dependiendo de sus límites la agrupación puede estar influida en varias formas (van Tongeren, 1987) razón por la cual no se recomiendan para la clasificación (Boesch, 1977). Este tipo de dato es producto de un muestreo cualitativo donde el componente cuantitativo ha sido manejado de forma grosera, un muestreo cuantitativo donde la precisión de los datos es dudosa, o simplemente una división basada en la experiencia. Por ejemplo, en el arrecife costero al N de La Habana, Cuba, son abundantes los corales *Montastrea cavernosa* y *Siderastrea radians*; *Dichocoenia stokesi* es común; y *Colpophyllia natans* es rara.

A diferencia de los datos de multiestado ordenado, los datos de *multiestado desordenado***,** o *sin secuencia lógica* no pueden ser organizados en una secuencia de grados del atributo. Un ejemplo sería la relación de una especie con el sustrato: infaunal, epifaunal y críptica. Evidentemente no habría criterios para codificar estas tres formas como 1, 2 y 3 pues no hay ninguna jerarquización. Esto se resuelve convirtiendo cada estado en un dato de presencia ausencia (por ejemplo: críptica (1), no críptica (0)) pero ello implica una sobreestimación del atributo "relación con el sustrato", por lo que la subdivisión ordinal de los datos puede ser subjetiva y crear discontinuidades que realmente no existen. Este tipo de dato ha quedado para otras disciplinas como la taxonomía numérica y no es de mayor aplicación ecológica como el anterior que si es posible su ordenación. Aún así, es una forma de hacer más refinada la información que de acuerdo a la práctica más común, nos conducirían, como veremos inmediatamente, al concepto de rango.

Si categorizamos los datos cualitativos hasta aquí vistos, considerando su escala de medición y sus propiedades tendríamos dos grupos, según describe Sharma (1996). Los datos de presencia-ausencia, estados excluyentes y multiestado desordenado que tienen una escala de medición nominal y los números son usados solo para categorizar, son inapropiados para cálculos de estadígrafos como la media y la desviación estándar y solo las estadísticas basadas en conteos como la moda o las distribuciones de frecuencias, son apropiados para ellos. El tipo de dato es, por tanto, *nominal***.** En los datos de multiestado ordenado existe una significación de jerarquías pero ésta es simplemente de escala ordinal pues la secuencia de categorías, no representa diferencias iguales del atributo medido. Los estadígrafos válidos para este tipo de datos son la moda, la mediana, las distribuciones de frecuencia y los métodos no paramétricos como la correlación por rangos (Sharma, 1996). El tipo de dato es llamado *ordinal*. Las variables que se miden usando escalas nominales u ordinales se refieren comúnmente como variables *no métricas***.**

*Datos de rango.-* Cuando al dato de multiestado cualitativo ordenado se le asigna un número, en vez de un estado jerárquico nominal (abundante o escaso, por ejemplo), entonces la secuencia de magnitud adquiere un carácter semicuantitativo y hablamos de rangos. El dato de rango gradúa la colección especie por especie y puede partir tanto de un criterio general de abundancia, según ya vimos, como de valores numéricos individuales que permitan diferenciar entre las especies abundantes, la primera, la segunda o la tercera.

Quizás pueda pensarse que si cada especie está caracterizada por un número representativo de su abundancia o proporción no es necesario asignar rangos, pues la clasificación puede proceder a partir de ellos. Esto es cierto, pero la gradación en rangos es una alternativa útil cuando la precisión de los valores es cuestionable por errores en el muestreo, número insuficiente de réplicas o diferencias notables en el esfuerzo, entre otras causas. El uso de rangos puede ser parte de una estrategia de muestreo donde se busque rapidez y se desee obtener un panorama general de la comunidad, con mayor ganancia informativa que con datos de multiestado codificados. En tal caso, el ordenamiento sucesivo de las especies en rangos mantiene el componente cualitativo de la información y aprovecha el cuantitativo que de otra forma se perdería.

La forma más común de asignar los rangos (Fig. 3.3) tomada de la estadística no paramétrica (Siegel, 1985), se basa simplemente en substituir los valores originales por números a partir del uno, de mayor a menor. Ante datos repetidos en la matriz original, como por ejemplo las especies C, D y E en la estación 3 que coinciden en el valor 5, se le asigna a cada una el rango que le corresponde en orden, en este caso 3, 4 y 5 pero el valor final de cada una será el promedio: (3 + 4 + 5)/3, o sea 4. Otro tanto ocurre con las especies E y F, en la estación 5.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 10 | 56 | 150 | 1 | 45 |
| B | 20 | 50 | 32 | 0 | 20 |
| C | 5 | 12 | 5 | 114 | 19 |
| D | 3 | 10 | 5 | 7 | 5 |
| E | 2 | 9 | 5 | 20 | 1 |
| F | 1 | 0 | 1 | 32 | 1 |

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 2 | 1 | 1 | 5 | 1 |
| B | 1 | 2 | 2 | 6 | 2 |
| C | 3 | 3 | 4 | 1 | 3 |
| D | 4 | 4 | 4 | 4 | 4 |
| E | 5 | 5 | 4 | 3 | 5,5 |
| F | 6 | 6 | 6 | 2 | 5,5 |

Figura 3.3. Matriz original de datos cuantitativos y matriz de rangos.

Este tipo de dato no debe usarse como norma habitual de trabajo, recuérdese que no representa el dato real e incluso valores iguales pueden recibir rangos diferentes o valores muy desiguales pueden coincidir en un mismo rango. Por ello, su empleo debe ser limitado a aquellas medidas de afinidad, que como veremos, se basan precisamente en datos de jerarquía.

Finalmente, vamos a referirnos a los datos derivados del empleo de *escalas de abundancia*, que si bien no son datos de rango en el sentido antes explicado, tienen también la característica de que un valor (del 1 en adelante) sustituye el dato primario; en estos casos siempre originalmente cuantitativo. Una escala de tal tipo (Tabla 3.5), cuya significación ecológica está avalada estadística y prácticamente, la ofrece Frontier (1969). Aunque el empleo de escalas de abundancia tiene como objetivo realizar muestreos rápidos y extensivos, no vemos ninguna objeción para que los datos así tratados puedan ser clasificados, sin olvidar que representan amplios intervalos.

Tabla 3.5. Escala de abundancias según Frontier (1969).

| Clases | Intervalos de abundancias |
|---|---|
| 1 | de 1 a 3 |
| 2 | de 4 a 18 |
| 3 | de 18 a 80 |
| 4 | de 80 a 350 |
| 5 | de 350 a 1500 |

**Datos cuantitativos**

Los datos *cuantitativos*, siempre de multiestado, son los llamados cardinales, magnitudes o cantidades dado que miden relaciones cuantitativas en sentido estricto (Crisci y López Armengol, 1983). Su obtención involucra, por tanto, una medición a lo largo de una escala, y una unidad de medida (Jobson, 1991). Si la unidad de medida es indivisible se dice que la variable o el tipo de dato es *discontinuo* o *discreto* (Jobson, 1991). Los datos discontinuos solo pueden expresarse por números enteros de ahí que su variabilidad sea discontinua, como es el caso del número de especies (o de cualquier taxa), aunque si este parámetro se estima por métodos como el de rarefacción de Sanders (1968), el número de especies para un tamaño de muestra dado puede ser un número fraccionario, caso en el cual caería artificialmente en la categoría de continuo.

Si la unidad de medida es infinitamente divisible de modo que al menos teóricamente la medición puede hacerse en unidades cada vez mas finas, se dice que el dato es *continuo* (Jobson, 1991). Los datos continuos expresan cualidades cuya variabilidad se distribuye en una escala continua, de ahí que su expresión pueda ser un número entero o fraccionario. Incluyen la mayoría de los parámetros ecológicos: abundancia, biomasa, cobertura, diversidad, dominancia, equitatividad; o la longitud, ancho o diámetro cuando se miden individuos de una población con propósitos de estudiar algún aspecto ecológico relacionado con su dinámica.

En el dato cuantitativo el atributo está representado por un número, que es expresión original de algún parámetro ecológico, bien sea estructural o funcional. En tales casos el análisis normal examinará cuánto se parecen las dos localidades en la composición, ya no solo cualitativa, sino también cuantitativa de sus especies; el análisis inverso, cuáles especies resultan más afines en sus patrones de distribución espacial o temporal.

Para la clasificación de los datos cualitativos aclarábamos que debían buscarse contrastes entre las entidades. Lo mismo es válido para los datos cuantitativos pero en este caso nos referimos a diferencias numéricas importantes que garanticen la formación de grupos, como se muestra en el ejemplo de la Tabla 3.6 que compara la abundancia de especies coralinas en ambientes limpios y contaminados (Herrera, 1991). Las estaciones limpias poseen una mayor representación de especies con dominancia de *Agaricia agaricites*, mientras que en las estaciones contaminadas el número de especies se reduce y dominan especies resistentes como *Siderastrea radians*. Ello se refleja en contrastes numéricos en la tabla que indican su idoneidad para la clasificación. Si los valores de la matriz son muy similares ello podría reflejar muestras procedentes de un biotopo muy homogéneo donde la creación de conjuntos puede tener poca significación ecológica y la afinidad varía en un estrecho intervalo. Si las diferencias numéricas entre los datos son de pequeña escala pero significativas para el agrupamiento del parámetro que se estudia deberán emplearse medidas de afinidad que ayuden a hacer evidentes dichas diferencias.

Distintos tipos de parámetros pueden brindar clasificaciones bien diferentes en lo cual influye, en primer lugar, su escala de variación. Por ejemplo, la heterogeneidad que se expresa en natios fluctúa entre 0 y 3; pero la cobertura porcentual entre 0 y 100 y la biomasa en gramos de peso húmedo

Tabla 3.6. Abundancia de especies de corales (colonias/m²) en nueve estaciones del arrecife costero al N de la Habana, Cuba, en un gradiente de contaminación al E y al O de la Bahía de la Habana y el Río Almendares, en comparación con áreas de referencia limpias en Playa Baracoa y Santa Cruz del Norte. Los subíndices del 1 al 4, indican el orden en el gradiente a partir de las fuentes contaminantes (según Herrera, 1991).

| | Playa | Río Almendares | | | Bahía de la Habana | | | | Sta. |
|---|---|---|---|---|---|---|---|---|---|
| Especies | Baracoa | $O^2$ | $O^1$ | $E^1$ | $O^4$ | $O^3$ | $O^2$ | $O^1$ | Cruz |
| *Madracis decactis* | 0.10 | 0.10 | 0 | 0.10 | 0 | 0 | 0 | 0.04 | 0 |
| *Diploria labyrinthiformis* | 0.10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| *D. clivosa* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| *D. strigosa* | 0.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| *Manicina areolata* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| *Montastraea cavernosa* | 0.95 | 0.30 | 1.40 | 0.70 | 0.30 | 0.06 | 0 | 0.04 | 0.22 |
| *M. annularis* | 0.25 | 1.10 | 0.30 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| *Stephanocoenia michelini* | 0.05 | 0 | 0.10 | 0.10 | 0 | 0.06 | 0 | 0.08 | 0 |
| *Isophyllia sinuosa* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.2 |
| *Dichocoenia stokesi* | 0.35 | 2.00 | 0.70 | 0.40 | 0.20 | 0.33 | 0 | 0.12 | 0.08 |
| *Meandrina meandrites* | 0.55 | 0.40 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0.06 |
| *Eusmilia fastigiata* | 0.20 | 0 | 0.30 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Agaricia agaricites* | 1.45 | 0.10 | 0 | 0.10 | 0 | 0 | 0 | 0.20 | 1.04 |
| *Siderastraea radians* | 1.00 | 0.80 | 2.50 | 3.15 | 3.30 | 1.06 | 2.10 | 1.56 | 0.12 |
| *Porites porites* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.04 |
| *P.astreoides* | 0.50 | 0.50 | 0.90 | 0 | 0 | 0 | 0 | 0 | 0.64 |
| Densidad total | 5.55 | 5.30 | 6.30 | 4.55 | 3.80 | 1.53 | 2.10 | 2.04 | 2.32 |

puede alcanzar valores mayores. Sin embargo, esta misma biomasa expresada en peso seco brindaría valores más moderados, con lo cual aclaramos que el cambio en la unidad de medida es también una fuente de variación. Este inconveniente se obvia seleccionando las variables más significativas a los efectos de la agrupación, estandarizándolas en la misma unidad (individuos/m² si se trata de abundancia, o gramos/m² si son datos de biomasa) para hacerlas comparables, o dándoles algún manejo estadístico.

Por otra parte, las diferencias entre clasificaciones efectuadas con distintos datos pueden tener una importante connotación teórica relacionada con lo que cada uno representa en la estructura o función de la comunidad que se estudia. Ignatiadis *et al*. (1992) encuentran resultados distintos al clasificar las comunidades fitoplanctónicas del Golfo de Saranicos (Mar Egeo) a partir de tres matrices: abundancia de especies (células/l); índices ecológicos (diversidad, equitatividad, dominancia, redundancia y riqueza); y parámetros relacionados con la biomasa (clorofila *a*, total de células, productividad primaria, tasa de asimilación y relación diatomeas/dinoflagelados).

La clasificación obtenida con abundancia ordena los grupos relacionados taxonómicamente y difiere de la obtenida con índices ecológicos, pues valores iguales de diversidad pueden corresponderse con composiciones de especies distintas. La discrepancia de estas clasificaciones con la de los parámetros relacionados con la biomasa ocurre pues aunque las muestras sean cuantitativamente similares en su estructura, la producción primaria o la tasa de asimilación están más influidos por el estado fisiológico de la célula y variables ambientales como la luz y la turbulencia (Ignatiadis *et al*., 1992).

El empleo de datos cuantitativos requiere de un mayor análisis para indagar por el cumplimiento de algunos requisítos mínimos. Sin embargo, estos requisitos no se corresponden con los de homocedasticidad, normalidad o linealidad tan importantes en las técnicas de inferencia estadística pues el análisis de grupos es una metodología para agrupar observaciones y como tal tiene fuertes propiedades matemáticas pero ningún fundamento estadístico. Al respecto, según Hair *et al.* (1995) hay que considerar: la *representatividad* de la muestra y su *multicolinealidad.*

En referencia a la representatividad muestral, dado que el investigador cuenta con un censo de la población para obtener grupos y desea con éste lograr una estructura, debe tener seguridad de que la muestra obtenida es realmente representativa de dicha población y los resultados, por tanto, generalizables. Cuando se emplean unidades muestrales los estimados de cualquier parámetro ecológico, bien sea densidad, biomasa, cobertura u otro, deben provenir de un número suficiente de observaciones o réplicas que garanticen la confiabilidad individual de cada dato, en cada una de las localidades o intervalos de tiempo. Además, debe tenerse en cuenta la posible influencia de las variaciones estacionales naturales o las diferencias locales debido a irregularidades o cambios dentro del mismo hábitat que influyan en la distribución de las especies. El análisis de grupos será solo tan bueno como lo sea la representatividad de la muestra (Hair *et al*., 1995).

En relación con la colinealidad aclaremos antes que este término define la fuerza de relación entre dos variables, que en el caso de tres o más se denomina multicolinealidad aunque a veces se emplea el término intercambiabilidad (Hair *et al*., 1995). Dos variables serán completamente colineales si su coeficiente de correlación r es igual a 1, mientras que un r igual a 0 indicaría una falta total de colinealidad. Como en la medida en que dos entidades sean muy colineales su aporte a la afinidad -medida como correlación- será muy alta, la multicolinealidad impone la unión conjunta de un gran número de entidades pues éstas son fuertemente ponderadas (Hair *et al.,* 1995) y tienden a dominar en el agrupamiento (Krzanowski y Marriott, 1966a) decidiendo de antemano la solución de la clasificación. Este requisito cobra mayor importancia cuando trabajamos con coeficientes de correlación como medidas de afinidad pues las restantes medidas no responden de igual forma a la colinealidad. Por otra parte, no creemos que esta propiedad sea una regla de los datos ecológicos como sí lo es de algunas variables morfométricas o estructurales que emplea la taxonomía numérica, donde de hecho la falta de multicolinealidad es una propiedad deseable en los caracteres a seleccionar (Sneath y Sokal, 1973).

La categorización de los datos cuantitativos también puede hacerse sobre la base de las relaciones entre los elementos que componen su escala de medición, subdividiéndolos en datos de *escala de intervalo* y de *escala de relación*. Las variables que emplean estos tipos de escalas se denominan *métricas* (Everitt y Dunn, 1991; Jobson, 1991). En el dato basado en escala de intervalo las diferencias entre puntos sucesivos de la escala son iguales, pero el punto cero es arbitrario (Everitt y Dunn, 1991) pues no tienen un valor de base natural (Sharma, 1996). Jobson (1991) cita como ejemplo la temperatura donde la medición procede a partir de un punto fijo y la relación entre dos medidas en °C no se mantiene si lo convertimos en °F. Con este tipo de dato pueden emplearse todos los métodos estadísticos salvo los basados en datos de escala de relación. La escala de relación constituye

la más versátil y tiene como propiedades que: a) dos valores a lo largo de la escala pueden ser expresados significativamente como una relación; b) la distancia entre puntos de la escala es significativa; c) los elementos a lo largo de la escala pueden ser ordenados de menor a mayor (Jobson, 1991). Estos datos constituyen el nivel más acabado de medición, permiten hacer comparaciones entre las diferencias y sus magnitudes relativas (Everitt y Dunn, 1991) y no hay restricciones a las pruebas estadísticas a emplear (Sharma, 1996).

**Reducción de los datos**

Los muestreos ecológicos pueden generar grandes matrices de datos por el número de estaciones, si se trata de muestreos muy extensivos en el espacio, pero el mayor volumen de información se debe fundamentalmente a las largas listas de especies que típicamente presentan varias dominantes y una larga «cola» de otras relativamente raras. El número de especies que se registra durante un muestreo depende del grupo que se trate. En los arrecifes coralinos de la plataforma SO de Cuba por ejemplo, se pueden encontrar en un muestreo 21 especies de corales, 34 de gorgonáceos, más de 70 de esponjas, entre 60 y 80 de moluscos y muchas más en poliquetos o peces, por solo mencionar algunos grupos (ver Alcolado, 1990). Quiere esto decir que siempre es conveniente analizar si la matriz original de datos que va a ser sometida a clasificación debe ser reducida, lo cual de modo general se realiza eliminando o fusionando la información de varias estaciones, o eliminando algunas especies. Para esto último, un primer criterio elemental sería la eliminación de aquellas cuya identificación taxonómica sea dudosa.

De las tres razones que comenta Boesch (1977) por las cuales se hace necesario reducir los datos: a) disminuir el número de cálculos y por tanto el costo; b) poder emplear ciertas estrategias clasificatorias y c) excluir datos de poca significación, esta última, por su connotación ecológica, debe ser siempre analizada aún cuando se cuenten con todas las facilidades computacionales. En el patrón de distribución típico de la composición cuantitativa y cualitativa de especies en una comunidad siempre existen varias especies raras en pequeño número, cuya probabilidad de aparición es baja. Por esta razón sus relaciones de coocurrencia pueden deberse más al azar que a requerimientos similares de hábitat por lo cual no brindan patrones precisos de distribución, al menos dentro de los marcos de un esfuerzo de muestreo razonable. Ello avala el criterio más usado para eliminar especies de una matriz: su baja frecuencia dentro de la muestra.

Otros criterios empleados han sido la eliminación de especies con poca contribución a la varianza cada cierto número de muestras (Bakus, 1990) o la exclusión de especies cuya abundancia o constancia caiga por debajo de determinado valor dentro de la información total de las localidades o intervalos de muestreo (Boesch, 1977). Aunque se han ensayado métodos más complicados, consideramos que los criterios de frecuencia, constancia y abundancia debidamente analizados en la matriz original de datos, pueden conducir al investigador que domina la ecología de su grupo a realizar las reducciones necesarias de especies.

No debe olvidarse que existen especies restringidas a determinados hábitats. Por ejemplo, de las varias zonas del arrecife de barrera el octocoralio *Gorgonia flabellum* se encuentra representada casi exclusivamente en la zona de embate por su alta resistencia al batimiento y puede estar ausente

o tener escasa abundancia en las restantes zonas. Considerar aisladamente su frecuencia o constancia respecto a todo el arrecife podría hacer pensar en su eliminación pero si consideramos su abundancia unido a los criterios sobre su zonación ecológica, podemos discernir que la misma es clave para la clasificación de la zona arrecifal pues tipifica su región más expuesta.

En relación con la reducción de la matriz eliminando localidades o intervalos de tiempo, Boesch (1977) señala como criterios más empleados la eliminación de muestras de dudosa calidad y la fusión de varias estaciones para lograr mayor homogeneidad. A esto último podríamos añadir que la unión de varias estaciones podría ser también una variante para incrementar el tamaño de muestra aunque tengamos que analizar la comunidad en un mayor contexto espacial. Por ejemplo, varias estaciones de una pradera marina de *Thalassia testudinum* pueden estar submuestradas individualmente pero en conjunto brindar un panorama claro de la composición del biotopo, siempre y cuando no existan diferencias importantes en sus características sedimentológicas e hidrológicas.

También pueden combinarse muestras de distintas épocas si se desea dilucidar los patrones espaciales sin contar con la información temporal, o por el contrario, combinar muestras de distintas localidades para analizar los patrones temporales. La reducción de datos no se practica, cuando tratamos con datos cualitativos donde todos tienen el mismo "valor" pero puede ser un paso necesario con los cuantitativos aunque advertimos del peligro que una reducción excesiva puede conducir a obtener clasificaciones artificialmente precisas.

**Transformación de los datos**

La transformación de los datos es un paso conocido de la estadística clásica cuando se requiere la normalización de los valores originales para poder aplicar determinadas pruebas paramétricas. Por definición, la *transformación* es una alteración del valor del atributo sin referencia al intervalo de valores de la población como un todo, o sea, que concierne a cada atributo individualmente y no a la totalidad de la muestra (Boesch, 1977). Al transformar se pretende ante todo reducir la importancia excesiva de las especies más abundantes o corregir la ausencia de éstas (Bakus, 1990). Una de las transformaciones más aplicadas es la logarítmica, bien sea en su expresión log X, o log (X+1) si existen ceros en la matriz de datos; otras también comunes son la de raíz cuadrada, cúbica y arcoseno.

Pero el dato original también puede ser modificado mediante *estandarización,* que se diferencia de la transformación propiamente dicha en que la alteración de los datos sí depende de alguna propiedad del ordenamiento de los valores como un todo. Son ejemplo de ella la conversión de los valores en porcentajes o proporciones, la expresión de los valores en unidades de desviación estándar y la estandarización doble. Al estandarizar se pretende reducir los valores a una escala común de comparación ante grupos de diferentes tamaños o grados (Bakus, 1990).

Boesch (1977) explica que se hace necesario transformar los datos cuando: a) existen grandes diferencias entre los valores, b) su distribución se aleja de la normalidad o c) el esfuerzo de muestreo no ha sido uniforme. Frecuentemente, en los datos cuantitativos unas pocas especies tienen valores

excesivamente altos y el resto muy bajos. Como algunos índices de afinidad son muy sesgados a los altos valores, de modo que las especies abundantes determinan el valor de afinidad, puede ser aconsejable una transformación (Fig. 3.4) que atenúe su contribución.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 123 | 982 | 478 | 50 | 27 |
| B | 67 | 10 | 13 | 1 | 0 |
| C | 15 | 35 | 234 | 5 | 0 |
| D | 2 | 1 | 3 | 1 | 0 |
| E | 1 | 0 | 1 | 79 | 8 |

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 4.9 | 9.9 | 7.8 | 3.7 | 3.0 |
| B | 4.1 | 2.1 | 2.3 | 1.0 | 0 |
| C | 2.5 | 3.3 | 6.2 | 1.7 | 0 |
| D | 1.2 | 1.0 | 1.4 | 1.0 | 0 |
| E | 1.0 | 0 | 1.0 | 4.3 | 2.0 |

Figura 3.4. Matriz original de datos cuantitativos y transformada mediante raíz cúbica.

Por otra parte algunos coeficientes derivados de la estadística, como la correlación, pueden emplearse como medidas de afinidad pero para ello se requiere que los datos se normalicen mediante alguna transformación. Finalmente, cuando las diferencias en el esfuerzo de muestreo entre dos localidades no permite la comparación directa de los datos, como ocurre por ejemplo con muestreos del bentos por recorrido (espacial o temporal) o con equipos de rastreo, es conveniente la estandarización respecto al total de estaciones, donde las abundancias de cada especie en una estación se suman y se dividen entre el total, con lo cual los datos quedan expresados en forma de proporciones, o porcentajes si se multiplican por 100 (Fig. 3.5).

De igual forma puede realizarse la estandarización por total de especies sumando las abundancias de cada una en todas las estaciones y dividiendo entre el total. Esta estandarización puede ponderar fuertemente las especies raras y subestimar las comunes por lo que se recomienda solo si la frecuencia de especies en la tabla no son muy diferentes (van Tongeren, 1987). Cuando estandarizamos por estaciones los porcentajes «valen» en el sentido de las proporciones de especies en la localidad (o sea por columnas), lo cual tiene por supuesto un significado ecológico.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 120 | 98 | 82 | 3 | 1 |
| B | 57 | 25 | 41 | 0 | 0 |
| C | 5 | 3 | 32 | 0 | 0 |
| D | 2 | 1 | 10 | 0 | 0 |
| E | 1 | 1 | 0 | 3 | 0 |

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 64.9 | 76.5 | 49.7 | 50.0 | 100 |
| B | 30.8 | 19.5 | 24.8 | 0 | 0 |
| C | 2.7 | 2.3 | 19.4 | 0 | 0 |
| D | 1.1 | 0.8 | 6.1 | 0 | 0 |
| E | 0.5 | 0.8 | 0 | 50.0 | 0 |

Figura 3.5. Matriz original de datos cuantitativos y matriz estandarizada por estaciones mediante porcentajes.

Los datos de rastreos o recorridos en distintas estaciones no deben ser estandarizados en porcentajes por especies (por filas) si los esfuerzos de muestreo son diferentes. Una estandarización porcentual por filas podría justificarse solo si los datos están referidos a la misma unidad. De cualquier manera, para estandarizar los datos mediante porcentajes o proporciones ambas localidades deben tener un

tamaño de muestra lo suficiente elevado que aproxime adecuadamente las proporciones calculadas. En el ejemplo de la Figura 3.5 el tamaño de muestra de las estaciones 4 y 5 no es adecuado para que sea válida esta estandarización.

Existen otros tipos señalados por Boesch (1977) como la estandarización doble simultánea que resuelve los problemas de escala en el análisis normal e inverso; y las estandarizaciones por máximo de estaciones y máximo de especies (van Tongeren, 1987), donde los valores individuales se dividen respectivamente entre el valor de la especie con abundancia máxima en una estación, o entre la abundancia máxima que alcanza una especie en todas las estaciones en que aparece.

**Tratamiento de datos mezclados**

Aunque a lo largo de este texto veremos matrices de estaciones/especies con datos de igual categoría, vamos a referirnos brevemente a cuando ocurren mezclas de varios tipos de datos en la matriz original para que el lector se sienta orientado para manejar su información adecuadamente. El tratamiento de datos mezclados puede variar según la disciplina que se trate pero Kaufman y Rousseeuw (1990) lo resumen en tres posibilidades.

Una primera y lógica aproximación es separarlos y realizar clasificaciones independientes para los tipos compatibles y relacionar después sus resultados. Si estos dan muy diferente puede ser más práctico la segunda: procesarlos juntos tratando todas las variables como cuantitativas, o por el contrario, y como tercera variante, reducirlo todo a datos cualitativos. Esto último tiene la ventaja de la sencillez pero la desventaja de sacrificar información potencialmente útil (Everitt, 1993).

Veamos esto con un ejemplo tomando una selección de datos de diferentes categorías de nuestro estudio de la langosta *Panulirus argus* y sus refugios naturales en los arrecifes del Suroccidental de Cuba (Tabla 3.7) cuyo objetivo era definir las regularidades de su conducta selectiva (Herrera *et al.,* 1991). Como datos cualitativos de estados excluyentes tenemos el sexo de la langosta refugiada: hembra o macho; y de presencia-ausencia la naturaleza del sustrato en la base del refugio, según se presente o no sedimento arenoso. El tipo de refugio: cuevas, huecos y solapas, representa al dato cualitativo de multiestado desordenado donde no existe una jerarquía y se convierten por tanto en datos de presencia-ausencia. El estadio de muda, proceso fisiológico que al transitar por una etapa

Tabla 3.7. Algunos parámetros cuantitativos y cualitativos de la langosta *Panulirus argus* y sus refugios en los arrecifes de la plataforma Suroccidental de Cuba.

| | Parámetros de la langosta | | | Parámetros del refugio | | | |
|---|---|---|---|---|---|---|---|
| Refugio | Sexo | Longitud total (cm) | Estadio de muda | Profundidad del refugio (cm) | Número de entradas | Sedimento en la base | Tipo de refugio |
| 1 | Macho | 30.2 | Intermuda | 57.6 | 1 | Si | Cueva |
| 2 | Hembra | 48.6 | Postmuda | 23.1 | 1 | No | Hueco |
| 3 | Macho | 21.8 | Postmuda | 11.5 | 1 | Si | Solapa |
| 4 | Hembra | 29.4 | Postmuda | 12.8 | 2 | Si | Hueco |
| 5 | Hembra | 45.2 | Premuda | 20.5 | 1 | No | Cueva |

de pre, inter y postmuda permite un orden jerárquico asignando números del 1 al 3; ejemplifica el dato de multiestado ordenado. Como datos cuantitativos continuos tenemos la longitud total de la langosta (LT) y la profundidad de su refugio (PR), parámetros cuya relación (PR/LT) se mantiene cercana a 2; y como ejemplo de datos cuantitativos discontinuos el número de entradas del refugio, generalmente 1 y ocasionalmente 2 aberturas.

En la Tabla 3.8 hemos tratado toda la información como datos cualitativos. Los datos cualitativos originales han sido codificados de acuerdo a su categoría pero los cuantitativos han sido convertidos en los primeros. Una variante de codificación podría ser la de convertir los datos cuantitativos continuos en datos de multiestado ordenado codificando las tallas en grandes (1), medianas (2) y pequeñas (3) según un intervalo previamente establecido y lo mismo para la profundidad del refugio. Sin embargo como en el proceso de conversión es importante mantener las relaciones como base de su posterior clasificación resulta más adecuado en este caso agruparlos en intervalos para convertirlos en datos de presencia-ausencia de modo que profundidades del refugio menores de 20 cm se corresponden en la tabla con longitudes totales menores de 30 cm, y viceversa.

Tabla 3.8 Conversión de la información de la Tabla 3.7 en datos cualitativos.

| | Sexo | | LT (cm) | | Estado de | Profundidad del refugio (cm) | | Número de entradas | | Sedimento en la base | | Tipo de refugio | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Refugio | Macho | Hembra | < 30 | > 30 | muda | < 20 | > 20 | 1 | 2 | Si | No | Cueva | Hueco | Solapa |
| 1 | 1 | 0 | 0 | 1 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 | 1 | 3 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 3 | 1 | 0 | 1 | 0 | 3 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 4 | 0 | 1 | 1 | 0 | 3 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 5 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |

En la Tabla 3.9 la información ha sido tratada cuantitativamente. Los datos cualitativos originales aportan el valor proveniente de su codificación bien sea 0 ó 1, ó entre 1 y 3; pues no tienen otro contenido de información. Los datos cuantitativos mantienen sus valores originales aunque alternativamente, para aliviar diferencias escalares pueden convertirse en datos semicuantitativos de rango, asignando números del 1 en adelante. De esta forma tenemos "números" en toda la tabla y la información se trata posteriormente como cualquier dato cuantitativo, incluido si necesitan alguna transformación (Crisci y López Armengol, 1983).

Tabla 3.9 Conversión de la información de la Tabla 3.7 en datos cuantitativos.

| | Sexo | | | Estado | Profundidad | Número | Sedimento | | Tipo de refugio | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Refugio | Macho | Hembra | LT (cm) | de muda | del refugio (cm) | de entradas | Si | No | Cueva | Hueco | Solapa |
| 1 | 1 | 0 | 30.2 | 2 | 57.6 | 1 | 1 | 0 | 1 | 0 | 0 |
| 2 | 0 | 1 | 48.6 | 3 | 23.1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 3 | 1 | 0 | 21.8 | 3 | 11.5 | 1 | 1 | 0 | 0 | 0 | 1 |
| 4 | 0 | 1 | 29.4 | 3 | 12.8 | 2 | 1 | 0 | 0 | 1 | 0 |
| 5 | 0 | 1 | 45.2 | 1 | 20.5 | 1 | 0 | 1 | 1 | 0 | 0 |

En este proceso de conversión los seis parámetros originales (Tabla 3.7) quedan convertidos en doce (Tabla 3.8) y once (Tabla 3.9) para los datos cualitativos y cuantitativos, respectivamente. En este incremento influyen sobre todo los datos cualitativos de multiestado ordenado que deben ser codificados individualmente como presencia-ausencia provocando una sobreestimación del parámetro "tipo de refugio".

Con este ejemplo sencillo solo pretendemos proveer al interesado con una pauta de trabajo, demostrar la libertad que se debe tener en el manejo de la información y la importancia de conocer los tipos de datos y sus relaciones. Ejemplos similares los brindan Crisci y Armengol (1983) para la taxonomía numérica.

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

*"Ni los hombres ni sus vidas se miden con el mismo rasero"*
*Miguel Montaigne*

# 4. MEDIDAS DE AFINIDAD

Por definición una *medida de afinidad* es una expresión matemática que permite resumir en un número el grado de relación entre dos entidades, sobre la base de la semejanza o la desigualdad entre la cualidad o la cantidad de sus atributos, o ambas. La selección de una medida de afinidad apropiada es tan importante que Legendre y Legendre (1979) señalan como corolario que la estructura derivada de la técnica de análisis será aquella de la matriz de afinidad y no necesariamente de toda la información de la matriz original de datos. Como criterios selectivos apuntan: la naturaleza del trabajo; las limitaciones matemáticas de cada medida; las propiedades de los métodos a los cuales será sometida la matriz de afinidad; y las disponibilidades de cálculo.

El criterio de qué se considera una entidad depende del tipo de clasificación. En la normal las entidades son las localidades o intervalos de tiempo y los atributos las especies que tipifican cada estación; mientras que en la inversa, las especies pasan a ser las entidades y los atributos las localidades o el tiempo. En cualquier caso la forma en que operan las medidas de afinidad puede ser igual, en primera instancia, aunque la escuela francesa (Legendre y Legendre, 1979) considera que el análisis inverso debe apoyarse preferentemente en medidas de correlación, concediendo a los datos una connotación más estadística. Boesch (1977) no hace esta distinción, pero sí aclara que al usar la correlación como afinidad deben aplicarse tests de significación sólo en el análisis inverso, y aún así, cuidando de que los datos cumplan los requisitos estadísticos necesarios.

Las medidas de afinidad se dividen en *cualitativas* y *cuantitativas* según empleen los tipos de datos que su denominación indique, aunque algunos datos cualitativos codificados se analizan a través de índices cuantitativos. También según su significación matemática pueden dividirse en índices de *distancia*, que varían entre 0 e $\alpha$; *correlación* (también llamados de dependencia), que varían entre -1 y 1; y *similitud* o *disimilitud* (llamados también de asociación) cuyo intervalo de variación está entre 0 y 1, ó entre 0 y 100, si los datos están expresados en porcentajes.

Disimilitud y distancia son análogas en el sentido de que dos entidades muy disímiles están muy distantes; en este caso un menor valor indica una mayor afinidad. Similitud, por su parte, coincide con la correlación en el sentido de que dos entidades muy similares están muy "correlacionadas", por tanto un mayor valor indica la mayor afinidad. En la literatura se manejan muchos nombres indistintamente y leemos términos como proximidad, parecido o semejanza. Incluso el vocablo similitud se emplea como denominación general de cualquier índice. Nosotros preferimos englobarlos como medidas de afinidad y emplear similitud con un sentido particular.

Para presentar ordenadamente los valores de afinidad calculados a partir de la matriz original de datos, que refleja la relación entre todas las entidades, se construye la matriz de afinidades, que es

siempre una matriz cuadrada, aunque solamente se representa una de sus mitades triangulares. Si partimos, por ejemplo, de una matriz original de datos de 20 estaciones y 50 especies (20 x 50), la matriz que resume la afinidad entre estaciones tendrá una dimensión de 20 x 20, las estaciones ocupando las filas y las columnas con lo cual se logra que cada una sea comparada consigo misma. Lo mismo puede decirse para las especies que en este ejemplo estaría representada por una matriz de 50 x 50. En todos los casos la diagonal de la matriz recoge la afinidad de una entidad con ella misma; tendrá siempre valores de 0 ó 1 ó 100, según la medida de afinidad que empleemos, y comúnmente se obvia en la representación gráfica.

La medida de afinidad es un elemento clave para simplificar las relaciones entre los datos primarios y facilitar su agrupamiento, pero en el caso particular de la clasificación de comunidades puede tener además una connotación ecológica especial relacionada con lo que se denomina la componente $\beta$ de la diversidad de especies (Southwood, 1994). Los ecólogos estructurales definen la diversidad de especies dentro de una comunidad o hábitat como diversidad $\alpha$ y la estiman cuantitativamente a través de índices ecológicos basados en el número de especies y/o la proporción de individuos entre ellas (por ejemplo el conocido índice de Shannon-Weaver). Cuando se trata de la diversidad entre hábitats o diversidad $\beta$, se puede alternativamente calcular y comparar los valores de algunos de estos índices ecológicos individualmente para cada hábitat o por otra parte emplear otro tipos de índices que comparen simultáneamente la composición de especies en ambos hábitats y resuma en un número el grado de relación, caso en el cual nos estaríamos refiriendo a una medida de afinidad.

La literatura recoge un sinnúmero de expresiones para medir la afinidad y el interesado puede acudir a los textos clásicos donde existen importantes revisiones (véanse, entre otros a Sokal y Sneath, 1963; Clifford y Stephenson, 1975; Boesch, 1977; Orlóci, 1978; Crisci y López Armengol, 1983). Las medidas que aquí presentaremos han sido seleccionadas bajo el criterio de: a) ofrecer una muestra de fórmulas representativas de uso común en ecología, b) reflejar con éstas, un espectro de propiedades matemáticas claves y c) enseñar aquellas que están avaladas por nuestra experiencia práctica.

**Medidas de afinidad cualitativas**

Cualitativamente las medidas de afinidad más empleadas son expresiones de similitud, cuya fórmula incluye algunos o todos de los cuatro elementos que hacen posible la comparación cualitativa de entidades, a saber: (a) los atributos que comparten (1,1); (b) los atributos presentes en la primera y no en la segunda (1,0); (c) los atributos presentes en la segunda y no en la primera (0,1); (d) los atributos ausentes en ambas (0,0). Convencionalmente estas cuatro letras (a, b, c y d) identifican dichos parámetros y suelen representarse en una tabla de dos entradas (Fig. 4.1), donde además se aclara cómo identificar las letras en el análisis normal y en el inverso.

Si hacemos un breve recorrido por algunos índices tropezamos con expresiones como las de Jaccard (1908): S= a/a+b+c, donde la fórmula se basa en los atributos compartidos y los que posee una entidad y otra, lo que la hace útil, según Boesch (1977), cuando hay muchos atributos positivos

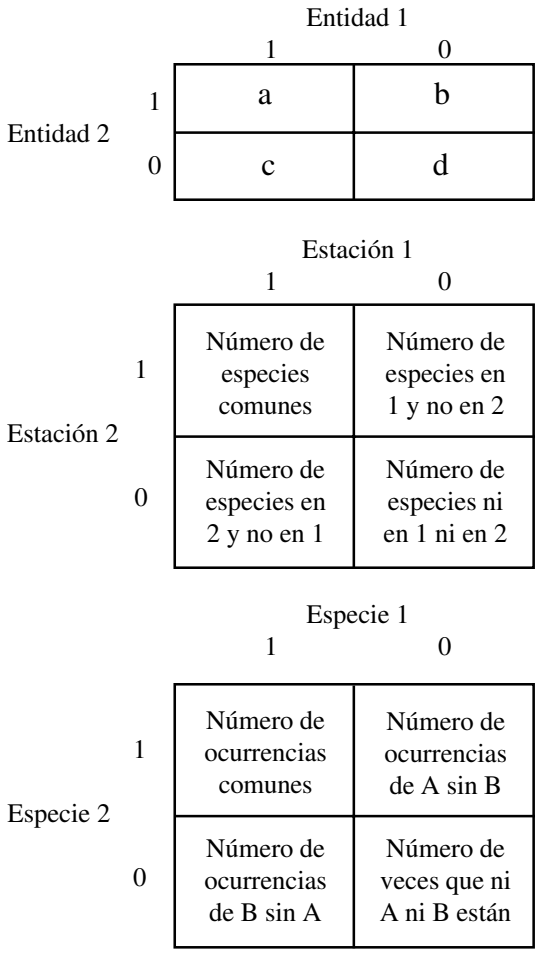



Figura 4.1. Tablas de contingencia de dos entradas mostrando los elementos a, b, c y d empleados en el cálculo de los índices cualitativos binarios (según Boesch, 1977).

compartidos. El índice de Sorensen (1948): S= 2a/2a+b+c; ya duplica la importancia de los atributos compartidos por lo que en condiciones de gran heterogeneidad de la matriz de datos cualitativos es útil para lograr comparaciones más efectivas entre colecciones ricas y pobres.

Por último, índices como el de apareamiento simple de Sokal y Michener (1958): S= a+d/n; o el de Sokal y Sneath (1963): S=2a+ 2d/2a+b+c+2d; incorporan los atributos negativos compartidos -que duplican su importancia en el segundo- por lo que se recomiendan cuando hay muchos ceros en la matriz debido a una alta fidelidad de conjuntos de atributos hacia determinadas entidades.

Para un mismo tipo de datos los diferentes índices cualitativos pueden dar valores muy distintos de acuerdo a las características de los datos y el tipo de coeficiente (Fig. 4.2). Según Everitt (1993) estas diferencias no serían de mayor importancia si los coeficientes fueran conjuntamente *monotónicos* en el sentido de que si los valores para diferentes pares de individuos calculados con un coeficiente se ordenaran en una serie monotónica -creciente o decreciente- los valores

**A.**

| Especies | Estaciones 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 |
| B | 0 | 0 | 0 | 1 |
| C | 0 | 0 | 0 | 1 |
| D | 0 | 0 | 0 | 1 |
| E | 1 | 1 | 0 | 1 |
| F | 1 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 |
| H | 0 | 1 | 1 | 1 |
| I | 1 | 1 | 0 | 0 |
| J | 0 | 0 | 0 | 0 |

**B.**

| Pares de estaciones | Medidas de afinidad Jaccard (1908) | Sorensen (1948) |
|---|---|---|
| 1-2 | 0.40 | 0.50 |
| 2-3 | 0.33 | 0.50 |
| 1-4 | 0.29 | 0.44 |
| 2-4 | 0.14 | 0.25 |
| 1-3 | 0.00 | 0.00 |
| 3-4 | 0.00 | 0.00 |

**C.**

| Pares de estaciones | Medidas de afinidad Sokal y Michener (1958) | Sokal y Sneath (1963) |
|---|---|---|
| 2-3 | 0.80 | 0.89 |
| 1-2 | 0.70 | 0.82 |
| 1-3 | 0.50 | 0.67 |
| 1-4 | 0.50 | 0.67 |
| 2-4 | 0.40 | 0.57 |
| 3-4 | 0.40 | 0.42 |

Figura 4.2. A: Matriz de datos cualitativos para 4 estaciones y 10 especies. B y C. Valores de similitud entre estaciones ordenados de forma decreciente, para dos índices cualitativos que no incluyen el elemento d en su fórmula y dos que sí la incluyen. En cada caso se indican los pares de estaciones que se comparan.

correspondientes para otro coeficiente estuvieran similarmente ordenados.Esto no ocurre necesariamente, particularmente entre índices que difieren en la inclusión de las ausencias conjuntas en su fórmula pero aquellos que son similares en los parámetros de su expresión matemática se ajustan más a un orden monotónico aunque con diferentes valores (Johnson y Wichern, 1992). La monotonicidad es importante porque algunos procedimientos para agrupar no se afectan si la medida de afinidad se cambia de forma tal que mantenga inalterable el orden relativo de los valores. Así, estrategias como el ligamiento simple y completo, que conoceremos en detalle más adelante, darían idénticos agrupamientos con los índices de Jaccard y Sorensen; o por otra parte con los de Sokal y Michener o Sokal y Sneath (Johnson y Wichern, 1992).

Aquellos índices que incluyen los ceros compartidos en su fórmula (entiéndase d) se denominan *invariantes* pues sus resultados no cambian cuando algunos o todos los atributos binarios se codifican diferente. Se recomienda su empleo con datos simétricos (estados excluyentes) donde tiene sentido ponderar los atributos negativos. Por otra parte los índices que no incluyen a d se refieren como *variantes* y se recomiendan para datos asimétricos (presencia-ausencia), donde como vimos el mayor peso estaba en los atributos positivos. Bajo este criterio, dado que los datos ecológicos de presencia-ausencia que son los más comunes, son asimétricos, ello llevaría a la conclusión de una alta similitud de haber muchos ceros en la matriz por lo que de preferencia deben emplearse índices variantes.

Kaufman y Rousseeuw (1990) en sus comentarios acerca del debate filosófico sobre la importancia de considerar o no las ausencias conjuntas, dicen que si bien desde el punto de vista matemático los coeficientes invariantes son más elegantes no puede existir un único coeficiente mejor porque se haga una distinción entre datos simétricos y asimétricos aunque la lógica apoya el empleo selectivo de los índices. Sokal y Sneath (1963) que hacen una extensa discusión sobre estos índices argumentan que no puede hacerse una regla rígida y rápida en relación con los valores negativos. Cada conjunto de datos debe ser considerado en sus méritos propios por el investigador familiarizado con su material (Everitt, 1993).

Nuevamente la solución a esta discusión debe darla el razonamiento ecológico. La no ocurrencia de una especie -asumiendo que no sea un problema de submuestreo- puede deberse a que se trate de una especie rara o con un tipo de dispersión espacial que no facilita su aparición dentro de los límites de un esfuerzo de muestreo razonable. Como no contribuyen esencialmente al patrón estructural de la comunidad tiene poco sentido considerarlas y muchas de ellas son de hecho eliminadas en el proceso de reducción de los datos. Precisamente este análisis de la información puede ayudar a darle cierta "simetría" al dato de presencia-ausencia al hacer que los ceros adquieran un valor. Además en algunos sistemas ecológicos como el litoral rocoso (Tabla 3.3) y es también el caso de los ambientes contaminados (Herrera, 1984) el patrón de zonación lo definen tanto las especies dominantes como la total ausencia de otras.

Existen otras fórmulas basadas solamente en dos elementos como los índices de Braun Blanquet y Simpson (Boesch, 1977), que dividen el número de ocurrencias conjuntas entre el número de especies de la lista más larga y más corta, respectivamente, y son útiles en zoogeografía para atenuar el efecto de diferencias en las listas de especies. Aquí presentaremos solo el Índice de Sorensen (1948) cuyo empleo nos ha demostrado su utilidad práctica y su adecuación a las más variadas tareas del quehacer ecológico.

*Índice de similitud de Sorensen.*- Como es común a las expresiones de similitud, este índice varia entre 0, entidades sin ningun atributo en común y 1, entidades idénticas. Se define por la expresión:

$$S = 2a/\ 2a + b + c$$

donde -en el análisis normal- a es el número de especies comunes y; b y c son el número de especies no compartidas en cada una de las estaciones o tiempos comparados. En el análisis inverso, a es el número de coocurrencias de las dos especies; y b y c son el número de apariciones no compartidas de cada una de las especies comparadas.

Tomando como punto de partida una matriz de datos cualitativos calcularíamos, comenzando por el análisis normal (Fig. 4.3) la similitud entre las estaciones 1 y 2. Sustituyendo los valores en la fórmula de Sorensen tenemos que la similitud entre las estaciones 1 y 2 es 0.75; valor que se colocará en la matriz de similitud en el lugar donde coinciden las dos entidades que se comparan.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 1 | 1 | 1 | 0 | 1 |
| B | 1 | 0 | 1 | 0 | 1 |
| C | 1 | 1 | 0 | 0 | 1 |
| D | 1 | 0 | 0 | 0 | 1 |
| E | 1 | 1 | 0 | 1 | 1 |

| Estaciones | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | |
| 1,00 | 0.75 | ? | ? | ? | 1 |
| | 1.00 | ? | ? | ? | 2 |
| | | 1.00 | ? | ? | 3 |
| | | | 1.00 | ? | 4 |
| | | | | 1.00 | 5 |

Figura 4.3. Cálculo de la similitud entre las estaciones 1 y 2. Datos: Las estaciones 1 y 2 comparten las especies A, C y E (a=3); la estación 1 tiene dos especies no compartidas (b=2); La estación 2 no tiene especies no compartidas (c=0). Por tanto: S= 2(3)/2(3)+2+0; S= 0.75.

En el análisis inverso (Fig. 4.4) calcularíamos la similitud entre las especies A y B. La similitud cualitativa entre las especies A y B es 0.86, valor que se ubicará en la matriz de similitud en el lugar donde coinciden ambas entidades.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 1 | 1 | 1 | 0 | 1 |
| B | 1 | 0 | 1 | 0 | 1 |
| C | 1 | 1 | 0 | 0 | 1 |
| D | 1 | 0 | 0 | 0 | 1 |
| E | 1 | 1 | 0 | 1 | 1 |

| Especies | | | | | |
|---|---|---|---|---|---|
| A | B | C | D | E | |
| 1.00 | 0.86 | ? | ? | ? | A |
| | 1.00 | ? | ? | ? | B |
| | | 1.00 | ? | ? | C |
| | | | 1.00 | ? | D |
| | | | | 1.00 | E |

Figura 4.4. Cálculo de la similitud entre las especies A y B. Datos: Las especies A y B concurren en las estaciones 1, 3 y 5 (a=3); la especie B nunca aparece sola (c=0); la especie A aparece sola una vez (b=1). Por tanto: S= 2(3)/2(3)+1+0; S=0.86.

Realizando los cálculos de la similitud de la estación 1 con la 2, 3, 4 y 5; de la 2 con 3, 4 y 5; de la 3 con 4 y 5; y de la 4 con la 5, se completa la matriz de similitud normal. La inversa será el resultado de calcular la similitud de A con B, C, D y E; de B con C, D y E; de C con D y E; y de D con E (Fig. 4.5). Llamamos la atención sobre algo que es común en este tipo de matrices: la repetición de valores de similitud, lo cual puede tener su influencia a la hora de los agrupamientos en el encadenamiento a un mismo nivel de varias entidades.

| 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|
| 1.00 | 0.75 | 0.57 | 0.33 | 1.00 | 1 |
| | 1.00 | 0.40 | 0.50 | 0.75 | 2 |
| | | 1.00 | 0.00 | 0.57 | 3 |
| | | | 1.00 | 0.33 | 4 |
| | | | | 1.00 | 5 |

| A | B | C | D | E | |
|---|---|---|---|---|---|
| 1.00 | 0.86 | 0.86 | 0.67 | 0.75 | A |
| | 1.00 | 0.67 | 0.80 | 0.57 | B |
| | | 1.00 | 0.80 | 0.86 | C |
| | | | 1.00 | 0.67 | D |
| | | | | 1.00 | E |

Figura 4.5. Matrices de similitud normal e inversa.

Este análisis es útil siempre y cuando existan en la matriz original de datos contrastes cualitativos notables entre entidades (como por ejemplo los que se observan en las Tablas 3.3 y 3.4), pues de otra forma, solamente estaríamos registrando una alta afinidad global que no aporta nada al proceso de clasificación. Si las especies analizadas tienen una distribución tan ubicua, que están en casi todas las estaciones, obviamente el dato cuantitativo pasa a ser esencial (Pielou, 1977).

**Medidas de afinidad cuantitativas**

Cuantitativamente las medidas de afinidad más empleadas son las distancias aunque algunos índices de disimilitud han ganado popularidad y la correlación se ha dejado para aplicaciones particulares. De modo general el cálculo se realiza comparando las magnitudes de cada atributo en las dos entidades involucradas considerando su orden en una tabla de dos entradas (Fig. 4.6).

| | Entidades | | | | | | |
|---|---|---|---|---|---|---|---|
| Atributos | 1 | 2 | 3 | . | . | . | k |
| 1 | $X_{11}$ | $X_{12}$ | $X_{13}$ | . | . | . | $X_{1K}$ |
| 2 | $X_{21}$ | $X_{22}$ | $X_{23}$ | . | . | . | $X_{2K}$ |
| 3 | $X_{31}$ | $X_{32}$ | $X_{33}$ | . | . | . | $X_{3K}$ |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| p | $X_{p1}$ | $X_{p2}$ | $X_{p3}$ | . | . | . | $X_{pk}$ |

Figura 4. 6. Tabla de dos entradas donde los elementos de las columnas varían desde j=1 hasta k y los de las filas desde i=1 hasta p.

Tal y como vimos con las medidas de afinidad cualitativas, algunas estrategias clasificatorias no cambian la estructura de grupos si se aplican índices de igual naturaleza y por tanto con propiedades de monotonicidad entre ellos, aunque para otras estrategias esto no se cumple. De las múltiples medidas de afinidad cuantitativas reportadas en la literatura solamente examinaremos cinco de ellas, con las cuales el interesado contará con una gama útil de expresiones con diferentes propiedades matemáticas entre las cuales podrá incluir otras que puedan ser de su interés.

**Medidas de distancia**

*Distancia euclidiana.*- Uno de los conceptos mas intuitivos de relación entre dos elementos es su distancia, que da una medida de su cercanía o alejamiento. De ahí que la distancia euclidiana sea, en esencia, una suma de las diferencias entre los valores de los atributos de cada entidad comparada, y no es más que una extensión simple en un espacio de varias dimensiones del conocido Teorema de Pitágoras (Pielou, 1984). Definida en su expresión más empleada por:

$$D = [\sum (X_{ij} - X_{ik})^2]^{1/2}$$

donde $X_{ij}$ y $X_{ik}$ identifican a los valores de los atributos de la especie i en las estaciones j y k que se comparan.

La forma de cálculo con la distancia euclidiana es operativamente similar a como ya vimos con el índice de Sorensen en el sentido de que cada estación y cada especie deben ser comparadas una a una, en el análisis normal e inverso, respectivamente. Partiendo de una matriz sencilla de datos cuantitativos, en el análisis normal comenzaríamos calculando la distancia entre las estaciones 1 y 2, como se indica en la Fig. 4.7. El valor obtenido se deposita en la matriz de distancias en el lugar donde coinciden ambas estaciones.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 20 | 10 | 14 | 50 | 72 |
| B | 7 | 13 | 2 | 32 | 10 |
| C | 18 | 7 | 23 | 10 | 2 |
| D | 0 | 9 | 0 | 5 | 0 |
| E | 2 | 1 | 1 | 1 | 13 |

| Estaciones | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | |
| 0 | 18,4 | ? | ? | ? | 1 |
| | 0 | ? | ? | ? | 2 |
| | | 0 | ? | ? | 3 |
| | | | 0 | ? | 4 |
| | | | | 0 | 5 |

Figura 4.7. Cálculo de la distancia euclidiana entre las estaciones 1 y 2. Se calculan las diferencias cuadráticas entre los valores de las estaciones 1 y 2 para cada especie: $(20 - 10)^2 = 100$; $(7 - 13)^2 = 36$; $(18 - 7)^2 = 121$; $(0 - 9)^2 = 81$; $(2 - 1)^2 = 1$. La suma de estos calculos es 339 cuya raíz cuadrada dará el valor de distancia, en este caso D= 18.4.

En el análisis inverso el calculo de la distancia se realiza ahora entre las especies A y B (Fig. 4.8) para las cuales se obtiene una distancia de 67,0 valor que se ubicará en la matriz inversa de distancias donde coinciden ambas entidades.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 20 | 10 | 14 | 50 | 72 |
| B | 7 | 13 | 2 | 32 | 10 |
| C | 18 | 7 | 23 | 10 | 2 |
| D | 0 | 9 | 0 | 5 | 0 |
| E | 2 | 1 | 1 | 1 | 13 |

| Especies | | | | | |
|---|---|---|---|---|---|
| A | B | C | D | E | |
| 0 | 67.0 | ? | ? | ? | A |
| | 0 | ? | ? | ? | B |
| | | 0 | ? | ? | C |
| | | | 0 | ? | D |
| | | | | 0 | E |

Figura 4.8. Cálculo de la distancia euclidiana entre las especies A y B. Se calculan las diferencias cuadráticas para cada estación: $(20 - 7)^2 = 169$; $(10 - 13)^2 = 9$; $(14 - 2)^2 = 144$; $(50 - 32)^2 = 324$; $(72 - 10)^2 = 3844$. La suma de estos calculos es 4490 cuya raíz cuadrada dará el valor de distancia, en este caso D= 67.0.

Calculando ordenadamente las distancias entre estaciones y especies, tal y como explicamos para el índice de Sorensen, obtendríamos finalmente las matrices de distancia normal e inversa que se muestran en la Fig. 4.9.

Estaciones

| 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|
| 0 | 18.4 | 9.3 | 40.2 | 55.6 | 1 |
| | 0 | 21.8 | 44.6 | 64.1 | 2 |
| | | 0 | 48.9 | 63.3 | 3 |
| | | | 0 | 34.6 | 4 |
| | | | | 0 | 5 |

Especies

| A | B | C | D | E | |
|---|---|---|---|---|---|
| 0 | 67.0 | 81.2 | 88.3 | 80.3 | A |
| | 0 | 33.8 | 29.9 | 33.8 | B |
| | | 0 | 29.8 | 31.3 | C |
| | | | 0 | 15.9 | D |
| | | | | 0 | E |

Figura 4.9. Matrices de distancias normal e inversa.

Como puede verse la fórmula de la distancia eleva al cuadrado las diferencias entre los atributos al comparar las entidades. Esto da lugar a que los atributos con altos valores sean exageradamente ponderados y se agudizan los problemas de escala entre los valores altos y bajos. En términos ecológicos esto implica que la distancia euclidiana sobreenfatiza la dominancia de los valores de las especies cuya abundancia es alta y puede dar lugar a una alta afinidad artificial entre entidades que no tienen en común muchos atributos (Boesch, 1977).

Expliquemos esto con un ejemplo. Supongamos que dos estaciones j y k, van a ser comparadas en su contenido cuantitativo de especies a través de la distancia euclidiana y descompongamos el aporte de las diferencias entre cada atributo, calculando su contribución porcentual a la suma cuya raíz cuadrada dará el valor final de distancia (Fig. 4.10). Como se observa, la diferencia entre los atributos correspondientes a la especie A (siendo uno de ellos alto) al ser elevada al cuadrado, aporta el 94.5% del valor de la sumatoria por lo que el cálculo de la distancia, en este caso, está basado prácticamente en una sola especie pues el resto contribuye aproximadamente solo en un 5.5 %. Por eso se dice que la distancia euclidiana sobreestima la influencia de los altos valores los cuales pueden llegar a dominar en el cálculo.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| | j | k | | | |
| Especies | $X_{ij}$ | $X_{ik}$ | $X_{ij}-X_{ik}$ | $(X_{ij}-X_{ik})^2$ | % |
| A | 80 | 20 | 60 | 3600 | 94.5 |
| B | 26 | 12 | 14 | 196 | 5.1 |
| C | 2 | 5 | -3 | 9 | 0.2 |
| D | 1 | 3 | -2 | 4 | 0.1 |
| E | 1 | 0 | 1 | 1 | 0.03 |
| | | | | 3810 | 100.0 |

Figura 4.10. Demostración del carácter sesgado hacia los altos valores de la distancia euclidiana.

Ante esta situación podríamos pensar no obstante, que el alto valor obtenido de distancia (DE = 61.7) es lógico ya que aún cuando las estaciones j y k no son muy distintas en los valores de las especies B, C, D, E y F, si lo son en la abundancia de la especie A. Pero esto no es todo. Comparemos ahora las dos estaciones mencionadas: j y k, con una tercera h (Fig. 4.11), bien diferente de las dos anteriores.

| | Estaciones | | |
|---|---|---|---|
| Especies | j | k | h |
| A | 80 | 20 | 0 |
| B | 26 | 12 | 1 |
| C | 2 | 5 | 10 |
| D | 1 | 3 | 30 |
| E | 1 | 0 | 21 |

| Estaciones | | | |
|---|---|---|---|
| j | h | k | |
| 0 | 61,7 | 91,2 | j |
| | 0 | 41.4 | h |
| | | 0 | k |

Figura 4.11. Matriz de datos originales y de distancias en el cálculo de la distancia entre las entidades j, k y h.

La matriz normal de distancias obtenida para estas tres estaciones muestra un menor valor entre k y h (DE = 41,4) indicando, a los efectos de nuestro análisis, que son mas afines entre sí que lo que lo son las entidades k y h con respecto a la j.

Sin embargo una ojeada a la matriz original de datos indica que la estructura de los datos de las estaciones h y k es bien diferente, incluso en su dominancia de especies, pero, comparativamente a la hora de decidir una agrupación a partir de esta matriz de distancias, las estaciones h y k, constituirían el primer grupo. Por esta razón es que se dice que la distancia euclidiana puede dar lugar a una alta afinidad artificial, en lo cual influye el hecho de que su intervalo de variación esté entre 0 e $\alpha$, lo cual solo permite establecer que dos entidades, están más cercanas entre sí que una tercera, pero no permite definir en que medida esta cercanía se corresponde con una alta afinidad, como sí ocurre con los índices que varían entre 0 y 1.

Por otra parte según Frontier (1969) en la medida en que la abundancia crece, comienza a variar regularmente entre intervalos mucho más amplios, recordemos su escala de abundancias cuando nos referimos a los datos de multiestado ordenado, codificados en rangos. Ello hace que la probabilidad de que exista una diferencia elevada entre dos valores altos sea mayor que entre dos valores pequeños de abundancia, lo que implica que la distancia euclidiana, bajo determinadas condiciones de los datos, puede conceder un papel determinante solo a las especies dominantes.

Veamos esto con un ejemplo (Fig. 4.12), empleando los valores extremos de la propia escala de Frontier (1969) (Tabla 3. 5), como atributos de una tabla hipotética, analizando la contribución porcentual de cada diferencia al valor final. Nótese como el aporte porcentual aumenta rápidamente en cada intervalo creciente de abundancia. Por esta razón el valor de la distancia euclidiana se ve afectado no solo por problemas de escala debido a atributos con altos y bajos valores, como vimos anteriormente, sino que también los altos valores conjuntos de las especies más abundantes, cuya diferencia debe ser regularmente mayor que las menos abundantes, inciden de manera determinante en el valor de la distancia. Quiere esto decir que una clasificación afectada por tales circunstancias brindaría agrupaciones que serían un reflejo predominantemente de las especies más abundantes.

| | Mínimo | Máximo | | | |
|---|---|---|---|---|---|
| Clases | $X_{ij}$ | $X_{ik}$ | $X_{ij}-X_{ik}$ | $(X_{ij}-X_{ik})^2$ | % |
| 5 | 350 | 1500 | -1200 | 1440000 | 94.9 |
| 4 | 80 | 350 | -270 | 72900 | 4.8 |
| 3 | 18 | 80 | -62 | 3844 | 0.3 |
| 2 | 4 | 18 | -14 | 196 | 0.01 |
| 1 | 1 | 3 | -2 | 4 | 0.00 |
| | | | | 1516944 | 100.0 |

Figura 4. 12. Cálculo de la distancia euclidiana con los valores extremos de la escala de Frontier (1969).

Aunque esta propiedad puede ser ventajosa cuando se trata de datos donde las diferencias en la estructura de las comunidades recaen fundamentalmente en varias especies dominantes, la distancia euclidiana puede resultar exagerada en este sentido. Por ello, aunque todas las medidas de distancia manifiestan en mayor o menor grado esta propiedad su efecto puede atenuarse modificando la fórmula original.

Consideremos, por ejemplo, que dividimos el valor de la distancia entre el número de pares comparados de modo que $D_p = 1/p[\Sigma(X_{ij}-X_{ik})^2]^{1/2}$ lo cual sería una distancia promedio donde la inclusión del factor p ayuda a que la diferencia no aumente indefinidamente en la medida que se incluyen nuevas variables. La llamada Distancia de Manhattan que se representa por $D_m = 1/p\ \Sigma|X_{ij}-X_{ik}|$ también emplea el factor de ponderación aunque elimina a todos los exponentes. Como vimos en el capítulo anterior una transformación o estandarización de los datos también podría contribuir a aliviar los problemas de escala, o si se quiere (y es más recomendable) se puede pensar en la búsqueda de índices menos sesgados como los que a continuación ofreceremos.

**Medidas de similitud (o disimilitud)**

*Índice de Bray-Curtis.-* El índice de Bray y Curtis (1957), es uno de los más ampliamente utilizados en la ecología cuantitativa actual y sus expresiones de similitud y disimilitud son:

$$S_{jk} = 2\Sigma \min(X_{ij}, X_{ik}) / \Sigma (X_{ij} + X_{ik})$$

$$D_{jk} = \Sigma |X_{ij} - X_{ik}| / \Sigma (X_{ij} + X_{ik})$$

Veamos un ejemplo sencillo de cálculo (Fig. 4.13) empleando la expresión de disimilitud que es la más usada. Partimos nuevamente de una matriz original de datos cuantitativos y calculamos la disimilitud entre las estaciones 1 y 2 para iniciar el análisis normal.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 20 | 10 | 14 | 50 | 72 |
| B | 7 | 13 | 2 | 32 | 10 |
| C | 18 | 7 | 23 | 10 | 2 |
| D | 0 | 9 | 0 | 5 | 0 |
| E | 2 | 1 | 1 | 1 | 13 |

| Estaciones | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | |
| 0 | 0.42 | ? | ? | ? | 1 |
| | 0 | ? | ? | ? | 2 |
| | | 0 | ? | ? | 3 |
| | | | 0 | ? | 4 |
| | | | | 0 | 5 |

Figura 4.13. Cálculo de la disimilitud de Bray-Curtis entre las estaciones 1 y 2. Para el numerador del índice se calcula el módulo de las diferencias para cada especie y se suman, o sea: $|20-10| + |7-13| + |18-7| + |0-9| + |2-1|$, donde $\Sigma|X_{ij}-X_{ik}| = 37$. Para el denominador del índice se suman los valores para cada especie, o sea: $(20+10) + (7+13) + (18+7) + (0+9) + (2+1)$, donde $\Sigma (X_{ij} + X_{ik}) = 87$. La relación 37/87 dará el valor de la disimilitud de Bray-Curtis, en este caso $D_{BC} = 0.42$.

Este índice concede aún un importante peso a los altos valores ya que en su expresión el numerador incluye la diferencia entre los atributos. Sin embargo, dado que la sumatoria de las diferencias no se eleva al cuadrado y posteriormente se divide entre la sumatoria de las sumas individuales, el índice de Bray-Curtis es una opción menos sesgada que la distancia euclidiana. Esto puede verse en el mismo conjunto de datos donde examinamos las propiedades de la distancia euclidiana (Fig. 4.10), comparando en este caso los valores que brinda el módulo de las diferencias en el numerador de la expresión de Bray-Curtis (Fig. 4.14). Nótese como el aporte porcentual de la diferencia entre los valores de la especie A disminuye a un 75 % en comparación con lo que significaba en la distancia, aunque sigue teniendo un peso importante.

| | Estaciones | | | |
|---|---|---|---|---|
| | j | k | | |
| Especies | $X_{ij}$ | $X_{ik}$ | $\lvert X_{ij}-X_{ik} \rvert$ | % |
| A | 80 | 20 | 60 | 75.0 |
| B | 26 | 12 | 14 | 23.3 |
| C | 2 | 5 | 3 | 5.0 |
| D | 1 | 3 | 2 | 2.5 |
| E | 1 | 0 | 1 | 1.3 |
| | | | | 100.0 |

Figura 4.14. Demostración del carácter menos sesgado del Indice de Bray-Curtis.

Σ

*Índice de Sanders.*- Cuando los valores estandarizados en porcentajes o proporciones se sustituyen en la fórmula de similitud de Bray-Curtis, llegamos a la siguiente expresión:

$$S_{jk} = \min (P_{ij} , P_{ik})$$

donde $P_{ij}$ y $P_{ik}$ son respectivamente los valores de los atributos (en porcentajes o proporciones) de las entidades que se comparan.

Esta expresión corresponde al índice de similitud porcentual de Sanders (1960) de gran popularidad en ecología marina, cuya forma de cálculo ejemplificaremos con dos estaciones (Fig. 4.15) donde los datos originales han sido estandarizados en forma de porcentajes. Como puede verse se trata solo de escoger el mínimo de los dos valores correspondientes al atributo que se compara y sumarlos. Para el cálculo de la similitud entre la estación 1 y 2, plantearíamos: $S_{jk} = 46.9 + 24.8 + 2.7 + 1.1 + 0$, o sea $S_{jk} = 75.5$. En su expresión de disimilitud, que es la más empleada, plantearíamos: $D_{jk} = 100 - S_{jk}$; $D_{jk} = 24.5$.

| | Estaciones | |
|---|---|---|
| Especies | 1 | 2 |
| A | 64.9 | 46.9 |
| B | 30.8 | 24.8 |
| C | 2.7 | 19.4 |
| D | 1.1 | 8.9 |
| E | 0,5 | 0.0 |

Figura 4.15. Cálculo del Índice de similitud de Sanders (1960) entre las estaciones 1 y 2.

Aparte de su utilidad general para la comparación de datos ecológicos, este índice resulta particularmente adecuado en los estudios de ambientes contaminados, por dos razones básicas (Herrera, 1984). En primer lugar, en los ambientes afectados por la contaminación, el efecto se traduce en una reducción de la densidad lo cual implica que al muestrear áreas contaminadas y limpias, con fines comparativos, los esfuerzos de muestreo sean necesariamente muy desiguales y los datos deban ser estandarizados, generalmente en forma de porcentajes. En segundo lugar, el desequilibrio que sufren las comunidades en ambientes contaminados hace que su estructura quede definida por un conjunto de especies dominantes, cuyo peso en la clasificación (debido al sesgo moderado que hereda este índice) ayuda a obtener subdivisiones claras de los distintos ambientes.

*Índice de Canberra.-* Como vimos al tratar la distancia euclidiana y el índice de Bray-Curtis, los atributos con altos valores tenían un gran peso en el valor de la afinidad, mientras que los de bajos valores prácticamente no tenían importancia. Para resolver esta desventaja es útil el índice de Canberra (Lance y Williams, 1966), definido en sus expresiones de similitud y disimilitud por:

$$S_{jk} = \Sigma \ (1/m)\ 2 \min (X_{ij}, X_{ik}) / (X_{ij} + X_{ik})$$

$$D_{jk} = (1/m) \ \Sigma \ [|X_{ij} - X_{ik}| / (X_{ij} + X_{ik})]$$

En este índice aparece un nuevo elemento en la fórmula: m, que no es más que el número de atributos considerados excluyendo las comparaciones de pares de ceros. El ejemplo de cálculo de este índice (Fig. 4.16) en su expresión de disimilitud -en el análisis normal- muestra que el índice de Canberra es el promedio de series de fracciones representativas del aporte entre entidades de cada atributo y lleva implícito, por tanto, una autoestandarización (Boesch, 1977). En este caso los altos valores contribuyen solo con una de las fracciones sumadas y no dominan en el coeficiente.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 |
| A | 20 | 10 | 14 | 50 | 72 |
| B | 7 | 13 | 2 | 32 | 10 |
| C | 18 | 7 | 23 | 10 | 2 |
| D | 1 | 9 | 0 | 5 | 0 |
| E | 2 | 1 | 1 | 1 | 13 |

| Estaciones | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | |
| 0 | 0.44 | ? | ? | ? | 1 |
| | 0 | ? | ? | ? | 2 |
| | | 0 | ? | ? | 3 |
| | | | 0 | ? | 4 |
| | | | | 0 | 5 |

Figura 4.16. Cálculo de la disimilitud de Canberra entre las estaciones 1 y 2. Para cada especie se calcula $|X_{ij}-X_{ik}| / (X_{ij}+X_{ik})$ y se suman, o sea: [|20-10| /(20+10)] + [|7-13| /(7+13)] + [|18-7| /(18+7)] + [|1-9| /(1+9)] + [|2-1| /(2+1)]. Esto es igual a [10/30] + [6/20] + [11/25] + [8/10] + [1/3], o sea, 0,33 + 0,30 + 0,44 + 0,80 + 0, 33. Esta suma da 2.20 que dividido entre el número de comparaciones sería 2,20/5 de donde $D_C$= 0.44.

Analicemos esto en el mismo conjunto de datos donde vimos las propiedades de la distancia euclidiana (Fig. 4.10) y el índice de Bray-Curtis (Fig. 4.14), calculando en este caso el aporte que realiza cada fracción a la sumatoria, que dividida entre m nos dará la disimilitud final.

Como se observa el peso exagerado de la diferencia entre los valores del atributo A desaparece y todas las fracciones contribuyen en la medida de sus diferencias (Fig. 4.17).

| | Estaciones | | | |
|---|---|---|---|---|
| | j | k | | |
| Especies | $X_{ij}$ | $X_{ik}$ | $\frac{\lvert X_{ij}-X_{ik}\rvert}{(X_{ij}+X_{ik})}$ | % |
| A | 80 | 20 | 0,60 | 23.3 |
| B | 26 | 12 | 0,37 | 14.4 |
| C | 2 | 5 | 0,43 | 16.7 |
| D | 1 | 3 | 0,50 | 19.5 |
| E | 1 | 0.2 | 0,67 | 26.1 |
| | | | | 100.0 |

Figura 4. 17. Demostración del carácter insesgado del Indice de Canberra hacia los altos valores.

En el ejemplo que acabamos de ver habrán notado que en el valor que corresponde a la especie E, en la estación k, habíamos puesto un cero en los cálculos del análisis normal de los dos índices anteriores (Figs. 4.10 y 4.14), mientras que aquí (Fig. 4.17) aparece el valor 0.2. La explicación es que cuando se emplea el índice de Canberra, la matriz original de datos debe ser transformada sustituyendo los ceros por un valor denominado "e", generalmente igual a 1/5 del menor valor de la matriz, diferente de cero. Esto se hace, pues cuando uno de los atributos es cero, la contribución de la fracción a la suma total es siempre 1. Al sustituir los ceros por un pequeño valor se garantiza una mayor contribución a la disimilitud cuando la diferencia entre los atributos es mayor, que cuando es más pequeña. Aclaremos esto con un ejemplo (Fig. 4.18) en una matriz hipotética con valores extremos. Si mantenemos los ceros en la matriz el aporte de todas las fracciones a la disimilitud es igual a 1, sin embargo, 1 y 0 son valores mucho menos disímiles entre sí que 100 y 0 ó 1000 y 0 por lo que el valor real de la disimilitud no está siendo reflejado.

| | Estaciones | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | | | |
| Especies | $X_{ij}$ | $X_{ik}$ | $\lvert X_{ij}-X_{ik}\rvert$ | $(X_{ij}+X_{ik})$ | $\frac{\lvert X_{ij}-X_{ik}\rvert}{(X_{ij}+X_{ik})}$ |
| A | 1000 | 0 | 1000 | 1000 | 1.00 |
| B | 100 | 0 | 100 | 100 | 1.00 |
| C | 10 | 0 | 10 | 10 | 1.00 |
| D | 1 | 0 | 1 | 1 | 1.00 |

Figura 4.18. Desglose por fracciones en el cálculo del índice de Canberra entre las estaciones 1 y 2 representadas por valores hipotéticos extremos.

Sustituyendo los ceros por un pequeño valor "e" (Fig. 4.19) en este caso igual a 0.2 se garantiza que en la medida que los valores van siendo más diferentes de cero su aporte a la disimilitud va siendo consecuentemente mayor.

| Especies | Estaciones 1 Xij | 2 Xik | \|Xij-Xik\| | (Xij+Xik) | \|Xij-Xik\| / (Xij+Xik) |
|---|---|---|---|---|---|
| A | 1000 | 0.2 | 999.8 | 1000.2 | 0.9996 |
| B | 100 | 0.2 | 99.8 | 100.2 | 0.9960 |
| C | 10 | 0.2 | 9.8 | 10.2 | 0.9608 |
| D | 1 | 0.2 | 0.8 | 1.2 | 0.6666 |

Figura 4.19. Papel del coeficiente "e" para reflejar las disimilitudes reales entre fracciones.

## Medidas de correlación

*Correlación lineal.*- Las medidas de correlación varían entre -1 y 1 y la más empleada como expresión de la afinidad es la correlación producto-momento proveniente de la estadística clásica:

$$Rjk = (Xij - Xj)(Xik - Xk) / (\sum (Xij - Xj)^2)^{1/2} (\sum (Xik - Xk)^2)^{1/2}$$

El empleo de la correlación como medida de afinidad ha sido muy criticada en la clasificación aunque los resultados son contradictorios (Everitt, 1993). Como desventajas se le señalan, que tiende a exagerar la contribución de los altos valores y puede dar patrones espurios de afinidad; y que pueden ocurrir correlaciones perfectas entre entidades muy desiguales (Boesch, 1977) lo cual se debe a que la correlación no se basa tanto en la magnitud de los valores sino en sus patrones (Hair *et al.,* 1995).

Esta última propiedad es obvia si comparamos tres estaciones a través de un coeficiente de correlación y la distancia euclidiana (Fig. 4.20). Las entidades 1 y 3 cuyo patrón de variación de los valores es idéntico tienen una alta correlación (r = 0.99) aun cuando sus valores están alejados una mayor distancia (D = 150.9). Por su parte las estaciones 2 y 3 guardan una baja correlación entre sí (r = 0.11) aunque su valor de distancia (D = 37.7) indica una mayor cercanía entre ellas. El empleo de correlación o distancia puede brindar entonces resultados no solo diferentes sino contrarios.

**A.**

| Especies | Estaciones 1 | 2 | 3 |
|---|---|---|---|
| A | 100 | 31 | 25 |
| B | 81 | 29 | 14 |
| C | 60 | 33 | 4 |
| D | 79 | 31 | 15 |
| E | 100 | 34 | 26.2 |

**B.**

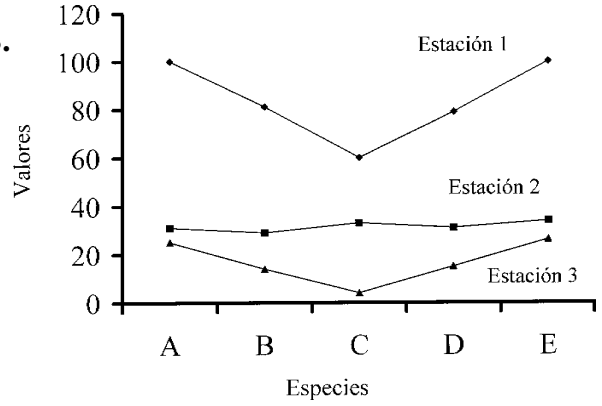



Figura 4. 20. **A.** Matriz de datos de 3 estaciones y 5 especies. **B**. Variación de los valores de especies en cada estación.

En favor de la correlación a veces se argumenta que su connotación estadística permite examinar la significación de la afinidad pero esto debe ser aplicado con precaución ya que los atributos de una matriz de estaciones y especies no constituyen siempre variables en un sentido estrictamente estadístico Algunos autores plantean que los coeficientes de distancia o similitud se emplean en el análisis normal mientras que los de correlación se emplean en el inverso (Johnson y Wichern, 1992) dándole así a los datos de la distribución de especies un sentido más estadístico que es la tónica de la escuela de Legendre y Legendre (1979).

Con datos binarios también es posible el empleo de la correlación producto momento cuya expresión en este caso sería:

$$r = ad\text{-}bc/[(a+b)(c+d)(a+c)(b+d)]^{1/2}$$

donde a, b, c y d se corresponden con las definiciones dadas en la Tabla de contingencia de la Fig. 4.2. Esta expresión se ha empleado en el análisis de especies donde además se le ha dado un sentido estadístico a las relaciones dado que el coeficiente de correlación está relacionado con el estadígrafo chi cuadrado ($r^2=X^2/n$) para examinar la independencia de dos variables (Johnson y Wichern, 1992).

*Correlación de Spearman.*- La correlación por rangos de Spearman, proveniente de la estadística no paramétrica (Siegel, 1985) puede ser una variante útil cuando se trabaja con datos de jerarquía. Se define por:

$$r = 1 - 6 \Sigma Di^2/ N(N^2 - 1)$$

donde Di son las diferencias entre los rangos de $X_{ij}$ y $X_{ik}$, y N es el número de pares de valores en los datos. Si deseamos comparar las estaciones 1 y 2, sustituyendo sus datos originales por rangos haríamos como en la Fig. 4.21. El empleo de datos de rango puede ser una alternativa para sacar provecho de datos cuantitativos, que aunque deficientes o incompletos, encierran un contenido mayor de información que la variante cualitativa.

| | Datos originales | | Datos de rangos | | D | $D^2$ |
|---|---|---|---|---|---|---|
| Especies | 1 | 2 | R1 | R2 | (R1- R2) | $(R1- R2)^2$ |
| A | 75 | 16 | 1 | 3 | -2 | 4 |
| B | 23 | 112 | 2 | 1 | 1 | 1 |
| C | 10 | 33 | 4 | 2 | 2 | 4 |
| D | 5 | 3 | 5 | 4,5 | 0,5 | 0.25 |
| E | 14 | 3 | 3 | 4,5 | -1,5 | 2.25 |
| F | 0 | 1 | 6 | 6 | 0 | 0 |
| | | | | | | 11.50 |

Figura 4.21. Cálculo de la correlación por rangos entre las estaciones 1 y 2, donde r = 6(11.5)/190 que da 0.36.

**Alternativas de empleo de la matriz de afinidad**

Una vez vistas las principales medidas de afinidad veamos dos formas en que la matriz de relaciones obtenida a partir de ellas puede ser empleada para la interpretación ecológica de los resultados, sin llegar necesariamente al paso más complejo, que ocupará nuestra próximo capítulo: la selección de uno o varios métodos agrupamiento.

*Diagrama de Trellis.*- En los inicios de la clasificación el diagrama de Trellis fue una forma de expresar las relaciones entre entidades en la propia matriz de afinidad. En esencia no es más que reordenar la matriz de afinidad haciendo coincidir los grupos de estaciones o especies más afines, de modo que empleando alguna simbología puedan expresarse de manera cuantitativa y gráfica las agrupaciones. Los pasos para confeccionar este tipo de diagrama se indican con un ejemplo sencillo en la Fig. 4.22.

Para lograr una representación adecuada pueden seguirse algunos criterios simples de análisis y ordenamiento de los datos, aunque Boesch (1977) comenta el empleo de métodos más complejos. Esta representación puede ser útil cuando se trata de matrices mas bien pequeñas donde la alternativa de un dendrograma pudiera parecer exagerada, pero para matrices grandes creemos más oportuno continuar la secuencia de pasos de la clasificación y elegir un método adecuado de agrupamiento, lo que puede permitir además hacer un análisis más refinado de los datos.

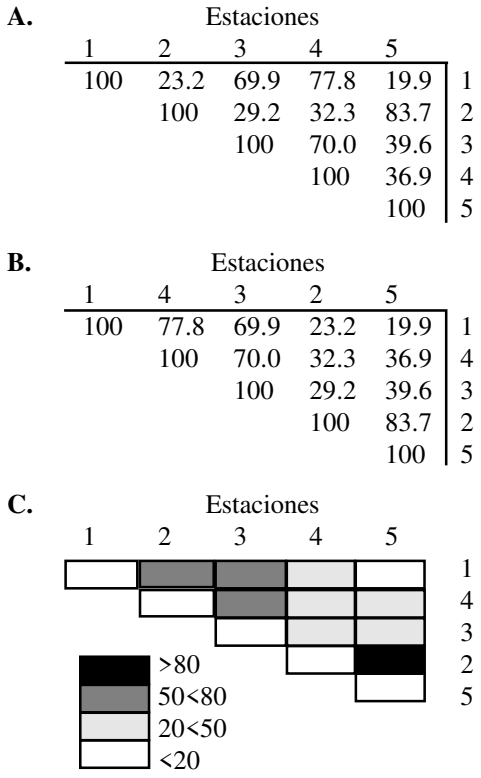



Figura 4. 22. **A**. Matriz de afinidad, **B**. Matriz reordenada, **C**. Diagrama de Trellis y escala de valores.

*Proyección de similaridad cenoclínica.*- Durante el proceso de agrupamiento la matriz de afinidad es totalmente transformada. Los valores originales, al ser comparados entidad-entidad, grupo-grupo, o grupo-entidad, son recalculados o seleccionados de acuerdo al algoritmo de clasificación empleado. Así, el resultado final viene a ser como un resumen de las afinidades globales entre los miembros de la matriz, representado generalmente en forma de árbol de clasificación. Sin embargo, la información original de la matriz de afinidad puede ser empleada para interpretar los resultados del estudio ecológico en alternativas como la *proyección de similaridad cenoclínica* (Boesch, 1977a), que como su nombre indica, no es más que proyectar o ubicar los valores de afinidad a lo largo de una cenoclina, entendiéndose como tal una región de gradiente o de variación de las condiciones ecológicas cuyos puntos de tránsito se conocen como ecotonos.

Para realizar este tipo de análisis se procede a estudiar la matriz de afinidad para evaluar si existe alguna combinación de entidades que revele un gradiente de cambio en los valores de afinidad en sentido espacial o temporal, que pueda ser reflejo de cambios ecológicos. Si las características de la matriz (en cuanto a dimensiones u ordenamiento de los valores) no permite ver claro las variaciones de la afinidad se procede a realizar tantas comparaciones como entidades existan, escogiendo en cada caso una de ellas, individualmente, para ser comparada con las restantes. En matrices pequeñas o cuando los valores tienen un ordenamiento natural puede ser fácil la selección de la entidad clave para la comparación sin necesidad de analizar todas las combinaciones, aunque esto último puede ser recomendable para elegir la opción más representativa.

En el ejemplo de la Figura 4.23, los valores de similitud porcentual muestran una tendencia de disminución al comparar la entidad 1 con las restantes (de izquierda a derecha) e inversamente, los valores de similitud porcentual muestran una tendencia de aumento al comparar la entidad 5 con las restantes (de abajo hacia arriba).

A partir de los valores de relación entre la entidad escogida se construye un gráfico de afinidad contra entidades como puede verse en la Figura 4.23. Al plotear las relaciones de la estación 1 y la 5, con las restantes, el valor correspondiente a la entidad seleccionada al ser comparada con ella misma, siempre tendrá el valor máximo o mínimo según sea similitud o correlación; o disimilitud o distancia, respectivamente. En el gráfico se observa un cambio en la similitud en el tránsito de las estaciones 3 a 4, que es claro tanto si se compara la 1 con las restantes, o la 5. Al comparar las variaciones de la afinidad en dos curvas, empleando un sentido del gradiente (por ejemplo la 1 con respecto a 2, 3, 4 y 5) y el opuesto (en el ejemplo la 5 con respecto a 4, 3, 2 y 1) se logra un punto de coincidencia que se corresponde con el cambio de la afinidad.

Estaciones

| 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|
| 100 | 77.8 | 69.9 | 23.2 | 20.9 | 1 |
| | 100 | 70.0 | 32.2 | 36.9 | 2 |
| | | 100 | 29.2 | 39.6 | 3 |
| | | | 100 | 83.7 | 4 |
| | | | | 100 | 5 |

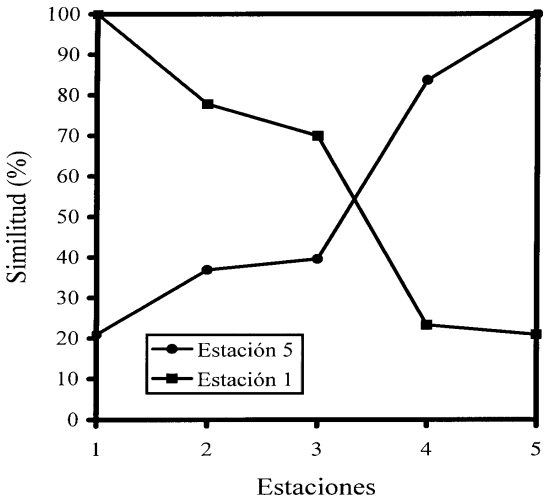



Figura 4.23. Izquierda. Matriz de similitud señalando las entidades que reflejan un gradiente de cambios en la afinidad. Derecha. Proyección de similaridad cenoclínica en la comparación de las estaciones 1 y 5 con las restantes. En la leyenda se indica la estación respecto a la cual se compara.

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

*"Cuanto simplifica, facilita"*
*José Martí*

# 5. MÉTODOS DE AGRUPAMIENTO

Los métodos de agrupamiento reciben diferentes denominaciones y subdivisiones según la forma en que operen y los resultados que brinden. De acuerdo a las definiciones de Boesch (1977) los que aquí presentamos son métodos *exclusivos:* pues una entidad aparece solo en un grupo; *intrínsecos:* pues la formación de los grupos se basa solamente en sus atributos; *jerárquicos:* pues en el proceso del método se optimiza una ruta entre las entidades individuales y todo el conjunto, por fusiones y fisiones progresivas de modo que los miembros de clases inferiores lo son también de las superiores; *aglomerativos* (o ascendentes): pues el agrupamiento procede por fusión progresiva comenzando por las entidades individuales y finalizando con la población completa; *politéticos:* como toda estrategia aglomerativa pues la medida de afinidad se aplica sobre todos los atributos considerando que un individuo se agrupa con aquel que más se le parece; y *combinatorios:* pues los valores de afinidad grupo-grupo o grupo-entidad pueden ser calculados sucesivamente de la matriz de afinidad entre entidades.

Lo anterior quiere decir que en el campo de la clasificación numérica existen también métodos de agrupamiento no exclusivos, extrínsecos, no jerárquicos, divisivos (disociativos o descendentes) monotéticos y no combinatorios, pero que no serán objeto de nuestro análisis. Al presente los métodos aglomerativos combinatorios se han revelado como los más útiles y fáciles de aplicar mientras que los divisivos, aunque atractivos teóricamente, están menos desarrollados y no han sido de amplia aplicación. Las ventajas y desventajas de cada uno de estos métodos son discutidas por Boesch (1977) y Pielou (1984), y aunque ésta última concede algunas ventajas a los métodos divisivos, la realidad, al menos por el momento, es que los métodos aglomerativos son los de más amplio uso en la ecología actual.

En su reciente revisión Krzanowski y Marriott (1996a) resumen en cuatro los múltiples métodos propuestos para análisis de grupos: partición óptima, mezclas de distribuciones fijas, métodos no paramétricos con estimación de la densidad local y métodos estrictamente jerárquicos; confirmando que estos últimos, que aparecieron en la literatura ecológica y taxonómica en la década del 50 continuan siendo el conjunto de técnicas de mayor aplicación.

**¿Cómo operan los métodos aglomerativos combinatorios?**

Tomemos como punto de partida una matriz de disimilitud o de distancias en la cual aparecen relacionadas tres entidades: h, i y j (Fig. 5.1). Si quisiéramos agruparlas según su grado de afinidad la primera unión sería sin dudas la de i y j, ya que poseen el menor valor de disimilitud o distancia; o sea son las menos disímiles o las menos distantes.

| h | i | j | |
|---|---|---|---|
| 0 | 0.90 | 0.76 | h |
| | 0 | 0.38 | i |
| | | 0 | j |

Figura 5.1. Matriz de disimilitud entre las entidades h, i y j.

Tenemos por tanto que i y j estarían unidas en un valor de 0.38 pero: ¿cuál es el valor de disimilitud o distancia que relaciona al grupo formado ij (que llamaremos k) con la entidad h? Siguiendo el ejemplo didáctico de Boesch (1977) esta situación podría ser representada geométricamente como se muestra en la Fig. 5.2. En correspondencia con los valores de la matriz de la Fig. 5.1, los segmentos indican diferentes distancias entre los puntos que representan las entidades. Las más cercanos son i y j ($D_{ij}$ = 0.38); le siguen h y j ($D_{hj}$ = 0.76) y las mas distantes son h e i ($D_{hi}$ = 0.90). Como las entidades i y j al fusionarse determinan el grupo k, cuya distancia o disimilitud al grupo h ($D_{hk}$ = ?) se desconoce; preguntamos entonces, ¿cúal es el valor de $D_{hk}$?

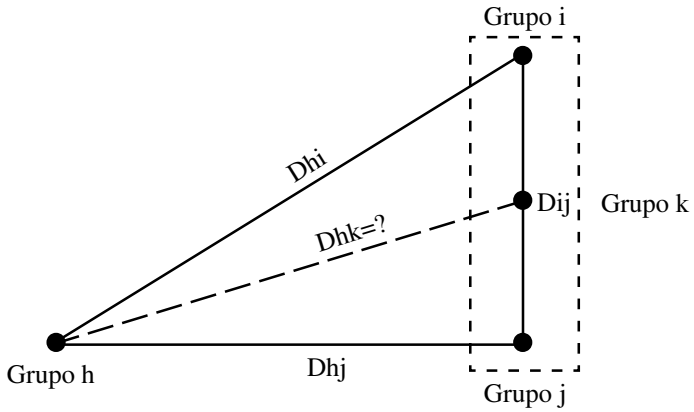



Figura 5. 2. Representación del cálculo de la distancia entre el Grupo h y uno nuevo k, según la ecuación combinatoria (tomado de Boesch, 1977).

En la matriz de afinidad reordenada de acuerdo a la agrupación formada y en un árbol de clasificación o dendrograma, representación final del proceso clasificatorio como veremos más adelante, esta interrogante se plantearía como en la Fig. 5.3.

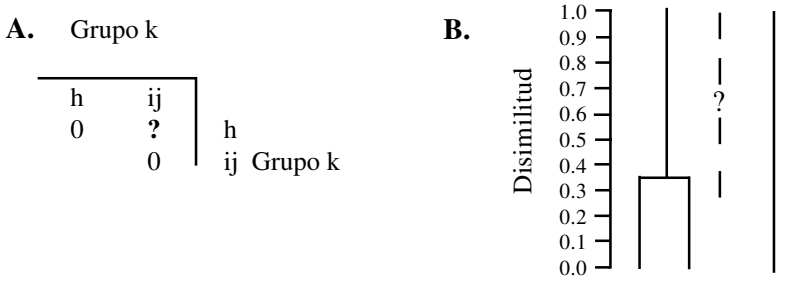



Figura 5.3. A. Valor de $D_{hk}$ a calcular en la matriz de afinidad. B. Interrogante gráfica: ¿a qué nivel de disimilitud se unirá el grupo k a la entidad h?

Lance y Williams (1966a) encontraron que para una variedad de estrategias combinatorias grupo-grupo o grupo-entidad la afinidad podía calcularse como variantes de una ecuación lineal simple de modo que:

$$D_{hk} = \alpha_i D_{hi} + \alpha_j D_{hj} + ß D_{ij} + \gamma | D_{hi} - D_{hj} |;$$

donde los parámetros $\alpha_i$, $\alpha_j$, ß y $\gamma$ determinan la naturaleza de la estrategia. Por eso, cuando leemos que una determinada clasificación se obtuvo por el método de promedio simple, ligamiento completo, u otro, lo único que se ha hecho es sustituir distintos valores de los coeficientes de la ecuación de Lance y Williams. Boesch (1977) cita ocho estrategias aglomerativas de las cuales extraemos cinco de las más empleadas para comentar sus propiedades y con las cuales realizaremos los cálculos de $D_{hk}$ del ejemplo de la Fig. 5.3, dando a los coeficientes los valores que le corresponden según se indica en la Tabla 5.1.

Tabla 5.1. Valores de los parámetros de la ecuación de Lance y Williams (1966a) para cinco técnicas de agrupamiento (tomado de Boesch, 1977).

| Método | αi | αj | γ | ß | Distorsión del espacio |
|---|---|---|---|---|---|
| Ligamiento simple | ½ | ½ | -½ | 0 | contractiva |
| Ligamiento completo | ½ | ½ | ½ | 0 | dilatadora |
| Promedio de grupos | ni/nk | nj/nk | 0 | 0 | conservativa |
| Promedio simple | ½ | ½ | 0 | 0 | conservativa |
| Estrategia flexible | ½(1-ß) | ½(1-ß) | | ß>0 | contractiva |
| | | | | ß=0 | conservativa |
| | | | | ß<0 | dilatadora |

*Ligamiento simple o vecino más cercano.*- Uno de los métodos aglomerativos jerárquicos más simples es como su nombre indica el del ligamiento simple ("single linkage") también conocido como el del vecino más cercano ("nearest neighbour") cuya característica distintiva es que la distancia entre grupos queda definida por el más cercano del par de individuos comparados. La pregunta de la Fig. 5.3 es solucionada en esta estrategia con el siguiente cálculo:

$D_{hk} = ½ (0.90) + ½ (0.76) + -½ |0.90 - 0.76|$
$D_{hk} = 0.76$

Como puede observarse en este método la afinidad entre dos entidades se define por el valor *mínimo*, por lo que la ecuación puede expresarse sencillamente como: $D_{hk} = \min (D_{hi}, D_{hj})$ Por ello en el ligamiento simple, en la medida en que un grupo crece, éste tiende a moverse muy cercanamente a otros grupos o entidades por lo que se plantea que tiene propiedades *contractivas* sobre el espacio (Boesch, 1977).

Esta propiedad se manifiesta en un fenómeno conocido como encadenamiento ("chaining") que se refiere a la tendencia del método a incorporar puntos intermedios a un grupo ya existente, mas que a iniciar un grupo nuevo. Por ello el ligamiento simple tiende a formar clasificaciónes con cierto desorden (Everitt y Dunn, 1991). El agrupamiento por valores mínimos implica que cada vez que

una entidad se adiciona a un grupo, la distancia de este grupo con respecto a los restantes o no cambia o se reduce, por lo que los grupos grandes crecen y se mantienen juntos mientras que los puntos aislados tienen a mantenerse libres (Krzanowski y Marriott, 1996a).

*Ligamiento completo o vecino más alejado.*- El método de ligamiento completo ("complete linkage") o del vecino más alejado ("furthest neighbour") es opuesto al del ligamiento simple en el sentido de que la distancia entre grupos se define aquí por el más distante del par de individuos comparados. En esta estrategia el valor requerido de $D_{hk}$ se calcula:

$D_{hk} = ½ (0.90) + ½ (0.76) + ½ |0.90 - 0.76|$
$D_{hk} = 0.90$

La afinidad entre dos entidades se define por el valor *máximo* por lo que la ecuación puede simplificarse como: $D_{hk}$= max ($D_{hi}$, $D_{hj}$). Por esta razón el ligamiento completo brinda resultados opuestos al ligamiento simple en el sentido de que en la medida que el grupo crece se escinde de determinados grupos o entidades, por lo que se plantea que tiene propiedades *dilatadoras* sobre el espacio. Mientras que el ligamiento simple tendia al encadenamiento, en el ligamiento completo ocurre tipicamente una agrupación intensa donde se forman varios grupos discretos. El agrupamiento por valores máximos implica que cada vez que una entidad se adiciona a un grupo la distancia de este grupo con respecto a los restantes o no cambia o crece. Cuanto más crezca un grupo menor será su tendencia a incorporar nuevas entidades lo que implica que en datos poco estructurados las entidades se agruparán en pequeños conjuntos que se combinarán a su vez en uno mayor (Krzanowski y Marriott, 1996a).

*Promedio simple.*- A diferencia de los métodos anteriores basados en la selección de valores extremos en el método de promedio simple ("simple average"), conocido también como de los pares de grupos no ponderados usando la media aritmetica ("unweighted pair group method using arithmetical averages" o UPGMA), la afinidad se define simplemente pòr su promedio. La interrogante planteada en la demostración se resuelve por tanto:

$D_{hk} = ½ (0.90) + ½ (0.76)$
$D_{hk} = 0.83$

Este método se considera como *conservativo* del espacio ya que introduce poca distorsión en las afinidades originales, propiedad que la hace una estrategia muy recomendada.

*Promedio de grupos.*- Similar en su concepto al anterior este método se conoce también como de los pares de grupos ponderados usando la media aritmetica ("weighted pair group method using arithmetical averages" o WPGMA). En este caso el cálculo de la nueva disimilitud se realiza:

$D_{hk} = ni/nk\ (0,90) + nj/nk\ (0.76)$
$D_{hk} = ½ (0.90) + ½ (0.76)$
$D_{hk} = 0.83$

En este método la afinidad entre grupos es también un *promedio,* solo que a diferencia del promedio simple donde se divide entre dos, en este caso los valores de afinidad se multiplican por el número de entidades involucradas en cada grupo y se dividen por el número total de entidades involucradas en la comparación. Se le considera también como *conservativo* sobre el espacio, con ligera distorsión de las afinidades originales y en la práctica sus resultados son muy similares a los del promedio simple.

*Estrategia flexible.* - Finalmente veamos una última variante de cálculo de $D_{hk}$ útil por su versatilidad, donde la disimilitud se calcula como:

$D_{hk}$ = [½(1 -(-0.25)](0.90) + [½(1 -(-0.25)](0.76) + (-0.25)(0.38)
$D_{hk}$= 0.47 + 0.56 - 0.09
$D_{hk}$ = 0.94

Debido a que los cambios en los métodos de agrupamiento implican solamente cambios en los parámetros de la ecuación de Lance y Williams (1966a), es claro que ello brinda infinitas posibilidades de estrategias aglomerativas. Por ello, dichos autores proponen una estrategia flexible asumiendo los siguientes criterios: $\alpha_i + \alpha j + ß = 1$; $\alpha_i = \alpha_j$; $ß < 1$ y $\gamma = 0$. En esta estrategia, $\alpha_i$ y $\alpha_j$ se ponen en función del coeficiente ß -denominado aquí coeficiente de intensidad del grupo- que al ser variable permite deliberadamente causar la distorsión sobre el espacio más apropiada para los propósitos de la investigación. Asi, en la medida en que ß se aleja de 0 la estrategia es *contractiva* y en la medida en que se acerca es *dilatadora* (Fig. 5.4).

Los valores de $D_{hk}$ calculados para varios valores de ß, negativos y positivos, son iguales a los del ligamiento simple para ß = 0.16; al del promedio simple si ß=0; y al del ligamiento completo para ß = -0.16, y se hacen mayores que 1 para valores muy bajos de ß negativos, lo cual en nuestro ejemplo ocurre a partir del valor -0.4. Ello puede parecer contradictorio dado que nuestra afinidad por definición varía entre 0 y 1 por lo que generalmente éstos valores se eluden (Fig. 5.4), situación que no ocurre con las distancias dado que estas varían entre 0 e $\alpha$.

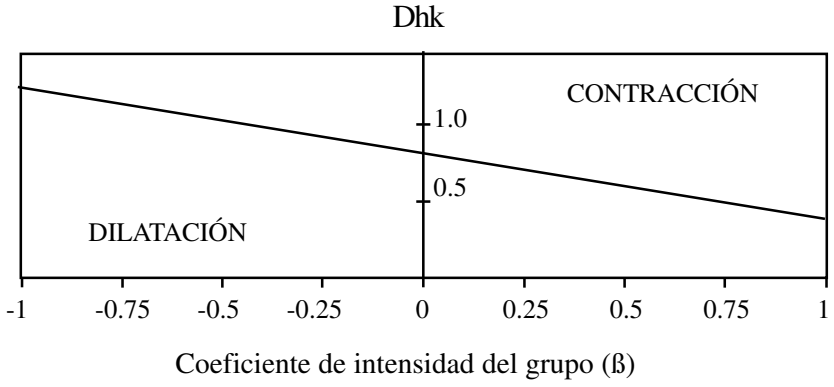



Figura 5.4. Variación de Dhk para diferentes valores de ß, según Dhk = 0.83 -0.45ß.

El valor de ß = -0.25 es usado convencionalmente con resultados satisfactorios pues brinda un agrupamiento intenso con moderada dilatación sobre el espacio, lo cual resulta útil para lograr una mejor definición de los conjuntos, aunque evitando los valores extremos, el interesado tendrá varias opciones válidas para jugar con sus resultados. Lance y Williams (1977) brindan un ejemplo gráfico que ha devenido en un clásico para ilustrar la influencia del coeficiente β (Fig. 5.5).

Otras estrategias aglomerativas señaladas por Boesch (1977) son la mediana, el centroide y la suma creciente de cuadrados pero que no tienen ventajas particulares. El interesado se puede dirigir además a Legendre y Legendre (1979) , Pielou (1984) o Everitt (1993).

**¿Cómo se hace un agrupamiento "a mano"?**

Intentar la clasificación de entidades a partir de una matriz de afinidad sin una computadora y el programa adecuado es algo realmente poco usual y nada recomendable, sobre todo si se trata de matrices grandes. Sin embargo, con el propósito de que el interesado sea capaz de hacerlo si lo necesita, además de que es un excelente ejercicio para comprender mejor como operan los métodos aglomerativos combinatorios, veamos un ejemplo sencillo de clasificación a partir de una matriz de afinidad hipotética (Figs. de la 5.6 a la 5.9).

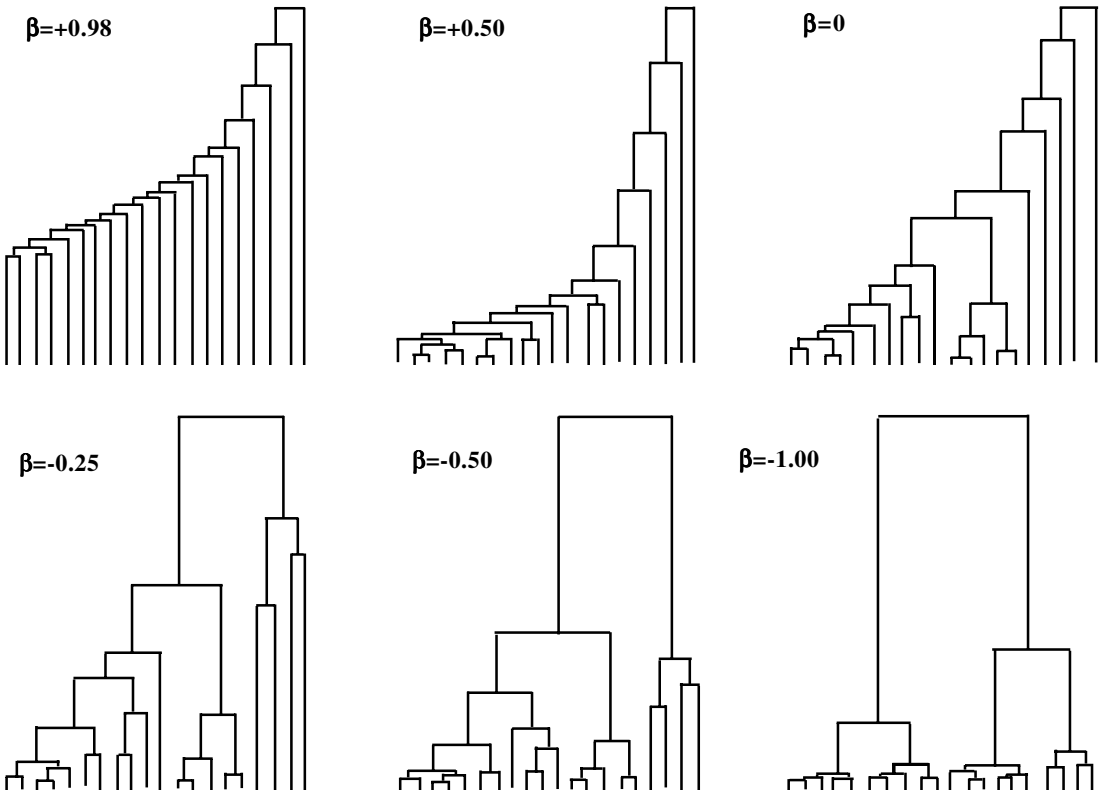



Figura 5.5. Influencia del valor de β sobre el espacio de los agrupamientos según Lance y Williams (1967).

Se impone como primer paso el exámen de la matriz para determinar cual será la primera unión que servirá de punto de partida al análisis restante. Recordemos que si se trata de una matriz de similitud o correlación, la primera unión corresponde al mayor valor de afinidad (máxima similitud o correlación entre entidades) pero si se trata de una matriz de disimilitud o distancia entonces la primera fusión corresponde al menor valor de la matriz (menor disimilitud o menor distancia). Para una matriz de disimilitud (Fig. 5.6A) entre cinco entidades, la primera unión corresponde a las entidades A y C con un valor de 0.23. Este valor será por tanto, el punto de partida de la clasificación. Al combinar las entidades A y C en un nuevo grupo queda ahora calcular cuáles serían los valores de disimilitud entre dicho grupo y las restantes entidades aun no agrupadas (Fig. 5.6B). Los valores de disimilitud entre las entidades que aún no han sido agrupadas mantiene sus valores originales.

**A.**

| | Entidades | | | | |
|---|---|---|---|---|---|
| A | B | C | D | E | |
| 0 | 0.38 | 0.23 | 0.72 | 0.69 | A |
| | 0 | 0,40 | 0.80 | 0.82 | B |
| | | 0 | 0.76 | 0.90 | C |
| | | | 0 | 0.36 | D |
| | | | | 0 | E |

**B.**

| | | | | |
|---|---|---|---|---|
| AC | B | D | E | |
| 0 | **?** | **?** | **?** | AC |
| | 0 | 0.80 | 0.82 | B |
| | | 0 | 0.36 | D |
| | | | 0 | E |

**C.**

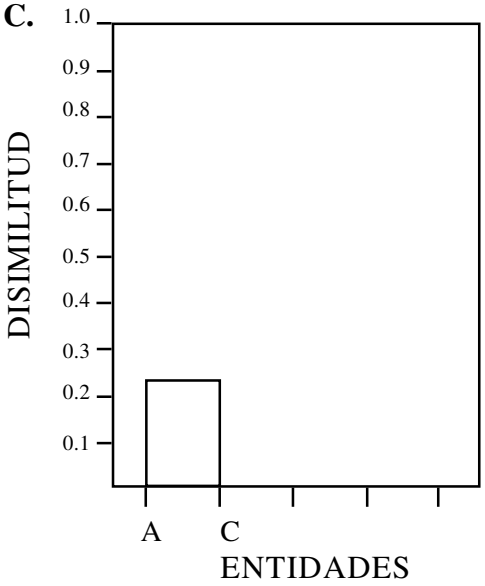



Figura 5.6. Clasificación de las entidades A, B, C, D y E. Primera etapa: fusión de las entidades A y C en un valor de 0.23. A. Matriz de afinidad original; B. Matriz de afinidad reordenada indicando los valores que deberán ser calculados; C. Representación gráfica del agrupamiento.

Como vimos anteriormente este cálculo se realiza a través de la ecuación de Lance y Williams, sustituyendo determinados coeficientes según la estrategia que se va a emplear, pero la situación no siempre es tan compleja ya que si se trata de los métodos de ligamiento completo, simple y promedio simple, se puede, sin necesidad de utilizar la ecuación combinatoria, buscar los nuevos valores con solo elegir los máximos, mínimos o promedios, respectivamente. Tomando como ejemplo el promedio simple, para los tres nuevos valores de disimilitud a calcular tenemos:

$$\text{Disimilitud AC-B} = \frac{\text{Disimilitud A - B + Disimilitud C - B}}{2} = \frac{0.38 + 0.40}{2} = 0.39$$

$$\text{Disimilitud AC-D} = \frac{\text{Disimilitud A - D + Disimilitud C - D}}{2} = \frac{0.72 + 0.76}{2} = 0.74$$

$$\text{Disimilitud AC-E} = \frac{\text{Disimilitud A - E + Disimilitud C - E}}{2} = \frac{0.69 + 0.90}{2} = 0.79$$

Tenemos entonces una segunda matriz (Fig. 5.7A) para escoger de nuevo el menor valor: 0.36 donde se unen las entidades D y E.

**A.**

| Entidades | | | | |
|---|---|---|---|---|
| AC | B | D | E | |
| 0 | 0.39 | 0.74 | 0.79 | AC |
| | 0 | 0.80 | 0.82 | B |
| | | 0 | 0.36 | D |
| | | | 0 | E |

**B.**

| | | | |
|---|---|---|---|
| AC | B | DE | |
| 0 | 0.39 | **?** | AC |
| | 0 | **?** | B |
| | | 0 | DE |

**C.**

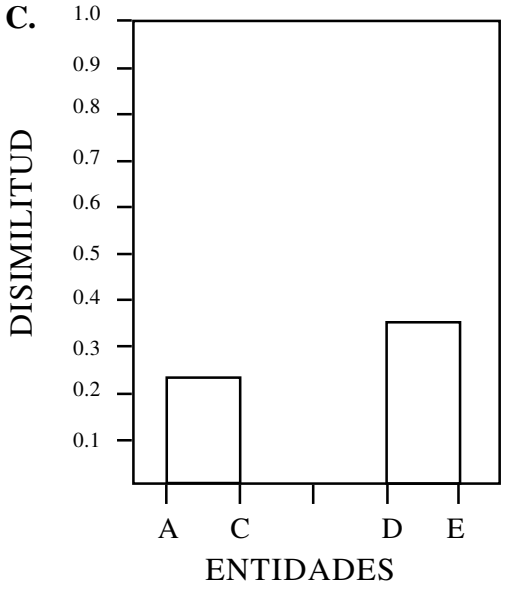



Figura 5.7. Clasificación de las entidades A, B, C, D y E. Segunda etapa: fusión de las entidades D y E en el valor 0.36. A. Matriz de afinidad reordenada en la primera etapa; B. Matriz de afinidad reordenada en la segunda etapa indicando los valores que deberán ser calculados; C. Reprentación gráfica del nuevo agrupamiento.

Tras la nueva unión hay valores de disimilitud desconocidos que deben ser calculados (Fig. 5.7B) en este caso la disimilitud entre el par anteriormente formado AC y el nuevo DE; y entre este último y la entidad libre B:

$$\text{Disimilitud DE-AC} = \frac{\text{Disimilitud D -AC} + \text{Disimilitud E- AC}}{2} = \frac{0.74 + 0.79}{2} = 0.76$$

$$\text{Disimilitud DE- B} = \frac{\text{Disimilitud D - B} + \text{Disimilitud E - B}}{2} = \frac{0.80 + 0.82}{2} = 0.81$$

Tenemos ahora una tercera matriz cuyo valor mínimo corresponde a la unión de AC y B, con 0.39 (Fig. 5.8A), por lo que nos queda calcular un último valor de disimilitud de acuerdo a:

$$\text{Disimilitud ACB-DE} = \frac{\text{Disimilitud ABC-D} + \text{Disimilitud ABC-E}}{2} = \frac{0.76 + 0.81}{2} = 0.78$$

Este valor ubicado en la cuarta y última matriz (Fig. 5.9A) indica el valor de afinidad al cual quedan relacionados todos los grupos. Ejemplos similares al aquí mostrado pueden encontrarse en los trabajos de Crisci y López Armengol (1983), Ludwig y Reynolds (1988), Krzanowski (1988) o Fielding (1999).

**A.**

| AC | B | DE | |
|---|---|---|---|
| 0 | 0.39 | 0.76 | AC |
| | 0 | 0.81 | B |
| | | 0 | DE |

**B.**

| ACB | DE | |
|---|---|---|
| 0 | **?** | ACB |
| | 0 | DE |

**C.**

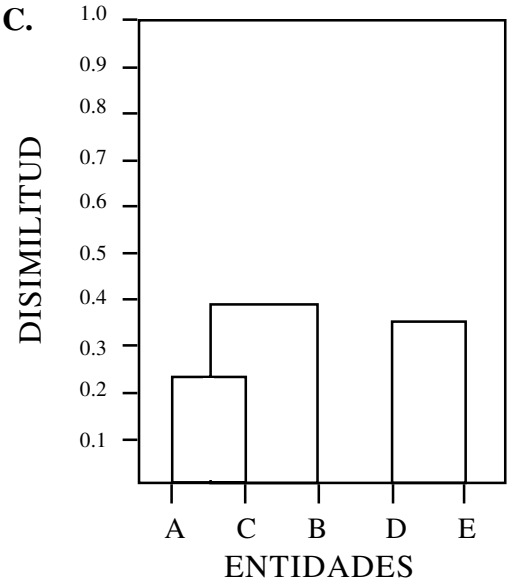



Figura 5.8. Clasificación de las entidades A, B, C, D y E. Tercera etapa: fusión de las entidades AC y B en el valor de 0.39. A. Matriz de afinidad reordenada en la segunda etapa; B. Matriz de afinidad reordenada en la tercera etapa indicando los nuevos valores que deberán ser calculados;, C. Reprentación gráfica del agrupamiento.

**A.**

| ACB | DE | |
|---|---|---|
| 0 | 0.78 | ACB |
| | 0 | DE |

**B.**

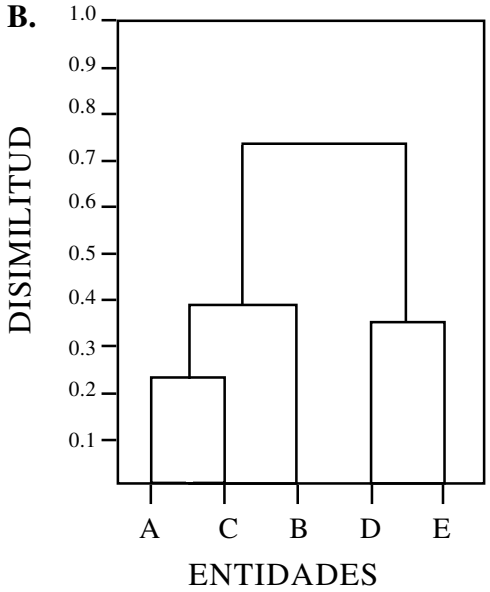



Figura 5.9. Clasificación de las entidades A, B, C, D y E. Cuarta y última etapa: fusión del grupo ACB y DE en un valor de 0.78 A. Matriz de afinidad reordenada en la tercera etapa; B. Representación gráfica de los agrupamientos.

Resumiendo tenemos que el orden sucesivo de los agrupamientos es el siguiente:

A con C en un valor de 0.23
D con E en un valor de 0.36
AC con B en un valor de 0.39
ACB con DE en un valor de 0.78

Para llevar estos valores al árbol de clasificación (como ya hemos ido viendo por pasos) se construye un eje de coordenadas donde los valores de disimilitud ocuparán el eje Y y las entidades el de las X. En el mismo se van representando la unión de las entidades y grupos según se vayan fusionando secuencialmente con líneas horizontales o nodos[2] que se corresponden con los valores de disimilitud a los cuales van ocurriendo dichas uniones, y con líneas verticales (internodos) cuya altura es igual a la afinidad entre entidades o grupos de ellas.

Una representación de tal tipo es lo que se conoce como dendrograma o árbol de clasificación (Fig. 5.10), que según Pielou (1984), más que un simple diagrama informativo es una ordenación bidimensional de un conjunto de datos representativo de sus interrrelaciones y en tal sentido es que debe ser analizado. Este brinda un cuadro completo del proceso de agrupamiento aunque no de la matriz de afinidad (Krzanowski y Marriott, 1996a). El dendrograma muestra una secuencia de fusiones o divisiones sucesivas que tienen lugar entre los grupos en la medida que el coeficiente de agrupamiento varía entre sus valores extremos. Hacia "abajo" todas las unidades forman grupos separados; hacia "arriba" todas caen en un solo grupo (Krzanowski, 1990).

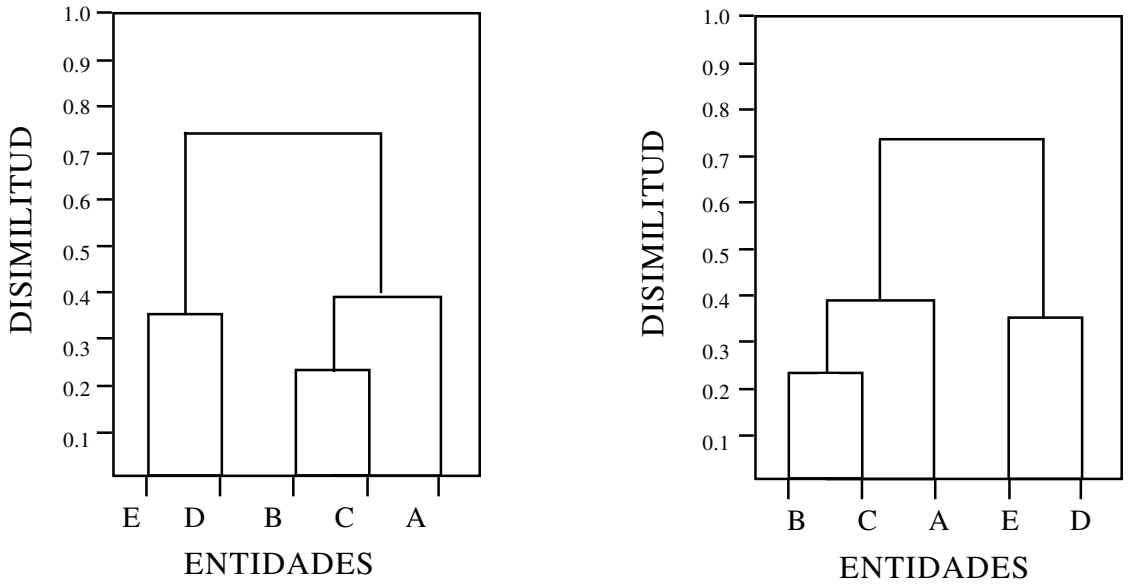



Figura 5.10. Intercambio de entidades en el eje del dendrograma.

El orden de las entidades en el eje X podrá notarse que es sencillamente opcional; éstas pueden ser intercambiadas sin que ello implique cambios en el significado del dendrograma (Fig. 5.10) En tal sentido Pielou (1984) y Krzanowski y Marriott (1996a) comparan el dendrograma con un «móvil» donde cada nodo puede girar libremente en el punto donde coincide con el internodo dibujado sobre él, lo cual es útil cuando se quiere dar a las entidades clasificadas un cierto orden que responda, por ejemplo, a un gradiente de cambio o a una orientación geográfica.

[2] La terminología de nodo e internodo corresponde a Pielou (1984) y será empleada aquí solo para facilitar nuestra explicación sobre el dendrograma. El término nodo en particular, será empleado con otro significado cuando nos refiramos al análisis nodal.

Cuando se trata de disimilitud o distancia la escala puede tener su valor cero en el origen de coordenadas pero para similitud la escala debe invertirse ya que a cada nueva jerarquía de unión le corresponde un menor valor, al igual que para la correlación. En esta última, incluso la escala puede ampliarse por debajo de cero, con valores negativos, dado que este coeficiente varia entre -1 y 1.

Los árboles de clasificación o dendrogramas son la representación más universalmente empleada para plasmar los resultados de los agrupamientos, pero no la única. De hecho Legendre y Legendre (1979) dividen éstas en los ya mencionados dendrogramas y en los llamados gráficos de encadenamiento, donde se representan con líneas o círculos las relaciones, pero que no tienen ventajas particulares en la facilitación de las interpretaciones.

Como otra alternativa de representación bidimensional de la matriz de afinidad, Krzanowski (1988) habla a favor del *árbol de espaciamiento mínimo* ("minimum spanning tree"), (Fig. 5.11) que no es mas que una red donde con ángulos arbitrarios se van espaciando los puntos mediante líneas rectas cuya longitud es proporcional a la disimilitud entre puntos. Esta representación es muy cercana a un dendrograma de ligamiento simple y las distancias calculadas no son más que las distancias umbrales del proceso aglomerativo de esta técnica (Chatfield y Collins, 1992).

Kaufman y Rousseeuw (1990) discuten varias alternativas gráficas y señalan como una que permite explorar fácilmente la estructura de los datos, la de *rótulo* ("banner") que consiste en una serie de barras y estrellas donde las primeras son repeticiones de las marcas de cada objeto y las segundas indican su relación (Fig. 5.12). Estas y otras alternativas brindarán al interesado opciones gráficas para presentar sus resultados aunque en el caso de matrices grandes consideramos que el árbol de clasificación es la representación más clara; las restantes son menos familiares y atractivas (Krzanowski y Marriott, 1996a).

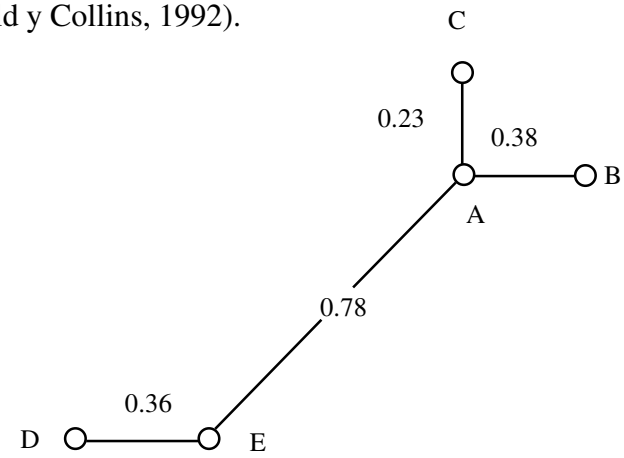



Figura 5.11. Arbol de espaciamiento mínimo para los agrupamientos obtenidos por el método de ligamiento simple a partir de la matriz de afinidad de la Figura 5.6A.

Para expresar los resultados de la taxonomía numérica Sneath y Sokal (1973) mencionan los modelos de bolas y barras, los diagramas de contorno y los mapas taxométricos aunque los fenogramas y cladogramas, nombres que reciben los dendrogramas en esta disciplina, son los más empleados.

**Una alternativa de agrupamiento de datos porcentuales**

Una vez vistos los principales métodos de agrupamiento veamos una variante sencilla de análisis y formación de grupos a partir de datos estandarizados en porcentajes, que dará al interesado una opción metodológica más para su trabajo práctico. En Pielou (1984) encontramos ejemplos donde se representan gráficamente los agrupamientos en un espacio bidimensional con un eje simple de


Figura 5.12. "Rótulo" de los agrupamientos obtenidos en la Figura 5.10.

coordenadas, tomando como datos originales los atributos de dos especies en un conjunto de estaciones. Esto representa el caso binario de clasificación que es el más sencillo (Hair *et al*., 1995) pero como generalmente contamos con datos de varias especies, o sea casos multivariados, tal representación en dos ejes no es posible a no ser que seamos capaces de resumir toda la información cuantitativa en dos conjuntos de valores.

Pongamos como ejemplo los datos porcentuales obtenidos en el estudio de las comunidades coralinas en trece perfiles del borde de la plataforma Suroccidental de Cuba (Herrera *et al.,* 1991), donde hemos resumido, con un sentido ecológico, los porcentajes en tres grupos: I. Porcentajes de *Montastraea annularis*, especie dominante en los arrecifes bien desarrollados; II. Porcentajes de *M. cavernosa, Siderastraea radians* y *Stephanocoenia michelini*, especies que comparten la dominancia en las zonas de poco desarrollo arrecifal; III. Porcentajes de las restantes especies que no presentan un patrón de dominancia marcado (Tabla 5.2).

Tabla 5.2. Porcentajes de tres grupos de especies de corales según su grado de tolerancia a algunos tensores naturales del arrecife, según datos de Herrera *et al.,* 1991).

| | Grupos | | |
|---|---|---|---|
| Estaciones | I | II | III |
| 1 | 11.9 | 61.5 | 26.6 |
| 2 | 8.8 | 51.3 | 39.9 |
| 3 | 14.1 | 55.9 | 30.0 |
| 4 | 33.1 | 20.4 | 46.5 |
| 5 | 57.1 | 14.2 | 28.7 |
| 6 | 17.6 | 47.7 | 34.7 |
| 7 | 31.3 | 39.8 | 28.9 |
| 8 | 28.2 | 37.9 | 33.9 |
| 9 | 1.5 | 71.2 | 27.3 |
| 10 | 27.8 | 41.3 | 30.9 |
| 11 | 48.8 | 27.4 | 23.8 |
| 12 | 32.7 | 27.2 | 40.1 |
| 13 | 5.8 | 57.8 | 36.4 |

Empleando la representación usada por Pielou (1984) los gráficos de la Fig. 5.13A, B y C, brindan parcialmente las relaciones entre los grupos relacionados, par a par. Veamos entonces una forma de relacionar los tres simultáneamente. Durante mucho tiempo los geólogos han empleado el conocido triángulo de Sheppard (1954) para explicar la composición de sus sedimentos sobre la base de los porcentajes de distintas fracciones granulométricas; también algunos estudios biológicos lo han incluído para la interpretación de sus datos.

Dado que este diagrama ofrece la posibilidad de plotear tres valores porcentuales correspondientes a un mismo conjunto de datos, es posible en una lista de especies, individualizar dos de las especies dominantes o dos conjuntos de ellas y analizarlas porcentualmente respecto al resto de la comunidad.

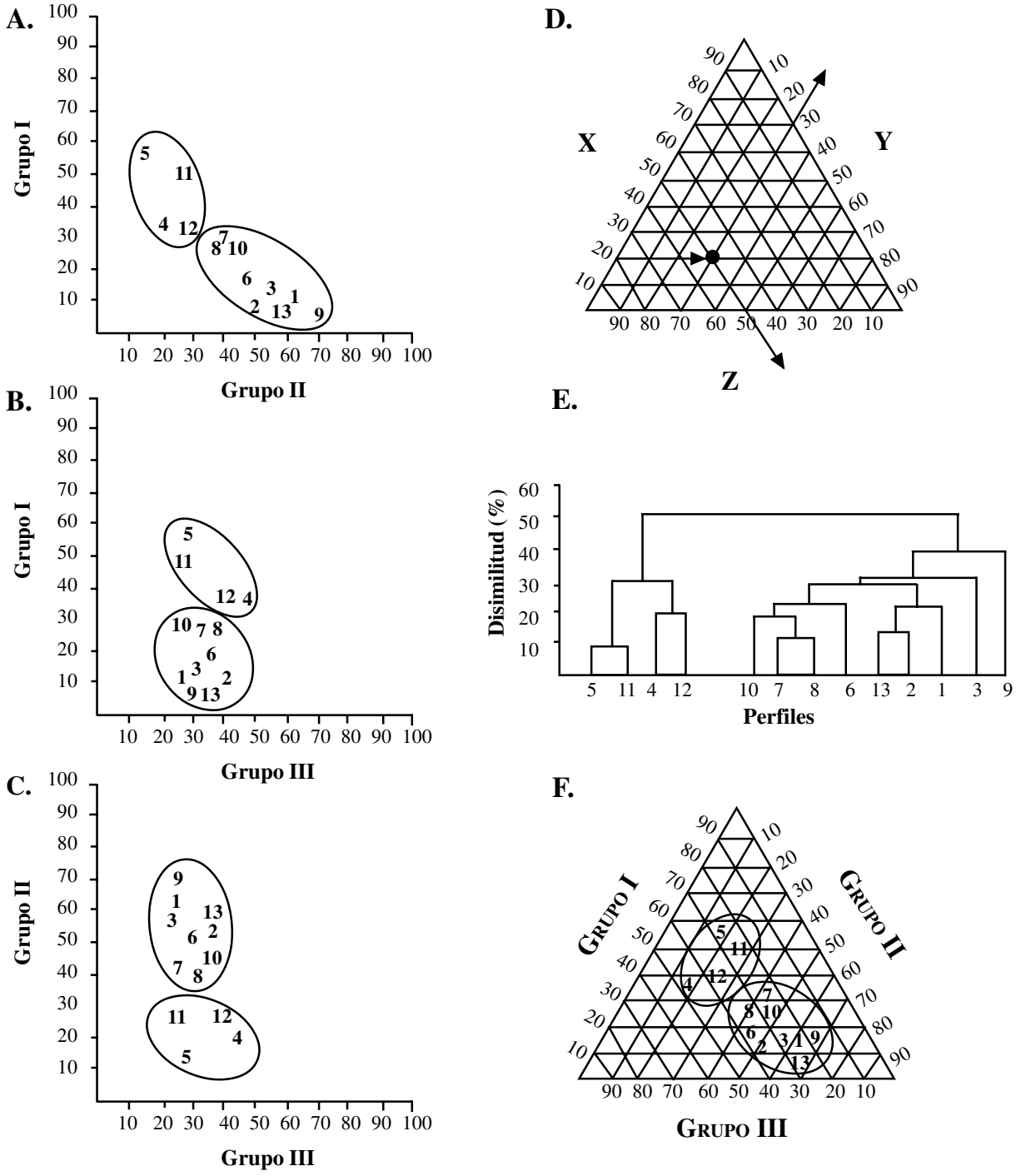



Figura 5.13. Análisis de los datos de veinticinco especies de corales en trece perfiles de los arrecifes de la plataforma Suroccidental de Cuba, organizadas en tres grupos (ver texto). **A, B y C.** Representación de las relaciones bivariadas entre grupos de especies. **D.** Ejemplo de ubicación de un punto con coordenadas X = 20; Y = 30 y Z = 50, en el diagrama triangular de Sheppard (1954). **E.** Dendrograma obtenido al clasificar los trece perfiles sobre la base de su composición coralina. **F.** Representación de los mismos datos en el diagrama triangular.

De esta forma se obtienen grupos en el triángulo que vendrán dados en lo fundamental por el aporte que realizan las especies que dominan en la comunidad. En la Fig. 5.13D se muestra un ejemplo de como trabajar con el triángulo de Sheppard (1954). Asumiendo un conjunto de datos donde los valores porcentuales correspondientes a las entidades X, Y y Z, son respectivamente 20, 30 y 50, la ubicación de los puntos, una vez elegido lo que se debe representar en cada eje, se realiza entrando horizontalmente por el eje izquierdo con el valor del elemento X hasta buscar la coincidencia con el valor en el eje de Z en las líneas inclinadas hacia la izquierda. Una vez determinado el punto de coincidencia de ambos valores, se asciende por las líneas inclinadas a la derecha hacia el eje de Y, completándose asi el 100%.

Para su empleo con propósitos de obtener grupos la matriz de datos originales debe ser analizada, particularmente en sus especies dominantes, y separarse dos conjuntos compuestos por los porcentajes de una o la suma de varias de ellas, para ocupar dos de los ejes del diagrama triangular, quedando el tercer eje destinado a la suma de los porcentajes de las restantes especies. Herrera (1992) encuentra que las agrupaciones así obtenidas resultan muy similares a las logradas al analizar los datos a través del índice de disimilitud porcentual de Sanders (1960) y empleando técnicas aglomerativas de promedio. La Fig. 5.13E muestra el dendrograma obtenido para los datos originales de 25 especies de corales en las trece estaciones y la Fig. 5.13F el resultado de la agrupación en el triángulo.

Aunque un elemento desventajoso en el triángulo es que no brinda numéricamente un valor del grado de afinidad entre los grupos, por otra parte presenta la ventaja de que en él están implícitas las causas que motivan la separación de los mismos (especies que los determinan), cosa que no ocurre en el árbol de clasificación que por sí solo no explica la influencia de los porcentajes de las distintas especies en el agrupamiento. Asi, en la Fig 5.13F es claro que el conjunto de las estaciones 4, 5, 11 y 12 representan al arrecife más desarrollado donde la dominancia corresponde a *M. annularis*. En este sentido el diagrama triangular refleja lo que solo un análisis nodal (relación entre las clasificaciones normal e inversa) podría dar. Esto constituye una importante ventaja pues el análisis nodal requiere de la clasificación de especies y estaciones y cuando se manejan datos estandarizados en porcentajes por columnas (estaciones) no siempre resulta adecuado realizar la clasificación por filas (especies), aunque esto último puede solucionarse analizando la matriz original de datos porcentuales y reagrupando casuísticamente las especies (Herrera, 1984).

No obstante, debemos aclarar, que la utilidad del diagrama triangular para visualizar de manera global las tendencias de agrupamiento de los datos, se reduce en la medida que se incrementa el número de especies dominantes, la visualización de los grupos en el diagrama triangular se hace más difícil (dado el espacio limitado que brinda el triángulo), la representación adquiere una mayor generalización y es preciso emplear la matriz de datos originales para su mayor precisión, razón por la cual solo se recomienda como una opción adicional, o si se quiere una representación gráfica novedosa de los datos.

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

*"Hay ocasiones y causas, porques y por qués para todas las cosas"*
*William Shakespeare*

# 6. INTERPRETACIÓN DE LAS CLASIFICACIONES

El objetivo de los métodos de clasificación es simplificar conjuntos complejos de datos por lo que los resultados que de ellos se obtienen no deben terminar en el dendrograma -como ocurre frecuentemente- sino que deben servir real y objetivamente para que el análisis ecológico proceda con mayor eficiencia. Por ello, los resultados deben ser analizados críticamente y considerar la posibilidad de su posterior refinamiento.

Una vez obtenido el dendrograma a partir de determinadas técnicas es importante analizar su concordancia con la matriz de afinidad original y posteriormente su estructura, pues de este modo podremos evaluar hasta qué punto la organización jerárquica obtenida es adecuada para los datos de la investigación. Tenemos entonces que llegada esta etapa deben responderse dos preguntas claves: ¿Se ajusta globalmente el dendrograma obtenido a la matriz de afinidad? ¿Cuál es la partición más óptima de la jerarquía? A éstas y otras interrogantes daremos respuesta en los epígrafes siguientes.

**Medidas de la bondad del ajuste**

Evaluar en qué medida el dendrograma refleja las relaciones originales de afinidad podría denominarse, empleando una expresión de la estadística clásica como la *bondad del ajuste*. Este paso es necesario, tanto si los datos originales se analizan con una sola técnica como con varias, que es lo más recomendable, pues con esto último (que podemos llamar comparación intraclasificatoria), podemos comparar y seleccionar las clasificaciones más adecuadas. Para este fin se han propuesto técnicas que permiten medir el ajuste del dendrograma a los valores de la matriz de afinidad, o lo que es lo mismo medir la distorsión, ya que las relaciones de afinidad son necesariamente distorsionadas al ser llevadas a una representación bidimensional (Crisci y López Armengol, 1983).

Aunque la comparación por examen visual de los árboles es válida, existen métodos cuantitativos basados fundamentalmente en la comparación de los valores de la matriz de afinidad con aquellos que han servido de base para la estructuración del dendrograma. El grado de discrepancia entre ambos valores puede considerarse una medida de la distorsión causada por la imposición de una estructura jerárquica (Krzanowslki y Marriott, 1996a). Algunas medidas, resumidas por estos autores y por Krzanowski (1990) son: la suma cuadrática de sus diferencias, la suma cuadrática ponderada, la correlación por rangos y la correlación cofenética, que es una de las más empleadas.

La *correlación cofenética* proviene de la taxonomía numérica (Sokal y Rolhf, 1962) y se basa, como su nombre indica, en calcular el grado de relación a través del coeficiente *r* de la correlación

producto-momento que aquí se denomina *coeficiente de correlación cofenético* CCC ("cophenetic correlation coefficient"). A partir de una matriz de afinidad (Fig. 6.1A) se dibuja el dendrograma (Fig. 6.1B), señalando en éste los valores de afinidad en los cuales han tenido lugar las distintas fusiones, según su orden consecutivo de creación. Con ellos, se confecciona una matriz de dimensiones similares a la de afinidad de la cual partimos que será la *matriz cofenética* (Fig. 6.1C), donde las relaciones entre cada entidad vendrán dadas por el valor de afinidad que le corresponda en la jerarquía de la clasificación.

En la matriz de la Fig. 6.1, los valores que relacionan a las entidades AC y DE, que forman grupos particulares es claro que son 0.23 y 0.36, respectivamente. Para ver los valores correspondientes a entidades más alejadas, por ejemplo A y E se asciende por el árbol y se toma el valor que corresponde al nodo superior de unión de dichas entidades, en este caso 0.78. El mismo valor corresponderá a la relación AD, lo cual es típico de las matrices cofenéticas, donde siempre habrá varios valores repetidos.

**A.**

| A | B | C | D | E | |
|---|---|---|---|---|---|
| 0 | 0.38 | 0.23 | 0.72 | 0.69 | A |
| | 0 | 0.40 | 0.80 | 0.82 | B |
| | | 0 | 0.76 | 0.90 | C |
| | | | 0 | 0.36 | D |
| | | | | 0 | E |

**C.**

| A | B | C | D | E | |
|---|---|---|---|---|---|
| 0 | 0.39 | 0.23 | 0.78 | 0.78 | A |
| | 0 | 0.39 | 0.78 | 0.78 | B |
| | | 0 | 0.78 | 0.78 | C |
| | | | 0 | 0.36 | D |
| | | | | 0 | E |

**B.**

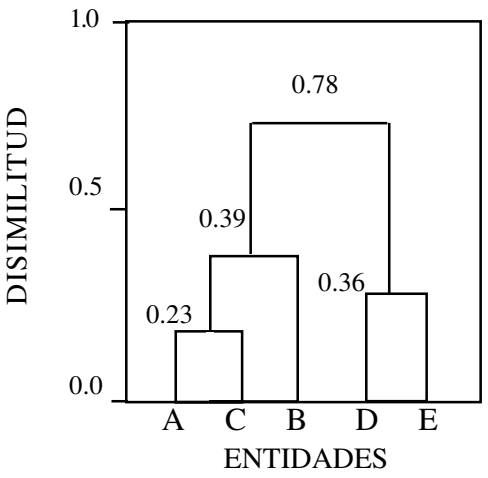



Figura 6.1. Pasos para el cálculo de la bondad de ajuste mediante la correlación cofenética A. Matriz de afinidad original; B. Dendrograma con los valores de disimilitud correspondientes a cada fusión; C. Matriz cofenética.

Empleando la correlación producto momento se relacionan los datos de la matriz de afinidad original con los valores que le corresponden en la matriz cofenética calculándose el coeficiente de correlación *r*. Una alta correlación entre matrices será señal de poca distorsión por lo que escogiendo los resultados de la técnica que brinde el mayor valor de *r* podemos considerar que las relaciones implícitas en la matriz de afinidad están siendo reflejadas con la mayor fidelidad. Según Sokal y Sneath (1963) valores superiores a 0.8 indican una buena representación de la matriz de afinidad por parte del dendrograma, aunque Rohlf (1970) alertó acerca de que aún valores de 0.90 no garantizaban necesariamente que el dendrograma resumiera adecuadamente las relaciones.

Los métodos de comparación intraclasificatoria se basan generalmente en métodos de correlación y de hecho, si solo nos interesa asumir que las afinidades observadas tienen una significación ordinal puede emplearse simplemente un coeficiente de correlación por rangos en lugar del CCC (Everitt y Dunn, 1991). De cualquier forma debe tenerse claro que un alto valor del coeficiente de correlación

cofenético es solo una medida de poca distorsión en la técnica empleada y no de una buena o mala clasificación (Crisci y López Armengol, 1983). El análisis de la bondad del ajuste o la comparación intraclasificatoria, si comparamos los resultados de varias técnicas, juzga ante todo aspectos metodológicos, para pasar después a la interpretación ecológica.

**Reglas de decisión**

Un dilema que enfrenta todo el que se encuentra ante los resultados de un árbol de clasificación es la determinación de grupos dentro de la jerarquía. Visto sobre un dendrograma la pregunta sería: ¿qué ramas del árbol pueden ser considerados «grupos» con una afinidad interna razonable? lo que llevado a un lenguaje más común sería: ¿dónde «cortar» el árbol para formar los grupos?

El dilema de una regla de decisión, que en inglés se denomina comúnmente "stopping rule", es clave ya que en su solución descansa la decisión correcta sobre nuestra estructura grupal. Básicamente pueden cometerse dos errores. El primero ocurre cuando la regla de decisión indica n grupos y de hecho hay menos, con lo cual se sobreestima el número de conjuntos. El segundo ocurre cuando la regla indica menos grupos con lo cual el número real de éstos se subestima. Aunque la severidad de ambos tipos de error cambia según el contexto del problema, los del segundo tipo pueden considerarse más serios debido a la pérdida de información implícita en la fusión errónea de grupos (Milligan y Cooper, 1985).

En principio la interpretación del dendrograma es un proceso visual simple donde juega un papel importante el conocimiento del sistema ecológico que se estudia. Hasta el carácter de quien interpreta es importante: las personas en extremo convencionales se aferran a su estructura de grupos como a una tabla de salvación. Sin sacar conclusiones *a priori*, aunque las intuyamos, reconoceremos primero los grupos mayores enlazados en los valores más bajos de similitud o correlación; o los más altos de disimilitud o distancia. De ellos derivaremos grupos, subgrupos, conjuntos o subconjuntos hasta definir las asociaciones más adecuadas, sin preocuparnos cuando una entidad constituya un grupo independiente. Es preferible que la técnica empleada aisle entidades con características propias o «extrañas» con respecto al resto a que queden mezcladas alterándonos la homogeneidad interna de su grupo y haciendo pasar inadvertidas sus particularidades.

Una dificultad en este tipo de análisis es que no hay una manera completamente satisfactoria para definir un grupo, aunque siempre tengamos una idea intuitiva de lo que éste significa (Chatfield y Collins, 1992). Precisamente Everitt (1993) considera que el término grupo, conjunto o clase se ha usado esencialmente de una manera intuitiva sin intentar darle una definición formal, lo cual puede no solo ser difícil sino equívoco. Por ejemplo, se ha sugerido como criterio para evaluar el significado de un grupo su valor de uso; si el criterio de grupo brinda una respuesta de valor para el investigador esto es suficiente.

Sin embargo, otros intentos por definir qué es un grupo emplean propiedades como la *cohesión interna y el aislamiento externo* lo cual está más cerca de la definición de clasificación que pretende de manera objetiva crear grupos muy homogéneos entre sí y bien diferentes de otros. Esto es lo que

nos dicen Hair *et al.* (1995) cuando explican que los grupos deben poseer una homogeneidad interna muy alta ("within cluster") y una heterogeneidad externa ("between cluster") también muy alta.

Un criterio elemental empleado para definir los grupos es dibujar una línea a lo largo del dendrograma, en un nivel dado de afinidad y considerar que todas las ramas que la crucen pueden ser considerados grupos independientes. Este criterio, conocido como *regla fija* debe establecer un nivel arbitrario e invariable de afinidad, asumiendo un cierto «nivel de significación». Bakus (1990) refiere que los ecólogos de latitudes templadas suelen usar un valor de 0.5 mientras que los del trópico, donde la diversidad de especies es mayor han usado 0.25 que da muchos grupos cada uno con pocas especies, aunque esto no constituye una generalidad. A manera de ejemplo, el dendrograma de la Fig. 6.2 muestra una subdivisión empleando como regla fija un valor de 0.5, que da como resultado la creación de ocho grupos, cuatro de los cuales (7, 9, 8 y 11) son entidades independientes.

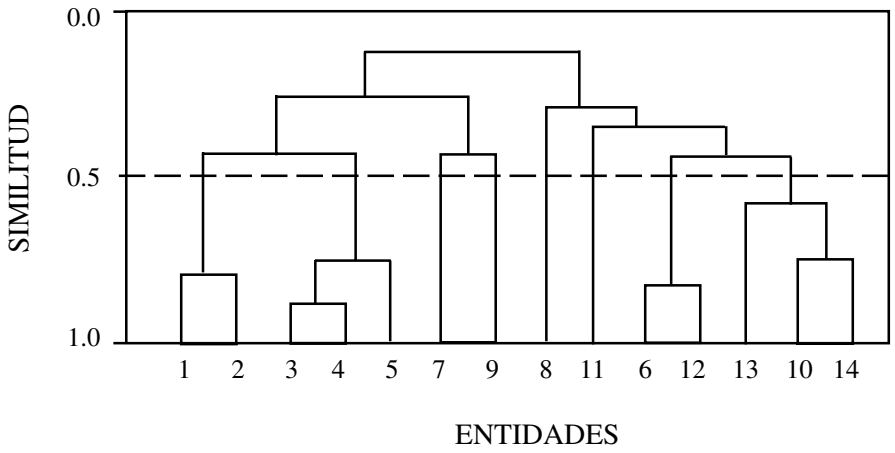



Figura 6.2. "Corte" para formar los grupos según una regla fija, donde se forman ocho grupos.

Un segundo criterio para la definición de grupos es la *regla variable* que no es más que el estudio del dendrograma, en consulta con la matriz original de datos (que en ocasiones se olvida que es en definitiva el punto de partida) para determinar grupos lógicos. Con este criterio dos grupos pueden ser considerados juntos a un nivel de afinidad mayor o menor que un tercero. El dendrograma de la Fig. 6.3 es el mismo de la Fig. 6.2 pero la división se ha realizado mediante una regla variable, delimitándose ahora seis agrupaciones, consideradas a distintos niveles de la jerarquía, donde solo dos (8 y 11) son entidades independientes.

El exámen del dendrograma para establecer una regla visual puede resultar más fácil si se plotean los valores de afinidad -en el ejemplo similitud- contra el número de fusiones en orden creciente. Si el número total de entidades a agrupar es n, en este caso n=14, el número de fusiones será n-1, o sea 13. Según Everitt y Dunn (1991) los cambios bruscos entre niveles de fusión adyacentes podrían ser indicativos de una solución de determinado número de grupos que estará determinado por el número que haya debajo del "corte"; punto donde presumiblemente la distancia entre las unidades del grupo se minimiza y la distancia entre grupos se hace máxima (Griffith y Amrhein, 1991). Ello resume esencialmente el método de la distancia entre grupos de Sharma (1996).

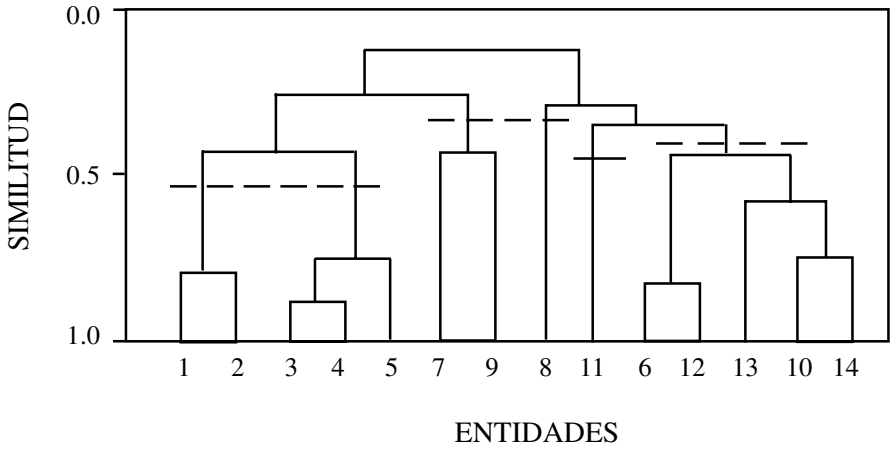



Figura 6.3. "Corte" para formar los grupos según una regla variable, donde se forman seis grupos.

Este ploteo (Fig. 6.4) para los datos hipotéticos del dendrograma de la Figura 6.3 refleja un salto de un valor de similitud de 0.6 a 0.3, a nivel de las fusiones 7 a 9, lo cual sugiere que el número de grupos a decidir debería buscarse a partir de este nivel de similitud que en el árbol de clasificación se corresponde con una solución entre seis y siete grupos.

Aunque se considera que la aplicación de una regla fija implica una menor subjetividad interpretativa de las clasificaciones esto es relativo, pues cuando el nivel de afinidad fijado cambia, pueden cambiar también las agrupaciones. Además, según Boesch (1977), la regla variable tiene dos ventajas básicas. En primer lugar algunas estrategias aglomerativas poseen una estrecha relación entre sus propiedades sobre la distorsión del espacio y el tamaño de los grupos, por tanto no hay justificación para una regla fija cuando la afinidad entre grupos y entidades depende precisamente del tamaño del grupo.

El segundo aspecto concierne a la naturaleza de los datos ecológicos y es particularmente importante en el análisis inverso. La mayoría de las matrices de datos incluyen especies abundantes y especies raras; se necesita por tanto una mayor afinidad para considerar los grupos de las primeras, que para las segundas, cuya probabilidad de ocurrencia es siempre baja.

Un elemento a considerar para definir si los conjuntos corresponden a diferencias reales entre los datos es el análisis de la escala de las agrupaciones. En el caso de un índice de similitud, por ejemplo, estaría justificada una estructura de grupos que variando entre 0 y 1 reflejara conjuntos unidos en valores contrastantes de afinidad en un intervalo amplio.

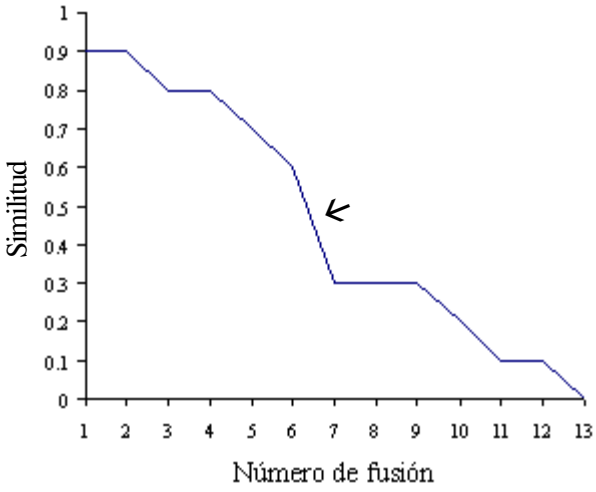



Figura 6.4. Variación de los valores de similitud en los niveles sucesivos de fusión en el dendrograma de la Fig. 6.3. La flecha indica el punto de cambio.

No sería lo mismo, si la escala global de variación de la afinidad entre los grupos integrantes del árbol estuviera entre 0.9 y 1, pues ello indicaría solamente que los datos son prácticamente iguales; o entre 0 y 0.1, que sería un reflejo de datos tan heterogéneos que ninguna estructura de grupos, aunque bien delimitados en el dendrograma, tendría una justificación lógica. El uso de un algoritmo de clasificación inevitablemente da como resultado una clasificación, independientemente de si las clases (de especies o estaciones) son conjuntos ecológicos reales o meramente fortuitos (Pielou, 1977). Esto es válido para cualquier otro tipo de índice, según su escala de variación, aunque reconocemos que en las distancias, al variar entre 0 e $\alpha$, este criterio es más difícil.

La estructura de grupos obtenida sugerirá, sin dudas, la mejor opción. Cuando los internodos del dendrograma son claramente de diferentes longitudes delimitando así grupos discretos bien diferenciados, esto debe reflejar la existencia de grupos naturales que pueden ser aislados sin arbitrariedad. Si esto no ocurre será necesario un análisis más cuidadoso para formarlos considerando que en tal caso podríamos estar haciendo una clasificación no natural, denominada *disección* (Pielou, 1984) que se practica cuando tenemos una población homogénea donde no hay agrupaciones naturales y aún así por razones prácticas deseamos dividirla en subgrupos. El número de subgrupos es arbitrario, como lo es el modo de obtenerlos (Chatfield y Collins, 1992) ya que los grupos se eligen por conveniencia sin buscar una solución óptima (Krzanowski y Marriott, 1996a). La selección de una regla de decisión es ante todo una cuestión de interpretación ecológica y en tal sentido la regla variable se ajusta más a la libertad de análisis que debe tener el investigador en la interpretación de sus resultados donde deben jugar un papel fundamental el sentido común, el juicio práctico y la fundamentación teórica del problema (Hair *et al.,* 1995).

Sin embargo, tanto la regla fija como la variable caen dentro de lo que se han denominado métodos "informales" para establecer el número de grupos, que aunque pueden ser suficientes cuando las diferencias entre conjuntos son contrastantes, llevan implícitas la posible influencia de lo que *a priori* se espera de los datos (Everitt, 1993). Existen, por tanto, aproximaciones más formales al problema de la determinación de los grupos que pretenden brindar una regla de decisión automática para eliminar los problemas de la subjetividad humana, de las cuales Sharma (1996) resume algunos puntos de vista.

Pero, sin dudas, el trabajo más importante al respecto, que aún continua siendo un clásico, es el de Milligan y Cooper (1985) que analizan treinta procedimientos provenientes de diferentes disciplinas, escogiendo métodos no restringidos a un tipo de dato en particular, no dependientes del método de agrupamiento y bien desarrollados matemáticamente, con lo cual podemos considerar a su trabajo un buen resumen de los mejores procedimientos registrados hace una década. Los autores concluyen que si bien éstos métodos pueden ser efectivos para determinar el número de grupos en los datos, los resultados varían de uno a otro en cuanto a efectividad y precisión, dependiendo de factores como el número de grupos y su grado de definición.

A través de Arabie y Hubert (1996) podemos conocer las recientes opiniones del propio G. Milligan al solicitársele un resumen del estado actual de las investigaciones en este campo: "Mi actual consejo es emplear dos o tres métodos de mi revisión de 1985; si se hallan resultados consistentes ya hay

apoyo sustancial para seleccionar los grupos; si hay un acuerdo parcial se debe optar por el mayor número de grupos y continuar el trabajo para confirmar cuáles deberían unirse; si no hay consistencia cualquier solución debe ser interpretable dentro del contexto del área de investigación y si no, exhorto a los investigadores a considerar la hipótesis de que no hay grupos en los datos".

Uno de los métodos mencionados por Everitt (1993) es el de Mojena (1977), ampliamente conocido y que por su sencillez puede servir de ejemplo de como proceden estas técnicas, muchas de las cuales son muy similares en sus procedimientos. A partir de los datos de las fusiones sucesivas de la matriz de afinidad se selecciona el número de grupos donde primero se satisfaga la desigualdad:

$$\alpha_{j+1} > \alpha + k\, S_{\alpha},$$

donde $\alpha_0$, $\alpha_1$, ...$\alpha_{n-1}$ son los niveles de fusión correspondientes a las etapas n, n-1,...,1 grupos, respectivamente; $\alpha$ es la media, $S_{\alpha}$ la desviación estándar y k una constante que según Mojena (1977) debe variar entre 2.75 y 3.50, aunque Milligan y Cooper (1985) hallaron que un valor de 1.25 era más adecuado.

Como se observa, este método se basa en la comparación de los valores parciales de afinidad respecto a la afinidad promedio. Otras aproximaciones como la que Sharma (1996) denominada RMSSTD del grupo ("root-mean-square total sample standard deviation") tienen igual principio pero comparan la desviación estándar o la varianza de los grupos en diferentes niveles respecto a estos parámetros para todos los datos buscando que la variación dentro de los grupos sea la menor posible. Altos valores de los estadígrafos de variabilidad indicarían la posible pérdida de homogeneidad del grupo.

Griffith y Amrhein (1991) argumentan con razón que si bien el aspecto numérico impone cierto grado de objetividad a la solución de grupos, existe un amplio componente subjetivo en la evaluación de los resultados de cualquier conjunto de datos relacionado con la selección de las variables, las medidas de afinidad y el algoritmo de clasificación, además de que muchas soluciones son dependientes de los datos y hasta del tipo de muestra.

**Reasignación**

Al analizar los resultados de la clasificación es frecuente que se detecte que algunas entidades están «fuera de grupo». Quiere esto decir que durante el proceso aglomerativo una entidad ha sido ubicada en un conjunto y sin embargo, de ser ubicada en otro el grupo resultante sería más homogéneo. Tal situación es a veces frustrante para el investigador que espera ver en el dendrograma sus ideas preconcebidas sobre la distribución de los datos, sin embargo no debe ser así. Si analizamos con detenimiento la secuencia de pasos de la clasificación, es claro que son varios los aspectos que pueden influir en la clasificación final.

En las técnicas jerárquicas que proceden por fusiones sucesivas las uniones son irrevocables de modo que cuando el algoritmo aglomerativo ha unido dos individuos no pueden ser posteriormente separados (Everitt, 1993). Por ello Kaufman y Rousseeuw (1990) comentan que los métodos jerárquicos sufren del defecto

de que nunca pueden reparar lo que han hecho en etapas previas. Por ello, ante entidades mal clasificadas se debe volver atrás para reconsiderar la calidad y naturaleza de los datos, las transformaciones efectuadas, los índices empleados (recordemos que estos últimos difieren en sus propiedades matemáticas) y por supuesto las propiedades de los propios métodos de agrupamiento. En el proceso aglomerativo una entidad puede ser captada para un grupo por parecerse solo a un miembro del mismo, quedando «atrapada» en el mismo por una fusión temprana. En tales casos puede pensarse sin temor, en su reasignación.

Los mayores inconvenientes en la interpretación surgen con los llamados grupos de entropía ("entropy groups") que constituyen grupos de objetos o individuos que no se ajustan a ninguna agrupación (Hair *et al*., 1995) y son representativos generalmente de datos aberrantes o extremos no detectados en el análisis previo de los datos. En tales casos debemos decidir si constituye un componente estructural válido de la muestra o si debe ser eliminado por su poca representatividad con lo cual el proceso clasificatorio debe reiniciarse sobre nuevas bases.

Claro está que eliminar o cambiar entidades de un grupo a otro puede atentar contra la objetividad del análisis. En este sentido dos criterios son importantes: el matemático, que nos permitirá definir qué particularidad aritmética del proceso ha hecho que la entidad dada halla sido ubicada en tal grupo; y el ecológico que con una perspectiva conceptual justificará porqué dicha entidad no corresponde esencialmente al grupo asignado. Si ambos criterios son bien manejados, la reasignación mediante inspección visual empleando la matriz de afinidad y ajustando los grupos a criterios conocidos sobre la influencia de los factores externos, puede ser practicada sin dudas. Una herramienta sencilla para la reasignación es el análisis nodal, que será tratado más adelante, dada su importancia no solo para reasignar sino también para interpretar.

Boesch (1977) comenta diferentes vías empleadas por varios autores para la reasignación de entidades pero reconoce que es imprescindible el desarrollo de nuevos y más objetivos métodos para este fin. Entre los mencionados está el análisis discriminante que consideramos uno de los más promisorios dado que permite asignar a cada entidad un valor probabilístico de admisión en el grupo propuesto así como excluir las entidades cuya probabilidad sea baja. En tal sentido, el análisis discriminante actúa como verificador de la homogeneidad de cada grupo y como reasignador de entidades, sobre la base de un criterio estadístico.

**Comparación interclasificatoria**

Cuando hablábamos de los datos cuantitativos, explicábamos que distintos tipos de parámetros ecológicos provenientes de un mismo muestreo pueden brindar clasificaciones diferentes. Bajo determinados propósitos podría ser de interés comparar los resultados entre clasificaciones. El grado de correspondencia entre clasificaciones diferentes de un mismo conjunto de organismos es conocida en taxonomía numérica como congruencia taxonómica, concepto que examina los resultados logrados con distintos caracteres (por ejemplo químicos y morfológicos). En ecología ello equivaldría a analizar en qué se parecen dos matrices de afinidad o dos dendrogramas obtenidos utilizando los datos de densidad y biomasa, por ejemplo y podríamos denominarlo congruencia ecológica.

Conceptualmente esta forma de evaluación de las clasificaciones en taxonomía parte de la premisa teórica de que diferentes caracteres pueden estar controlados por un mismo conjunto de genes y por tanto es de interés examinar si brindan una estructura clasificatoria similar. En ecología, los distintos parámetros comunitarios que sirven de punto de partida a la clasificación generalmente reflejan diferencialmente la influencia de determinados factores bióticos y abióticos y no están necesariamente correlacionados en la magnitud de sus valores, si bien todos responden a una situación ecológica global. Sin embargo, las diferentes clasificaciones pueden brindar información complementaria para evaluar la estabilidad de los resultados (Ignatiadis *et al*., 1992).

La evaluación de la congruencia ecológica podría arrojar luz sobre la significación de distintos parámetros en la distribución de la comunidad pero su mayor aplicación podría estar en la comparación de clasificaciones entre variables bióticas y abióticas. Este tipo de evaluación se realiza tanto en las matrices de afinidad como en los dendrogramas, para lo cual sirve su comparación visual, métodos de correlación como los ya explicados; o algunos índices que se basan en relaciones entre el número de objetos que se agrupan juntos en las dos clasificaciones comparadas y el número total (Everitt, 1993).

De cualquier forma, al considerar las diferencias entre dendrogramas basados en diferentes medidas de afinidad, la comparación no se puede basar en diferencias menores pues los patrones de relaciones del árbol son solo una representación aproximada de los valores de la matriz de afinidad en la cual se basa y pequeñas variaciones son suficientes para alterar las uniones (Boyce, 1969). Digby y Kempton (1991) proponen la comparación de las matrices de datos reordenadas según los resultados de las clasificaciones. La comparación interclasificatoria puede hacerse más compleja si variamos las medidas de afinidad y las técnicas de agrupamiento buscando comparaciones más representativas pero si se cuenta con las facilidades de cálculo es preferible emplear métodos de ordenamiento.

**Evaluación de las diferencias entre los grupos**

El objetivo de los métodos de clasificación es brindar una generalización hipotética sobre la estructura de los datos multivariados. Por ello, más que técnicas para el examen de diferencias de hipótesis - más cercanas a la estadística- son esencialmente métodos matemáticos que se usan cuando no tenemos ninguna hipótesis *a priori* y estamos en la fase exploratoria de nuestra investigación (Statistica, 2000). No obstante, puede existir interés de establecer estadísticamente, la realidad de los agrupamientos, y Boesch (1977) comenta algunas aproximaciones.

Por su procedencia de la estadística clásica los coeficientes de correlación al ser empleados como medidas de afinidad, han devenido también en tests de comparación aunque ya discutimos lo inapropiado que esto puede resultar. También, una vez construidos los grupos o creada una hipótesis de agrupación, las posibles diferencias estadísticas entre los valores originales que componen cada conjunto pueden ser evaluados con tests de significación, paramétricos o no paramétricos, mediante la adecuada transformación de los datos, si es necesario. Van Tongeren (1987) recomienda la U de Mann-Whitney como método de distribución libre. Para éste y otros métodos no parámetricos el lector puede consultar a Siegel (1985), que constituye un clásico del tema.

Al margen de cualquier prueba estadística pueden calcularse simplemente los estadígrafos clásicos con fines descriptivos. Así, los grupos pueden ser comparados en sus valores medios, el nivel de sobrelapamiento de sus valores máximos y mínimos o de sus intervalos de confianza, si se calculan la varianza y la desviación estándar, que asimismo nos darán un índice del grado de homogeneidad de nuestros grupos al ser estimadores de la dispersión.

Everitt y Dunn (1991) comentan algunas aproximaciones gráficas que sirven para evaluar la cohesión interna de los grupos, propiedad muy deseable dentro de la clasificación. En esencia estos métodos gráficos se basan en analizar los datos originales, crudos o aplicando algún estadígrafo, dentro de los grupos propuestos. Una prueba estadística valiosa para la evaluación de diferencias entre grupos es el análisis discriminante, cuya utilidad para la reasignación ya discutimos.

En el presente no están desarrollados métodos estadísticos dirigidos específicamente a examinar las diferencias entre grupos, por lo que en primera instancia este juicio corresponde al investigador. Un análisis previo de la matriz de datos (ver Tablas 3.3 y 3.4 en el Capítulo 3) ya podría indicar si su estructura merece ser sometida a un análisis de clasificación, de la misma forma que la escala de variación de la afinidad puede dar un criterio de diferencias entre grupos, como comentamos al hablar de las reglas de decisión.

**Relación de las clasificaciones con factores externos**

En los estudios ecológicos el interés de formar grupos a partir de la matriz original de datos no es un hecho formal sino que responde siempre a ciertas consideraciones acerca de cómo distintos factores ambientales están haciendo variar la estructura de las comunidades, bien sea en sentido espacial o temporal. Las vías para relacionar los resultados de la clasificación con factores externos (entiéndase variables bióticas o abióticas) están limitadas solamente por la imaginación del investigador (Boesch, 1977).

Criterios sencillos como la mapificación de los grupos para analizarlos visual o estadísticamente respecto a los factores ambientales, los gráficos conjuntos del dendrograma y los valores de las variables correspondientes a cada grupo o la ubicación bajo el árbol de un esquema que resuma el gradiente de variación del factor determinante de la clasificación, devienen en herramientas interpretativas sencillas que además tienen un alto valor para la expresión gráfica de resultados, como ejemplificaremos en nuestro último capítulo.

Una variante útil, lo es también relacionar paralelamente la clasificación de los datos de la comunidad y la de aquellos factores bióticos o abióticos asociados al estudio, para comparar posteriormente la estructura de ambos árboles. Clarke y Ainsworth (1993) proponen calcular la disimilitud de Bray-Curtis con los datos de la estructura de la comunidad y la distancia euclidiana con las variables ambientales, y posteriormente calcular la correlación por rangos entre ambas matrices sometiendo estos resultados al ordenamiento por escalado multidimensional. Para este fin, el interesado puede acudir a varios textos que resumen métodos de presentación de resultados (Digby y Kempton, 1991; Everitt, 1993; Hair *et al*., 1995).

**Analisis nodal**

La matriz original de datos es portadora de una doble información que puede ser revelada realizando el análisis normal e inverso, los cuales brindan aisladamente, información sobre la estructura de grupos de estaciones y especies, respectivamente. Sin embargo, un importante salto en la interpretación de la información se logra cuando relacionamos ambas a través del análisis nodal (ver Boesch, 1977), cuya ejecución ejemplificaremos en la Fig. 6.5.

**A.**

| | Estaciones | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Especies | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| A | 30 | 5 | 42 | 0 | 51 | 0 | 0 | 0 |
| B | 0 | 0 | 0 | 5 | 0 | 2 | 3 | 0 |
| C | 2 | 1 | 1 | 1 | 0 | 0 | 2 | 0 |
| D | 25 | 0 | 22 | 0 | 30 | 0 | 0 | 2 |
| E | 0 | 0 | 0 | 9 | 0 | 7 | 5 | 0 |
| F | 0 | 15 | 0 | 0 | 1 | 0 | 0 | 13 |
| G | 3 | 0 | 1 | 1 | 1 | 0 | 3 | 3 |
| H | 0 | 1 | 0 | 1 | 0 | 8 | 6 | 0 |
| I | 0 | 10 | 0 | 1 | 0 | 0 | 0 | 9 |
| J | 0 | 0 | 0 | 10 | 0 | 5 | 2 | 0 |
| K | 0 | 0 | 0 | 2 | 0 | 1 | 11 | 0 |
| L | 0 | 0 | 0 | 7 | 0 | 5 | 4 | 0 |

**B.**

| | Estaciones | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Especies | 1 | 3 | 5 | 2 | 8 | 4 | 6 | 7 |
| A | 30 | 42 | 51 | 5 | 0 | 0 | 0 | 0 |
| D | 25 | 22 | 30 | 0 | 2 | 0 | 0 | 0 |
| F | 0 | 0 | 1 | 15 | 13 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 10 | 9 | 1 | 0 | 0 |
| J | 0 | 0 | 0 | 0 | 0 | 10 | 5 | 2 |
| E | 0 | 0 | 0 | 0 | 0 | 9 | 7 | 5 |
| L | 0 | 0 | 0 | 0 | 0 | 7 | 5 | 4 |
| B | 0 | 0 | 0 | 0 | 0 | 5 | 2 | 3 |
| H | 0 | 0 | 0 | 1 | 0 | 1 | 8 | 6 |
| K | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 11 |
| C | 2 | 1 | 0 | 1 | 0 | 1 | 0 | 2 |
| G | 3 | 1 | 1 | 0 | 3 | 1 | 0 | 3 |

**C.**

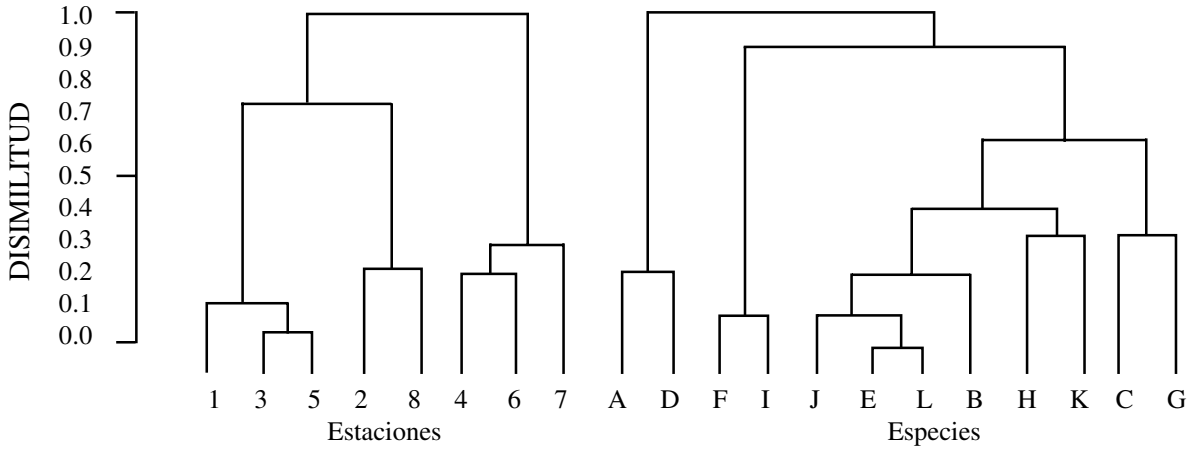



Figura 6.5. **A**. Matriz original de datos. **B.** Matriz de datos reordenada sobre la base de los grupos obtenidos en las clasificaciones. **C.** Dendrogramas de las clasificaciones normal (estaciones) e inversa (especies).

Procedemos entonces a reordenar la matriz original de datos (Fig. 6.5A) sobre la base de las nuevas agrupaciones opbtenidas (Fig. 6.5.C), delimitando con líneas horizontales y verticales a través de la tabla, la extensión de los grupos. Este paso definirá los nodos que no son mas que pequeñas matrices

dentro de la matriz total donde coinciden exactamente los datos correspondientes a un grupo de especies con un grupo de estaciones. En el ejemplo de la Fig. 6.5B se han delimitado doce nodos. El primero, por ejemplo, relaciona el grupo de las especies A y D con el grupo de las estaciones 1, 3 y 5 donde es claro que la unión responde a una alta abundancia de estas dos especies. El segundo a la derecha, relaciona el mismo grupo de especies con el grupo de estaciones 2 y 8, donde la abundancia de éstas es menor, y el tercero con las estaciones 4, 6 y 7, donde este grupo de especies está ausente.

Con un propósito gráfico la matriz original así subdividida puede ser llevada a un rectángulo con las dimensiones adecuadas y las divisiones correspondientes que representen las relaciones grupos de especies-grupos de estaciones. Como las dimensiones de cada nodo se ajustan al número de entidades de cada grupo de estaciones y especies, ya brinda de entrada una información sobre la riqueza de especies en los conjuntos de estaciones. Dentro de cada nodo, cuya información cuantitativa y cualitativa se conoce, pueden calcularse diferentes índices que expresen el comportamiento de los grupos de especies en los grupos de estaciones. Tal es el caso de la *constancia nodal***,** definida como:

$$C_{ij} = A_{ij}/N_i N_j,$$

donde $A_{ij}$ es el número de ocurrencias de los miembros del grupo i de especies en el grupo j de estaciones, y $N_i$ y $N_j$ son respectivamente el número de entidades de cada grupo. En otras palabras, la constancia no es más que el número de ocurrencias reales dentro del nodo, dividido entre todas las posibles ocurrencias si el nodo hubiera estado «lleno». Este índice cualitativo varia entre 0, cuando ninguna de las especies del grupo considerado está en el grupo de estaciones que se analiza; y 1 cuando todas las especies están representadas. Multiplicando este valor por 100 la constancia puede ser expresada en porcentajes.

Veamos ahora un ejemplo del cálculo de la constancia pero antes aclaremos que para ello es útil representar en los nodos la composición cualitativa de la tabla (con cruces), pues este índice opera con la información cualitativa (Fig. 6.6). Por ejemplo, calculemos las constancias para los tres nodos superiores, que tienen en común el valor de $N_j = 2$, ya que en los tres, el grupo de especies está compuesto por la A y la D, y varían en los valores de número de ocurrencias y número de estaciones de cada grupo:

| Nodo izquierdo | Nodo central | Nodo derecho |
|---|---|---|
| $A_{ij} = 6$ | $A_{ij} = 2$ | $A_{ij} = 0$ |
| $N_j = 3$ | $N_j = 2$ | $N_j = 3$ |
| $C_{ij}=6/3x2$ | $C_{ij}=2/2x2$ | $C_{ij}=0/2x3$ |
| $C_{ij} = 1$ | $C_{ij} = 0.5$ | $C_{ij} = 0$ |

| | Estaciones | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Especies | 1 | 3 | 5 | 2 | 8 | 4 | 6 | 7 |
| A | X | X | X | X | | | | |
| D | X | X | X | | X | | | |
| F | | | X | X | X | | | |
| I | | | | X | X | X | | |
| J | | | | | | X | X | X |
| E | | | | | | X | X | X |
| L | | | | | | X | X | X |
| B | | | | | | X | X | X |
| H | | | | X | | X | X | X |
| K | | | | | | X | X | X |
| C | X | X | | X | | X | | X |
| G | X | X | X | | X | X | | X |

Figura 6.6. Representación de la presencia-ausencia de las especies en la matriz de datos reordenada.

El cálculo de todas las constancias nos daría el gráfico de constancia nodal de la Fig. 6.7 donde se han representado gráficamente los valores del índice calculado dentro de los nodos con una leyenda que explica sus intervalos de variación. Generalmente se establece un gradiente de colores donde los valores más bajos tienen los tonos más claros y los más altos los más oscuros. Al valor de 1, o 100 si la constancia se expresa en porcentajes, se le asignaría el color negro.

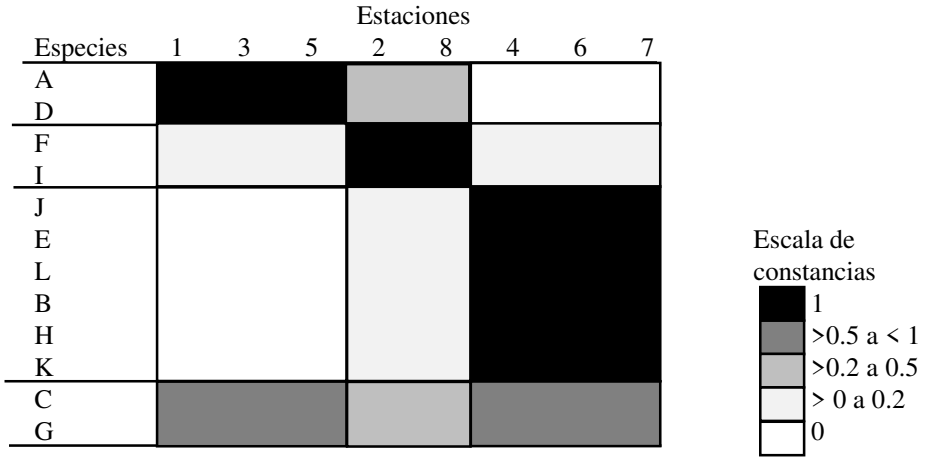



Figura 6.7. Representación gráfica de las constancias calculados a partir de la información de la Fig. 6.6.

Si consideramos ahora no solo el comportamiento del grupo de especies en su grupo de estaciones sino que lo analizamos a lo largo de todos los grupos de estaciones calcularíamos entonces la *fidelidad* que da una medida del grado en el cual las especies "seleccionan" o están limitadas a determinadas estaciones. La fidelidad se obtiene dividiendo la constancia para un nodo entre la constancia para todos los conjuntos de estaciones. o sea:

$$F_{ij} = A_{ij} / N_i N_j / \Sigma A_{ij} / N_i \Sigma EN_j$$
$$F_{ij} = A_{ij} \Sigma N_j / N_j \Sigma A_{ij}$$

El análisis nodal puede ser un último eslabón del proceso clasificatorio que permite extraer el máximo de información de la matriz original de datos y complementar además la interpretación aislada de las clasificaciones normal e inversa. Recordemos su importancia para la reasignación ya que al encerrar de manera global la información de toda la tabla, y de manera aislada la subdivisión de los valores originales por grupos, permite un análisis más objetivo acerca de la distribución de las entidades.

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

*"Gris es, amigo, toda teoría; verde el árbol dorado de la vida"*
*Juan Wolfgang Goethe*

# 7. LA CLASIFICACIÓN EN LA PRÁCTICA

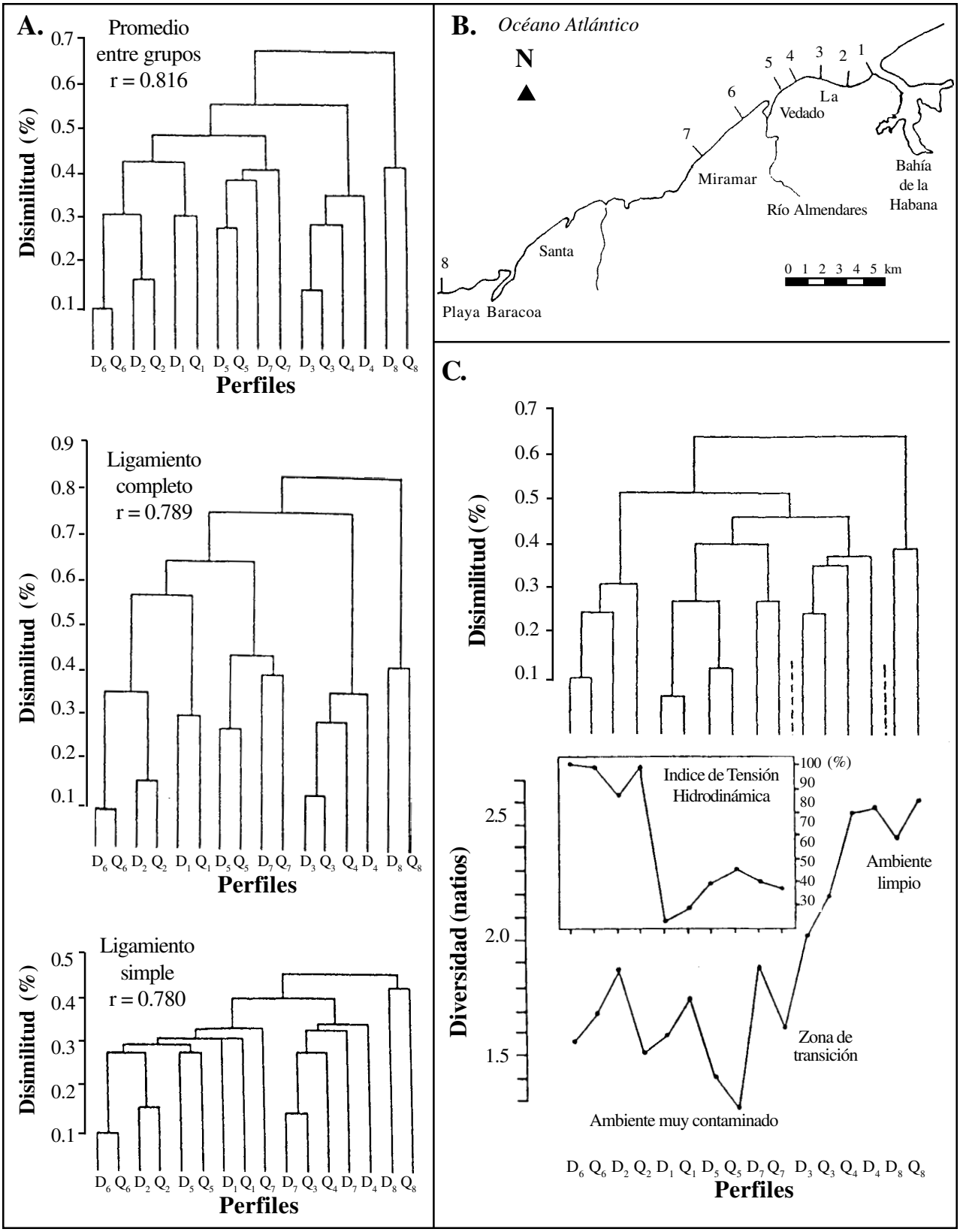
No vamos a terminar sin que veamos juntos, ya todos más conocedores del tema, algunos ejemplos de ecología marina donde se han aplicado las distintas técnicas aquí discutidas. En ellos verán representados prácticamente no solo lo dicho en la secuencia de pasos, sino también diferentes maneras de cómo ajustar los métodos a distintas situaciones, diversas alternativas de análisis, formas gráficas de representar y comparar, y explicaciones ecológicas derivadas de los resultados de las clasificaciones.

**Ejemplo 1.** *Estructura ecológica de las comunidades de gorgonáceos en un gradiente de contaminación en los arrecifes coralinos del litoral de La Habana, Cuba.*

Para estudiar los efectos de la contaminación sobre el arrecife costero del litoral Norte de la Habana, afectado por la carga contaminante proveniente de la Bahía de la Habana y del Río Almendares, Herrera y Alcolado (1983) evaluaron la estructura ecológica de las comunidades de gorgonáceos en 10 y 15 m de profundidad, en ocho perfiles (Fig. 7.1B), representativos de ambientes limpios y contaminados. Estandarizando en proporciones los datos originales de número de individuos/ especie, -debido al método de muestreo empleado (recorrido isobático) y por la desigualdad en los esfuerzos de muestreo entre estaciones limpias y contaminadas (por las diferencias de abundancia)-; se calculó la disimilitud de Sanders (1960) y se realizaron los agrupamientos por los métodos de ligamiento simple, completo y promedio entre grupos; para cuyos resultados se calculó el coeficiente de correlación cofenético.

Ante todo detengámonos en los tres dendrogramas (Fig. 7.1A) para comentar lo discutido acerca de las propiedades de cada estrategia. Gráficamente es claro el efecto contractivo del ligamiento simple que no llega a brindar definiciones tan claras como los otros dos métodos, cuyos resultados son muy similares. La estrategia de promedio brindó el mayor valor del coeficiente de correlación cofenético por lo que decidimos escogerlo y reordenar las entidades mediante giros en el internodo que llega al valor 0.55 de disimilitud, obteniendo así un orden de los perfiles acorde al gradiente de severidad de la contaminación (Fig. 7.1C).

Figura 7.1. **A**. Dendrogramas obtenidos con tres técnicas de agrupamiento en la clasificación de los datos Ü porcentuales de las comunidades de gorgonáceos en 10 (D) y 15 (Q) metros de profundidad, en ocho perfiles de los arrecifes de la costa Norte de la Habana, Cuba. El valor r es el coeficiente de correlación cofenético. **B.** Ubicación de los perfiles en el área de estudio en un gradiente de contaminación respecto a la Bahía de la Habana y el Río Almendares, dos importantes fuentes de contaminación orgánica del litoral habanero. **C.** Dendrograma de promedio entre grupos, reordenado según el gradiente de contaminación y en relación con el índice de tensión hidrodinámica y la diversidad.


A.
Promedio entre grupos r = 0.816
Disimilitud (%)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
$D_6$ $Q_6$ $D_2$ $Q_2$ $D_1$ $Q_1$ $D_5$ $Q_5$ $D_7$ $Q_7$ $D_3$ $Q_3$ $Q_4$ $D_4$ $D_8$ $Q_8$
Perfiles
Ligamiento completo r = 0.789
Disimilitud (%)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
$D_6$ $Q_6$ $D_2$ $Q_2$ $D_1$ $Q_1$ $D_5$ $Q_5$ $D_7$ $Q_7$ $D_3$ $Q_3$ $Q_4$ $D_4$ $D_8$ $Q_8$
Perfiles
Ligamiento simple r = 0.780
Disimilitud (%)
0.5
0.4
0.3
0.2
0.1
$D_6$ $Q_6$ $D_2$ $Q_2$ $D_5$ $Q_5$ $D_1$ $Q_1$ $Q_7$ $D_7$ $Q_3$ $Q_4$ $D_7$ $D_4$ $D_8$ $Q_8$
Perfiles
B.
Océano Atlántico
N
1
2
3
4
5
6
7
8
La Vedado
Miramar
Río Almendares
Bahía de la Habana
Santa
Playa Baracoa
0 1 2 3 4 5 km
C.
Disimilitud (%)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Indice de Tensión Hidrodinámica
100 (%)
90
80
70
60
50
40
30
Diversidad (natios)
2.5
2.0
1.5
Ambiente limpio
Zona de transición
Ambiente muy contaminado
$D_6$ $Q_6$ $D_2$ $Q_2$ $D_1$ $Q_1$ $D_5$ $Q_5$ $D_7$ $Q_7$ $D_3$ $Q_3$ $Q_4$ $D_4$ $D_8$ $Q_8$
Perfiles

En este gradiente las profundidades correspondientes a un mismo perfil: 10 y 15 m, guardaron siempre una disimilitud entre sí inferior a la diferencia entre perfiles indicando que los cambios comunitarios se manifiestan fundamentalmente más en sentido horizontal (distancia a la fuente contaminante) que en la vertical (batimétricamente). Por ello en adelante nos referiremos solamente a los perfiles por su número. Los dos primeros grupos; formado uno por los perfiles más cercanos a la bahía (1 y 2), y el río (6); y el otro por perfiles algo más alejados del puerto (3 y 4); definen el sector más contaminado, donde los fenómenos de sedimentación, turbidez, incremento de la cobertura vegetal del fondo, entre otros factores, imponen condiciones extremas al arrecife que se refleja en la estructura de la comunidad coralina. En esta zona, los valores de diversidad son los más bajos y dominan especies adaptadas morfológica y fisiológicamente para tolerar el ambiente tensado.

La subdivisión en dos conjuntos, aún cuando todas las estaciones son representativas de condiciones ambientales muy desfavorables, quedó explicada cuando analizamos los valores del índice de tensión hidrodinámica, pues debido a cambios en la orientación de la costa el batimiento es muy intenso en los perfiles 3 y 4, y moderado o bajo en las restantes. Esto hace que del conjunto de especies resistentes a la contaminación dominen en los perfiles 3 y 4 aquellas tolerantes a regímenes hidrodinámicos severos (*Plexaura flexuosa, Eunicea mammosa*, *E. tourneforti* **y** *E. calyculata*) mientras que en los perfiles 1, 2 y 6, especies representativas de ambientes poco agitados como *Plexaura homomalla* forma *kuekenthali* devienen en las indicadoras de contaminación.

En los restantes perfiles la diversidad aumenta sugiriendo condiciones más favorables y se definen claramente dos conjuntos: uno formado por perfiles más alejados en los gradientes de la bahía (5) y el río (7) donde aún prevalecen, aunque con menores valores especies relacionadas con la contaminación; y el otro (8) mucho más distante ubicado en un área limpia con asociaciones típicas de arrecifes costeros naturales. La clasificación de las comunidades de gorgonáceos brindó elementos claros acerca de la zonación ecológica del litoral, la dirección Oeste del gradiente, en concordancia con el sistema de circulación costero y la extensión de los efectos.

**Ejemplo 2.** *Estudio de la comunidad de corales escleractíneos en los arrecifes coralinos del borde de la plataforma Suroccidental de Cuba.*

Como parte del estudio del hábitat de la langosta *Panulirus argus* (Herrera *et al*., 1991) en el borde de la plataforma Suroccidental de Cuba, se estudió la comunidad de corales en catorce perfiles (Fig. 7.2B) mediante identificación y conteo *in situ*, en recorridos isobáticos. Con igual método se estudiaron, solo en diez perfiles, las esponjas y los gorgonáceos, pero para estos últimos se cortaron 250 ramas de las distintas colonias para su identificación en el laboratorio.

Los datos de corales y esponjas fueron estandarizados en porcentajes y calculada la afinidad con el índice de disimilitud de Sanders (1960). Con los gorgonáceos ocurrió que las ramas se fragmentaron durante su conservación y aunque eran obvias las tendencias de abundancia, no era posible determinar el número exacto de individuos por especies para el cálculo de porcentajes que requiere este índice. Una alternativa era emplear solamente la información de presencia-ausencia, pero con el interés de no

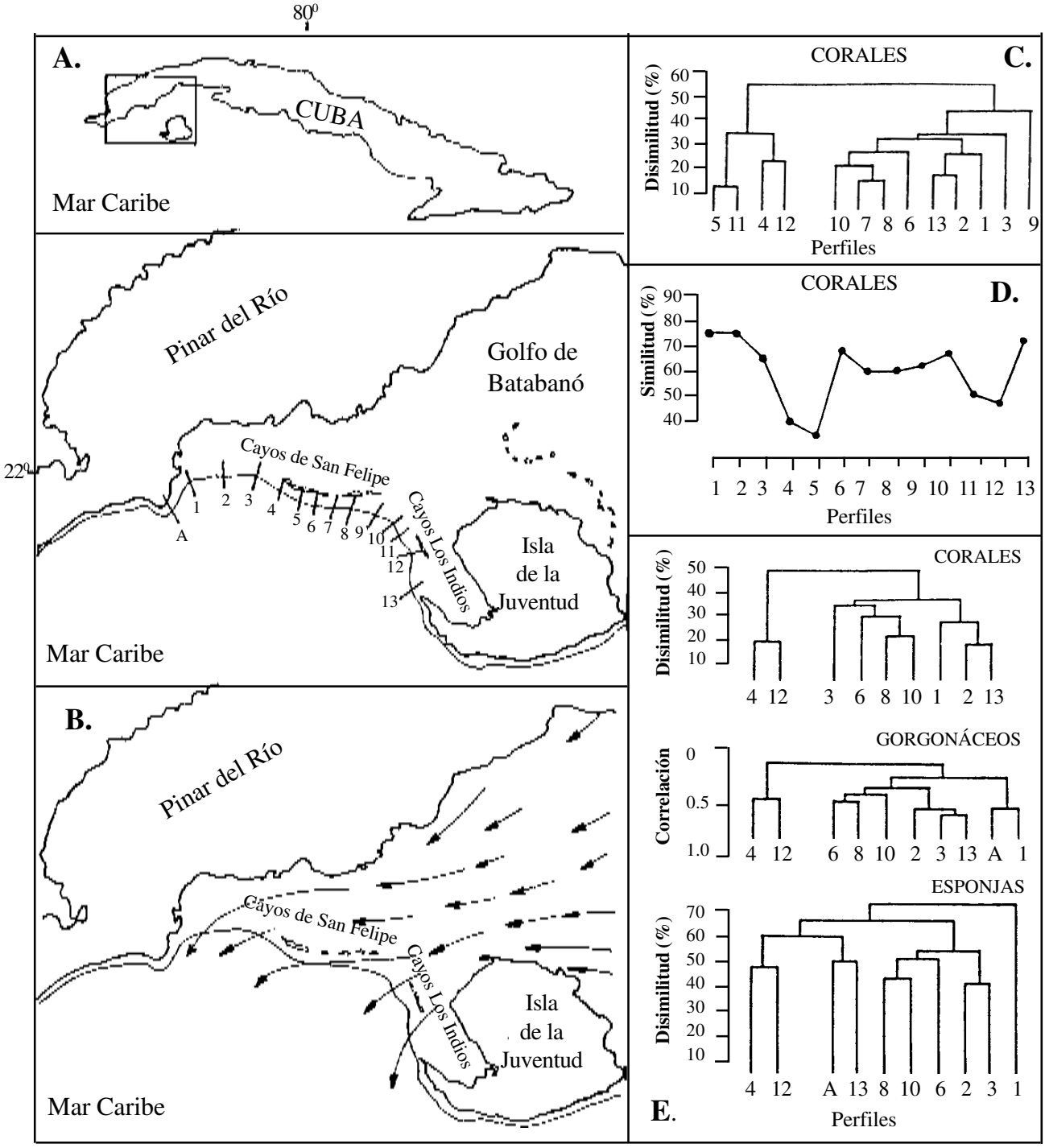



Figura 7.2. **A.** Perfiles de muestreo de los arrecifes coralinos en el borde de la plataforma Suroccidental de Cuba. **B.** Sistema de corrientes de la macrolaguna del Golfo de Batabanó. **C.** Dendrograma con los datos porcentuales de las comunidades de corales mostrando la separación de perfiles protegidos (izquierda) e influídos por las aguas interiores (derecha). **D.** Proyección de similaridad cenoclínica al comparar el Perfil 1 con los restantes. **E.** Comparación interclasificatoria de los árboles de los tres grupos sésiles.

desaprovechar el componente cuantitativo decidimos asignar rangos a las especies y emplear como medida de afinidad la correlación de Spearman. La técnica de agrupamiento empleada fue la de promedio simple.

Una primera clasificación de la comunidad coralina separó el borde de la plataforma en dos sectores (Fig. 7.2C): uno que comprendía cuatro perfiles protegidos tras los costados oceánicos de cierta parte de las cayerías (4, 5, 11 y 12); y otro que agrupaba a los nueve perfiles restantes (1, 2, 3, 6, 7, 8, 9, 10 y 13), situados en sectores sin cayos o donde éstos, por su posición, permitían el libre contacto de las aguas interiores del Golfo de Batabanó con el océano, de acuerdo al sistema de corrientes imperante (Fig. 7.2B). La proyección de similaridad cenoclínica (Fig. 7.2D), donde se compararon todos los perfiles respecto al 2 (representativo del sector abierto) mostró también, que en un tránsito por el borde de la plataforma, de Oesta a Este, existía un cambio en la comunidad coralina en los perfiles 7 y 8, al Suroeste de la Cayería de San Felipe, y en el 14 y 15 al Sur Suroeste de la Cayería de Los Indios.

La comparación interclasificatoria de los dendrogramas de los tres taxones estudiados (Fig. 7.2E) confirmó que los cambios que se observaban en la comunidad coralina se repetían en otros importantes miembros de la biocenosis sésil (esponjas y gorgonáceos), lo que sin duda estaba reflejando diferencias sectoriales en el borde de la plataforma que concernían al complejo arrecifal como un todo.

Los resultados de la clasificación fueron analizados a la luz del sistema de circulación del Golfo de Batabanó (Fig. 7.2B) que por efecto de los vientos del Este y el Noreste presentan predominantemente una componente con sentido Oeste de forma tal que las aguas interiores alcanzan el borde oceánico acarreando consigo altas concentraciones de nutrientes, materia orgánica y partículas suspendidas. Esta influencia se reduce o desaparece en aquellos sectores donde las cayerías constituyen una barrera natural a la influencia de las aguas interiores.

El análisis de los datos ecológicos obtenidos junto a las numerosas observaciones de buceo realizadas directamente y desde el sumergible ARGUS, mostraron en las zonas protegidas un alto desarrollo arrecifal con dominancia del morfotipo profundo de *Montastraea annularis* que creciendo en forma de platos, junto a otras especies brindaba una cobertura del fondo de entre 50 y 66%. En las zonas de influencia no existe desarrollo arrecifal y los corales, representados por pequeñas incrustaciones de *Siderastraea radians* o *Montastraea cavernosa*, contribuyen poco a la cobertura del fondo (entre 4 y 14%) que en su mayor parte está cubierto de sedimento y/o vegetación, reflejando el efecto de las aguas interiores que promueven la turbidez, la abrasión por las partículas arrastradas por las fuertes corrientes y en ciertas zonas los fenómenos de sedimentación.

El estudio de la fauna arrecifal mediante técnicas de clasificación numérica brindó un panorama claro de cómo la influencia de las aguas interiores de la macrolaguna, incide decisivamente en la estructura del arrecife del borde de la plataforma Suroccidental de Cuba, delimitando zonas ecológicas particulares.

**Ejemplo 3.** *Tipificación de biotopos en la Bahía de Cárdenas en Cuba, a través de la estructura ecológica de sus comunidades de bivalvos.*

Al emplear las comunidades de moluscos bivalvos como bioindicadores de las condiciones ambientales imperantes en la Bahía de Cárdenas, Herrera y Espinosa (1988) realizaron la clasificación normal de veintitrés estaciones, empleando los datos de número de individuos de cuarenta y tres especies estandarizados en proporciones, empleando el índice de Sanders (1960) y con la técnica de promedio simple.

Los resultados del dendrograma (Fig.7.3A), junto con los datos de granulometría de los sedimentos y la presencia y tipo de vegetación, permiteron delinear los principales biotopos (Fig. 7.3C.), que aunque estrictamente delimitados en la Figura, los cambios entre zonas se manifiestan mas bien como transiciones graduales con variaciones paulatinas en los porcentajes de especies. Así, *Brachidontes modiolus* domina en el biotopo de fango microalevrítico con elevada contaminación orgánica, para dar paso a *Chione cancellata* que predomina en toda la región de fango microalevrítico y mantiene porcentajes importantes en el de fango alevrítico arcilloso, donde se incrementa la importancia de *Tagelus divisus* y *Corbula caribaea*.

En el biotopo de fango microalevrítico con *Thalassia testudinum*, éstas dos últimas especies definen aún más su dominancia y comienza a ser más abundante *Tellina alternata*. En los fondos arenosos, *Codakia orbiculata* y *Trachycardium muricatus* dominan en la pradera de fanerógamas para ser sustituídos por *Macrocalista maculata* en la zona arenosa sin vegetación.

Esta sucesión que se manifiesta igualmente en los índices ecológicos de diversidad, equitatividad y dominancia es un reflejo de preferencias por el sustrato según las adaptaciones de las especies y su grado de tolerancia dentro del acusado gradiente del régimen hidrológico que impera entre el interior de la bahía y la región oceánica colindante, en virtud del carácter de estuario negativo de este acuatorio. En el orden cualitativo, dado que algunas estaciones individuales no cumplían el requisito de tamaño de muestra, se unieron todas las correspondientes a un mismo biotopo para ser analizadas por técnicas de promedio con el índice de Sorensen (Fig. 7.3B), observándose que la composición cualitativa está relacionada más con el tipo de sedimento que con la vegetación.

Aprovechando el valor bioindicativo de la fauna de bivalvos, y su relación con los factores ambientales determinantes de su distribución, la clasificación de sus comunidades en la Bahía de Cárdenas permitió tipificar los biotopos fundamentales, que fueron corroborados posteriormente por otros autores mediante fotointerpretación, con resultados sorprendentemente similares.

**Ejemplo 4**. *Reclasificación de la biodiversidad coralina caribeña incluyendo los datos de la Hispaniola en la matriz de Chiappone et al. (1996).*

El presente ejemplo tiene el propósito de ilustrar el valor de repetibilidad de la clasificación numérica. Chiappone *et al.* (1966), para su clasificación de las comunidades coralinas del Caribe y el Atlántico, confeccionan una matriz de presencia-ausencia con datos tomados de la literatura para unas veinte

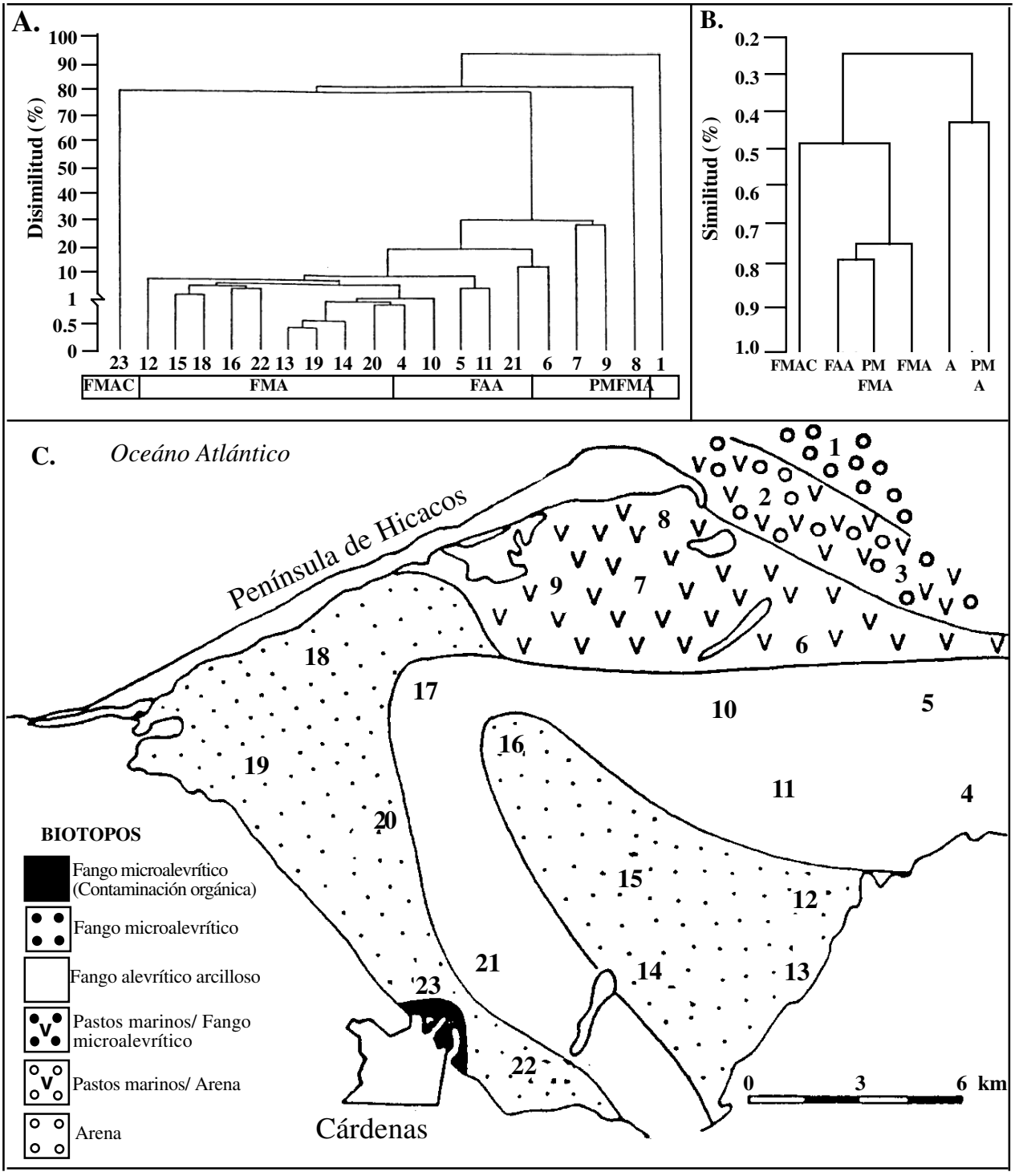



Figura 7.3. Clasificación normal de veintitrés estaciones de la Bahía de Cárdenas en Cuba, considerando los datos de las especies de moluscos bivalvos. **A**. Dendrograma con datos porcentuales. **B**. Dendrograma con datos cualitativos **C.** Mapa de biotopos en la Bahía de Cárdenas obtenido a partir de las clasificaciones. Las letras indican: FMAC: Zona contaminada de fango microalevrítico; FMA: Fango microalevrítico; FAA: Fango alevrítico arcilloso; PMFMA: Pastos marinos sobre fango microalevrítico; PMA: Pastos marinos sobre arena; A: Arena.

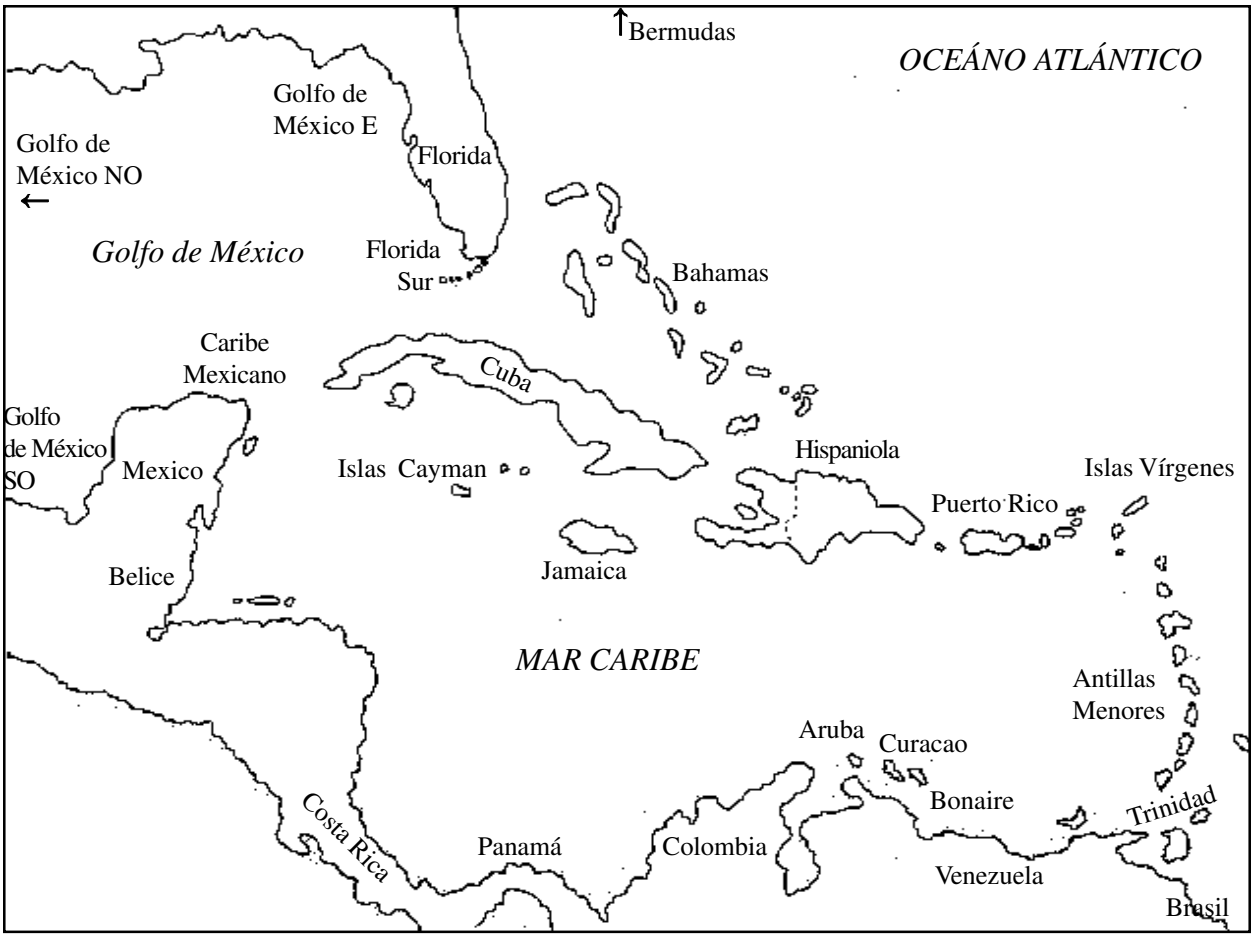



Figura 7.4. Mapa indicando las veinte localidades consideradas por Chiappone *et al.* (1996) en su clasificación de los datos de presencia-ausencia de las especies coralinas de varios sistemas arrecifales atlánticos. En el presente ejemplo se añaden los datos de los arrecifes de la Hispaniola.

localidades (Fig. 7.4), donde no incluyen la Hispaniola. A esta matriz, compuesta por 65 especies y veinte localidades añadimos una nueva columna correspondiente a nuestra recopilación sobre la fauna coralina de la Hispaniola que incluye unas 48 especies (de las listadas por los autores) y realizamos nuevamente la agrupación empleando sus mismos métodos: el índice de similitud de Jaccard, expresado en porcentajes, y la estrategia de promedio de grupos para los agrupamientos, que fueron representados gráficamente en un dendrograma.

Si comparamos el dendrograma original de Chiappone *et al.* (1996) (Fig. 7.5A) con el obtenido tras incorporar la información de la Hispaniola (Fig. 7.5B) vemos que los cinco grupos de localidades, indicadores de las afinidades en la distribución de las especies coralinas, se mantienen en ambos. Así podríamos subdividir ambos dendrogramas en: I. Grupo donde quedan aislados los sitios del Golfo de México, II. Grupo de localidades caribeñas, donde quedó incluida la Hispaniola, que forman un núcleo de cerca de un 70% de afinidad, III. Grupo que incluye solo a Bermudas, IV. Grupo formado por Venezuela y Trinidad y V.. Grupo de Brasil. Comparando con el grupo de localidades caribeñas, la similaridad se va haciendo menor hacia las mayores latitudes del Golfo de

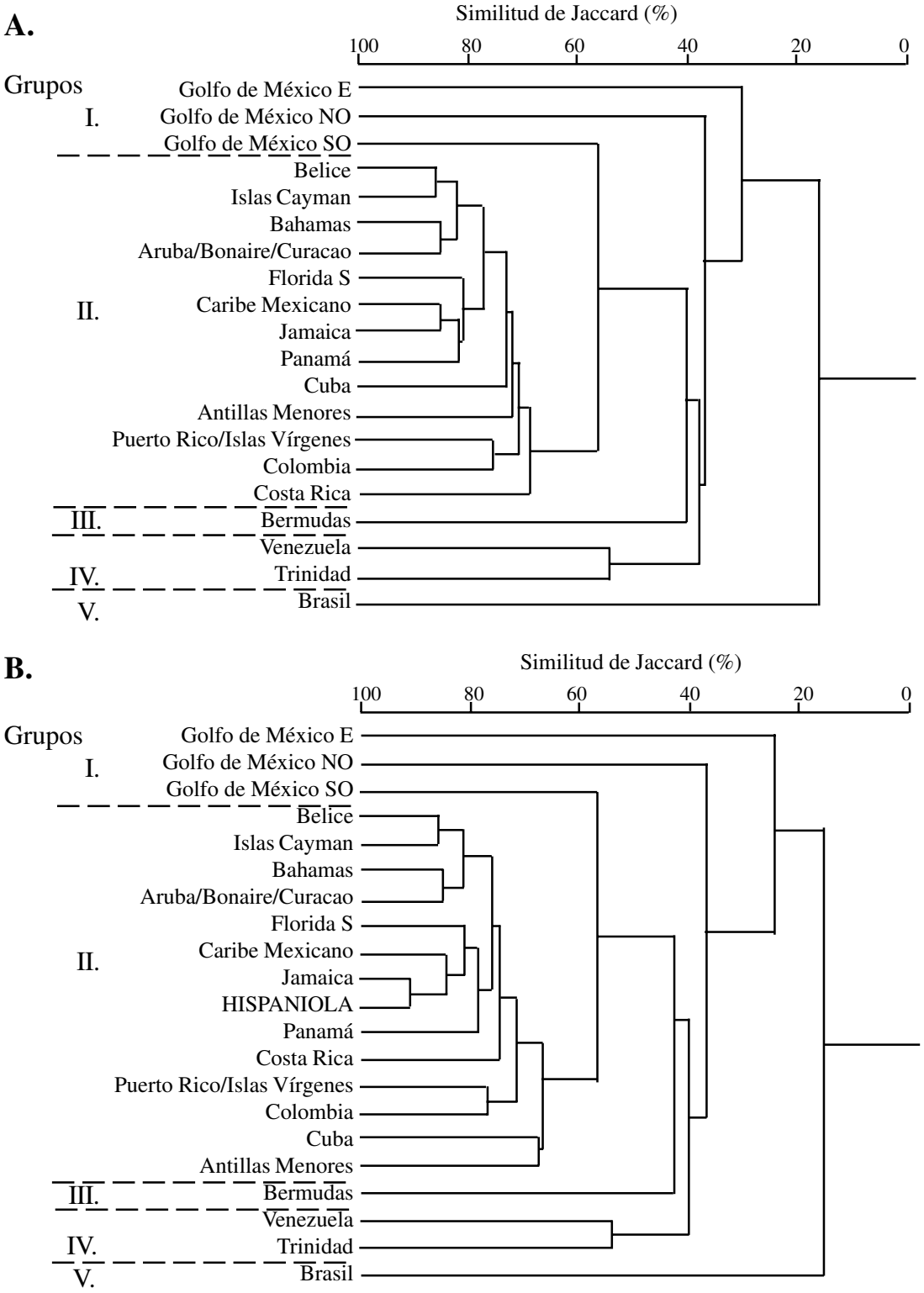



Figura 7.5 A. Dendrograma obtenido por Chiappone *et al.* (1996) en su clasificación de los datos de presencia-ausencia de especies coralinas de varios sistemas arrecifales atlánticos. B. Reclasificación de los datos de los autores incorporando los datos de presencia-ausencia de las especies coralinas de la Hispaniola.

México y Bermudas o hacia latitudes mas bajas como Venezuela, Trinidad y Brasil. Estos resultados coinciden con los criterios de subdivisión de provincias zoogeográficas en la región del Caribe y el Atlántico (Schuhmacher, 1978) y pueden hacerse evidentes cuando se hace la proyección de similaridad cenoclínica (Fig. 7.6), donde se compara la similitud entre la Hispaniola y las restantes localidades.

La Provincia indoocidental del Caribe, incluye de Norte a Sur a la Florida y el Golfo de México, el Archipiélago de las Bahamas, las Antillas Mayores y Menores, el Caribe Central, las costas de Centroamérica y parte de las de Suramérica (Wood, 1983), pero su área más representativa es el Mar del Caribe (Achituv y Dubinsky, 1990), donde se registra la mayor riqueza de especies (Chiappone *et al.,* 1996). Esta riqueza disminuye en la costa del Golfo de México y hacia los extremos de distribución particularmente en la Provincia arrecifal brasileña, que comparativamente con los arrecifes caribeños, es pobre en especies coralinas y con un gran endemismo (Schuhmacher, 1978). De acuerdo a nuestros resultados, dentro de las Antillas Mayores, la biodiversidad coralina de la Hispaniola guarda una afinidad de 78% con Puerto Rico, 80% con Cuba y 93% con Jamaica, que es uno de los países donde los arrecifes han sido mejor estudiados.

Con el presente ejemplo hemos destacado que cuando se conocen las técnicas empleadas por otros autores es posible repetir los mismos procedimienentos y como en este caso incorporar nuevos datos que deseen ser comparados. El dendrograma de la Fig. 7.5.B complementa el obtenido por los autores del trabajo original al introducir información de una de las islas más importantes de la Antillas Mayores cuya biodiversidad coralina ha sino bien estudiada.

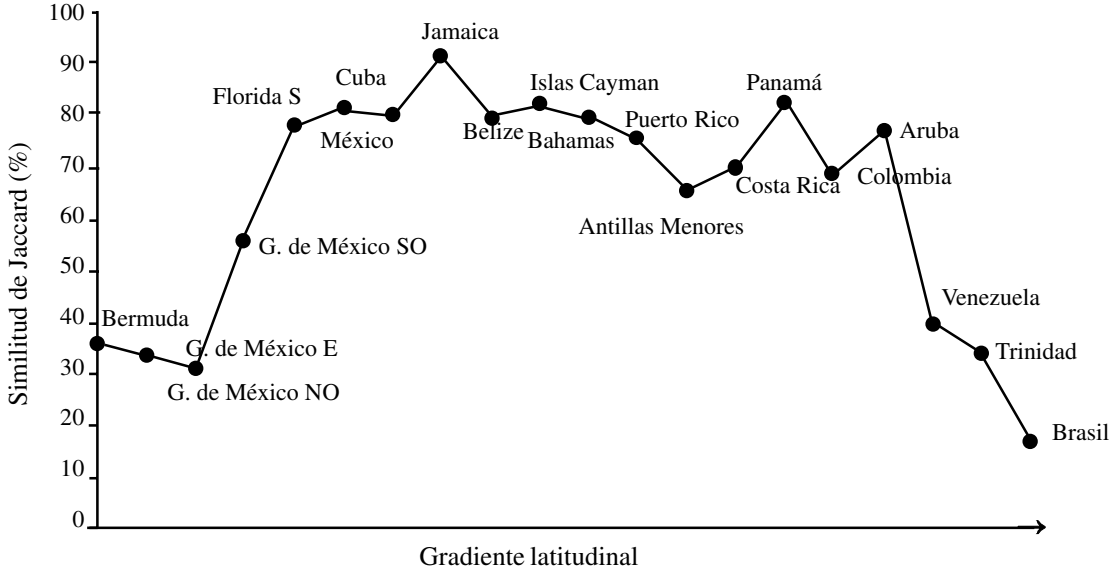



Figura 7.6. Proyección de similaridad cenoclínica comparando la lista de especies de corales escleractíneos de la Hispaniola con las de veinte localidades de la región arrecifal atlántica.

**Ejemplo 5.** *Clasificación de datos de las pesquerias de Samaná bajo el concepto de los complejos ecológicos de pesca.*

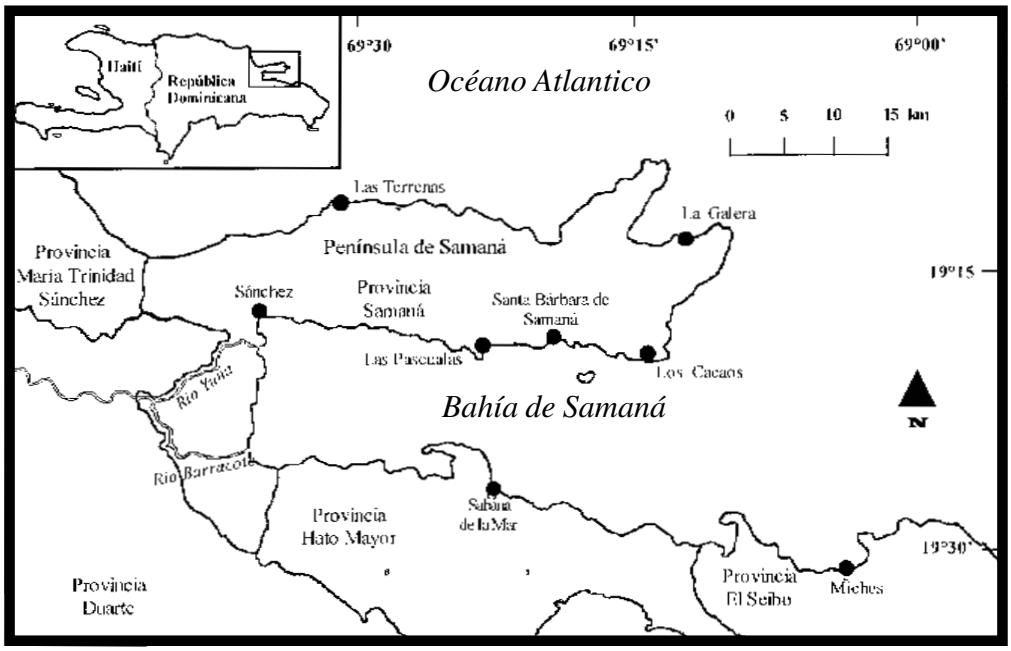
La clasificación de las familias de peces y crustáceos representadas en las pesquerías y los sitios de desembarco de la región de Samaná y el análisis de estos resultados a la luz del concepto de los complejos ecológicos de pesca de Baisre (1985)[3], brinda un interesante ejemplo de la aplicación de este tipo de técnicas. La información aquí manejada, sobre las capturas en ocho sitios de desembarco, proviene de los resultados de Sang *et al.* (1997) (ver Tabla 3.4). En el análisis normal, clasificamos la información cualitativa (Fig. 7.7.A) con el índice de Sorensen (1948) y la información porcentual (Fig. 7.7.B) a través del índice de Sanders (1960). En el análisis inverso, clasificamos los datos cualitativos por familias (ver Tabla 3.4) representadas en las capturas (Fig. 7.7C) con el índice de Sorensen (1948). La estrategia de agrupamiento empleada en ambos casos fue la de promedio simple.

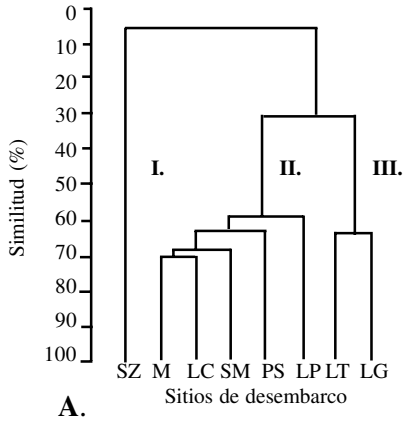
Los datos originales sobre 56 familias de peces y 10 de invertebrados fueron reducidos aplicando el criterio de eliminación de aquellas familias con constancias menores del 25%, haciendo la salvedad para las especies de alta fidelidad a ambientes específicos. Asi, al analizar la información por sitios de desembarco, se obtuvieron tres grupos de localidades (Fig. 7.7A y B) que se identificaron, con áreas donde predominan en los desembarcos capturas de los complejos ecológicos del litoral estuarino (Sánchez), pastos marinos-arrecifes coralinos (Las Pascualas, Samaná, Los Cacaos, Sabana de la Mar y Miches) y aguas oceánicas (La Galera y Las Terrenas). Asimismo el análisis de las familias permitió identificar cuales son aquellas asociadas a determinados complejos (Fig. 7.7C).

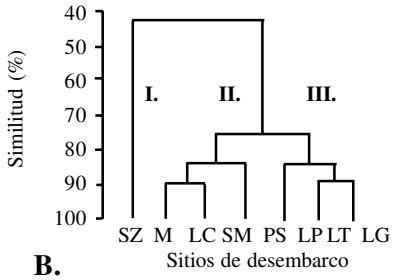
La composición y diversidad de las especies de las pesquerías guardan estrecha relación con los ambientes de las áreas de pesca (Fig. 7.8). En las localidades más internas se encuentran representadas familias de crustáceos infaunales como Penaeidae, típicas demersales estuarinas como Centropomidae y Mugilidae, o pelágicas como Engraulidae. En los sitios más hacia el océano se incrementa la representación de familias asociadas a la unidad ecológica que forman los manglares de borde junto a los pastos marinos y los arrecifes coralinos con familias demersales neríticas arrecifales como Serranidae, Holocentridae o Scaridae. En los lugares donde se realizan pesquerías de mar abierto están representadas especies pelágicas del complejo de las aguas oceánicas de las familias Coryphaenidae y Scombridae. Otras familias demersales como Lutjanidae, Haemulidae o Scianidae tienen una amplia distribución en todos los complejos al igual que las pelágicas neríticas como Carangidae y Sphyraenidae. El análisis nodal de la Fig. 7.8 revela que estas transiciones se observan como un gradiente de cambio del interior de la bahía, donde predominan condiciones estuarinas, hacia la zona externa donde el desarrollo de los arrecifes coralinos se va incrementando hacia la región oceánica.

Estos resultados constituyen un importante aporte en el interés de complementar la organización de nuestros recursos pesqueros, basada actualmente solo en criterios comerciales, desde una perspectiva ecológica que encamine el ordenamiento pesquero sobre bases científicas y permita crear unidades lógicas para el seguimiento estadístico de las capturas y la estandarización del esfuerzo pesquero.

[3]Baisre (1985) introduce el concepto de los complejos ecológicos, subdividiendo las pesquerías en aquellas correspondientes al complejo: litoral estuarino, pastos marinos-arrecifes coralinos y aguas oceánicas. Esto tiene un alto valor práctico pues subdivide los recursos pesqueros dentro de unidades naturales que facilitan su manejo y control.



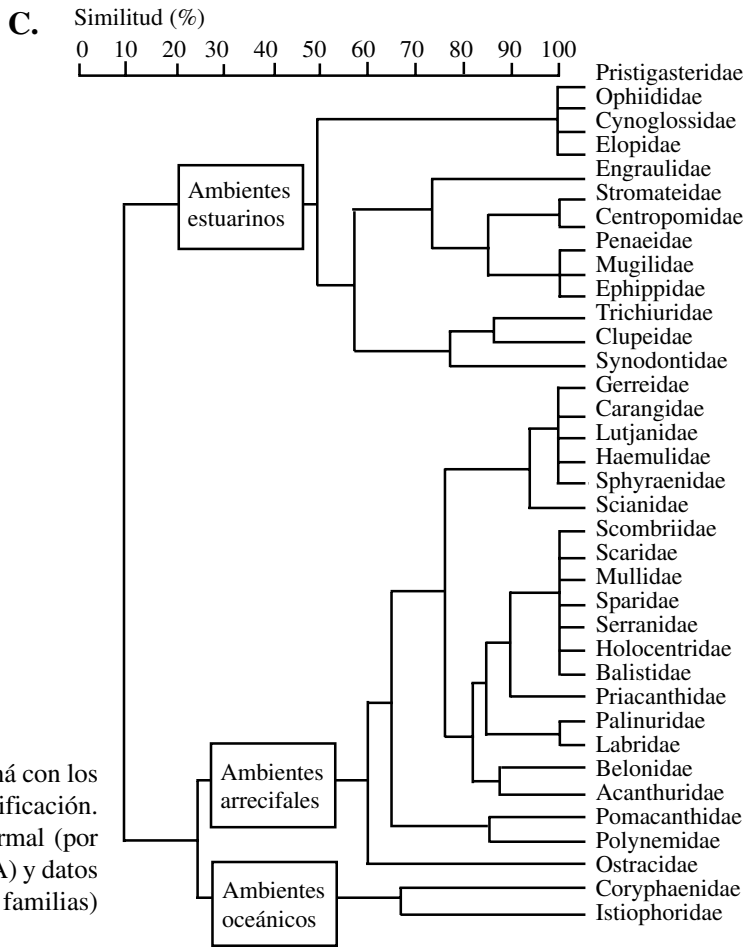




Figura 7.7. Arriba. Mapa de la región de Samaná con los ocho sitios de desembarco empleados en la clasificación. Abajo. Dendrogramas obtenidos en el análisis normal (por sitios de desembarco) con datos cuantitativos (A) y datos porcentuales (B); y en el análisis inverso (por familias) con datos cualitativos (C).

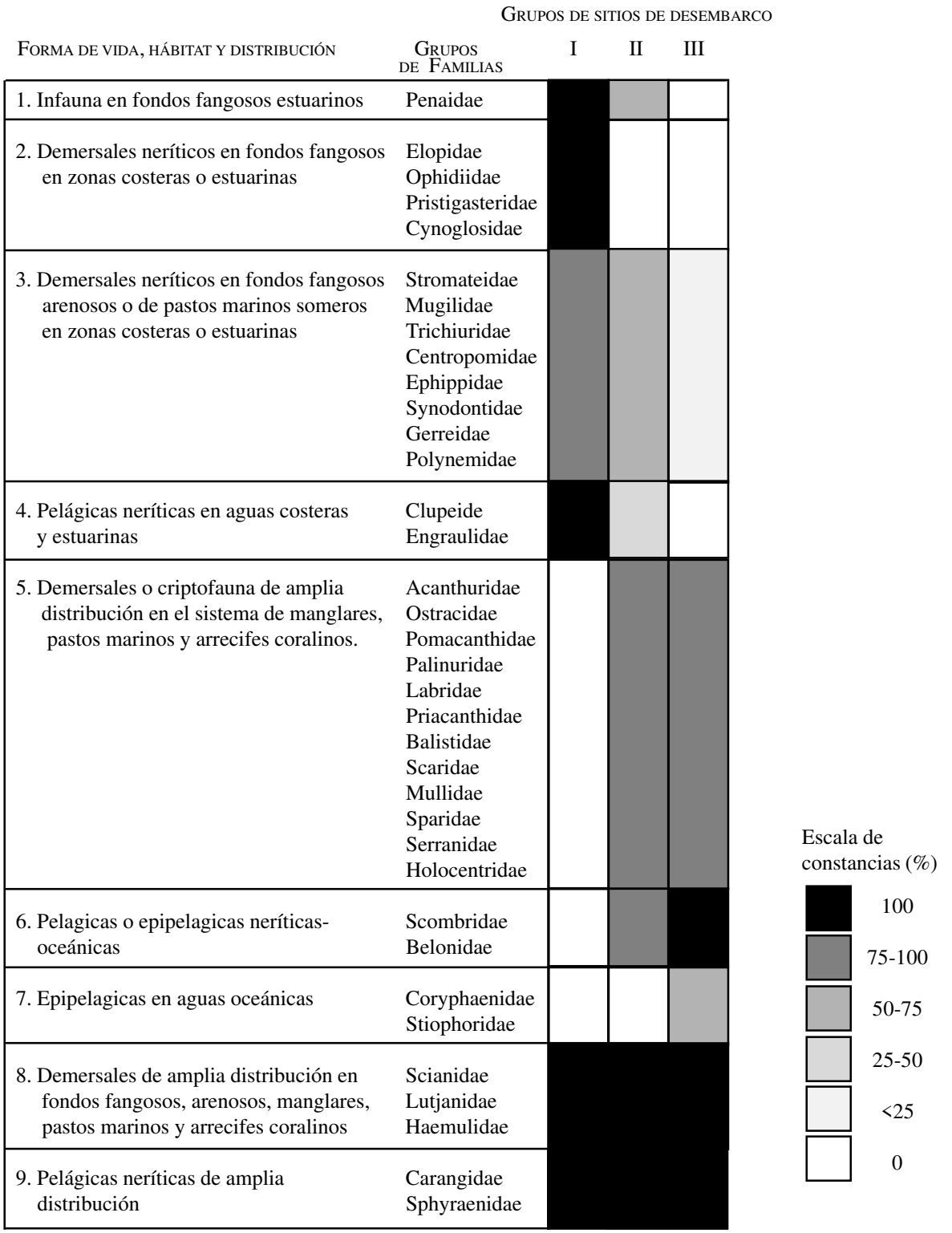



Figura 7.8. Gráfico de constancia nodal para nueve grupos de familias de peces y crustáceos de las pesquerías de Samaná (números arábigos) y tres grupos de sitios de desembarco (números romanos), obtenidos mediante análisis jerárquico de los datos cuantitativos y cualitativos (ver Figura 7.8).

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

*"Yo soñaba en clasificar..."*
*Dulce María Loynaz*

# 8. A MODO DE CONCLUSIÓN

Con lo hasta aquí expuesto queda claro que la clasificación numérica es ante todo un método y como tal debe ser conocido y aplicado en todos sus pasos. Para ello, el interesado contará con los principios básicos y los ejemplos aquí discutidos que le abrirán las puertas para acceder a los clásicos de la materia, muchos de ellos referenciados al final de este trabajo.

Además, como tratamos con una disciplina compleja y dinámica sujeta sin dudas a nuevos cambios y aportes toda nueva literatura sobre el tema será de gran utilidad en nuestra actualización y para ello, el enfoque que hemos mantenido a lo largo del texto, permitirá que cualquier nuevo hallazgo pueda ser ubicado fácilmente en el orden lógico de todo el proceso clasificatorio, con lo cual lo metodológico habrá cumplido su cometido.

Quien así lo entienda podrá contar con una herramienta interpretativa más, que al aplicarla deberá evitar introducir sesgos con sus preconcepciones de la realidad ecológica. Una disciplina que tuvo su origen en la práctica, en lo que algunos han llamado clasificaciones subjetivas -que tal vez no lo eran tanto-, no debe tornarse en algo dogmático, al contar con un cuerpo metodológico como la conocemos hoy, que la hace más objetiva. En la ecología que estudiamos, que es sencillamente una parte de la misma vida, la clasificación depende de muchos puntos de vista y muchas cosas pueden resultar dificiles de encasillar, bajo un nivel de conocimiento dado. Si esto le ocurre asimile la filosofía que recoge magistralmente este poema del Premio Cervantes de 1992, la poetisa de América, Dulce María Loynaz, cuando nos confiesa:

Yo soñaba en clasificar
el Bien y el Mal, como los sabios
clasifican las mariposas:
Yo soñaba en clavar el Bien y el Mal
en el obscuro terciopelo
de una vitrina de cristal...
Debajo de la mariposa
blanca, un letrero que dijera: «EL BIEN».
Debajo de la mariposa
negra, un letrero que dijera: «EL MAL».
Pero la mariposa blanca
era el mal...¡Y entre mis dos mariposas,
volaban verdes, aúreas, infinitas,
todas las mariposas de la tierra!....

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

# 9. REFERENCIAS

Alcolado, P. M. (editor) 1990. El bentos de la macrolaguna del Golfo de Batabanó. Editorial Academia, La Habana, 161 pp.

Anderberg, M. R. 1973. Cluster Analysis for Applications. Academic Press, New York, 359 pp.

Anderson, T. W. 1984. An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, New York, 665 pp.

Arabie, P y L. Hubert 1996. Advances in cluster analysis relevant to marketing research. En: From Data to Knowledge. W. Gaul y D. Pfeifer, eds., Springer, Berlin, pp. 3-19.

Baisre, J. A. 1985. Los complejos ecológicos de pesca: definición e importancia en la administración de las pesquerías cubanas. FAO Fish. Rep., 327, Suppl.: 251-272.

Bakus, G. J. 1990. Quantitative Ecology and Marine Biology. A. A. Balkema, Rotterdam, pp. 63-78

Boesch, D. F. 1977. Application of numerical classification in ecological investigations of water pollution. Ecological Res. Ser., EPA-600/3-77-033, 115 pp.

Boesch, D. F. 1977a. A new look at the zonation of benthos along the estuarine gradient. En: Ecology of Marine Benthos. Editor B. C. Coull, University of Arizona Press, 513 pp.

Boyce, A. J. 1969. Mapping diversity: a comparative study of some numerical methods. En: Numerical Taxonomy. Editor A. J. Cole, Academic Press, pp. 1-31.

Braun-Blanquet, J. 1979. Fitosociología. H. Blume Ediciones, Madrid, 820 pp.

Bray, R. J. y J. T. Curtis 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr., 27: 325-349.

Chatfield, C. y A. J. Collins 1992. Cluster Analysis. En: Introduction to Multivariate Analysis. Chapman & Hall, Londres, pp. 212-230.

Chiappone, M., K. M. Sullivan y C. Lott 1996. Hermatypic scleractinean corals of the Southeastern Bahamas: a comparison to western atlantic reef systems. Carib. J. Sci., 32(1): 1-13.

Clarke, K. R. y M. Ainsworth 1993. A method of linking multivariate community structure to environmental variables. Mar. Ecol. Prog. Ser., 92(3): 205-219.

Clifford, H.T. y W. Stephenson 1975. An Introduction to Numerical Classification. Academic Press, New York, 229 pp.

Crisci, J. V. y M. F. López Armengol 1983. Introducción a la Teoría y Práctica de la Taxonomía Numérica. Serie de Biología, Monografía No. 26, Secretaría General de la Organización de Estados Americanos, Programa Regional de Desarrollo Científico y Tecnológico, 132 pp.

Digby, P. G. N. y R. A. Kempton 1991. Multivariate Analysis of Ecological Communities. Chapman & Hall, Londres, 206 pp.

Dunn, G. y B. S. Everitt 1982. An Introduction to Mathematical Taxonomy. Cambridge University Press, Cambridge, 152 pp.

Esbensen, K., S. Schonkopf y T. Midtgaard 1994. Multivariate Analysis in Practice. Camo A.S., Noruega, 361 pp.

Everitt, B. S. y G. Dunn 1991. Cluster Analysis. En: Applied Multivariate Data Analysis. John Wiley & Sons Inc., New York, pp. 99-124.

Everitt, B. S. 1993. Cluster Analysis. John Wiley & Sons Inc., New York, 170 pp.

Fielding A. H. 1999. Cluster Analysis, a web-based tutorial. http://149.170.199.144/multivar/ca.htm

Frontier, S. 1969. Sur une méthode d'analyze faunistique rapide du zooplancton. J. Exp. Mar. Biol. Ecol., 3: 18-26.

Greig-Smith, P. 1983. Quantitative Plan Ecology. Blackwell, Oxford, 514 pp.

Griffith, D. A. y C. G. Amrhein 1991. Cluster analysis: an introduction to objects grouping techniques. En: Statistical Analysis for Geographers. Prentice Hall, New Jersey, pp. 423-438.

Hair, T. F. Jr., R. E. Anderson, R. L. Tatham y W. C. Black 1995. Cluster Analysis. En: Multivariate Data Analysis with Readings. Prentice-Hall, New Jersey, pp. 420-483.

Harris, R. J. 1985. A primer of multivariate statistics. Academic Press, New York, 576 pp.

Herrera, A. y P.M. Alcolado 1983. Efectos de la contaminación sobre las comunidades de gorgonáceos al Oeste de la Bahía de la Habana. Cien. Biol., 10:69-86.

Herrera, A. 1984. Clasificación numérica de las comunidades de gorgonáceos al Oeste de la Bahía de la Habana. Cien. Biol., 12: 105-124.

Herrera, A., del Valle, R. y N. del Castillo 1987. Aplicación de métodos de clasificación numérica al estudio ecológico del litoral rocoso. Rep. Invest. Inst. Oceanol., 70: 1-17.

Herrera A. y J. Espinosa 1988. Características de la fauna de bivalvos en la Bahía de Cárdenas. Rep. Invest. Inst. Oceanol., 17: 1- 21.

Herrera, A. 1991. Efectos de la contaminación sobre la estructura ecológica de los arrecifes coralinos en el litoral habanero. Tesis presentada en opción al grado científico de Doctor en Ciencias. Academia de Ciencias, La Habana, Cuba, 110 pp.

Herrera, A., G. Gotera, D. Ibarzábal, G. González, R. Brito y E. Díaz 1991. Ecología de los arrecifes del borde de la plataforma SO de Cuba y su relación con la langosta *Panulirus argus*. Rev. Invest. Mar., 12(1-3): 163-171.

Herrera, A. 1992. Un método rápido para el análisis e interpretación de datos porcentuales. Cien. Biol., 24:1-10.

Herrera, A., P. Alcolado y P. García-Parrado. 1997. Estructura ecológica de las comunidades de gorgonáceos en el arrecife de barrera del Rincón de Guanabo. Avicennia, 6/7: 73-85.

Ignatiadis, L., K. Pagov y V. Gialamas 1992. Multivariate analysis of phytoplanktonic parameters: a sample study. J. Exp. Mar. Biol. Ecol., 160(1): 103-114.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaudoise Sci. Nat., 44: 223-230.

Jobson, J. D. 1991. Applied Multivariate Data Analysis I. Regression and Experimental Design. Springer-Verlag, New York, 621 pp.

Johnson, R. A. y D. W. Michern 1992. Clustering. En: Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey, pp. 573-627.

Kaufman, L. y P. J. Rousseeuw 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc., Nueva York, 342 pp.

Krzanowski, W. J. 1990. Principles of Multivariate Analysis: a User´s Perspective. Clarendon Press, Oxford, 563 pp.

Krzanowski, W. J. y F. H. C. Marriott 1996. Initial Data Analysis. En: Multivariate Analysis 1. Distributions, Ordination and Inference. Edward Arnold, New York, pp. 43-74.

Krzanowski, W. J. y F. H. C. Marriott 1996a. Cluster Analysis. En: Multivariate Analysis 2. Classification, Covariance Structures and Repeated Measurements. Edward Arnold, London, pp. 61-94.

Lance, G. N. y W. T. Williams 1966. Computer programs for classification. Proc. ANCCAC Conference, Canberra, May 1966, Paper 12/3.

Lance, G. N. y W .T. Williams 1966a. A generalized sorting strategy for computer classifications. Nature, 212:218

Legendre, L. y P. Legendre 1979. Ecologie Numerique. Les Presses de L'Universite du Quebec, 247 pp.

Ludwig, J. A. y J. F. Reynolds 1988. Statistical Ecology. John Wiley and Sons, New York, 337 pp.

Manly, B. F. J. 1995. Cluster Analysis. En: Multivariate Statistical Methods A Primer. Chapman & Hall, Londres, pp. 128-145.

Milligan, G. W. y M. C. Cooper 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2): 159-179.

Mojena, R. 1977. Hierarchical grouping methods and stopping rules: An evaluation. Computer. J. 20, 359- 363.

Morrison, D. F. 1990. Multivariate Statistical Methods. McGraw-Hill, New York, 495 pp.

Neff, N. A. y L. F. Marcus 1980. Cluster Analysis, Numerical Cladistics and Tree Analysis. En: A Survey of Multivariate Methods for Systematics. New York: Privately Published, 243 pp.

Orlóci, L. 1978. Multivariate Analysis in Vegetation Research. Dr.W. Junk B.V., Publishers, The Hague, Boston, 451 pp.

Pielou, E. C. 1977. Mathematical Ecology. John Wiley & Sons, Nueva York, 385 pp.

Pielou, E. C. 1984. The Interpretation of Ecological Data. John Wiley & Sons, New York, 263 pp.

Popper, R. y H. Heymann 1996. Analyzing differences among products and panelists by multidimensional scaling. En: Multivariate Analysis of Data in Sensory Science. Editores T. Noes y E. Risvik, Elsevier Science B. V., Amsterdam, p. 159- 184

Rencher, A. C. 1995. Methods of Multivariate Analysis. John Wiley & Sons, Inc., New York, 627 pp.

Sanders, H. L. 1960. Benthic studies in Buzzards Bay. III. The structure of the soft bottom community. Limnol. Oceanogr., 5: 138-153.

Sanders, H. L. 1968. Marine benthic diversity: a comparative study. Amer. Nat., 102(925): 243-282

Sang, L., D. León, M. Silva and V. King 1997. Diversidad y composición de los desembarcos de la pesca artesanal en la región de Samaná. Proyecto de Conservación y Manejo de la Biodiversidad en la Zona Costera de la República Dominicana GEF-PNUD/ONAPLAN, 52 pp.

Sharma, S. 1996. Cluster Analysis. En: Applied Multivariate Techniques. John Wiley & Sons Inc., New York, 493 pp.

Sheppard, F. R. 1954. Nomenclature based on sand-silt-clay relations. J. Sediment. Petrol., 24(3): 51-158.

Siegel, S. 1985. Estadística No Paramétrica Aplicada a las Ciencias de la Conducta. Editorial Trillas, México, 344 pp.

Sneath, P. H. A. y R. R. Sokal 1973. Numerical Taxonomy. W. H. Freeman & Co., San Francisco, 575 pp.

Sokal, R. R. y C. D. Michener 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull., 38: 1409-1438.

Sokal, R.R. y F. J. Rohlf 1962. The comparison of dendrograms by objective methods. Taxon, 11:33-40.

Sokal, R. R. y P. H. A. Sneath 1963. Principles of Numerical Taxonomy. W. H. Freeman & Co., San Francisco, 359 pp.

Sorensen, T. 1948. A method of stablishing groups of equal amplitude in plant sociology based on similarity of species content and its applications to analysis of the vegetation on Danish commons. Biol. Skr., 5: 1-34.

Southwood, T. R. E. 1994. Diversity, species packing and habitat description. En: Ecological Methods. Chapman & Hall, London, pp. 420-455.

Statistica 2000. Cluster analysis. En: Electronic Textbook. StatSoft. Inc. http://www.statsoftinc.com/textbook/stcluan.html#h

Stuessy, T. F. 1990. Plant Taxonomy. Columbia University Press, New York, 514 pp.

Van Tongeren, O. F. R. 1987. Cluster Analysis. En: Data Analysis in Community and Landscape Ecology. Editores Jongman, R. H. G., ter Braak, C. J. F. y O. F. R. Van Tongeren, Pudoc Wageningen, the Netherlands, pp. 174-203.

**Tomado de: Herrera, Alejandro 2000. La clasificación numérica y su aplicación en la ecología. Universidad INTEC/Programa EcoMar, Inc. Editorial Sanmenycar, Santo Domingo, 121 pp.**

**Esta primera edición de**
**La clasificación numérica y su aplicación en la ecología**
**se terminó de imprimir en diciembre del año 2000**
**en los talleres gráficos de Impresora Sammerycar C. por A.**
**Santo Domingo, República Dominicana**

**Acerca del Autor.**

Alejandro Herrera Moreno nació en La Habana, Cuba en 1952 y se graduó de Licenciado en Ciencias Biológicas en la Universidad de la Habana en 1976 con una especialidad en Biología Marina. Desde esa fecha trabajó como investigador en el Instituto de Oceanología de la Academia de Ciencias y colaboró con el Centro de Investigaciones Marinas (CIM) de la Universidad de la Habana y el Centro de Investigaciones Pesqueras (CIP) del Ministerio de la Industria Pesquera. Alcanzó el grado de Investigador Titular y de Doctor en Ciencias Biológicas con la presentación de la Tesis sobre *Impactos de la contaminación marina sobre los arrecifes coralinos del litoral habanero*. Estableció nexos de colaboración científica y docente con numerosas instituciones de investigación de América y Europa, donde trabajó en Proyectos de Investigación e impartió numerosos Cursos de Postgrado en disciplinas, como ecología del bentos marino, arrecifes coralinos, contaminación marina, pesquerías de la langosta *Panulirus argus*, clasificación numérica y evaluación de impacto ambiental. En República Dominicana se ha desempeñado como Consultor de Organismos Internacionales como el Programa de Naciones Unidas para el Desarrollo (PNUD) y el Fondo de las Naciones Unidas para la Infancia (UNICEF); y Organismos de Cooperación Internacional como la Cooperación Técnica Alemana (GTZ) y la Agencia de Cooperación Internacional de Japón (JICA). Contribuyó, como Asesor en Gestión de Calidad del Ambiente Marino, a la creación y desarrollo del Instituto Nacional de Protección Ambiental (INPRA) y la actual Subsecretaría de Gestión Ambiental, de la Secretaría de Recursos Naturales y Medio Ambiente de la República Dominicana. Actualmente es el Presidente del Programa Ecomar, Inc. en la República Dominicana. El autor ha publicado más de ochenta trabajos especializados y de periodismo científico y cuenta con tres libros de literatura y educación ambiental para niños, el último de ellos publicado en República Dominicana bajo el título *El Mar para los Niños*. En la presente obra: *La clasificación numérica y su aplicación a la ecología*, plasma un trabajo de varios años de investigación y docencia en esta novedosa disciplina de la clasificación y ofrece a los estudiantes e investigadores de Nuestra América, de manera didáctica sus experiencias, con la certeza de que "cada alumno que progresa es un maestro que nace."