

“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 α ; *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 β de la diversidad de especies (Southwood, 1994). Los ecólogos estructurales definen la diversidad de especies dentro de una comunidad o hábitat como diversidad α 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 β , 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

		Entidad 1	
		1	0
Entidad 2	1	a	b
	0	c	d

		Estación 1	
		1	0
Estación 2	1	Número de especies comunes	Número de especies en 1 y no en 2
	0	Número de especies en 2 y no en 1	Número de especies ni en 1 ni en 2

		Especie 1	
		1	0
Especie 2	1	Número de ocurrencias comunes	Número de ocurrencias de A sin B
	0	Número de ocurrencias de B sin A	Número de veces que ni A ni B están

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.					B.			C.		
Especies	Estaciones				Pares de estaciones	Medidas de afinidad		Pares de estaciones	Medidas de afinidad	
	1	2	3	4		Jaccard (1908)	Sorensen (1948)		Sokal y Michener (1958)	Sokal y Sneath (1963)
A	1	0	0	1	1-2	0.40	0.50	2-3	0.80	0.89
B	0	0	0	1	2-3	0.33	0.50	1-2	0.70	0.82
C	0	0	0	1	1-4	0.29	0.44	1-3	0.50	0.67
D	0	0	0	1	2-4	0.14	0.25	1-4	0.50	0.67
E	1	1	0	1	1-3	0.00	0.00	2-4	0.40	0.57
F	1	0	0	0	3-4	0.00	0.00	3-4	0.40	0.42
G	0	0	0	0						
H	0	1	1	1						
I	1	1	0	0						
J	0	0	0	0						

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 monótono aunque con diferentes valores (Johnson y Wichern, 1992). La monotonidad 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 varía entre 0, entidades sin ningún 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.

Especies	Estaciones				
	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	1,00	1,00	?	?	?	1	
		0,75	?	?	?	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	
A	1.00	0.86	?	?	?	A
B		1.00	?	?	?	B
C			1.00	?	?	C
D				1.00	?	D
E					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	1.00	0.75	0.57	0.33	1.00	1
2		1.00	0.40	0.50	0.75	2
3			1.00	0.00	0.57	3
4				1.00	0.33	4
5					1.00	5

	A	B	C	D	E	
A	1.00	0.86	0.86	0.67	0.75	A
B		1.00	0.67	0.80	0.57	B
C			1.00	0.80	0.86	C
D				1.00	0.67	D
E					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).

Atributos	Entidades						
	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 más 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.

Especies	Estaciones				
	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.

Especies	Estaciones				
	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).

Explicemos 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.

Especies	Estaciones		$X_{ij} - X_{ik}$	$(X_{ij} - X_{ik})^2$	%
	j X_{ij}	k X_{ik}			
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.

Especies	Estaciones		
	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
	0	41,4
		0

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 más 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 y α , 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.

Clases	Mínimo	Máximo	$X_{ij} - X_{ik}$	$(X_{ij} - X_{ik})^2$	%
	X_{ij}	X_{ik}			
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 $D_p = 1/p[\sum(X_{ij}-X_{ik})^2]^{1/2}$ 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 $D_m = 1/p \sum |X_{ij}-X_{ik}|$ 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:

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

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

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.

Especies	Estaciones				
	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	
1	0						1
2		0.42	?	?	?		2
3			0	?	?		3
4				0	?		4
5					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 $\sum |X_{ij}-X_{ik}| = 37$. 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 $\sum (X_{ij} + X_{ik}) = 87$. 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.

Especies	Estaciones		X _{ij} -X _{ik}	%
	j	k		
A	X _{ij} 80	X _{ik} 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 Índice de Bray-Curtis.

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.

Especies	Estaciones	
	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:

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

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

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.

Especies	Estaciones				
	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	
	1	0	0.44	?	?	?	1
	2		0	?	?	?	2
	3			0	?	?	3
	4				0	?	4
	5					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).

Especies	Estaciones		$\frac{ X_{ij}-X_{ik} }{(X_{ij}+X_{ik})}$	%
	j	k		
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 Índice 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.

Especies	Estaciones		$ X_{ij}-X_{ik} $	$(X_{ij}+X_{ik})$	$\frac{ X_{ij}-X_{ik} }{(X_{ij}+X_{ik})}$
	1	2			
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		X _{ij} -X _{ik}	(X _{ij} +X _{ik})	$\frac{ X_{ij}-X_{ik} }{(X_{ij}+X_{ik})}$
	1	2			
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:

$$R_{jk} = \frac{(X_{ij} - X_j)(X_{ik} - X_k)}{(\sum (X_{ij} - X_j)^2)^{1/2} (\sum (X_{ik} - X_k)^2)^{1/2}}$$

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.

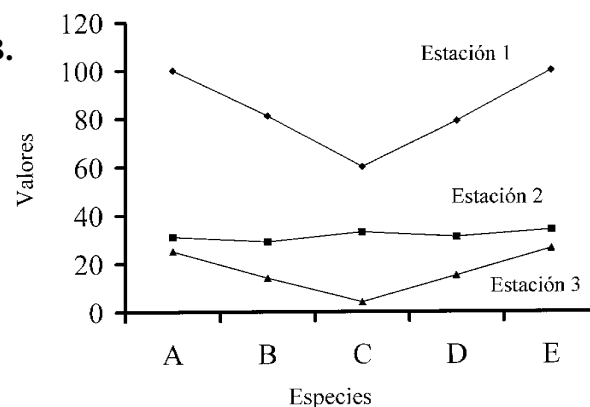


Figura 4.20. A. Matriz de datos de 3 estaciones y 5 especies. B. Variación de los valores de especies en cada estación.

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:

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

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:

$$r = 1 - \frac{6 \sum D_i^2}{N(N^2 - 1)}$$

donde D_i 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.

Especies	Datos originales		Datos de rangos		D	D ²
	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 $r = \frac{6(11.5)}{190}$ 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 nuestro 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.

Proyección de similitud 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 similitud 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.

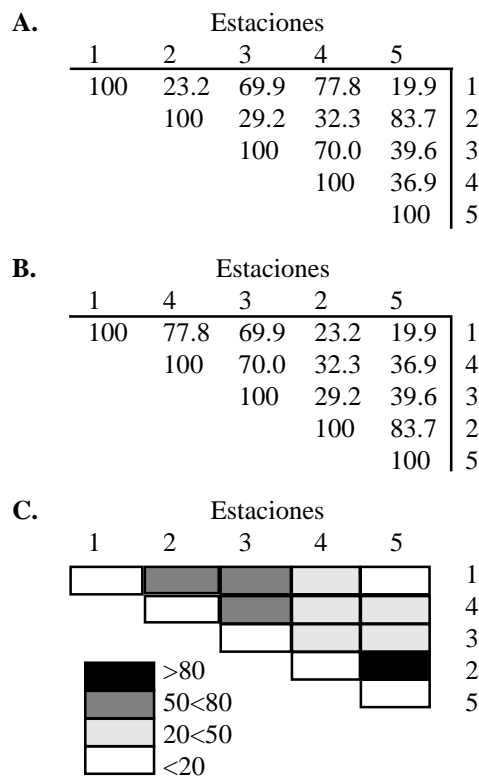


Figura 4. 22. **A.** Matriz de afinidad, **B.** Matriz reordenada, **C.** Diagrama de Trellis y escala de valores.

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.

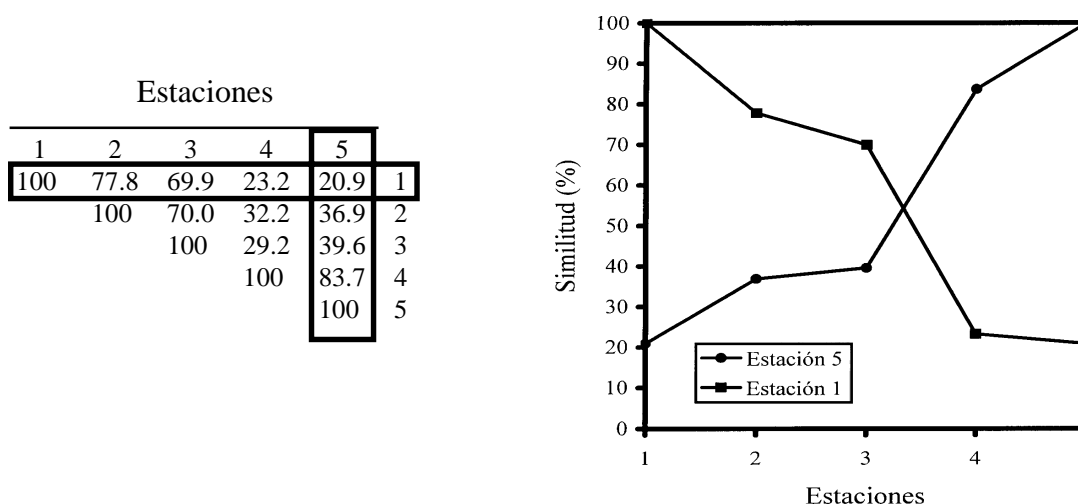


Figura 4.23. Izquierda. Matriz de similitud señalando las entidades que reflejan un gradiente de cambios en la afinidad. Derecha. Proyección de similaridad cenocónica en la comparación de las estaciones 1 y 5 con las restantes. En la leyenda se indica la estación respecto a la cual se compara.