

# *Análisis de los datos en ciencias humanas y naturales*

Héctor B. LAHITTE  
Jorge FRANGI

El objeto de este trabajo \* es dar un panorama de los métodos objetivos que se utilizan para describir, analizar y sintetizar la complejidad de las relaciones que existen en las Ciencias Naturales, incluyendo las Ciencias del Hombre.

La difusión actual de los medios informativos y las facilidades de cálculo y de memorización que se han desarrollado en la mayor parte de las disciplinas científicas o técnicas, es en gran medida producto de las voluminosas masas de datos recogidos a partir de la observación. En general, desde el punto de vista metodológico, un trabajo de investigación consiste en los siguientes pasos:

1. Definición de los temas y objetivos de la investigación.
2. Selección de los métodos de campo o laboratorio para la toma de muestras.
3. Selección de los métodos de síntesis de datos.
4. Verificación de los resultados obtenidos.

Quizá con excepción de los puntos 1 y 2, que son por lo general bien específicos de cada disciplina científica y aun más de cada una de las especialidades dentro de ellas, la aplicación de los métodos de síntesis y de verificación son de interés general para las diversas ramas de las Ciencias Naturales, porque en definitiva todas pretenden un mismo objetivo general cual es la descripción, análisis y síntesis

\* Ha colaborado en la presentación del mismo el señor D. Omar J. Sómma.

de un conjunto de caracteres y objetos provenientes de un universo corrientemente complejo que es necesario simplificar objetivamente para extraer conclusiones válidas.

Quienes los utilizan se encuentra ahora confrontados a vastas tablas de datos muy ricos en información pero difícilmente explotables en un primer punto de vista que intente conclusiones rápidas. Después de algunos años, sin embargo, han aparecido un gran número de publicaciones relativas tanto a las ciencias de la naturaleza como a las ciencias humanas, en las que se presentan los datos y/o resultados ordenados según gráficos o dendrogramas difíciles de descifrar por un lector no iniciado<sup>1</sup>. Estas representaciones (o modelo) tienen justamente la ambición —aunque suene paradójico— de dar una descripción sintética y clara de extensas tablas o cuadros con valores numéricos. Este conjunto de técnicas son las que en Francia se llaman de «análisis de datos» (y que en la bibliografía de habla inglesa incluyen los métodos de clasificación, taxonomía numérica y ordenamiento). Esta terminología puede parecer poco clara, en tanto se trata de designar en última instancia un tipo de síntesis; pero el término de análisis evoca más bien una actividad menos ambiciosa y más laboriosa (queda claro que síntesis no significa interpretación). Además es importante recordar que estos términos están fuertemente connotados en estadística a tal punto que algunos investigadores prefieren hablar de «estadísticas descriptivas multidimensionales», designación que tiene la ventaja de ser más analítica que la anterior y de simultáneamente referirse o aludir a la llamada «estadística elemental» (histogramas, diagramas triangulares, etc.). Otros autores prefieren hablar simplemente de métodos de descripción por cuadros o tablas.

El uso de estos métodos es contemporáneo a la aparición y empleo de las computadoras a tal punto que un gran número de técnicas de análisis de datos dependen de éstas. Sin embargo, en lo que concierne a los principios fundamentales no podemos considerarlos más que como una parte de la estadística aplicada (la que comprende no sólo estudios de tipo probabilísticos, sino también métodos no probabilísticos: forma de representación de datos y generación de hipótesis).

Cabe destacar que, si bien el objetivo de este trabajo es explicar sólo someramente el proceso de recolección y análisis de información,

<sup>1</sup> Por ejemplo, en los estudios sinicológicos existe una enorme variedad de técnicas de campo para poder evaluar algún parámetro de importancia de las especies presentes en una comunidad. Esas técnicas, como el lector comprenderá, dependen de numerosas variables, entre las que se encuentran, objetivos, parámetros a estimar, tiempo disponible o costo de muestreo, características inherentes a las especies a estudiar (tipo biológico, morfología, etc.).

Los datos fitosociológicos por lo general consisten en listas de especies con su valor de importancia y datos del ambiente (tipo de suelo, exposición, altitud, etcétera), registrados en un sitio o lugar. Ese conjunto de información se conoce con el nombre de relevamiento o censo; siendo la muestra el conjunto de relevamientos efectuados en un área.

se hace necesario indicar algunos aspectos relacionados con la toma de muestras o datos. La aplicación de técnicas estadísticas a diversas disciplinas normalmente se hace con la convicción de que aquéllas le dan al estudio un carácter «objetivo» y científicamente más respetable. No obstante, si bien es cierto que la aplicación de estas técnicas eliminan el bias personal y brindan resultados reproducibles (siempre que se aplique el mismo método al mismo conjunto de datos), no es menos cierto que es frecuente que se presenten al investigador situaciones complejas que le obligan a tomar decisiones subjetivas. Por ejemplo, en el caso de trabajos en el campo o en el laboratorio, quién no ha debido decidir, en base a su experiencia y conocimientos previos, acerca de la naturaleza de los datos que debe tomar y qué técnica de campo y de análisis de laboratorio va a aplicar para obtener la información deseada. Lo que es particularmente importante es ver en qué medida las decisiones subjetivas que deben tomarse son relevantes para el sistema que se estudia y por ende cuál será su influencia sobre los resultados que se obtendrán.

Si hacemos un poco de historia podríamos reconocer a R. A. Fisher (1940) como uno de los fundadores del análisis de los datos, en especial en lo que a análisis de correspondencia se refiere. Algo semejante sucede con el célebre tratado de estadística de M. G. Kendall y A. Stuart (1961), en el que encontramos la base de las operaciones del análisis canónico de tablas de contingencia: método o técnica elaborada para calcular en definitiva los parámetros a probar la hipótesis de independencia entre líneas y columnas de una tabla. Hoy, con la aparición de la obra de J. C. Gardin, se ponen en evidencia todas las propiedades (fundamentalmente algebraicas) del método; se suponen hipótesis de proximidad, distancia o similitud, entre las asociaciones posibles existentes entre líneas y columnas de una tabla. Mientras tanto, la computadora sigue su desarrollo haciendo posible tales descripciones sobre todo cuando es muy vasto el vínculo de datos que deban manipularse. La práctica de esas descripciones ha suscitado conjeturas nuevas y modificado los centros de interés del especialista. Casi podríamos afirmar sin temor a equivocarnos que la posición de Benzécri es el punto de referencia y partida de todos los especialistas franceses en análisis de datos. Este autor opina que la estadística debe honrarse de ser una ciencia experimental que tiene todo que aprender de la naturaleza. Sin embargo, el análisis de correspondencia puede ser considerado como una rama del llamado «análisis factorial» (cuya historia es aún más antigua).

A principios de siglo, los psicólogos buscaron ir más allá de las notas obtenidas producto de los *test* aplicados a numerosos objetos. La búsqueda de esta variable explicativa (factor de aptitud, por ejemplo) fue tal vez el comienzo de la definición y búsqueda del análisis de factores (C. Sperman, 1904). Todo comenzó con la definición de

diferentes parámetros o factores (memoria, inteligencia, etc.) no observables directamente, pero sí susceptibles de explicación en el sentido estadístico del término. En estos primeros momentos se trataba más bien de resumir para explicar la información más que descripciones que permitieran el análisis propiamente dicho. Luego el análisis factorial tal como fue aplicado por los psicólogos se fue enriqueciendo con los aportes, por ejemplo, de H. Votelling, quien es de los primeros en desarrollar el análisis de los componentes principales. Sería importante aquí señalar un detalle, esto es: la autonomía de los métodos elaborados, probablemente producto de la especificidad de los dominios de aplicación. Retomando lo antedicho diremos que el modelo *a priori* con que se manejan los psicólogos al hacer análisis factorial no es más que un simple pretexto que permite explotar un universo «multidimensional» de datos. Son estos investigadores los pioneros de la explotación de los datos «multidimensionales», no sólo en Francia, sino también en Estados Unidos (J. D. Carrol, J. D. Kruskal, R. Shepard, etc.) quienes llamaron a esto «Multidimensional Scaling».

Estos autores proponen técnicas de representación por tablas, o cuadros, esencialmente construidas con datos provenientes de la observación de las experiencias psicológicas, según técnicas hipotéticas y apriorísticas.

El primer trabajo para analizar es, obviamente, el recoger los datos (por encuestas, cuestionarios, sondajes, etc.), para luego representarlos bajo la forma de «tablas rectangulares», «cuadros de doble entrada», «matriz primaria de datos» o «tabla bruta».

En la práctica, estas tablas corresponden a un campo de aplicación muy vasto, que toca distintos ámbitos de las ciencias, tanto humanas, sociales como naturales.

En Antropología, por ejemplo, las columnas de tales cuadros o tablas representan los objetos (los individuos, representaciones, países, etc.) y las líneas representan los parámetros característicos de esos objetos. En los trabajos fitosociológicos las columnas corresponden a los relevamientos (objetos) en tanto en las filas se anotan las especies o caracteres ambientales. Si se trata de especies, en la intersección de filas y columnas se ponen el valor de importancia de cada una de ellas (densidad, frecuencia, cobertura, biomasa, por ejemplo) en los relevamientos correspondientes.

	OBJETOS	1	2	3	n
PROPIEDADES	cara	x	x	x	
	tocado	x	x	x	
(En Arqueología: partes, etc.; en Ecología: parámetros, etc.)	patas	x	x	x	
	etc.				

En Taxonomía vegetal y animal las columnas corresponden a cada individuo considerado y las filas a características morfológicas, fisiológicas, etc., de las mismas.

Una vez que los datos experimentales han sido recogidos (según la técnica específica y explícita de la disciplina en cuestión)<sup>2</sup> y los valores de los parámetros definidos y anotados, pueden ser construidas diversas tablas a  $p$  columnas y  $n$  líneas —si  $n$  objetos son observados—. La posibilidad de cuantificar los parámetros nos permite hablar de «tablas de medida». Las tablas contienen valores numéricos para expresar el ejemplo. En estos casos podemos transformar el código numérico de un parámetro que será, naturalmente, cuantitativo (representaciones, por ejemplo) y construir un código cualitativo ordinal (de 0 a 20 representaciones por centímetro cuadrado, etc., será codificado como 1, 2, 3, etc.) o nominal (las mismas clases serán codificadas, respectivamente, como 100, 010, 001)); sólo la presencia o ausencia de un carácter es lo que en este caso se toma en cuenta.

		<u>Zona a</u>	<u>Zona b</u>	<u>Zona c</u>	<u>Zona n</u>
Representaciones	0-20	x	x		
por cm <sup>2</sup>	20-30				
	30-40			x	
	... n				

Otro tipo de datos corresponde al caso en el que podemos calcular el número de ocurrencias simultáneas de una observación, con uno de los parámetros que la caracterizan, por ejemplo, la aparición de determinados materiales (cerámica, madera, metal, etc.) en un sitio: los objetos son los distintos sitios y los parámetros los materiales, para evitar confusiones diremos que éstas son las que se llaman «tablas o cuadros de frecuencia o contingencia».

		<u>Sitio a</u>	<u>Sitio b</u>	<u>Sitio n</u>
Materiales:	cerámica	x		
	madera	x	x	
	lítico		x	
	etc.			

<sup>2</sup> Es importante aclarar que existen dos aproximaciones —dependiendo de la naturaleza y cantidad de información del universo en cuestión—, una que con códigos ya construidos recoge la información en el campo (primera normalización del lenguaje). Otra que a partir del dato mismo y según un comentario de él elabora códigos artificiales, a partir de los cuales comienza la descripción propiamente dicha que en cualquier caso necesita de la unidad de referencia elegida. La primera aproximación es de mayor uso en aquellas disciplinas con caracteres analíticos históricos y convencionalmente establecidos. La segunda, en cambio, es frecuente en el conjunto de las Ciencias Sociales, Humanas y del Hombre y aún en ramas de la Biología, cuya esfera experimental intenta definir sistemas descriptivos convencionales a partir de unidades de referencia, cuyos caracteres no son determinables *a priori*. (Ver gráfico 1).

Cuando, por ejemplo, tratamos de conocer el orden de preferencia respecto de un determinado objeto (y/o situación) construimos las que se llaman «tablas de preferencia».

		<u>Color a</u>	<u>Color b</u>	<u>Color n</u>
Representaciones:	1	x		
	2		x	
	3	x		
	n			

En tales tablas o cuadros los parámetros son cualitativos ordinales. Cuando evaluamos la similitud o la semejanza entre cada pareja de objetos constituimos para presentar los datos las llamadas «tablas de proximidad» o «matrices de similitud».

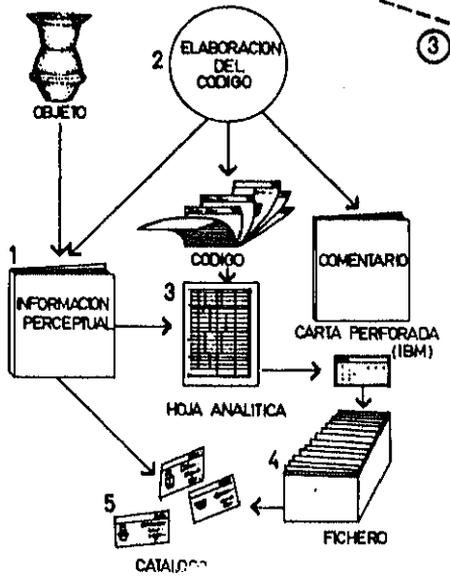
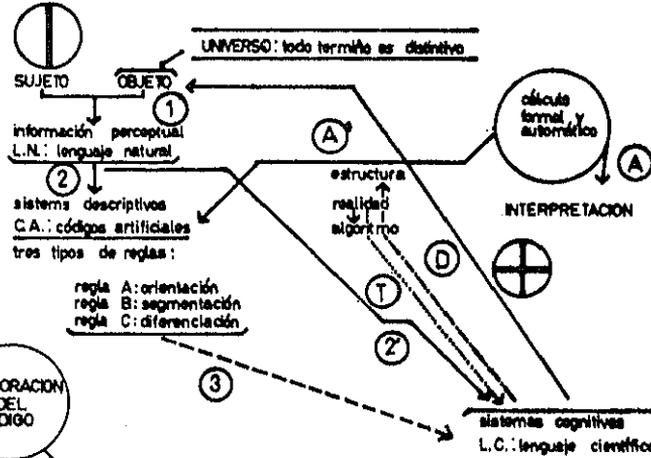
	Representaciones:	<u>1</u>	<u>2</u>	<u>3</u>	<u>n</u>
Representaciones:	1	10	3	5	
	2	3	10	5	
	3	5	5	10	
	n				

Por ejemplo, si el objetivo es dar cuenta de las representaciones (en el caso de la Arqueología), o censos (en el caso de la Ecología), y queremos comparar unas contra otras, cada uno de los casilleros correspondientes a un tipo de relación de dos objetos contendrá una nota entre 1 y 100 ó 0 y 1 que mide las semejanzas teóricas entre ellas que, como veremos más adelante, se obtienen mediante la aplicación de ciertos índices. Esta nota puede ser un medio que se obtiene con la ayuda no necesariamente de un inventario, sino como el resultado o producto de un análisis de objetos (o textos, en el caso de análisis de contenido semántico, etc.).

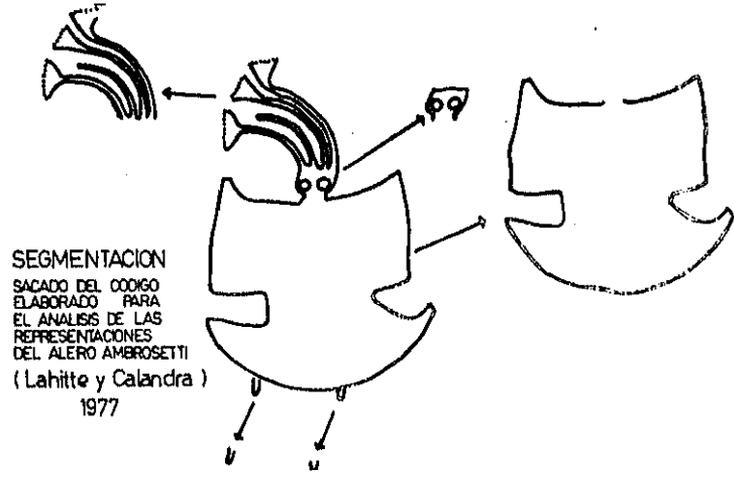
Encontramos también el caso en el cual queremos registrar «calidad» entre individuos o materiales. Podemos construir entonces una tabla cuadrada en la que se inscriben en líneas y columnas cada uno de los individuos o materiales, siendo la intersección de la línea y la columna una franja formada por los individuos o materiales y en los cuales se anotan, según un juicio establecido convencionalmente, lo que juzgamos como inferior, equivalente o superior,  $-1$ ,  $0$ ,  $1$ , respectivamente, obteniéndose así los llamados «cuadros de dominancia».

	Representaciones:	<u>1</u>	<u>2</u>	<u>3</u>	<u>n</u>
Representaciones:	1	0	-1	1	
	2	-1	0	1	
	3	1	1	0	
	n				

TRES REINOS DE CUANTIFICACION SEMIOLOGICA



CONSTRUCCION DE UN REPERTORIO  
(Gardin 1968-70)  
(Lahitte 1971)



Finalmente, si para un parámetro determinado los objetos no pueden tomar más que dos valores: 1 y 0, todo o nada, datos binarios, estamos en presencia de una «tabla de incidencia».

		Representaciones			
		1.	2.	3.	n.
Representaciones	1.	0	1	0	
	2.	1	0	1	
	3.	1	0	1	
	n.				

Esta etapa de elaboración reconoce tres inconvenientes: la ineptitud para el tratamiento de conjunto de datos muy vastos; su interpretación visual no es generalmente fiable, en tanto depende de cada observador, y, finalmente, su poder de síntesis es muy débil.

#### ELABORACIÓN DE LA MATRIZ SECUNDARIA

Debido a las dificultades que surgen al tratar de comparar visualmente o manualmente (*peek-a-boo*, *port-a-punch*, por ejemplo), la similitud entre objetos o la semejanza de distribución entre caracteres, es conveniente aplicar índices o coeficientes numéricos que sintetizen en forma objetiva y en un único valor numérico o en un punto del espacio esa similitud relativa (afinidad, similitud, asociación, correlación, proximidad, correspondencia, o sus complementos, disimilitud, diferencia, distancia). Estos coeficientes son aplicados en diversas disciplinas de la ciencia, como la sistemática vegetal y animal, la Arqueología, la Antropología, la Ecología, etc. Ellos reciben distintos nombres tales como índices de similitud, índices de asociación, coeficientes de comunidad, etc. Se pueden citar como los más usados el coeficiente de Jaccard, el coeficiente de correlación puntual, *test* de  $\chi^2$ , *test* de correlación para datos cuantitativos, coeficiente de «simple matching», etc.

Debe tenerse en consideración que diferentes *tests*, si bien pueden ser monotónicos, suelen conducir a resultados distintos de una misma asociación o similitud que se está probando; por ello es importante la elección del índice o, aún mejor, comparar los resultados obtenidos por diferentes coeficientes.

El cálculo del o de los índices seleccionados a todos los pares posibles de objetos o de caracteres presentes en la matriz original de datos permite construir una nueva matriz elaborada o secundaria que puede ser de «tipo Q» (matriz de similitudes entre objetos) o de «tipo R» (matriz de similitudes entre caracteres). En la práctica, muchas veces es conveniente construir la matriz secundaria con las medidas de distancia, que son el complemento de las medidas de similitud.

La matriz secundaria en el sustrato básico a partir del cual las diversas técnicas de descripción y síntesis tratan de extraer relaciones (McIntosh, 1967; Whittaker, 1967). Veamos a continuación algunas de esas técnicas.

#### TRANSCRIPCIÓN O REORDENAMIENTO DE MATRICES

Si la matriz bruta se construyó sin un cierto orden o patrón, los datos de la matriz secundaria serán simplemente una reducción de la anterior pero no presentarán orden alguno.

Las matrices pueden reordenarse sobre la base de un criterio externo a la misma. Para un ejemplo fitosociológico, podemos disponer los censos en un orden creciente de altitud, latitud, temperatura, salinidad, humedad en general, datos de fundamental importancia que facilitan la interpretación del fenómeno estudiado. Aún más, los valores de los índices pueden agruparse en rangos que se representan mediante tramas, tonos o diseños. Estos se utilizan para reemplazar los valores numéricos en la matriz, la cual resulta entonces con una gran expresividad gráfica y permite reconocer zonas o núcleos de diferente similitud (véase, en páginas posteriores: semiología gráfica de Bertin).

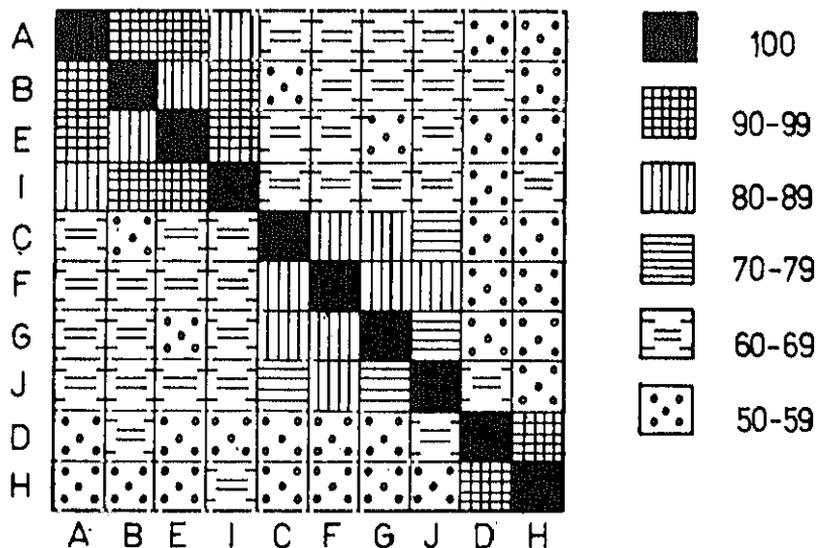
Otro mecanismo consiste en hacer el reordenamiento sobre criterios internos. Mediante prueba y error se tratan de disponer los valores más altos de similitud cercanos a la diagonal de la matriz. Una vez realizado esto, los índices pueden reemplazarse por tramas o tonos de intensidad decreciente cuando disminuye la similitud (véase Sokal y Sneath, 1963; ver gráfico 2).

Este tipo de matrices reordenadas es común entre psicólogos, sistemáticos y ecólogos.

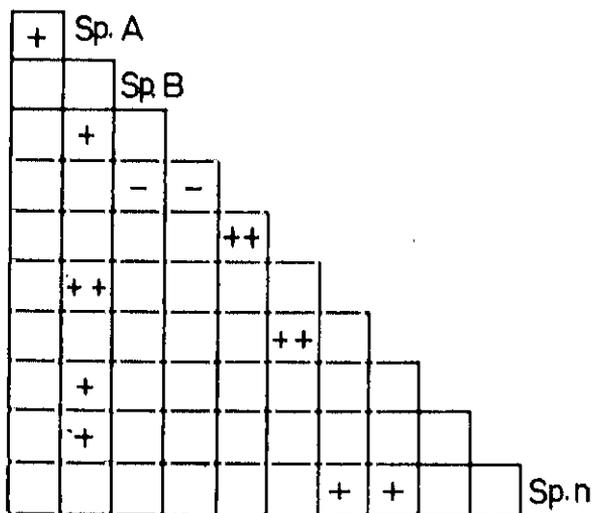
El reordenamiento de las matrices puede hacerse visualmente, pero la operación depende entonces de la interpretación del investigador y es por demás complicada si la masa de datos es muy grande. Existen, sin embargo, diversas técnicas objetivas de reordenamiento de matrices que dan resultados reproducibles (Sorensen, 1948; Clausen, 1957).

Esas matrices ordenadas reciben también el nombre de «diagramas de trellis».

Se puede representar sistemáticamente los parámetros  $a_1, a_2, a_n$  correspondiente a un objeto  $b$  por las series de Fourier en el que las amplitudes armónicas son los valores de los parámetros. Esto permite transcribir las columnas de una tabla en fórmulas matemáticas. Este método en realidad fue propuesto en 1968 por Speet quien utiliza la homología entre los datos obtenidos por un electrocardiograma, electroencefalograma y las características del sonido; a cada tipo de curva (los objetos) hacemos corresponder un sonido particular, por lo que dos objetos próximos se corresponden con dos sonidos vecinos, a



MATRIZ REORDENADA MEDIANTE DIFERENTES TRAMAS PARA DISTINTOS PORCENTAJES DE SIMILITUD. (tomada de SOKAL y SNEATH - 1963 )



MATRIZ DE ASOCIACION INTERESPECIFICA O ESCALOGRAMA DONDE SE INDICAN LAS ASOCIACIONES POSITIVAS Y NEGATIVAS SIGNIFICATIVAS.

partir de los cuales establecemos proximidades y clasificaciones; en nuestros términos sería un principio de clasificación para la elaboración de una tabla bruta.

Los métodos gráficos o visuales de J. Bertin merecen una mención aparte. En efecto, el autor de la *Semiología gráfica* (1967) propone un método de tratamiento de tablas más elaboradas que una simple representación o transcripción. Después de haber codificado cada caso de una tabla o cuadro por gamas de grises de diferentes valores (cuyos extremos son el negro y el blanco), las permutaciones manuales de los casos permiten aproximar los objetos de mayor similitud (sobre las líneas: permutaciones sobre los objetos; sobre las columnas: permutaciones sobre los parámetros).

### MÉTODOS DE CLASIFICACIÓN

Las técnicas clasificatorias estadísticas se conocen bajo el nombre genérico de «Taxonomía numérica». Esta puede definirse como la evaluación numérica de la afinidad o similitud entre unidades taxonómicas y el agrupamiento de estas unidades en taxa sobre la base de sus afinidades (Sokal y Sneath, 1963).

En Botánica puede tratarse, por ejemplo, de poner en evidencia los subespacios de un mismo taxón a partir de una tabla de datos donde las plantas están caracterizadas por un cierto número de medidas. En Arqueología o en otras disciplinas en las que se tratan del reconocimiento de formas, la clasificación automática puede ser utilizada para obtener tipologías. Más generalmente podemos decir que la clasificación es la base de todo conocimiento. El lenguaje, sin ir más lejos, es también una forma de clasificación: cada concepto expresa una clase de ocurrencia de hecho del mismo tipo. Por extensión, en inteligencia artificial la clasificación es considerada como un procedimiento derivado de las computadoras en el cual se da una información de orden semántico que no aparece en la tabla inicial en forma clara. La información dada por una clasificación se sitúa, en efecto, a nivel semántico: no se trata de atender a un resultado verdadero o falso, probable o improbable, sino más bien aprovechable o desechable.

Nos preguntamos: desechable, ¿por qué?

Por múltiples razones: los reagrupamientos, cuando aparecen, permiten la construcción de nuevas hipótesis. Los reagrupamientos que no existen refuerzan la posibilidad de poder separar los parámetros utilizados. Las clases obtenidas y sus imbricaciones aseguran una vía concisa y estructurada de datos. Las clases significativas entrañan la definición de funciones de decisiones, permitiendo atribuir un nuevo objeto a las clases más próximas.

La clasificación de elementos se usa en tantas disciplinas, que ello explica el gran desarrollo y la variedad de técnicas clasificatorias de que se dispone hoy día.

Considerando la gran diversidad de técnicas clasificatorias existentes se indicarán primero las características fundamentales que hacen a la operación de los métodos (Lambert y Dale, 1964).

Uno de los objetivos frecuentemente perseguidos es ubicar a un conjunto de elementos (cerámicos, viviendas, materiales, especies animales o vegetales, asociaciones vegetales, etc.) a clases o jerarquías de un sistema taxonómico preexistente conocido y aceptado. Estas técnicas se conocen bajo el nombre de «análisis discriminante». No obstante, por lo general, el investigador desea sacar el máximo provecho de la información que tiene y trata de crear a partir de ella el marco sistemático apropiado. Por otra parte, muchas veces se carece de un esquema de clasificación y es entonces necesario recurrir a estas últimas técnicas citadas.

grupos de objetos o caracteres a distintos niveles de homogeneidad,

La clasificación puede ser jerárquica, o sea, la misma establece generándose unidades cada vez mayores de manera análoga a las especies, géneros, familias, etc., de los sistemas de clasificación biológica. O bien puede ser reticulada, en donde los grupos no se reúnen en unidades mayores o menores, sino que pertenecen a un único nivel jerárquico.

De acuerdo con la forma en que proceden los métodos para la elaboración de los grupos, la clasificación puede ser divisiva o aglomerativa. Los métodos divisivos parten del total de caracteres u objetos (o sea, de la máxima información) y en base principalmente a las diferencias, van subdividiendo ese total en grupos sucesivamente más pequeños. Los métodos aglomerativos parten de cada unidad individual (carácter u objeto) de mínimo contenido de información y los van combinando en grupos en base a sus similitudes, hasta que finalmente se obtiene un único grupo. Los métodos divisivos son, por lo general, más rápidos debido a que se puede parar el proceso de subdivisión cuando se llega a un nivel deseado (que, por lo general, no es cada individuo clasificado). Los métodos aglomerativos suelen requerir mayor tiempo de computación, ya que se parte de cada individuo de clasificación y es necesario concluir con el análisis para obtener las jerarquías superiores.

Considerando el número de atributos o parámetros que intervienen en la clasificación, el método puede ser monotético, si los grupos se combinan o separan por la presencia de uno solo de esos atributos; o bien politéticos, si se usa más de un atributo para ese fin.

Los métodos politéticos dan como resultado clasificaciones más informativas y estables (Lambert y Dale, 1964).

Los métodos pueden ser autoestructurados si establecen las relaciones entre los relevamientos a partir de una matriz de tipo Q, para generar los grupos. O bien métodos de estructura transpuesta que emplean las relaciones entre caracteres establecidos en una matriz de tipo R para obtener las relaciones entre los objetos.

Por último, se pueden asignar factores de peso para dar mayor importancia a aquellos caracteres u objetos que se estiman más relevantes.

Generalizando, se puede señalar que un método debe operar usando la mayor cantidad de información disponible con el objeto de maximizar las diferencias entre los grupos (o minimizar las diferencias entre los grupos) dentro de los límites y restricciones que imponen las posibilidades de cálculo y costo.

Los principales métodos de clasificación automática también difieren según el tipo de información que resulta de su aplicación. Todos estos métodos, como ya señalamos, se fundan sobre la elección de un índice de semejanza. Vemos en un ejemplo sencillo como se procede para acceder a una clasificación:

Sean dos objetos  $a$  y  $b$  caracterizados por un número finito de parámetros cuyo índice de semejanza es en este ejemplo la distancia euclidiana cuadrática:

$d_1(a, b)$  = suma de los cuadrados de las diferencias de los valores tomados por  $a$  y  $b$  sobre cada parámetro;

otros índices pueden ser, por ejemplo:

$d_2(a, b)$  = suma de los cuadrados de las diferencias ponderadas por la suma de todos los valores tomados por  $a$  y  $b$  sobre cada parámetro.

En lenguaje matemático sería:

$$d_1(a, b) = \sum_{j=1}^h (a_j - b_j)^2$$

$$d_2(a, b) = \sum_{j=1}^h \frac{1}{p(j)} (a_j - b_j)^2$$

donde  $p(j)$  es la suma de los valores tomados para los parámetros  $j$ .

Retomado lo expuesto —clasificaciones jerárquicas— podemos decir que una vez definido un índice de similitud existen numerosos criterios matemáticos que permiten extraer clases homogéneas y bien

separadas. Podemos decir, por ejemplo, que dos clases son «buenas» si dos objetos en una misma clase se asemejan más que dos objetos tomados en clases diferentes. En base a los coeficientes obtenidos y volcados en una matriz secundaria podemos representar proximidades mediante un «árbol» o dendrograma (gráfico 3).

El inconveniente que encontramos es que calcular manualmente una relación jerárquica de más de 20 ó 30 objetos resulta un trabajo realmente engorroso, es necesario el uso de computadoras (cuyo empleo comienza en 1950), que permiten trabajos con muchos objetos en muy poco tiempo.

Desde hace treinta años, aproximadamente, han aparecido cantidad de publicaciones referidas al tema y un sinnúmero de programas que permiten la construcción de jerarquías y la salida en poco tiempo de cálculos computados. Cada aproximación que aparece después de esta época ha permitido esclarecer de más en más el análisis de los datos.

Los métodos de «partición» presentan la ventaja de permitir el tratamiento de tablas o cuadros de grandes dimensiones (se puede clasificar, por ejemplo, 10.000 objetos que hayan reunido ciertas condiciones y obtener una tipología); su objetivo, como hemos señalado, es el dar clases de objetos bien incorporados, bien separados o de intersección vacía.

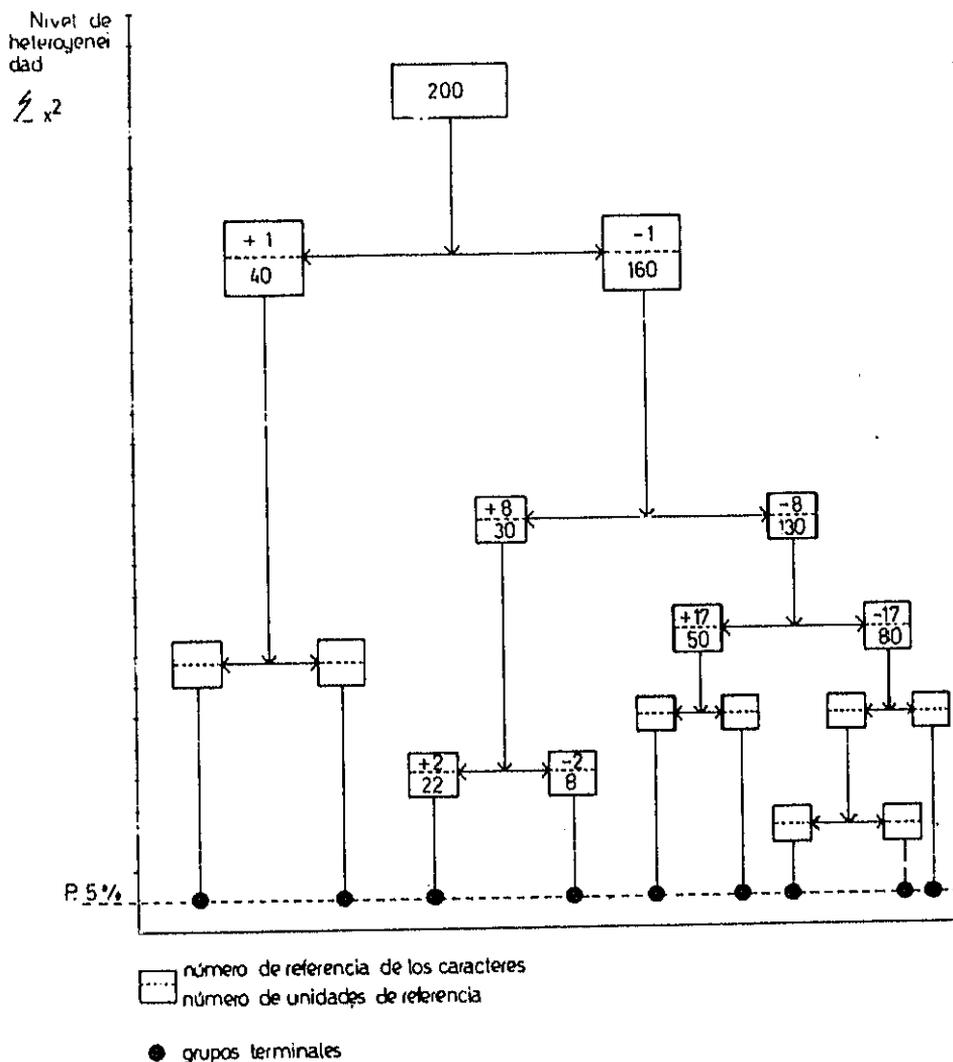
En fitosociología, uno de los métodos de clasificación más usados es el de análisis de asociación normal de Williams y Lambert (1959, 1960), que indicaremos mediante un diagrama de flujo (gráfico 4).

El resultado de la aplicación del método a un conjunto de relevamientos de vegetación se expresa en un dendrograma, que muestra la marcha del proceso de clasificación jerárquica en grupos que se obtienen en cada subdivisión y el grado de heterogeneidad de cada grupo.

El método es jerárquico porque da grupos a niveles diferentes de heterogeneidad, es divisivo porque parte del total de los censos, monotético cualitativo porque usa sólo datos de presencia-ausencia de especies, y de estructura transpuesta porque parte de un espacio de tipo R (matriz de asociación interespecífica) para clasificar grupos de relevamientos.

Los métodos de asociación interespecífica, aplicados en Ecología, tienen como premisa que si dos o más especies están asociadas entre sí, esto indica la acción de algún factor importante (permanente o de pulso, inerte, biológico o antrópico) que es causa de la asociación. Por ello la jerarquía facilita la interpretación y la búsqueda de aquellos factores que actúan a cada nivel.

Como ejemplo del método aglomerativo politético puede verse el método de análisis de información de Williams, Lambert y Lance (1966).



EJEMPLO HIPOTETICO DE DENDROGRAMA DE SALIDA DEL ANALISIS DE ASOCIACION NORMAL

## MÉTODOS DE ORDENAMIENTO

Los métodos de clasificación tienen como finalidad clasificar unidades de referencia. Esto es posible si se considera que la naturaleza está compuesta por unidades integradas, definidas y discretas que pueden ser combinadas para formar clases o tipos abstractos que son reflejo de entidades naturales en el mundo real. Si esto es válido, o sea, si las entidades forman parte de la estructura de la naturaleza y no únicamente parte de la estructura de un sistema de clasificación, es lógico que apliquemos técnicas clasificatorias.

Si, en cambio, suponemos y observamos que no existen entidades bien definidas, que hay una transición gradual entre las mismas, o, dicho en otras palabras, que la naturaleza o parte de ella cambia continuamente y no puede diferenciarse, excepto arbitrariamente, grupos de entidades asociadas (concepto de *continuum*), la clasificación resultante no será natural aunque pueda ser útil y necesaria.

Ordenamiento es el proceso de distribución de las muestras en relación a uno o más gradientes o ejes de variación (Coodall, 1954, citado en Whittaker, 1967). Sin duda la secuencia teórica antes evocada y el concepto de *continuum* son aplicables a un conjunto de disciplinas tal como Ecología, Antropología, Etnología, Geología Histórica, etc. En Antropología o Arqueología, donde, por ejemplo, pueden existir contactos culturales más o menos importantes, se pueden reconocer influencias de los patrones de comportamiento en la elaboración de viviendas, cerámica, etc., en los cuales se hace evidente el *continuum* cultural. En Ecología estas ideas fueron señaladas por diversos autores desde hace medio siglo (Gleason, 1926; Ramensky, 1926) y han dado lugar a lo que hoy se conoce como «teoría individualista» (Whittaker, 1970). Si participamos de esta idea podemos generalizar que el reconocimiento de grupos de objetos o caracteres será en tales casos arbitraria, ya que son extraídas subjetivamente de una variación continua de los mismos a lo largo de un gradiente (ambiental, cultural, temporal, etc.). (Véase, para distintas especialidades, los trabajos de Gleason, 1926; Ramensky, 1926; Whittaker, 1967; Lévi-Strauss, 1949).

Según Pielou (1969), como métodos para sintetizar los resultados de un muestreo, el ordenamiento tiene dos criterios arbitrarios para definir las clases y no es necesario suponer que las distintas clases (si existe alguna) están jerárquicamente relacionadas.

La forma de selección o construcción de los ejes da como resultado dos tipos básicamente diferentes de ordenamiento:

1. Ordenamiento directo o análisis directo de gradientes: en el que representan el rango de variación de factores ambientales (o externos) medidos o estimados en las distintas unidades de muestreo bajo análisis.

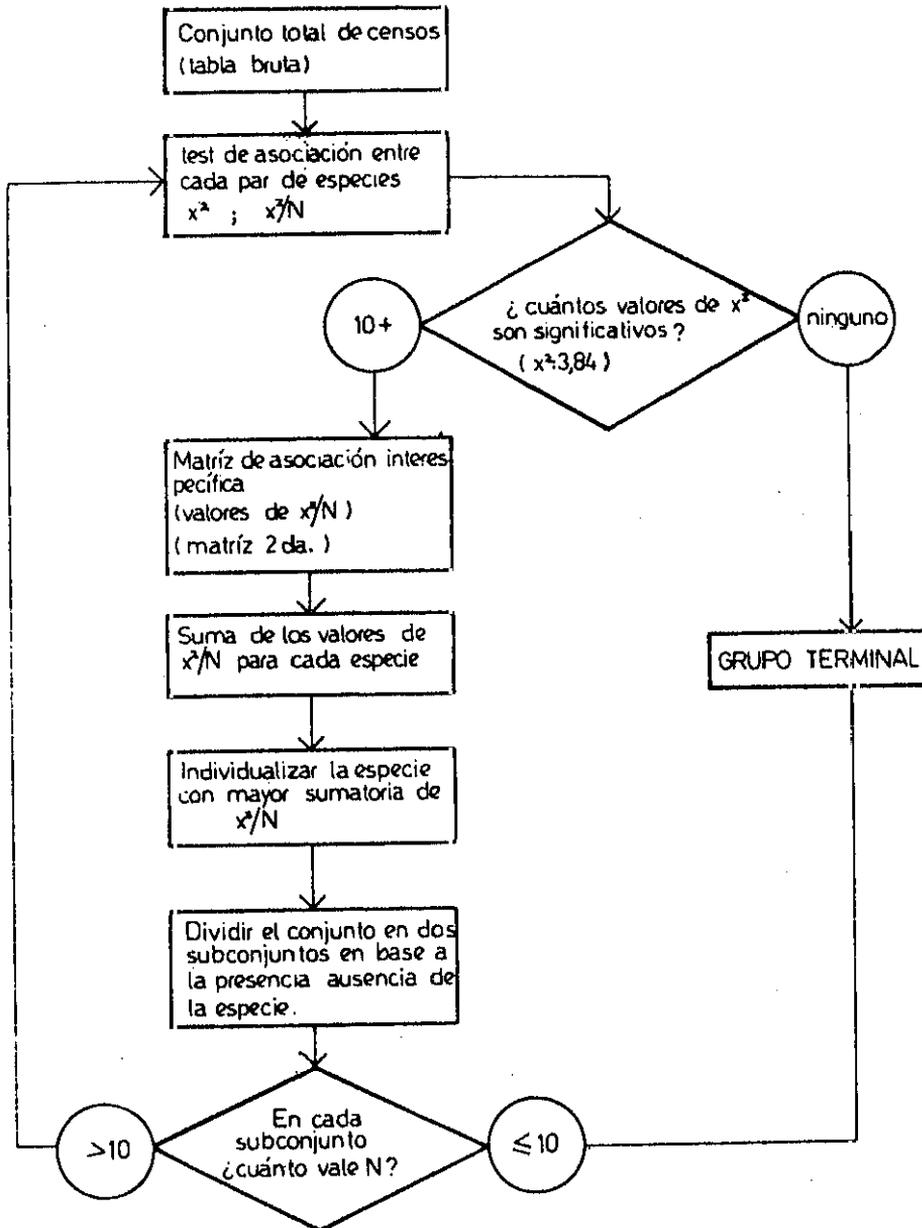


DIAGRAMA DE FLUJO DE ANALISIS DE ASOCIACION NORMAL DE WILLIAMS Y LAMBERT TOMADO Y MODIFICADO DE MONASTERIO Y SARMIENTO.

2. Ordenamiento comparativo o análisis indirecto de gradiente: en el que los ejes son construcciones matemáticas efectuadas en base a las distancias, diferencias o disimilitud existente entre los propios objetos que estamos ordenando.

Las técnicas de análisis directo de gradiente son de gran uso en Ecología. Pueden emplearse seleccionando, por ejemplo, en una montaña un gradiente altitudinal o bien de exposición, pendiente, etc., a lo largo del cual se efectúan relevamientos. Posteriormente esta información de vegetación y ambiente se vuelca sobre gráficos de distribución de especies o de comunidades (gráfico 5). Uno de los autores que más ha trabajado con este tipo de ordenamientos es R. H. Whittaker.

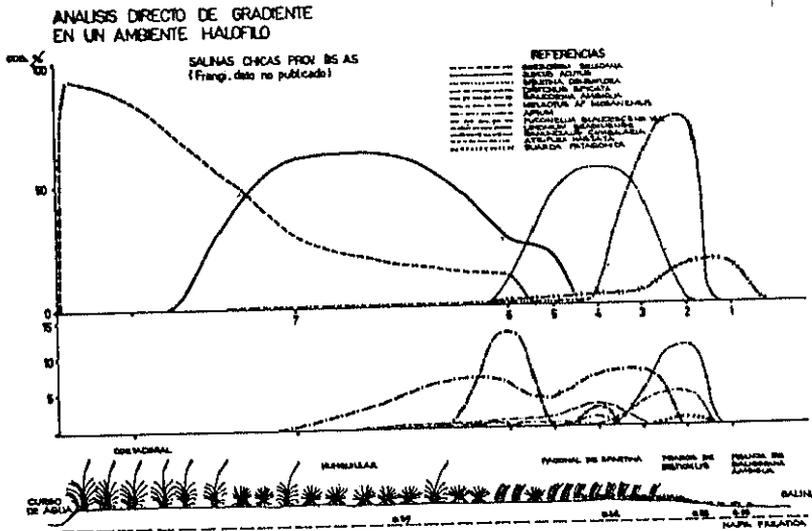
Entre las técnicas de ordenamiento indirecto de gradiente más simples, pero de muy buenos resultados en Ecología, se encuentra el método de ordenamiento simple u ordenamiento polar de Bray y Curtis (1957). En el diagrama de flujo adjunto se indica cómo opera el método cuando se aplica a un conjunto de relevamientos para ordenarlos en un sistema biaxial. El diagrama está simplificado y se remite al lector a la bibliografía (Bray y Curtis, 1957; Beals, 1960; Newsome y Dix, 1968) para una explicación del método, aplicaciones, modificaciones, fórmulas involucradas y *test* de ordenamiento. El resultado de este método se expresa en gráficos como el de la representación presentada aquí (gráfico 6).

Cuando esta técnica es aplicada a problemas fitoecológicos existen algunas suposiciones de gran interés para el ecólogo. Una de ellas es que el o los ejes resultantes de ordenamiento de las distancias están relacionadas con variables activas y selectivas del medio. Ordenamiento y plantas ayudan así al ecólogo a detectar cuáles son las variables simples o complejas que están correlacionadas con el patrón observado de la vegetación. Eventualmente se puede hacer un ordenamiento en tres dimensiones, pero esto suele complicar la representación.

#### EL ANÁLISIS FACTORIAL

El análisis factorial es un método general de análisis de datos cuyo nombre se extrae de los trabajos que durante años realizaron los psicólogos y que actualmente ponen en marcha ecólogos, antropólogos, lingüistas<sup>3</sup>. Como hemos comentado anteriormente, ellos en efecto han desarrollado una labor fundante intentando formalizar en un modelo general los resultados obtenidos para un individuo *i* según un *test j*, lo que sería igual a una combinación lineal de parámetros representando el valor tomado por el factor K para el individuo *i*.

<sup>3</sup> DAGNELIE, P., «Contribution á l'étude des communautés végétales par l'analyse factorielle». *Bull. Serv., Carte Phytogeogr.*, B, 5, 7-71, 93-195, 1960.



Fórmula:

$$x_{ij} = \sum_{k=1}^p i_k f_{jk} + b_{ij}$$

los estadísticos se han abocado a descubrir métodos que permitan la explotación de este modelo ligado y ligando el valor de un dato a la combinación de muchos factores no observables. Es importante recordar que existen problemas geométricos simples subyacentes. Se trata de reconstruir las numerosas coordenadas de un punto a partir de un número reducido de ellas (valores de los factores) sobre una marca apropiada. Es un problema que intentan resolver los dos métodos de análisis factorial descriptivo (que serían los actualmente más utilizados en Francia), el «análisis en correspondencia» y el «análisis en componentes principales»<sup>4</sup>.

Sería imposible exponer en este artículo los aspectos técnicos de tales métodos (cuya explicación y aplicación podemos ver en el trabajo realizado en el alero Ambrosetti —Lahitte y Calandra, 1977-78—); sólo intentaremos dar un bosquejo de los mismos, para luego dar algunos ejemplos de aplicación que dan una idea de las posibilidades de estos «instrumentos de observación».

Una tabla o cuadro rectangular de datos a  $n$  líneas y  $p$  columnas puede ser representada por los puntos de un espacio a  $p$  o a  $n$  dimensiones (de la misma forma que un punto en un espacio a tres dimensiones es caracterizado por una serie de tres coordenadas, un punto en un espacio a  $p$  dimensiones será un conjunto a  $p$  números). Las  $n$  líneas de una tabla o cuadro estarán representadas por  $n$  puntos en un espacio a  $p$  dimensiones. Entre esos puntos es posible evaluar las distancias elegidas de manera tal que podemos describir las similitudes entre tales líneas. Evidentemente, la visualización de la nube de puntos y el conjunto de similitudes entre líneas no puede realizarse en nuestro caso más que en un espacio a tres dimensiones (gráfico 7).

Intentaremos, por tanto, obtener un gráfico, imagen de la nube de puntos del espacio a  $n$  dimensiones, representando la nube por un plano esperando que las proximidades relativas entre los puntos y las líneas de la tabla no sean deformadas por esta operación. Naturalmente, la geometría de la nube de puntos, la distancia utilizada y el criterio de ajuste varían según las técnicas; ellas deben ser elegidas de forma tal que den a los resultados una forma de lectura simple.

<sup>4</sup> Es importante tener en cuenta que este método presenta una serie de inconvenientes —tal como la experiencia lo demuestra— cuando lo referimos a datos de naturaleza biológica. De ahí que deben tomarse precauciones para la interpretación de los ejes en términos de factores biológicos, ya que éstos se constituyen en base a criterios puramente matemáticos y a partir de un postulado que implica que cada componente es función lineal de la especie (situación que no se corresponde con la situación real-natural).

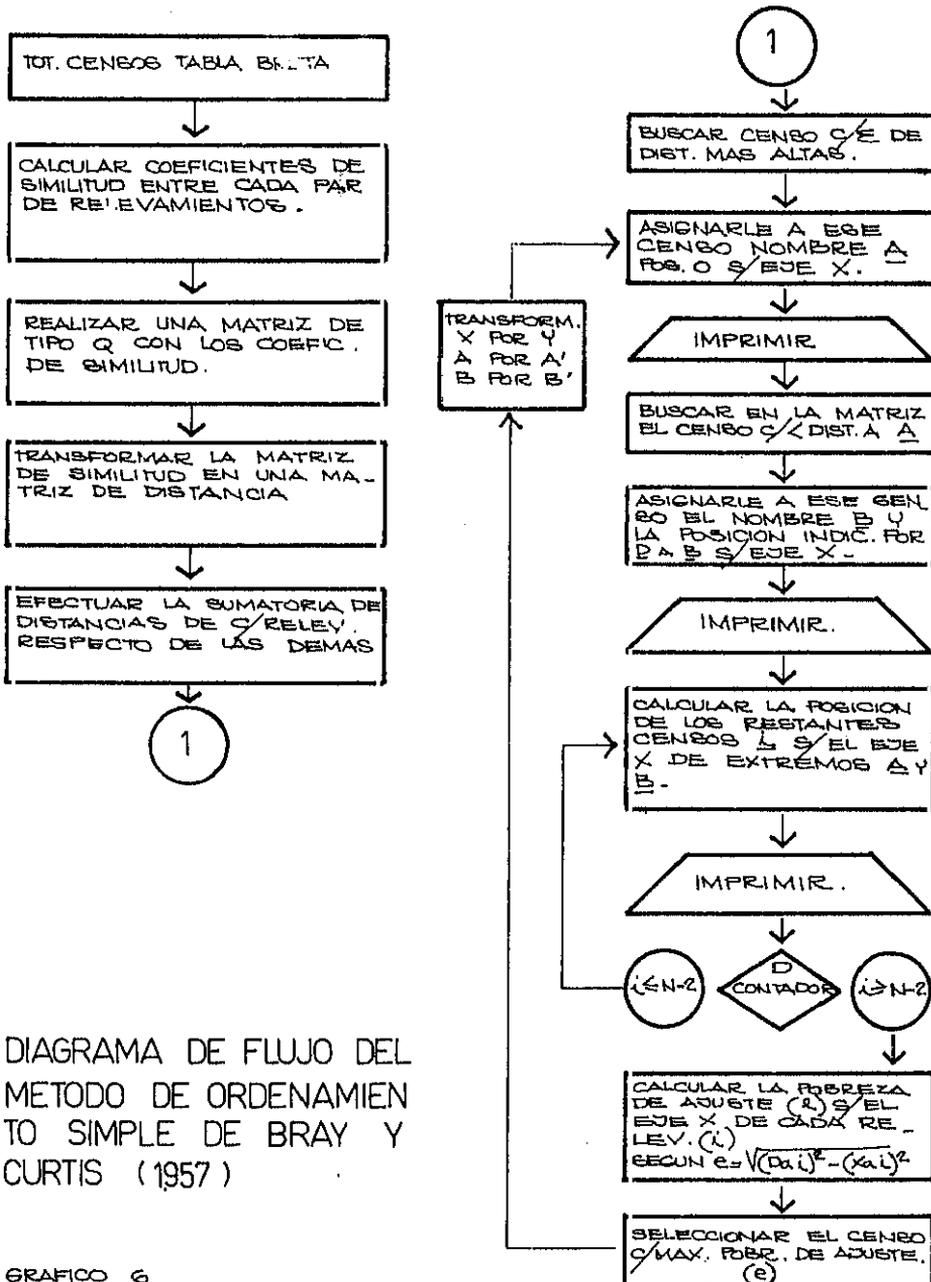


DIAGRAMA DE FLUJO DEL METODO DE ORDENAMIENTO SIMPLE DE BRAY Y CURTIS (1957)

El primer método «análisis en componentes principales» se refiere particularmente a las tablas de medidas relativas a los individuos. Tomemos, por ejemplo, una tabla compuesta por líneas que representan actividades cotidianas para una cierta cantidad de individuos, para los cuales hemos medido, por medio de un cuestionario y/o encuesta apropiada, el tiempo consagrado a cada una de las actividades; tema y forma de análisis que puede resultar de vital importancia para los antropólogos que intentan sistematizar las secuencias de actividades domésticas (por ejemplo, un día de vida), o como en Ecología el ordenamiento de los censos, en dos o más especies representadas por los ejes. El análisis en «componentes principales» pone en juego una distancia que permite interpretar las proximidades existentes entre las variables en término de correlación. De esta forma se obtiene una representación plana de la «matriz de correlación» relativa a las actividades o especies (tabla simétrica que contiene coeficientes de correlación calculados para cada pareja de variables). El eje de las abscisas o «primer factor» es el espacio a una dimensión que se ajusta mejor (en el sentido de los «menores cuadrados»), la sombra o nube de puntos. El eje de las ordenadas o «segundo factor» permite engendrar el mejor subespacio a dos dimensiones, siempre según el mismo criterio.

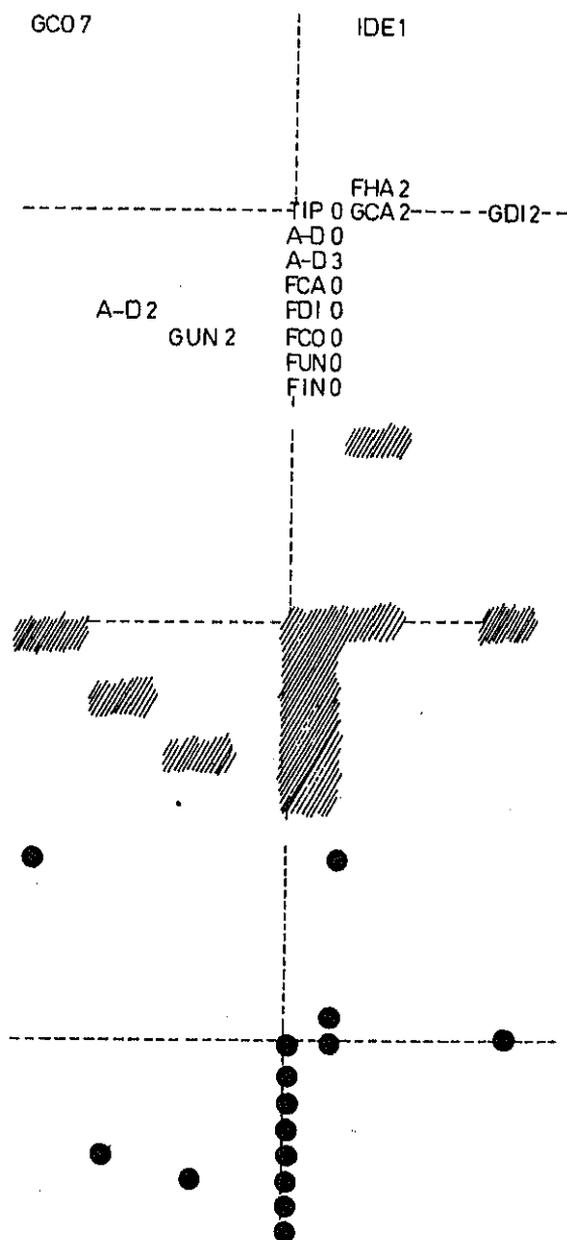
Podemos así extraer ejes hasta llegar a un procedimiento de detención (el más corriente: una simulación por computadora) que nos advierte que la información aportado por el último eje extraído es del orden del tamaño de las fluctuaciones de la muestra, los errores de medida, etc. Las operaciones efectuadas sobre los valores numéricos, tomados como punto de partida, permiten un cierto tipo de reglas de lectura:

1. Que dos o más actividades o especies son las que dan lugar a correlaciones.
2. Que los puntos representativos sean próximos los unos de los otros.
3. Que ellos estén alejados o apartados de su origen.

Por el contrario, dos actividades sirven como correlaciones negativas (que varían en sentido inverso) y corresponderían a los puntos que están en general excepcionalmente alejados o apartados.

El análisis de correspondencia tiene un empleo más general que el método precedente debido particularmente a la claridad de sus reglas de interpretación.

Las clásicas «tablas cruzadas», llamadas «tablas de contingencia» o «dependencia», son las que mejor se presentan a cualquier tipo de análisis. Consideremos, por ejemplo, una tabla a  $n$  líneas y  $p$  columnas



EL ANALISIS FACTORIAL APLICADO A LAS REPRESENTACIONES DEL ALERO AMBROSETTI. CIBOIS-LAHITTE, H-CALANDRA, H- (dato no publicado-oct. 1977)

GRAFICO 7

en la que se cruzan  $x$  categorías y  $x$  causas de distintas clases. En la intersección de una línea y una columna de esa tabla encontraremos el número de individuos que verifican las características correspondientes. El análisis de correspondencia nos va a descubrir las similitudes existentes entre perfiles-líneas, o sea, una característica está diferenciada por una serie de  $x$  números que describen cada una de las partes de cada una de las causas en el total de la categoría. De la misma forma, una causa se describe o reconoce por una serie de  $x$  porcentajes que representan su perfil. Otra forma es la llamada «representación simultánea» de las «líneas» y las «columnas» de la tabla en el plano de los primeros factores de análisis de correspondencia; esta figura es el mejor subespacio a dos dimensiones (en el sentido de criterio de ajuste) susceptible de restituir las proximidades entre puntos-perfiles, similitudes calculadas luego o según una fórmula adaptada a la tabla de contingencia (distancia llamada de «*khi-deux*»). A pesar de poseer características muy sugestivas, esta figura debe ser interpretada según reglas estrictas ligadas o dependientes del modo de construcción de este tipo de representación. También las proximidades entre puntos tienen más sentido que los puntos que se alejan del origen de los ejes. Las proximidades observables entre puntos-líneas o dos puntos-columna se interpretan ahora en términos de «similitud de perfiles». Para los puntos situados en la periferia del gráfico es posible interpretar la posición de un punto-columna, por ejemplo, en relación a todos los puntos-línea (y viceversa). De la misma forma podemos tener una idea del perfil de las categorías y de las causas examinando la posición relativa de este último punto y sus diferentes categorías de actividad. Otros parámetros de interpretación más delicada permiten afinar las distintas reglas de lectura y apreciar la cualidad de la representación.

Es claro que la lectura de estos gráficos necesita de una gran experiencia y tal vez de un alto grado de prudencia. Pero el especialista en estadística puede extraer una gran masa de información de manera sistemática.

No es raro que los ejes extraídos (sobre todo los primeros) exhiban dimensiones subyacentes al fenómeno estudiado, conocidas y por consecuencia identificables. Como en todos los procesos complejos de observación la fase de interpretación demanda a la vez una gran masa de instrumentos de análisis como también un conocimiento profundo del campo de estudio. Esta etapa es normalmente facilitada por un procedimiento que se designa como «adjunción de variables suplementarias» o también «pasivas» (en oposición a las llamadas «variables analizadas o activas»), que consiste en hacer aparecer sobre la tipología obtenida la nueva información.

Una vez elegido el método de tratamiento y los datos incorporados a la computadora, las operaciones que siguen son automáticas: ellas

dan siempre un resultado. Así resulta útil distinguir en la estructura obtenida y representada un resultado: aquello que es imputable a una hipotética estructura real de los datos. Es igualmente útil apreciar el efecto de los errores instrumentales, es decir, los rasgos inherentes a los principios mismos de las técnicas empleadas. La incidencia de ciertos tipos de errores pueden ser descubiertos por las simulaciones. El principio de prueba de validación es el siguiente: simulamos datos, tomamos nuevos datos arbitrarios que pueden tener una cierta relación de vecindad con los datos reales. Por ejemplo, elegir los datos perturbadores o alterados por los errores o las permutaciones introducidas voluntariamente y elegidas en función de nuestro conocimiento del fenómeno estudiado. La simulación permite también poner a prueba las hipótesis de trabajo adaptadas a los problemas que se plantea el investigador; mientras que los procedimientos de estadística clásica no pueden poner a prueba más que las hipótesis que dan lugar a una formulación matemática, lo que restringe evidentemente el campo de posibilidades.

Los datos simulados son analizados de la misma forma que los datos reales y luego los diferentes resultados son comparados. Podemos así observar según el caso la estabilidad de la representación obtenida o por el contrario su fragilidad.

La evaluación de los errores llamados «instrumentales», es decir, introducidos por el método mismo, se realizan con la ayuda de los «procedimientos de contraste». Por esto, luego de una primera etapa se analiza la tabla para observar una estructura conocida, luego en una segunda fase comparamos la estructura obtenida por el programa de análisis en el ejemplo estudiado. Los diversos métodos de análisis precedente nos permiten representar de forma práctica los datos teniendo en cuenta su similitud o diferencia de semejanza (obviamente según un código que implique la puesta en evidencia de los rasgos o datos tabulados).

Las técnicas de análisis multivariado reposan generalmente en el hecho que los valores de sus variables se distribuyen según la ley de Laplace-Gauss (representada por la famosa curva). Con la ley indicada o «ley normal a una dimensión», podemos encontrar en la naturaleza la ley «normal multidimensional», hipótesis necesariamente especificada por el tratamiento operativo de una población caracterizada por muchas variables. El análisis de los datos permite revelar las inferencias dudosas. Se contenta en efecto con ayudar a los investigadores a observar la realidad multidimensional sin requerir la ayuda de hipótesis complejas o conclusiones erróneas. Este nuevo instrumento de observación debe su existencia a la aparición de las computadoras, siendo sus fundamentos los principios con que comenzamos a exponer este trabajo.

## TÉCNICAS DE PLEXUS O GRAPHOS

Con los datos de la matriz de similitud o distancia es posible construir una representación gráfica que expresa las relaciones entre los objetos o caracteres.

Se confecciona así un modelo espacial en la que objetos o caracteres están indicados por puntos separados por distancias o conectados por líneas de distinto tipo que indican la similitud relativa entre aquellos. Como ya se ha señalado, la representación en un plano de un espacio multidimensional que ejemplifica las relaciones entre objetos o caracteres es siempre complicada e implica la pérdida de algún grado de información. Por lo general, se utilizan representaciones bidimensionales o a lo sumo tridimensionales.

A partir de estos diagramas es posible efectuar una clasificación, si bien es cierto que ella dependerá del grado de asociación positiva, proximidad o similitud entre los elementos que conforman cada grupo representado y de la disimilitud, asociación negativa o distancia entre los elementos de grupos diferentes. Por otra parte, la formación de «nubes» de puntos más definidas o dispersas influyen también en la ubicación de los límites entre los grupos, si dicha ubicación se hace en forma subjetiva.

Dichos grupos o «clusters» también pueden reconocerse por métodos objetivos (ver Sorensen, 1948; Hopkins, 1957; Araos, Sarmiento y Monasterio, 1971).

Este tipo de representaciones con puntos y líneas es tema de estudio de la teoría de los grafos. Los grafos más utilizados en Ciencias Naturales son los «no dirigidos», donde las líneas que unen los puntos no indican una dirección. Este tipo de grafos se pueden transformar en «grafos dirigidos», utilizando flechas que indiquen el carácter u objeto más similar al considerado (Flamant, Bergé; Lahite y Calandra).

De igual forma que es posible utilizar diagramas de plexus para clasificar, es posible emplearlos para realizar un ordenamiento; y aun a través del ordenamiento es posible, por ejemplo, en Ecología, relacionar variables del medio, distribuciones de especies y características de las comunidades (McIntosh, 1973).

Las técnicas de matrices y plexus disminuyen su eficacia a medida que aumenta el número de caracteres u objetos, el rango de variación de los agrupamientos y el número de direcciones de las interacciones del medio y los grupos.

Como conclusión podemos decir que los métodos estadísticos de análisis de datos, sin pretensión de indicar que las aquí esbozadas se aplican a todos los problemas de las Ciencias Naturales, constituyen herramientas objetivas de análisis y síntesis. Las distintas aproximaciones (clasificación y ordenamiento), no son mutuamente excluyentes,

lo son complementarias. Cada una de ellas es particularmente útil para distintos fines. Como la identificación de tipos (la identificación y/o combinación de caracteres y/o propiedades en un todo armónico. Ejemplo, tipos cerámicos, tipo de vegetación, conjuntos habitacionales, estratos geológicos, tipos de suelos, etc.) mediante una clasificación que facilita la interpretación del fenómeno, objeto, área o hecho en cuestión.

Asimismo, el ordenamiento es de gran utilidad para la posterior interpretación de variaciones de unidades de referencia en áreas o situaciones donde los límites de contacto son imprecisos, indefinidos y de gran cobertura.

## BIBLIOGRAFIA

- ANDERBERG, M. R.:  
1973 *Cluster analysis for applications*. Academia Press. USA.
- ARAOZ, J.; SARMIENTO, G., y MONASTERIO, M.:  
1971 «An essay in the use of association and dissociation measures in phytosociological classification». *J. Ecol.*, 59 (1): 39-50.
- BEALS, E.:  
1960 «Forest Bird Communities in the Apostle Islands of Wisconsin». *Wilson Bull.*, 72: 156-181.
- BENZECRI, J. B.:  
1963 *L'analyse des données*. Dunod, París.
- BERGE, C. de:  
1963 *Theorie des graphes et ses applications*. Dunod, París.
- BERTIER, P., y BOUROCHÉ, J. M.:  
1975 *Analyse des données multidimensionnelles*. PUF, París.
- BORDAZ van H., y BORDAZ, J.:  
1970 *A computer assisted pattern recognition method of classification on seriation applied to archaeological material*. CNRS, París.
- BORILLO, M.:  
1970 *La verification des hypotheses en archeologie*. CNRS, París.
- BRAY, J. R., y CURTIS, J. T.:  
1957 «An ordination of Upland Forest Communities of Southern Wisconsin». *Ecol. Monog.*, 27: 325-349.
- CIBOIS, Ph.:  
1978 *L'analyse factorielle? mais c'est tres simple*. Ed. CNRS, París.
- CLAUSEN, J. J.:  
1957 «A comparison of some methods of establishing plantcommunity patterns». *Bot. Tidsskr.*, 53: 253-278.  
1957 «A phytosociological ordination of the conifer swamps of Wisconsin». *Ecology*, 38: 226-233.
- CLARKE, D.:  
1968 *Analytical archaeology*. Methuern & Co., Londres.  
1973 *Models in archaeology*, Londres.
- COOLEY, W., y LHINES, P. R.:  
1971 *Multivariate data analysis*.
- COWGIL, G.:  
1969 «Archaeology applications of factor, cluster and proximity analysis». *American Antiquity*, 33, Salt Lake City.

- DAGNELIE, P.:  
1960 «Contribution a l'étude des communautés végétales par l'analyse factorielle». *BULL, Serv. Carte phytogéogra. B.*, 5: 7-71, 93-195.
- FERNÁNDEZ DE LA VEGA, Wenceslao:  
1970 «A propos d'une formulation mathématique des problèmes de filiation de manuscrits», ed. CNRS, cap. 30-71, Marsella.
- FLAMANT, C.:  
1972 «L'analyse de similitud». *C. R. Operationelle*, 4-2, Bruselas.
- GARDIN, J. C.:  
1970 *Archaeologie et calculateur*. CADA, CNRS, París.
- GLHASON, H. A.:  
1926 «The individualistic concept of the plant association». *Bull. Torrey Bot. Club*, 53: 7-26.
- GOODALL, D. W.:  
1954 «Vegetational classification and vegetational continua». *Angew. Pfl. Soziol.*, 1: 168-82.  
1963 «The continuum and the individualistic association». *Vegetatio*, 11: 297-316.
- HARTIGAN, J. A.:  
1975 *Clustering algorithms*.
- HOPKINS, B.:  
1957 «Pattern in the Plant community». *J. Ecol.*, 45: 451-463.
- IHM, P.:  
1960 «Détermination automatique des objets de l'Age du Bronze et le problème de la découverte automatique de groupes culturelles». *EPHE*, 6°, París.
- KSHISSAGAR, M.:  
1972 *Multivariate analysis*. Marcel Dekker.
- LAHITTE, H., y CALANDRA, H.:  
1977 «Codificación arqueológica de las viviendas del período formativo, sus representaciones tipológicas sobre tarjetas perforadas». *Rev. del Centenario*, pp. 173-195. Museo de La Plata.
- LAHITTE, H., y D'ANTONI, H.:  
1972 «Proyecto de código para el análisis descriptivo y documental en Palinología». *Rev. de la CIC*, de R. Negro, serie 1.
- LAHITTE, H.:  
1977 «A panorama of theoretical archaeology». *Current Antr.*, Chicago.
- LAMBERT, J. M., y DALE, M. B.:  
1964 «The use of statistics in Phytosociology». *Adv. in Ecol. Res.* (JBCragg, edit.), 2, Academic Press, Nueva York.
- LEBBART, P. J. Fenelon:  
1975 *Statistique et informatique appliquée*, Dunod, París.
- LERMAN, J. C.:  
1970 *Les bases de la classification automatique*. Gauthier, Villers.
- MARANDA, P.:  
1966 «Analyse quantitative et qualitative de mythes sur ordinateur». *Bol. CNRS*, París-Marsella.
- McINTOSH, R. P.:  
1967 «The continuum concept of vegetation». *Bot. Rev.*, 33: 130-187.  
1973 «Matrix and Plexus techniques», en *Handbook of Vegetation Science*. Part V., W. Junk, The Hague.
- MONASTERIO, M., y SARMIENTO, G.:  
«Análisis ecológico y fitosociológico de la Sabana en la estación biológica de los Llanos». *Bol. Soc. Venez. Cs. Nat.*, 27 (113-114): 477-524.

PAGES, R.:

1959 «L'analyse codée». *Chiffres*, n.º 2.

RAMENSKY, L. G.:

1926 «Die Grundgesetz massigketiten im Aufbau der Vegetationsdecke». *Botan. Centralblatt*, NF, 7: 453-455.

ROUANET, H., y LEPINE, D.:

1967 «A propos de l'analyse des domies Selon Benzecri». *Annee psychologique*.

STONE, J., y cols.:

1962 «The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information». *Behavioral Science*.

SOKAL y SNEATH:

1963 *Principles of numerical taxonomy*, Freeman, S. Fco.

VICKERY, B. C.:

1961 *On retrieval system theory*. Butter W., Londres.

WHITTAKER, R. H.:

1956 «Vegetation of the Great Smoky Mountains». *Ecol. Mon.*

1970 *Communities and Ecosystems*. Macmillan, N. Y.

WILLIAMS y LAMBERT:

1959- *Multivariate methods in plant ecology*.

1966

WILLIAMS, LAMBERT y LANCE:

1966 «Multivariate methods in plant ecology V, Similarity analysis and information analysis, *J Ecol.*, 54: 427-445.

WHITTAKER, R. H.:

1967 «Gradient Analysis y Vegetation», *Biol. Rev.* 42: 207-264.