1 ejem. 16

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Facultad de Ciencias

ALGUNOS METODOS JERARQUICOS Y OTROS SUBDOMINANTES DE TAXONOMIA NUMERICA

Tesis presentada por

LUZ MARIA MARTINEZ MALO

para la obtención de la licenciatura

en matemáticas

México D. F.

Diciembre de 1979

INDICE

INTRODUCCION

La idea de este trabajo surge de la gran necesidad que existe de aplicar las diferentes ramas matemáticas a ciencias sociales; pero principalmente de la carencia de una teoría que sea accesible a los investigadores sociales, ya que, en general, las matemáticas que pueden aplicar se exponen dentro de marcos matemáticos exclusivamente. Por esta razón en este trabajo se procura, hasta donde es posible, poner al alcance de todas las personas interesadas, independientemente de su especialidad, la teoría de clasificación que se basa en Taxonomía Numérica. También se presenta, en capítulos posteriores, la misma teoría formalizada únicamente con conceptos matemáticos; de esta forma esta exposición se dirige a todas aquellas personas que precisen de la aplicación de técnicas taxonómicas, ya sea como herramienta para manejar su información, o como material de documentación matemática.

Este trabajo se basa esencialmente en los conceptos presentados por Nicholas Jardine y Robin Sibson en su libro "Mathematical Taxonomy", del año 1971. Se expone una idea general acerca de los diferentes métodos que utilizan estos autores para construir sistemas de clasificación a partir de Coeficientes de Disimilaridad, señalando las ventajas y desventajas de su uso. También se pretende formalizar el desarrollo matemático de esta teoría. La presentación está hecha de la siguiente manera: los capítulos 1, 2 y 5, así como el apéndice A, son accesibles a todo lector. En el capítulo 1 se exponen algunos conceptos elementales de clasificación; el de Coeficiente de Disimilaridad, que es en el que se basa toda la teoría de Taxonomía Numérica; el de Dendrograma, gráfica que se utiliza para representar sistemas de clasificación; y por último algunos métodos de clasificación Jerárquica: el de Conexión Simple (Single Linkage), el de Conexión Completa (Complete Linkage), y ciertos Métodos Promedio. El capítulo 2 trata de métodos más generalizados, los llamados Métodos B_k, que son, prin-

cipalmente, uno de los objetivos de este trabajo. El capítulo 5 presenta muy brevemente algunas de las ideas básicas que deben regir la aplicación de los distintos sistemas de clasificación. El apéndice A muestra las diferentes opciones para elegir los Coeficientes de Disimilaridad o Similaridad adecuados, así como algunos Métodos Promedio. Los capítulos 3 y 4 corresponden a la formalización matemática de los conceptos expuestos en los capítulos 1 y 2 respectivamente. El apéndice B establece la equivalencia entre las definiciones dadas por N. Jardine y R. Sibson en el libro mencionado anteriormente, con las utilizadas en este trabajo para los mismos propósitos. A pesar de que para los lectores con intereses meramente matemáticos puede resultar repetitivo el leer todos los capítulos, se recomienda que así lo hagan, ya que podría ser de utilidad para lograr una mejor comprensión de la teoría.

Por último, esta tesis no pretende obtener una visión completa de Taxonomía Numérica, ya que existe un gran número de trabajos que sugieren múltiples posibilidades en el mismo campo.

Se hace explícito un agradecimiento al Dr. Jaime Litvak King, director del Instituto de Investigaciones Antropológicas de la Universidad Nacional Autónoma de México, por las facilidades que se prestaron para la realización de este trabajo; a los actuarios Arturo López y Manuel Román, a las matemáticos Guadalupe Ibargüengoitia y Margarita Chávez, por haber revisado la tesis; al matemático Alejandro Sierra, por su ayuda y por las discusiones que sirvieron para aclarar muchos puntos; y principalmente al matemático Guillermo Espinosa, quien dirigió el desarrollo del trabajo, por su interés, paciencia, y por sus valiosas aportaciones.

CAPITULO 1

1.1 Clasificación

Una de las formas más comunes para asimilar los objetos e ideas que nos rodean en la vida diaria, es el ordenamiento de ellos en grupos o clases, de tal manera que resulte más sencillo su manejo. Así, se entiende por *clasificación* el proceso de ordenar objetos en grupos o clases, de acuerdo a las relaciones que presenten con base en sus características. En cuanto a las maneras de establecer criterios para relacionar a los objetos que se quieren clasificar, existen diferentes enfoques:[1]

Relación Fenética. Se relacionan los objetos de acuerdo al parecido o similaridad que presentan en el momento de su estudio, sin tomar en consideración el origen de ese parecido o el desarrollo que éste siguió en el pasado. Es decir, se relacionan los objetos con base en su fenotipo.

Relación Cladística. Relación que expresa el parecido de los objetos de acuerdo a sus antecedentes o ancestros comunes; es decir, de acuerdo a su desarrollo evolutivo. Se ilustra por medio de árboles o ramificaciones, y se estudia cuántas ramas hay, cuál rama proviene de otra y en qué secuencia.

Relación Cronológica. Está basada en relaciones con respecto a una escala evolutiva de tiempo. Este tipo de relación es, por ejemplo, la utilizada en un proceso de seriación en arqueología (ordenamiento cronológico del desarrollo de hechos u obje-

[1] Véase "Numerical Taxonomy", de Peter H. A. Sneath y Robert R. Sokal, 1973.

tos, para la construcción de hipótesis).

Relación Filética o Filogenética. Relación que incluye a las relaciones fenéticas, cladísticas y cronológicas (aunque algunos autores sólo consideran a las cladísticas).

Al estudio de la clasificación es a lo que se le conoce con el nombre de *Taxonomía*, incluyendo a la identificación, o sea al proceso de asignar objetos no reconocidos en las clases establecidas una vez hecha la clasificación. *Taxononomía Numérica* es el agrupamiento por métodos numéricos (ya sean matemáticos, estadísticos, etc.) de los objetos en consideración. Para la utilización de estos métodos, se requiere transformar la información original de los objetos en cantidades numéricas.

Las relaciones en las que se basa la Taxonomía son principalmente fenéticas, debido a que es difícil poseer un criterio que combine todas las consideraciones. Además, se piensa que las relaciones fenéticas son aquellas que dan como consecuencia clasificaciones satisfactorias debido a las siguientes razones:[2]

1. La clasificación por relaciones fenéticas es accesible a todos los grupos, mientras que la clasificación filogenética o simplemente la cladística, requieren de ciertos conocimientos o inferencias históricas sobre los caminos evolutivos de los objetos.
2. En la mayoría de los casos, los restos fósiles rela-

[2]Véase "Numerical Taxonomy", de Peter H. A. Sneath y Robert R. Sokal, 1973.

cionados con los objetos no son disponibles, y aunque lo fueran, deben ser interpretados con criterios fenéticos.

Por otro lado, los grupos formados de acuerdo a cualquiera de los criterios anteriores pueden ser de dos tipos, *Monotéticos* y *Politéticos*[3]. Los frupos monotéticos son los formados por todos aquellos elementos que poseen un único conjunto de características en común, y el poseerlas todas es equivalente a ser miembro de ese grupo. Los grupos politéticos son los constituídos por objetos que poseen "un buen número" de las características de un conjunto dado, aunque no necesariamente todas ellas. Ya que los métodos numéricos de clasificación tienden a combinar y a dar mayor peso a aquellas características cuyo valor numérico es más grande, los grupos taxonómicos son grupos politéticos.

En este trabajo se introducen algunos de los métodos más usados en Taxonomía Numérica, y también se presentan otros métodos menos usuales, que son, en parte, el objeto de este estudio. Después de exponer los conceptos intuitivos de clasificación y grupo taxonómico, a continuación se tratan de explicar estos mismos conceptos, de tal manera que correspondan hasta donde es posible, con su definición matemática.

Existen diversas formas de interpretar el concepto de clasificación. Las *Clasificaciones No Jerárquicas* son aque-

[3] Véase "Numerical Taxonomy", de Peter H. A. Sneath y Robert R. Sokal, 1973.

llas en las que se obtiene una "partición" del conjunto de objetos en grupos ajenos (*i.e.* no existe ningún objeto que pertenezca a grupos diferentes), de tal manera que todo objeto forma parte de algún grupo. En las *Clasificaciones Jerárquicas* también se obtienen particiones del conjunto de objetos; sin embargo, es posible conseguir una sucesión de particiones diferentes con base en qué tan "fina" se quiera hacer la clasificación. En esta sucesión de agrupaciones se puede obtener desde una partición en la que los grupos estén formados por cada uno de los objetos de manera individual, hasta una partición formada por un solo conjunto constituído por todos los objetos que se van a clasificar. Esto sucede pasando por niveles intermedios en los que los grupos se van combinando y creciendo hasta llegar a formar uno sólo. Esta idea se puede ilustrar con el siguiente ejemplo: supóngase que se quiere hacer una clasificación de un cierto grupo de personas; si lo que se desea es obtener grupos de individuos que pertenezcan a la misma especie, obviamente se tendrá un solo grupo formado por todas las personas; si además se quiere que los grupos estén constituídos por individuos de la misma especie y el mismo sexo, se obtendrán dos grupos ajenos. Si se desea tomar en consideración la especie, el sexo y la misma estatura, el número de grupos resultantes tenderá a aumentar; de esta manera, a medida que se imponen nuevas características, las restricciones para la formación de grupos son mayores y se llega al punto en que se obtienen grupos formados por cada u

na de las personas estudiadas, como elementos únicos. Las *Clasificaciones Numéricamente Estratificadas* son agrupaciones menos restringidas, ya que la sucesión de conjuntos, formada con la misma idea que en el caso anterior, puede estar constituída por grupos que se traslapen entre sí; es decir, un objeto puede pertenecer a dos o más grupos diferentes. Se imponen condiciones sobre el número de elementos en el traslape de grupos, o sobre el "diámetro" del traslape[4]. Este trabajo se concentra en presentar algunos métodos Jerárquicos y otros Numéricamente Estratificados.

1.2 Coeficientes de Disimilaridad

A partir de esta sección, todas las consideraciones que se hacen son en relación a lo que la Taxonomía Numérica se refiere, ya que todos los métodos que se exponen son solamente numéricos.

El paso más importante en Taxonomía Numérica es la estimación del parecido entre los objetos que se van a clasificar. La primera etapa consiste en la extracción de la información sobre las características que se tomarán en consideración. La elección de estas características está basada en el tipo de relaciones[5] entre objetos, que para el caso convenga. Esta información original puede tomarse de textos o directamente de

[4] Véase la nota 3 del capítulo 2.
[5] Véase la sección 1.1 de este capítulo.

de los objetos mediante métodos de análisis objetivo, y puede ser un conocimiento de tipo presencia-ausencia (si se posee o no se posee la característica en consideración), o se puede tratar de una tabla de abundancias o porcentajes (en qué grado o cantidad se poseen las características; es decir, de tipo distribucional). Por ejemplo, si se estuvieran analizando ciertos tiestos u otros objetos arqueológicos, se podría contar con tablas de información original como las siguientes:[6]

componente químico / tiesto	K_2O	MgO	CaO	FeO_3	Al_2O_3	SiO_2
I	1.70	2.80	4.60	5.10	15.80	52.90
II	0.45	3.50	4.45	6.00	16.75	63.55
III	0.90	2.70	4.50	5.95	14.70	59.70
IV	0.80	3.30	3.85	5.70	14.55	58.80
V	1.00	3.10	3.95	6.40	16.20	64.00
VI	1.20	6.10	4.30	5.90	16.00	60.50
VII	1.05	5.40	4.20	6.25	15.80	55.30
VIII	1.55	2.30	4.00	5.40	17.05	59.40

tabla 1

característica / pieza	cafe alisado	café pulido	café estacado	negro pulido	rojo pulido	naranja pulido	rojo café	blanco fugitivo
A	1	1	1	0	0	0	0	0
B	1	1	0	1	0	0	1	1
C	1	0	0	0	0	0	1	0
D	1	1	0	1	1	0	0	1
E	0	0	0	1	1	0	1	1
F	1	1	1	1	0	0	0	1
G	1	1	1	1	0	1	0	1
H	1	1	0	1	1	0	0	0
I	0	1	1	0	1	0	0	1

tabla 2

[6] Tablas obtenidas de estudios de las arqueólogas Angela Minzoni y María del Carmen Serra.

En la tabla 1 la información original consiste en el porcentaje por unidad analizada, de los diferentes componentes químicos del material con que están hechos los tiestos. En la tabla 2 las piezas arqueológicas están analizadas de acuerdo a características cromáticas; si el objeto posee la característica, se anota un 1, si no la posee se asigna un 0.

Para calcular la disimilaridad (o similaridad) entre las parejas de objetos es necesario contar con información numérica acerca de ellos. En un sistema de clasificación un *Coeficiente de Disimilaridad (Similaridad)* podría pensarse como la asignación de un valor (generalmente numérico) que mida en cierto sentido, las diferencias (o parecidos) entre cada par de objetos. A continuación se habla solamente de coeficientes de disimilaridad, aunque todas las afirmaciones que aquí se hacen también tienen validez para los coeficientes de similaridad, es decir, para los casos en los que se trabaja con parecidos. A pesar de que generalmente se cuenta con coeficientes de disimilaridad numéricos, puede suceder que estas disimilitudes se expresen por medio de un rango u orden comparativo no numérico; sin embargo, los métodos aquí utilizados no parten de coeficientes de este tipo.

La manera de calcular estos coeficientes es muy diversa y depende del tipo de tabla de información original que se tenga. Algunos de los coeficientes más usados son:[7]

1. Coeficiente de Gower

[7] Véase el apéndice A, sección A.1.

2. Métrica de Minkowski
3. Distancia de Manhattan
4. Distancia Euclidiana
5. Coeficiente de Jaccard
6. Coeficiente de Dice
7. Poporción de Presencias y Ausencias Comunes
8. Presencias y Ausencias Comunes
9. Presencias Comunes

Por ejemplo, en el caso de la tabla 1, la Distancia de Manhattan (d) entre los tiestos III y IV se calcula así:

$$d(III,IV)=|0.90-1.20|+|2.70-6.10|+|4.50-4.30|+$$
$$|5.95-5.90|+|14.70-16.00|+|59.70-60.50|$$
$$=0.30+3.40+0.20+0.05+1.30+0.80$$
$$=6.05$$

Análogamente se procede con el resto de los objetos, y así se obtiene la siguiente tabla (o matriz) de disimilaridades entre los tiestos:

tiestos

tiestos	I	II	III	IV	V	VI	VII	VIII
I	0	14.60	9.75	9.90	14.45	12.70	7.20	9.30
II	14.60	0	7.25	8.40	2.85	7.40	12.20	7.80
III	9.75	7.25	0	2.65	7.30	6.05	8.95	4.75
IV	9.90	8.40	2.65	0	8.05	7.00	8.00	5.30
V	14.45	2.85	7.30	8.05	0	7.75	11.85	7.85
VI	12.70	7.40	6.05	7.00	7.75	0	6.70	7.10
VII	7.20	12.20	8.95	8.00	11.85	6.70	0	10.00
VIII	9.30	7.80	4.75	5.30	7.85	7.10	10.00	0

tabla 3

En la casilla correspondiente al cuarto renglón y a la quinta columna, se encuentra el coeficiente de disimilaridad entre los objetos IV y V, y así sucesivamente. En la diagonal de la matriz está la disimilaridad entre un objeto y él mismo; por lo que en estos lugares se obtienen ceros. Este tipo de tabla es simétrica, es decir, la disimilaridad entre el objeto "X" y el "Y" es la misma que la que hay entre el objeto "Y" y el "X".

Utilizando ahora el coeficiente de Presencias y Ausencias Comunes con la información contenida en la tabla 2, el coeficiente (s) de similaridad entre las piezas D y E se obtiene sumando al número de características presentes en ambos objetos, el número de características ausentes en éstos; es decir:

$$s(D, E)=5$$

La tabla de similaridades completa es la siguiente:

piezas

piezas	A	B	C	D	E	F	G	H	I
A	8	4	5	4	1	6	5	5	5
B	4	8	5	6	5	6	5	5	3
C	5	5	8	3	4	3	2	4	2
D	4	6	3	8	5	6	5	7	5
E	1	5	4	5	8	3	2	4	4
F	6	6	3	6	3	8	7	5	5
G	5	5	2	5	2	7	8	4	4
H	5	5	4	7	4	5	4	8	4
I	5	3	2	5	4	5	4	4	8

tabla 4

En este caso en la diagonal se encuentra el número máximo de coincidencias que puede haber, que es el número total de caracs

terísticas, ya que un objeto coincide consigo mismo en todas ellas.

La importancia de los coeficientes de disimilaridad radica en que son el dato básico que requieren los diversos métodos de Taxonomía Numérica.

1.3 Dendrogramas

Algunas clasificaciones de tipo Jerárquico son representadas por medio de unas gráficas llamadas *Dendrogramas*. Los dendrogramas son un tipo de árboles cuyos vértices terminales se asocian a los objetos del conjunto que se va a clasificar, y en cuyas ramificaciones se marcan valores numéricos que miden la disimilaridad entre los objetos. Los dendrogramas ofrecen una buena visualización de los agrupamientos de los objetos con base en sus disimilaridades. Los métodos Jerárquicos requieren de tablas de coeficientes de disimilaridad, como dato de entrada, y producen dendrogramas como el de la siguiente figura.

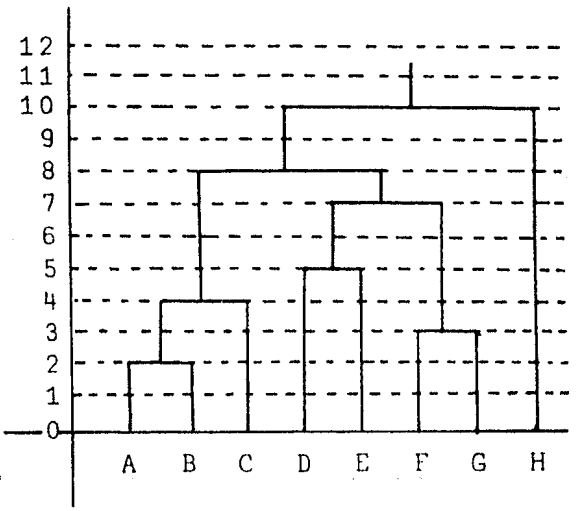


figura 1

El conjunto de objetos que se quiere clasificar es el fromado por A, B, C, D, E, F, G y H. La escala de valores (niveles) de la izquierda mide las disimilaridades entre los objetos. Al leer estas gráficas desde su parte inferior, los grupos de objetos se van constituyendo a medida que las ramas del árbol se unen; las ramas correspondientes a los objetos más parecidos son las primeras en unirse, y mientras el parecido disminuye se unen los demás objetos. La manera de efectuar las lecturas de un dendrograma es la siguiente: se recorre una línea horizontal sobre la gráfica, comenzando desde su parte inferior, y en sentido ascendente se va moviendo a lo largo de todo el dendrograma. Para cada movimiento de la línea se tiene una familia de grupos en la clasificación, grupos que están representados por las uniones de las ramas. Si para una determinada posición de la línea todavía no se unen algunas ramas del árbol, entonces la familia de grupos está constituída por conjuntos cuyos únicos elementos son cada uno de los objetos. A medida que la línea se mueve hacia arriba, las uniones de las ramas comienzan a aparecer y los grupos anteriormente formados se combinan con otros para formar grupos más grandes. De esta manera, las familias de grupos en posiciones altas de la línea están constituídas por un menor número de grupos de más elementos. Para el dendrograma de la figura 1 se muestra a continuación el procedimiento descrito para algunas posiciones de la línea. En la figura 2 cada objeto forma un grupo para esa posición de la línea, hay 8 grupos, uno por cada objeto.

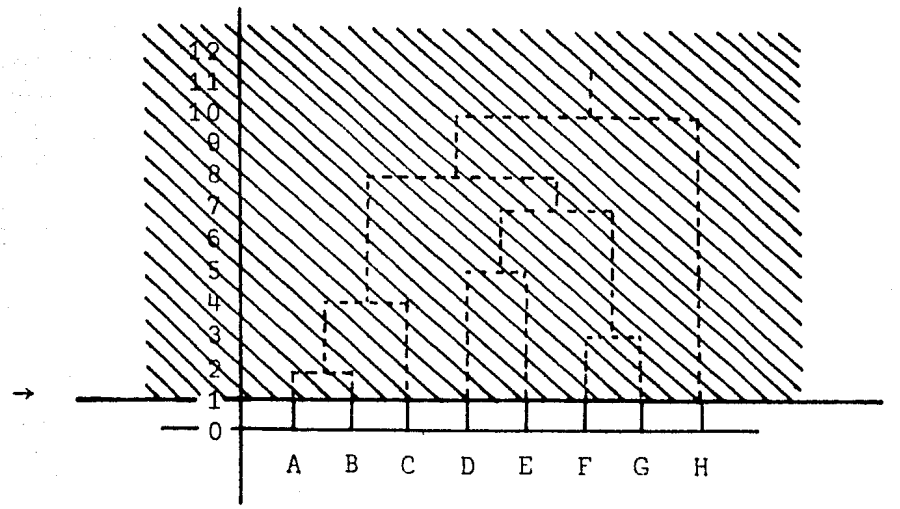


figura 2

Figura 3: en este caso hay 5 grupos, el formado por A, B y C; el de F y G; y D, E y H en un grupo cada uno.

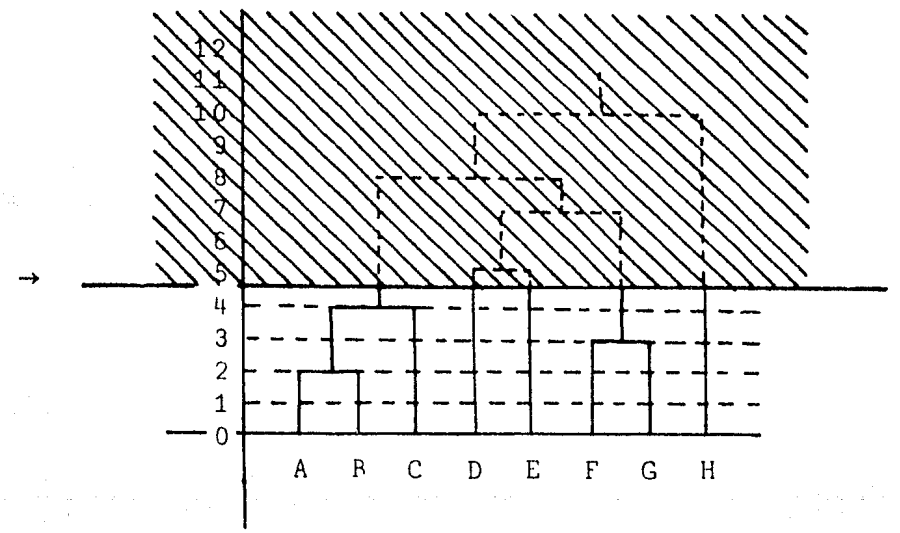


figura 3

Para un nivel un poco menor a 10, como en la figura 4, ya sólo existen 2 grupos, uno grande formado por todos los objetos sin incluir a H, y H en un grupo aisladamente. Este objeto es el más diferente de los restantes, ya que es el último en juntarse a todos para formar una familia de un solo grupo a partir

del nivel de asociación 10.

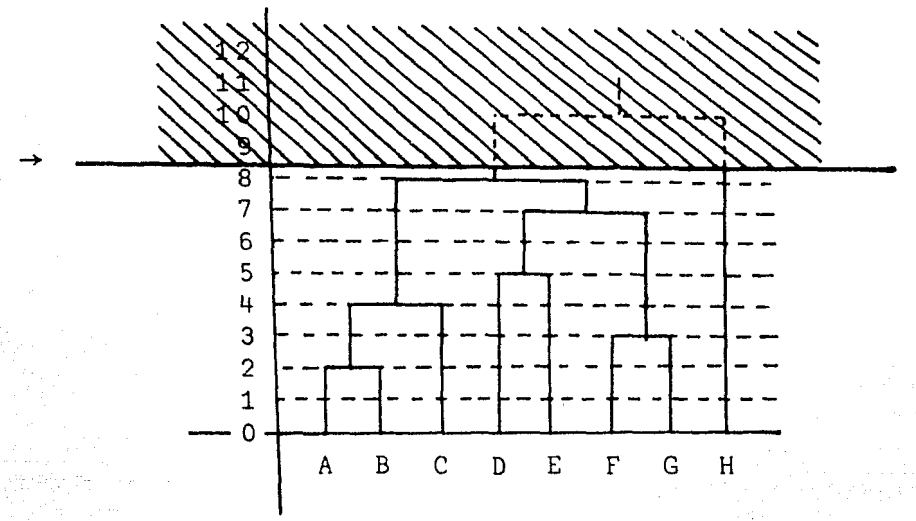


figura 4

Dado un dendrograma como el de la figura 1, es posible obtener la tabla de disimilaridades entre objetos o grupos de objetos, si se define el coeficiente de disimilaridad entre una pareja como el nivel más pequeño tal que esa pareja se encuentra unida en un mismo grupo. Por ejemplo, la pareja más parecida en el dendrograma anterior es la formada por A y B, cuya medida de disimilaridad es de 2 unidades; la disimilitud entre los objetos F y G es 3; entre el grupo formado por A y B, y el objeto C, el coeficiente es 4; así como la disimilaridad entre los grupos constituídos por A, B y C, y por D, E, F y G, es 8. La tabla completa para este conjunto de objetos es la número 5.

Al efectuar las lecturas del dendrograma en los diferentes niveles de asociación, lo que se obtiene son las familias de conjuntos que determinan los grupos de la clasificación. Ca

objetos

objetos	A	B	C	D	E	F	G	H
A	0	2	4	8	8	8	8	10
B	2	0	4	8	8	8	8	10
C	4	4	0	8	8	8	8	10
D	8	8	8	0	5	7	7	10
E	8	8	8	5	0	7	7	10
F	8	8	8	7	7	0	3	10
G	8	8	8	7	7	3	0	10
H	10	10	10	10	10	10	10	0

tabla 5

da uno de los objetos se encuentra presente sólo en uno de estos conjuntos para cada nivel, es decir, se trata de conjuntos ajenos. En niveles mayores, estos conjuntos pueden ser agrupados a su vez en conjuntos más grandes, de tal manera que finalmente todos los objetos están clasificados en una forma jerárquica. Esto quiere decir que si dos objetos están unidos en un determinado nivel, entonces lo siguen estando para niveles mayores. Como se puede observar en el ejemplo anterior, las familias de grupos para cada nivel se forman por conjuntos de objetos que se encuentran unidos en ese nivel o en niveles menores; así se obtiene entonces una sucesión de familias de grupos crecientes, que para este caso particular, comienza en el nivel 0, donde cada uno de los objetos constituye un grupo individualmente, y termina en el nivel 10, donde todos los objetos forman un único grupo.

De manera inversa, si lo que se tiene es una tabla de coeficientes de disimilaridad, como la tabla 5, es posible regresar a su dendrograma original, ya que cada número en la tabla

marca el nivel de asociación de las ramas correspondientes a las diferentes parejas de objetos. Se ordenan los números de la tabla, de menores a mayores y se selecciona el menor (que puede no ser único) para considerarlo como el primer nivel de asociación en el dendrograma. Todas las parejas de objetos que tengan este mismo coeficiente de disimilaridad se unen para formar grupos en este nivel; las ramas que todavía no se unen corresponden a objetos que permanecen aislados en un grupo (es decir, que sus coeficientes con respecto a los demás objetos son mayores que este primer nivel de asociación). De esta manera se obtiene la familia de grupos determinada por el nivel correspondiente al primer número escogido, y con ésto se construye el segmento inferior del dendrograma. Este razonamiento se repite tantas veces como números diferentes haya en la tabla de coeficientes de disimilaridad, y en cada paso se construye un segmento más del dendrograma, sobre el construído en el paso anterior. Supóngase que se tiene la tabla 5; los siguientes pasos ilustran con gráficas el proceso de construcción del dendrograma, los puntos representan a los objetos y las líneas a las uniones de dichos objetos.

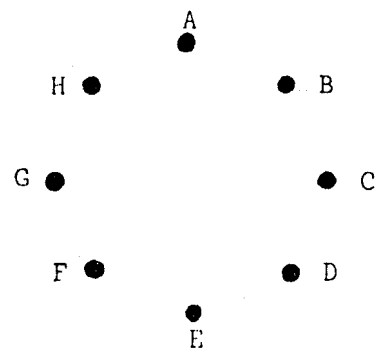


figura 5

Figura 5: 1. Se selecciona el número más pequeño de la tabla, el 0, que corresponde a los lugares de la diagonal de la matriz, puesto que un objeto se parece más a sí mismo que a cualquier otro. En esta figura están representados 8 grupos, uno por cada objeto.

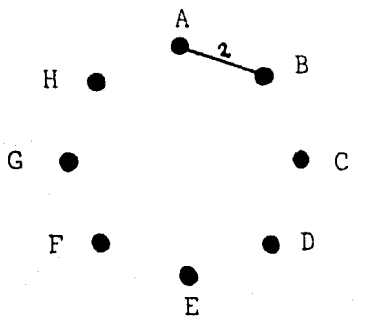


figura 6

Figura6: 2. El número más pequeño inmediato es el 2, correspondiente a la disimilitud entre los objetos A y B, que forman así un grupo.

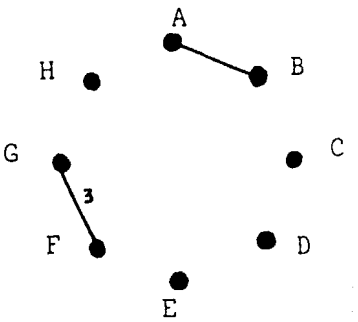


figura 7

Figura 7: 3. El siguiente número es el 3, que mide la disimilaridad entre F y G. Hay 6 grupos: A y B; C; D; E; F y G; y H.

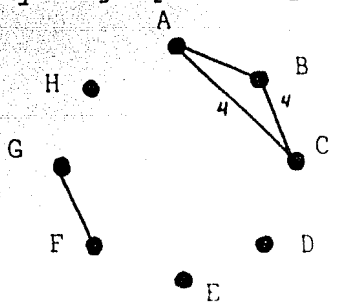


figura 8

Figura 8: 4. Se selecciona el número 4, que es el coeficiente entre A y C, y B y C. La familia de grupos es de 5 elementos: A, B y C; D; E; F y G; y H.

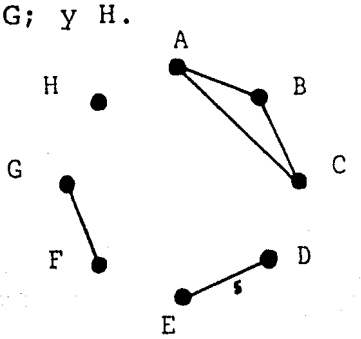


figura 9

Figura 9: 5. La disimilaridad entre E y D es 5; ahora hay 4 grupos.

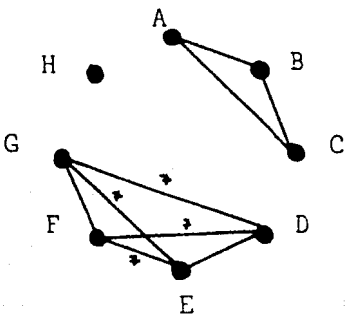


figura 10

Figura 10: 6. La disimilaridad entre D y F, D y G, E y F, y E y G es 7. Los grupos son 3: A, B y C; D, E, F y G; y H.

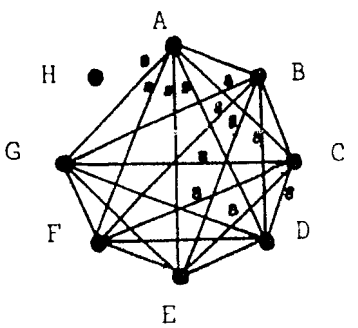


figura 11

Figura 11: 7. El siguiente número es el 8. Sólo falta el objeto H para formar un solo grupo con todos los demás elementos.

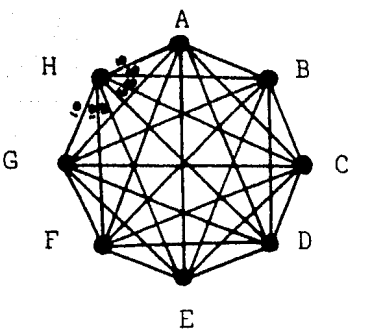


figura 12

Figura 12: 8. El número mayor es el 10, que mide la disimilitud entre H y el resto de los objetos. Esta es la gráfica completa, todos los objetos forman un solo grupo. Al ir de esta manera construyendo las familias de grupos para cada nivel, se obtiene el mismo dendrograma de la figura 1.

El proceso recién descrito no necesariamente puede llevarse a cabo. Considérese el siguiente ejemplo:

objetos

objetos	J	K	L	M	N
J	0	6	5	7	8
K	6	0	4	9	8
L	5	4	0	8	7
M	7	9	8	0	7
N	8	8	7	7	0

tabla 6

Al tratar de construir el dendrograma correspondiente a esta tabla de disimilaridades, surge un problema cuando se llega al nivel de asociación 5 y se trata de buscar la familia de grupos correspondiente. En este nivel el objeto J se une en un grupo con el objeto L, y este último ya está en un mismo grupo junto con el K; sin embargo la pareja J y K no está en ese

grupo todavía, ya que su coeficiente es 6 y por lo tanto no es posible construir esta familia de grupos, y, en consecuencia, el dendrograma tampoco. La gráfica correspondiente al nivel 5 es la siguiente:

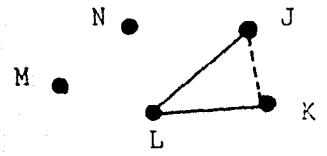


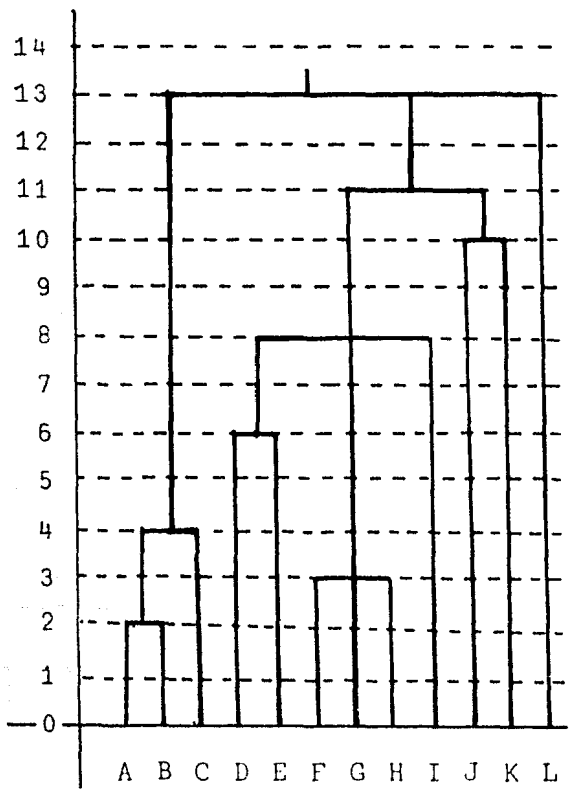
figura 14

En este nivel sólo aparecen las aristas JL y LK, y el grupo no está bien constituído como los de las gráficas anteriores, ya que la arista JK aparece hasta el nivel 6. Es posible demostrar[8] que para evitar casos como éste, se requiere de tablas que cumplan con la propiedad de que si se toman todas las posibles ternas de objetos del conjunto que se va a clasificar y los coeficientes de disimilaridad correspondientes a los pares de objetos que las forman, entonces el coeficiente más grande en cada una de las ternas se tiene que repetir por lo menos dos veces en cada una de ellas. Si una tabla cumple con esta propiedad, se dice que satisface la *Condición de Ultrametría*, y a los coeficientes de disimilaridad se les llama *Coeficientes Ultramétricos*.

También es posible demostrar[9] que dada una tabla ultramétrica, se le puede asociar un único dendrograma que la represente. Inversamente, dado un dendrograma, éste define a una ú-

[8] Véase el capítulo 3, sección 3.3.
[9] Véase el capítulo 3, sección 3.3.

nica tabla de disimilaridades ultramétrica. Debido a esta correspondencia, existe una identificación entre los dendrogramas y las tablas ultramétricas (figura 15). Los diferentes métodos de clasificación son maneras distintas de transformar tablas de coeficientes de disimilaridad dadas en tablas de coeficientes de disimilaridad ultramétricos, que se identifican con los dendrogramas; ésto se hace deformando la información original de alguna manera. Es, entonces, a partir de las tablas de disimilaridad ultramétricas que se construyen las familias de grupos para cada uno de los diferentes niveles.



	A	B	C	D	E	F	G	H	I	J	K	L
A	0	2	4	13	13	13	13	13	13	13	13	13
B		0	4	13	13	13	13	13	13	13	13	13
C			0	13	13	13	13	13	13	13	13	13
D				0	6	8	8	8	8	11	11	13
E					0	8	8	8	8	11	11	13
F						0	3	3	8	11	11	13
G							0	3	8	11	11	13
H								0	8	11	11	13
I									0	11	11	13
J										0	10	13
K											0	13
L												0

Figura 15

1.4 Métodos Jerárquicos

La Taxonomía Numérica parte de una matriz de disimilaridades,

con base en la cual, al aplicar los diferentes métodos, los grupos se forman, modificándose la estructura de esta matriz original. Uno de los métodos Jerárquicos más conocidos y tal vez el más sencillo es el *Método de Conexión Simple*. La manera en que transforma a una tabla de disimilaridades en otra ultramétrica es la siguiente: se busca el coeficiente más pequeño de la tabla; la pareja de objetos que lo posee se une para formar un grupo. La forma de recalcular la disimilaridad entre este nuevo grupo y el resto de los objetos, consiste en escoger el coeficiente más chico existente entre cada elemento del nuevo grupo y el objeto en consideración. Este principio se repite, y en cada paso después que dos grupos se han unido para formar uno nuevo, la disimilaridad entre éste y algún otro grupo se determina de igual manera que en el primer paso. Los grupos formados por este proceso son los que se obtienen en el dendrograma correspondiente al método de Conexión Simple, y los niveles donde se unen los diferentes objetos o grupos de objetos son los coeficientes de disimilaridad que se seleccionaron en cada paso del proceso. En las diferentes etapas o pasos, para la formación de grupos en un determinado nivel "X", se requiere solamente que exista una cadena entre los objetos cuyo coeficiente sea menor o igual a "X".

El siguiente es un ejemplo de la aplicación del método de Conexión Simple a una matriz de disimilaridades. En él se han enumerado los pasos en la formación de grupos con el objeto de observar la manera en que la matriz original va deformando su

estructura hasta convertirse en una matriz ultramétrica. También se ilustra con gráficas este mismo proceso; las dos explicaciones son equivalentes.

Sean A, B, C, D, E, F, G y H los objetos que se van a clasificar. La tabla de coeficientes de disimilaridad dada es la siguiente:

objetos

objetos	A	B	C	D	E	F	G	H
A	0	3	4	5	10	7	15	10
B		0	3	2	6	8	13	20
C			0	1	5	5	6	8
D				0	4	6	10	12
E					0	3	6	9
F						0	5	4
G							0	3
H								0

tabla 7

1. Se elige el coeficiente más pequeño, en este caso es el 0, correspondiente a los lugares en la diagonal de la tabla. Cada objeto forma un grupo de manera individual.

objetos

objetos	A	B	R	E	F	G	H
A	0	3	4	10	7	15	10
B		0	2	6	8	13	20
R			0	4	5	6	8
E				0	3	6	9
F					0	5	4
G						0	3
H							0

tabla 8

2. El siguiente es el correspondiente a C y D, que forman el grupo "R"; se calcula la disimilaridad entre el grupo nuevo y los demás objetos, de la manera descrita anteriormente.

objetos

objetos	A	S	E	F	G	H
A	0	3	10	7	15	10
S		0	4	5	6	8
E			0	3	6	9
F				0	5	4
G					0	3
H						0

tabla 9

3. El coeficiente más pequeño es el 2, que corresponde a B y R que se unen para formar el grupo "S".

objetos

objetos	T	E	F	G	H
T	0	4	5	6	8
E		0	3	6	9
F			0	5	4
G				0	3
H					0

tabla 10

4. Ahora el número siguiente es el 3, que corresponde a las parejas A y S, E y F, y G y H. En este caso se considera primero al de A y S, que se agrupan en "T"; sin embargo, el orden en que se tomen en cuenta las parejas de objetos o grupos de objetos, no altera los grupos resultantes para este método, aunque existen otros en los que sucede lo contrario.

objetos

objetos	T	U	G	H
T	0	4	6	8
U		0	5	4
G			0	3
H				0

tabla 11

E y F forman en este mismo nivel el grupo "U".

objetos

objetos	T	U	V
T	0	4	6
U		0	4
V			0

tabla 12

Para este coeficiente, por último, G y H forman "V".

objetos

objetos	X	V
X	0	4
V		0

tabla 13

5. Considerando primero al coeficiente de T y U, se forma el grupo "X", quedando así solamente 2 grupos.

objetos

objetos	Z
Z	0

tabla 14

Por último los grupos "X" y "V" forman "Z", grupo que contiene a todos los objetos como sus elementos.

En el siguiente proceso con gráficas, los grupos están representados por líneas alrededor de los puntos; hay tantas gráficas como números diferentes haya en la tabla original, hasta que en un determinado nivel todos los objetos formen parte de un solo grupo.

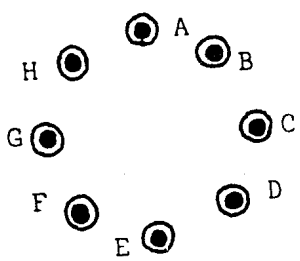


figura 16

1. A nivel 0, cada objeto forma un grupo.

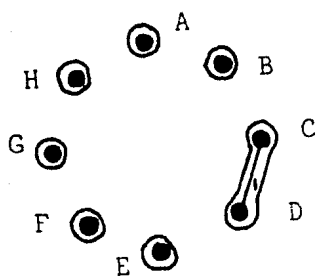


figura 17

2. A nivel 1, aparece la arista DC.

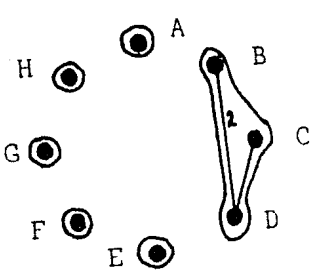


figura 18

3. La arista BD surge a nivel 2.

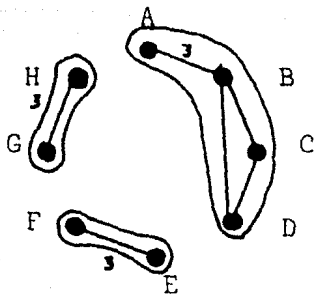


figura 19

4. A nivel 3 aparecen AB, BC, EF y GH. Ahora hay 3 grupos.

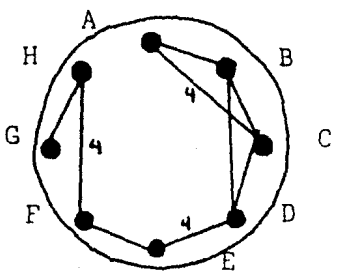


figura 20

5. A nivel 4 se añaden las aristas AC, DE y FH.

Este método considera a las componentes de la gráfica que se encuentran conectadas, como las familias de grupos formadas en cada nivel. El dendrograma y la tabla ultramétrica resultantes son:

objetos

objetos	A	B	C	D	E	F	G	H
A	0	3	3	3	4	4	4	4
B		0	2	2	4	4	4	4
C			0	1	4	4	4	4
D				0	4	4	4	4
E					0	3	4	4
F						0	4	4
G							0	3
H								0

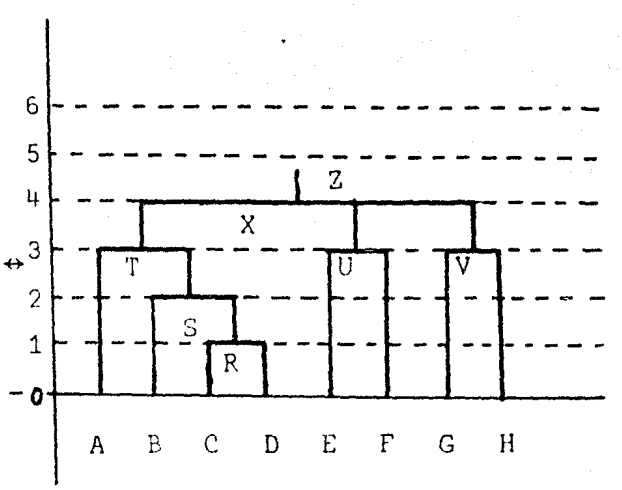


figura 21

De esta manera, el método de Conexión Simple produce grupos grandes en los que los objetos pertenecientes a dos grupos diferentes pueden tener un parecido mayor entre ellos que el que poseen objetos distintos de un mismo grupo. Este defecto llamado "encadenamiento", hace que este método no sea muy utilizado; sin embargo, posee la ventaja de que si en un determinado paso del proceso se encuentran dos o más parejas de objetos cuyos coeficientes de disimilaridad sean iguales, entonces el orden en que se consideran no altera el resultado final en la formación de grupos. Puede demostrarse[10] que este método tiene la propiedad de ser un método continuo; es decir, si se altera con pequeños cambios la información contenida en la tabla original, entonces se producen solamente pequeños cambios en la tabla ultramétrica resultante, y también, por consiguiente, en el dendrograma que la representa. En el método de Conexión Simple, los grupos se constituyen al considerar "la mínima condición necesaria" para unir un objeto de algún grupo, a algún otro objeto perteneciente al mismo grupo.

Otro método Jerárquico parecido al de Conexión Simple en cuanto a su sencillez, es el método llamado de *Conexión Completa*. En este método también se busca el coeficiente de disimilaridad más pequeño entre las posibles parejas de objetos; estos objetos se unen formando un grupo, y al recalcular la disimilaridad entre este nuevo grupo y los demás objetos, lo que se

[10] Véase el capítulo 3, sección 3.4.

busca ahora es el coeficiente más grande entre los elementos del nuevo grupo y el objeto en consideración de los restantes. Es decir, en cada paso después de que dos grupos se han mezclado, la disimilaridad entre este nuevo que han formado y otro, está basada en "la máxima condición necesaria" para unir a todos los objetos de un grupo a cualquier otro objeto. Este método tiene la desventaja de que el orden de entrada de los datos, por el contrario del método de Conexión Simple, altera el resultado final en las familias de grupos (en el caso de existir coeficientes de disimilaridad iguales), además no posee la característica de ser un método continuo. Para ilustrar el efecto de la aplicación de el método de Conexión Completa a la misma tabla (7) de disimilaridades dada para el método expuesto anteriormente, se procede de la misma manera.

objetos

objetos	A	B	C	D	E	F	G	H
A	0	3	4	5	10	7	15	10
B		0	3	2	6	8	13	20
C			0	1	5	5	6	8
D				0	4	6	10	12
E					0	3	6	9
F						0	5	4
G							0	3
H								0

tabla 7

1. El coeficiente más pequeño es el correspondiente a C y D, que forman el grupo "R".

objetos

objetos	A	B	R	E	F	G	H
A	0	3	5	10	7	15	10
B		0	3	6	8	13	20
R			0	5	6	10	12
E				0	3	6	9
F					0	5	4
G						0	3
H							0

tabla 15

2. A y B forman el grupo "S". En este caso puede comprobarse que si se escogiera en otro orden el coeficiente que se repite, los dendrogramas resultantes serían diferentes.

objetos

objetos	S	R	E	F	G	H
S	0	5	10	8	15	20
R		0	5	6	10	12
E			0	3	6	9
F				0	5	4
G					0	3
H						0

tabla 16

3. "T" se forma con E y F.

objetos

objetos	S	R	T	G	H
S	0	5	10	15	20
R		0	6	10	12
T			0	6	9
G				0	3
H					0

tabla 17

4. G y H forman el grupo "U".

objetos

objetos	S	R	T	U
S	0	5	10	20
R		0	6	12
T			0	9
U				0

tabla 18

5. El siguiente grupo, "V", es formado por S y R.

objetos

objetos	V	T	U
V	0	10	20
T		0	9
U			0

tabla 19

6. "X" se forma por T y U.

objetos

objetos	V	X
V	0	20
X		0

tabla 20

7. Por último, V y X forman "Z".

objetos

objetos	Z
Z	0

tabla 21

La secuencia de gráficas para este caso es la siguiente:

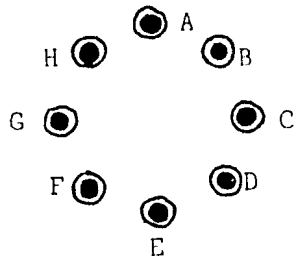


figura 22

1. A nivel 0 cada objeto forma un grupo.

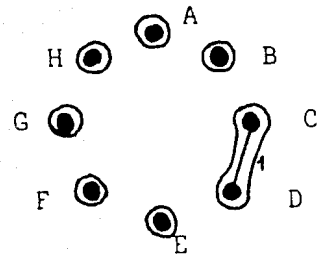


figura 23

2. A nivel 1 aparece la arista DC.

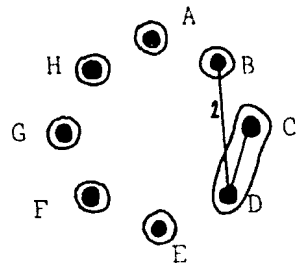


figura 24

3. En el nivel 2 aparece la arista BD.

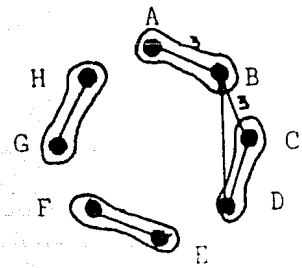


figura 25

4. A nivel 3 aparecen AB, BC, EF y GH.

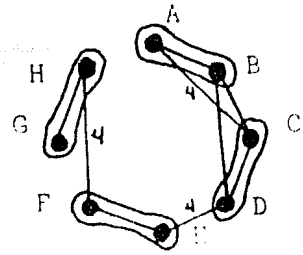


figura 26

5. A nivel 4 se agregan AC, DE y FH.

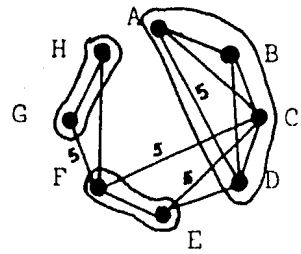


figura 27

6. AD, CE, CF y FG se añaden a nivel 5.

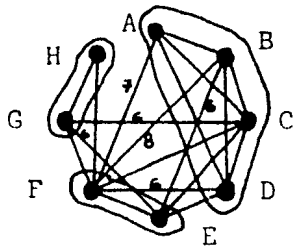


figura 28

7. BE, CG, DF y EG se agregan a nivel 6, AF a nivel 7, y BF a nivel 8.

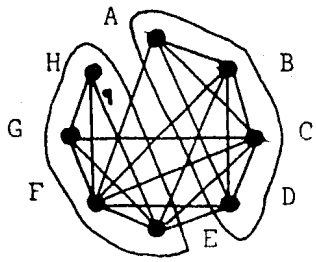


figura 29

8. A nivel 9 aparece EH.

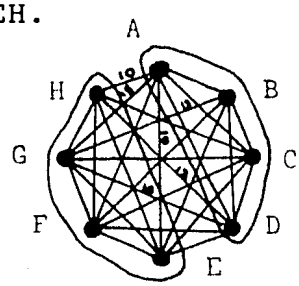


figura 30

9. A nivel 10 se añaden AE, AH y DG. DH aparece a nivel 12; BG a nivel 13 y AG a nivel 15.

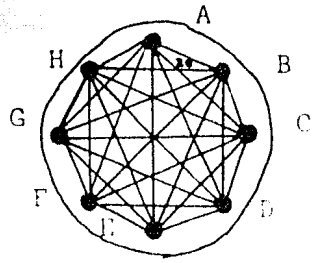


figura 31

10. POr último, en el nivel 20 aparece BH.

A diferencia del método de Conexión Simple, en este método los grupos ilustrados en las gráficas, son las componentes completas (*i.e.* no basta que los puntos estén conectados por algunas líneas, sino que tienen que estar presentes todos los posibles enlaces entre esos puntos) que no se intersectan entre ellas. El dendrograma y la tabla ultramétrica resultantes son:

objetos

objetos	A	B	C	D	E	F	G	H
A	0	3	5	5	20	20	20	20
B		0	5	5	20	20	20	20
C			0	1	20	20	20	20
D				0	20	20	20	20
E					0	3	9	9
F						0	9	9
G							0	3
H								0

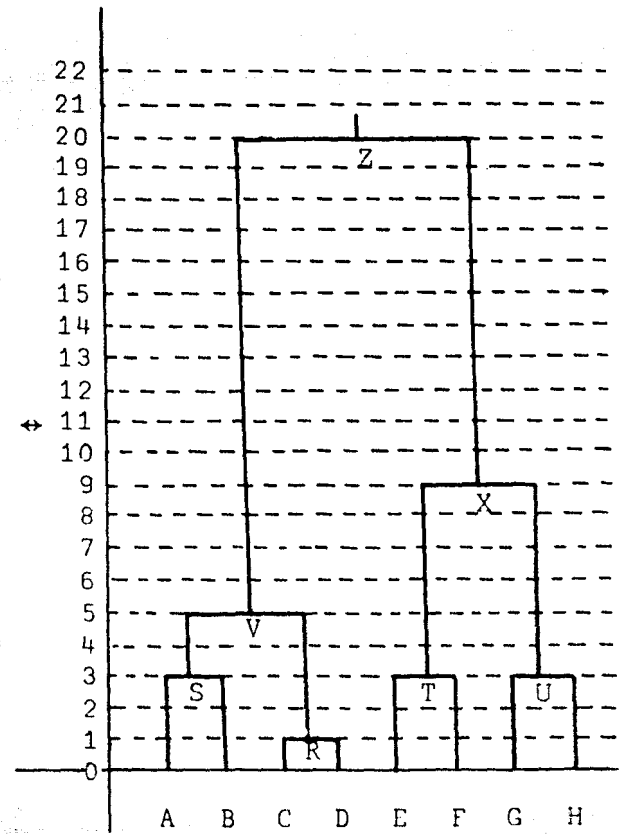


figura 32

Este método, por el contrario del método de Conexión Simple, no produce grupos tan grandes, sino grupos más compactos que se van uniendo entre sí de una manera más estricta en niveles de disimilaridad más altos. Este método no produce encade-

namiento.

En contraste con los dos métodos expuestos anteriormente, que presentan posibilidades extremas en las condiciones para la formación de grupos, existe otro tipo de métodos Jerárquicos, los *Métodos Promedio*, que determinan a los grupos mediante consideraciones promedio en la recalculación de las disimilaridades entre los nuevos grupos. Las diversas maneras de hacer estas recalculaciones dan lugar a los diferentes métodos Promedio. Entre los principales se encuentran los llamados: *Método Promedio entre grupos*, *Método Promedio dentro de grupos*, *Promedio pesado dentro de grupos* e *Incremento al Promedio pesado dentro de grupos*.[11]

La manera de recalcular las disimilaridades, en general, se hace evaluando un promedio entre las disimilaridades de los elementos del nuevo grupo y las de los elementos del grupo en consideración; o bien, calculando la disimilaridad promedio entre el nuevo grupo y los anteriormente formados (ésto puede hacerse dando pesos diferentes de acuerdo a la importancia en tamaño, de un grupo determinado). Al igual que el método de Conexión Completa, estos métodos poseen la desventaja de que el orden de entrada de los datos iguales de la matriz de disimilaridades original, altera los resultados finales, y además son métodos discontinuos; sin embargo poseen la conveniencia de que no están sujetos a valores extremos para el establecimiento de los grupos. Los resultados en los dendrogramas generalmente

[11] Véase el apéndice A, sección A.2.

son parecidos entre sí al aplicar los diferentes métodos Promedio, y en menor grado a aquellos obtenidos por los métodos de Conexión Simple y Conexión Completa.

1.5 Otros métodos Jerárquicos

Una idea diferente en el concepto de clasificación Jerárquica es la que se conoce con el nombre de *Triadas*[12]. Este proceso consiste en obtener todas las posibles ternas de objetos del conjunto que se quiere clasificar, y para cada una de ellas se decide cuáles son los elementos más parecidos entre sí. Este procedimiento podría ser numérico, pero no necesariamente, ya que el razonamiento que se hace puede basarse en un criterio intuitivo del usuario. En este proceso se construyen "árboles" de datos, obtenidos con base en el parecido en las diferentes ternas, mediante algoritmos que calculan un promedio de ocurrencias en que un cierto objeto se parece más a otro, que a un tercero. Sin embargo, este método tiene un gran inconveniente, ya que además de la obtención de una lista de las posibles ternas de objetos, se requiere de la comparación y evaluación del parecido entre cada una de ellas y entre cada uno de sus elementos. El número de combinaciones de M elementos tomados por ternas es:

$$\frac{M(M-1)\ (M-2)}{6},$$

número que aumenta considerablemente a medida que crece el nú-

[12] Véase "Clustering Algorithms", de John A. Hartigan, 1975.

mero de objetos. A pesar de lo anterior, la idea de las triadas es importante ya que las ternas son los conjuntos más pequeños para basar una comparación de parecidos entre objetos. Este trabajo no trata de estos métodos, sino que se concentra en los métodos Jerárquicos presentados en este capítulo, y en otros Numéricamente Estratificados, presentados en el siguiente.

CAPITULO 2

2.1 Métodos Numéricamente Estratificados

En el capítulo anterior se señalaron las ventajas y desventajas en el manejo de los diversos métodos de clasificación Jerárquica. El método de Conexión Simple tiene el defecto de encadenamiento, defecto que el método de Conexión Completa y los métodos Promedio solucionan; sin embargo estos últimos también poseen la desventaja de ser discontinuos. Lo que se busca es tratar de evitar estas limitaciones mediante un sistema más general de clasificación.

La idea principal de los *Métodos Numéricamente Estratificados*, como se señala en el capítulo anterior, es permitir que los grupos formados en cada nivel se traslapen entre ellos, es decir, un objeto puede pertenecer a dos o más grupos al mismo tiempo. Esta idea hace que estos métodos trabajen con consideraciones más generales que los métodos Jerárquicos, en los que un objeto puede pertenecer solamente a un grupo para un determinado nivel en la clasificación. Esta idea puede ser de mucha utilidad en la práctica, ya que en general un objeto puede identificarse con varios grupos a la vez y no estrictamente con uno sólo.

Es posible hacer variar el grado de traslape permitido entre los grupos al hacer una clasificación. A medida que existe un mayor traslape entre los grupos la información original es alterada en menor grado; sin embargo la interpretación de los resultados es cada vez más difícil a medida que este grado au-

menta[1]. Las restricciones en el traslape pueden ser de diversos tipos[2]. Las restricciones *Absolutas* son, por ejemplo, restricciones en el número de elementos en el traslape; las restricciones *Internas* son, por ejemplo, restricciones en el diámetro del traslape[3]. Al escoger diferentes grados de traslape se dá lugar a una serie de clasificaciones (una por cada grado escogido), de las cuales la que permite el traslape mínimo equivale al método de Conexión Simple de clasificación Jerárquica; la que permite el traslape máximo equivale a la obtención en el resultado final, de la tabla original sin modificación alguna.

La construcción de sistemas de clasificación Jerárquica está basada en tablas de coeficientes de disimilaridad como datos de información original. Los métodos Numéricamente Estratificados también parten de estas tablas para transformarlas en otro tipo de tablas ultramétricas que se identifican con unas gráficas correspondientes a la generalización del concepto de dendrograma. Esta generalización recibe el nombre de *k-Dendrograma* (el significado de la "k" se explica en la sección siguiente). También se trata de un tipo de árbol cuyos vértices terminales representan al conjunto de objetos que se va a clasificar, y en cuyas ramificaciones se asocian los valores que mide la disimilitud entre los objetos. Un k-dendrograma

[1] Esto puede observarse en el ejemplo de este capítulo.
[2] Véase "Mathematical Taxonomy", de N. Jardine y R. Sibson, 1971.
[3] El diámetro de un conjunto de objetos es el valor mayor de los coeficientes de disimilaridad entre cualquier pareja de elementos de ese conjunto.

es un árbol como el siguiente:

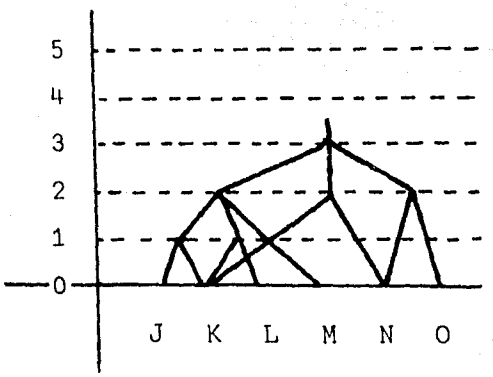


figura 1

El conjunto de objetos que se van a clasificar es el constituído por J, K, L, M, N y O. Los grupos también se van formando a medida que las ramas de árbol se unen; primero se juntan las ramas correspondientes a los objetos de mayor parecido. La manera de leer los k-dendrogramas es exactamente igual a la de los dendrogramas. El siguiente ejemplo muestra algunas lecturas para el k-dendrograma anterior.

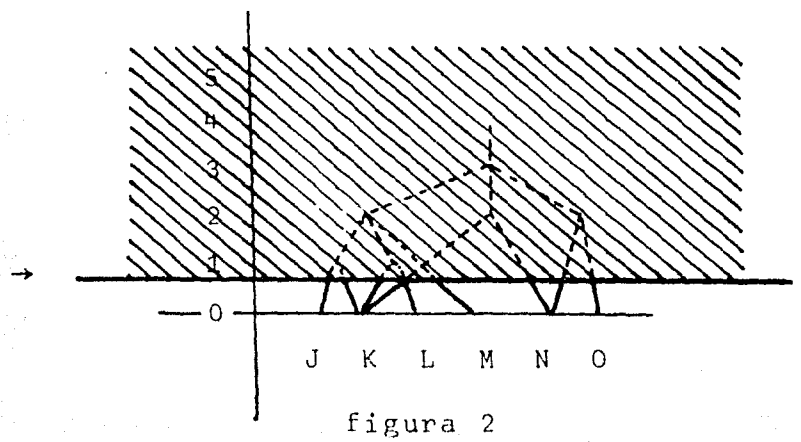


figura 2

Cada objeto constituye un solo grupo de manera individual, para los niveles de asociación menores al nivel 1. Hay 6 grupos, uno por cada objeto.

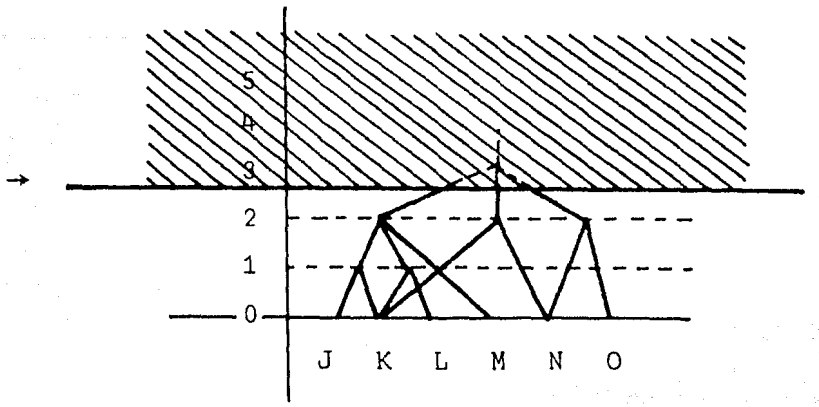


figura 3

En este nivel hay 3 grupos: J, K, L y M; K y N; y N y O. Como puede observarse, el objeto K pertenece a dos grupos diferentes, lo mismo que el objeto N. Para un nivel un poco mayor a 3 todos los objetos ya se encuentran unidos formando un solo grupo.

Definiendo de igual forma la disimilaridad entre una pareja de objetos, como el nivel más pequeño tal que esa pareja se encuentra ya unida en un mismo grupo, es posible dado un k-dendrograma como el anterior, obtener la tabla de disimilaridades entre objetos y grupos de objetos. La tabla completa para el k-dendrograma anterior es la siguiente:

	J	K	L	M	N	O
J	0	1	2	2	3	3
K		0	1	2	2	3
L			0	2	3	3
M				0	3	3
N					0	2
O						0

tabla 1

Para cada nivel de asociación se obtienen familias de grupos,

que en este caso no son ajenos entre sí. De igual manera que para los dendrogramas, estas familias están formadas por los conjuntos de objetos ya unidos en un determinado nivel o en niveles menores. De la misma forma en que se ilustró el procedimiento inverso de pasar de tablas de disimilaridad a dendrogramas, a continuación, también con gráficas, se muestra el proceso para pasar a k-dendrogramas. En este caso los grupos formados en cada nivel son las componentes completas para su gráfica correspondiente, es decir, los subconjuntos más grandes de puntos de la gráfica que tienen todos los posibles enlaces entre ellos (permitiéndose el traslape). Los grupos están representados por líneas delgadas alrededor de los puntos.

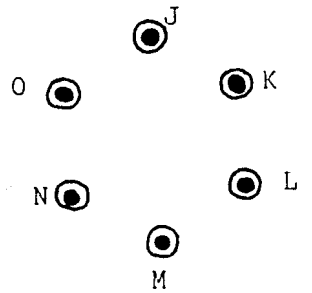


figura 4

1. A nivel 0 cada objeto forma un grupo.

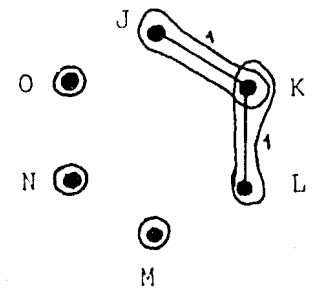


figura 5

2. Se selecciona ahora el coeficiente cuyo valor es 1; hay 5 grupos: J yK; K y L; y M, N y O aisladamente en un grupo cada uno.

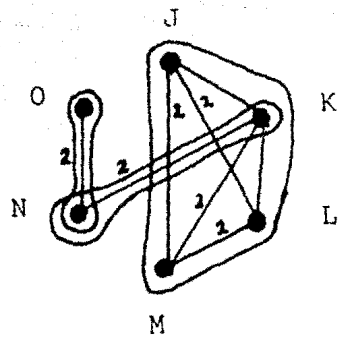


figura 6

3. Aparecen las aristas cuyo nivel es 2. Hay 3 grupos: J, K, L y M; K y N; y N y O.

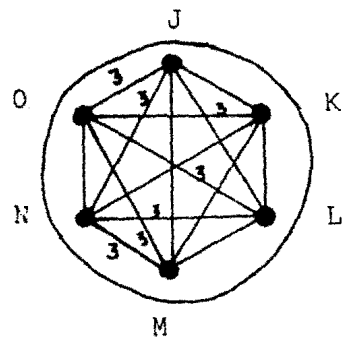


figura 7

4 Todos los objetos forman un grupo a nivel 3.

La condición de ultrametría también se generaliza en la condición de *k-Ultrametría* , y los coeficientes de disimilaridad que la satisfacen son los *k-Ultramétricos*. Es posible demostrar[4] que dada una tabla de coeficientes de disimilaridad k-ultramétricos se le puede asociar un solo k-dendrograma que la represente, e inversamente. Esto dá lugar a la correspondencia e identificación entre tablas k-ultramétricas y los k-dendrogramas. Algunos métodos Numéricamente Estratificados son formas diferentes de transformar tablas de coeficientes de disimilaridad dadas en tablas de coeficientes k-ultramétricos.

[4] Véase el capítulo 4, sección 4.2.

2.2 Los métodos B_k

Entre los métodos Numéricamente Estratificados se encuentran los llamados B_k. Estos métodos se basan en una restricción absoluta en el traslape permitido entre los grupos formados; es decir, son restricciones en el número de elementos en el traslape. La letra k puede tomar diferentes valores: desde 1 hasta N-1, donde N es el número de objetos del conjunto que se va a clasificar, y cada valor distinto dá lugar a un método diferente. Al utilizar un cierto valor de k, la restricción que se hace es que el traslape entre grupos no contiene más de k-1 elementos. Así, B_1 no permite traslape alguno entre sus grupos; B_2 permite 1 elemento solamente; B_3 permite 2 elementos en el traslape.

Al utilizar el método B_1, donde los grupos resultantes son ajenos, lo que se hace es aplicar el método de Conexión Simple; al ir aumentando el valor de k se va permitiendo un mayor número de elementos en la intersección, hasta que k toma el valor N-1 o más, lo que se obtiene como resultado final es la tabla de disimilaridades original[5]. De esta manera, al ir variando los valores de k, desde 1 hasta N-1, cada vez se va deformando en menor grado la información original, hasta que en el último se obtiene esta misma. Así, el usuario cuenta con la ventaja de poder escoger el grado de deformación que desea que la tabla original adquiera; también, si se calculan primero las tablas correspondientes a los valores mayores de k, en-

[5]Véase el capítulo 4, sección 4.3.

tonces se pueden obtener las tablas para valores menores a partir de las primeras y no de la original; ésto significa ahorro del trabajo posterior para el usuario. Además de estas ventajas, también se trata de métodos continuos, es decir, a pequeños cambios en los datos de información original, corresponden pequeños cambios en los resultados.

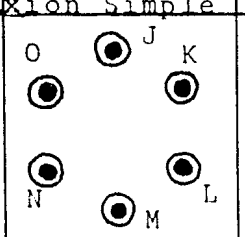
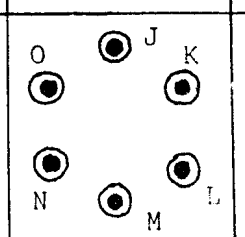
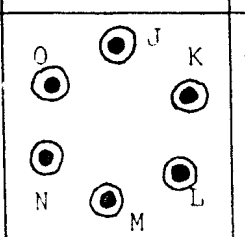
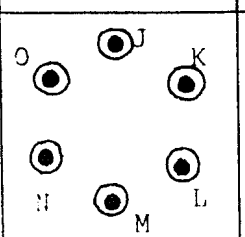
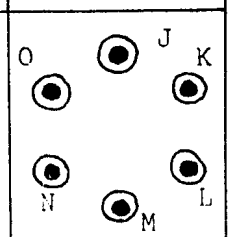
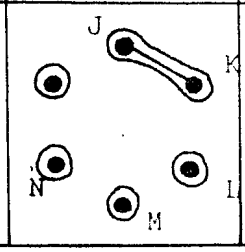
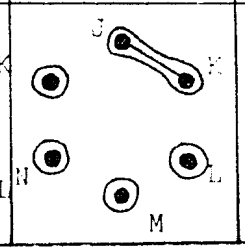
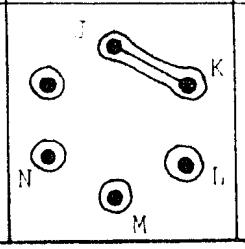
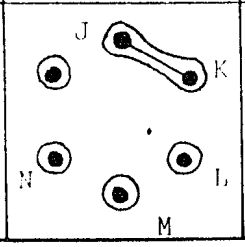
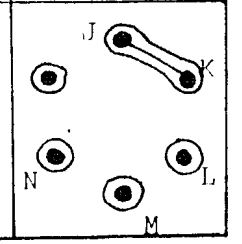
La manera en que trabajan estos métodos, para un determinado valor de k, y pensando en gráficas cuyos puntos representan a los objetos, y cuyas aristas unen aquellos puntos cuyos coeficientes de disimilaridad son menores o iguales a un determinado nivel, es la siguiente: se busca el coeficiente más chico de la tabla y se colocan las aristas correspondientes; se buscan las componentes completas más grandes (si hay una contenida en otra, se escoge la mayor) para esa gráfica; cuando dos de estos conjuntos de puntos tienen por lo menos k puntos en común (en su traslape), se toman como un solo conjunto, completando las aristas que falten para constituir una componente completa de la gráfica. Enseguida se vuelven a considerar las nuevas componentes completas de la gráfica actual hasta que ya no se necesiten agregar nuevas aristas. Las componentes completas así formadas, constituyen las familias de grupos para el nivel en consideración. Después se pasa a buscar el coeficiente más pequeño siguiente y se procede de la misma manera hasta que haya un solo grupo en donde todos los puntos de la gráfica formen una sola componente completa. Ejemplo: sea la siguiente tabla la matriz de disimilaridades dada.

objetos

objetos	J	K	L	M	N	O
J	0	1	3	3	5	6
K		0	2	4	3	5
L			0	3	4	4
M				0	4	4
N					0	2
O						0

tabla 2

El conjunto de objetos que se quiere clasificar es el formado por: J, K, L, M, N y O. Los diferentes métodos que se aplican son: B_1, B_2, B_3, B_4 y B_5; desde que no se permite traslape alguno entre los grupos formados, hasta permitir familias de grupos que pueden tener 4 elementos en la intersección. La siguiente tabla muestra con gráficas las familias de grupos formadas a cada nivel, para los diferentes métodos usados; los grupos se señalan igual que anteriormente.

método / nivel	B_1 = Conexión Simple	B_2	B_3	B_4	B_5
0					
1					

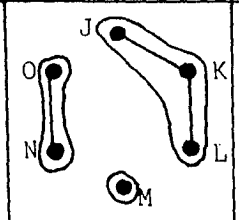
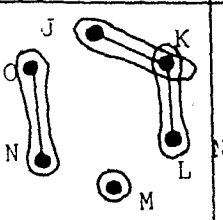
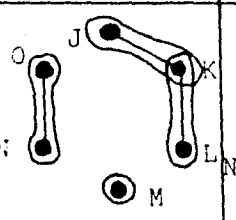
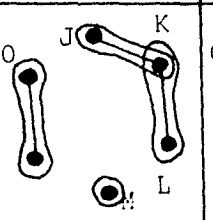
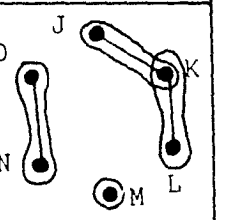
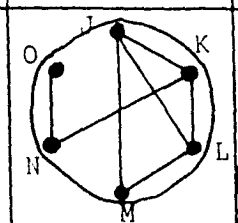
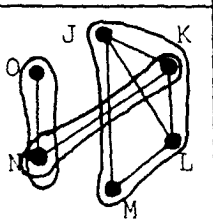
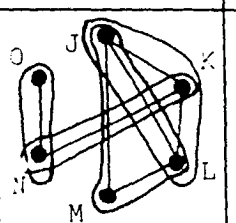
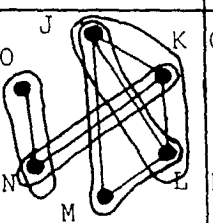
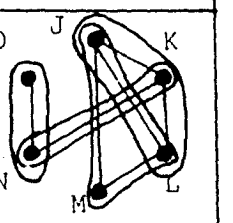
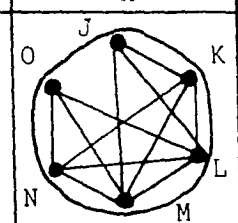
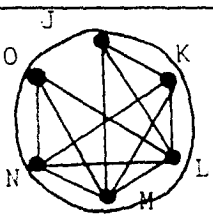
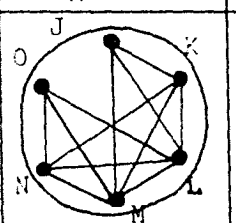
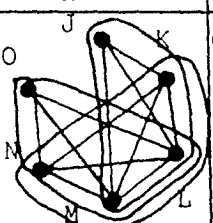
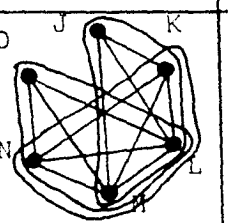
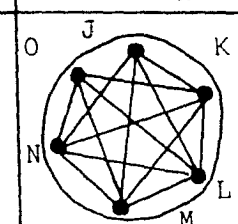
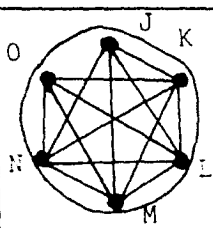
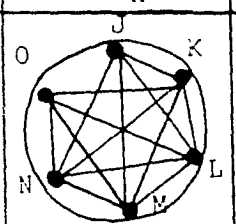
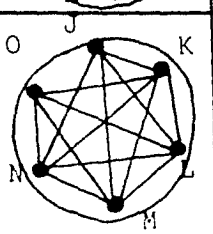
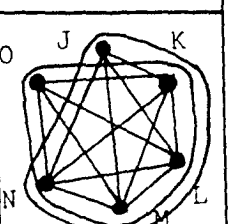
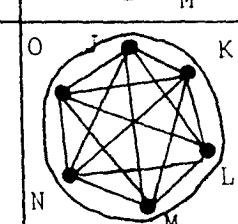
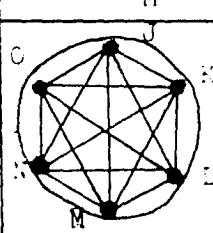
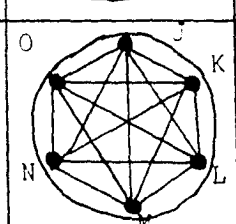
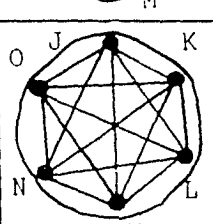
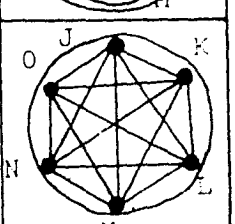
método / nivel	B_1 = Conexión Simple	B_2	B_3	B_4	B_5
2					
3					
4					
5					
6					

tabla 3

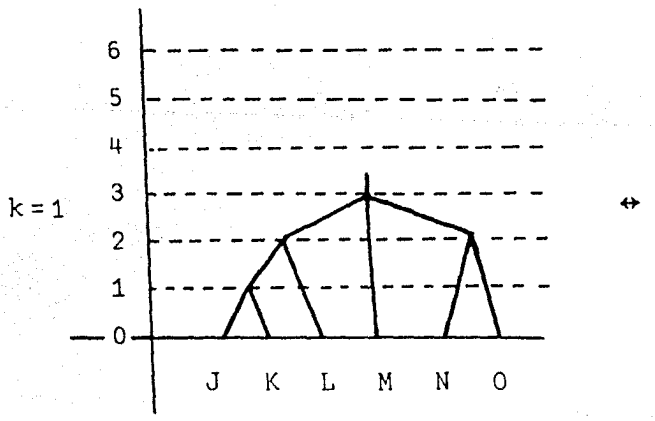
Para facilitar las lecturas de las familias de grupos formadas a cada nivel, se muestra la siguiente tabla:

método / nivel	B_1	B_2	B_3	B_4	B_5
0	(J) (K) (L) (M) (N) (O)	(J) (K) (L) (M) (N) (O)	(J) (K) (L) (M) (N) (O)	(J) (K) (L) (M) (N) (O	(J) (K) (L) (M) (N) (O)

nivel / partición	B_1	B_2	B_3	B_4	B_5
1	(JK) (L) (M) (N) (O)	(JK) (L) (M) (N) (O)	(JK) (L) (M) (N) (O)	(JK) (L)(M) (N) (O)	(JK) (L) (M) (N) (O)
2	(JKL) (M) (NO)	(JK) (KL)(M) (NO)	(JK) (KL)(M) (NO)	(JK) (KL)(M) (NO)	(JK) (KL)(M) (NO)
3	(JKLMNO)	(JKLM) (KN) (NO)	(JKL) (JLM) (KN) (NO)	(JKL) (JLM) (KN) (NO)	(JKL) (JLM) (KN) (NO)
4	(JKLMNO)	(JKLMNO)	(JKLMNO)	(JKLM)(KLMN) (LMNO)	(JKLM)(KLMN) (LMNO)
5	(JKLMNO)	(JKLMNO)	(JKLMNO)	(JKLMNO)	(JKLMN) (KLMNO)
6	(JKLMNO)	(JKLMNO)	(JKLMNO)	(JKLMNO)	(JKLMNO)

tabla 4

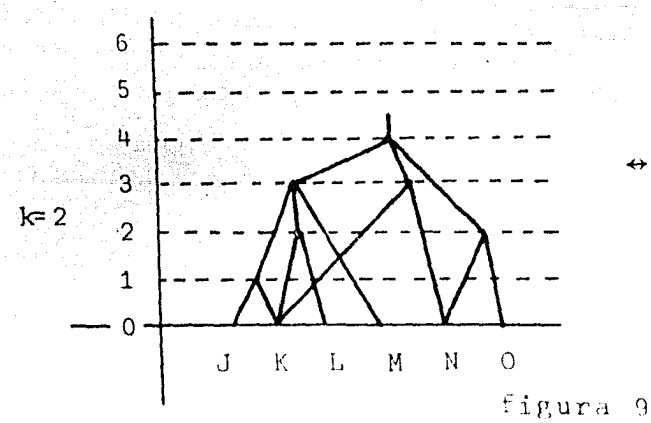
Los k-dendrogramas y tablas de disimilaridades resultantes, para los diferentes valores de k, son:



↔

	J	K	L	M	N	O
J	0	1	2	3	3	3
K		0	2	3	3	3
L			0	3	3	3
M				0	3	3
N					0	2
O						0

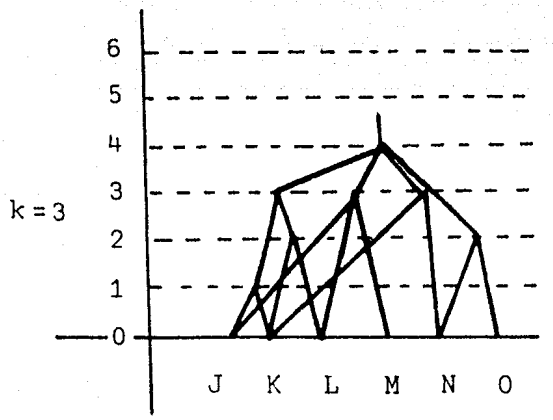
figura 8



↔

	J	K	L	M	N	O
J	0	1	3	3	4	4
K		0	2	3	3	4
L			0	3	4	4
M				0	4	4
N					0	2
O						0

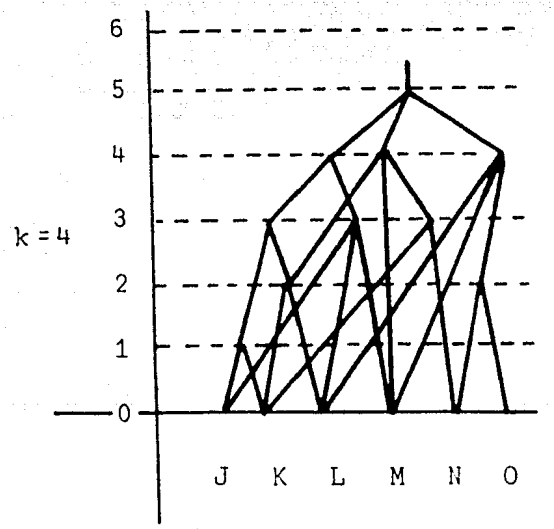
figura 9



↔

	J	K	L	M	N	O
J	0	1	3	3	4	4
K		0	2	4	3	4
L			0	3	4	4
M				0	4	4
N					0	2
O						0

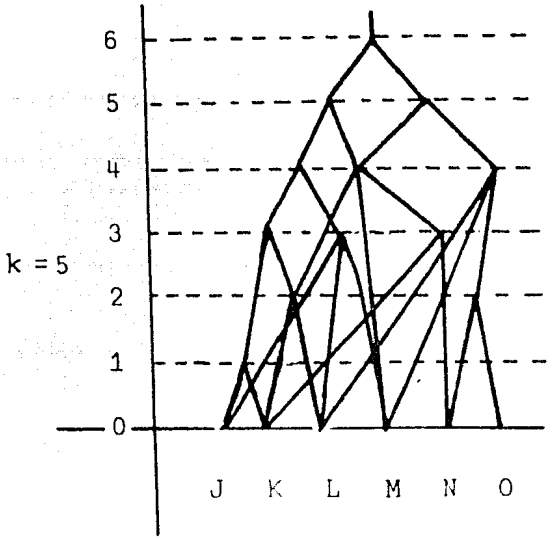
figura 10



↔

	J	K	L	M	N	O
J	0	1	3	3	5	5
K		0	2	4	3	5
L			0	3	4	4
M				0	4	4
N					0	2
O						0

figura 11



↔

	J	K	L	M	N	O
J	0	1	3	3	5	6
K		0	2	4	3	5
L			0	3	4	4
M				0	4	4
N					0	2
O						0

figura 12

La última tabla, que corresponde al método B_5, coincide con la tabla de disimilaridades que se tiene originalmente, y la primera corresponde a la tabla obtenida por el método de Conexión Simple. Estas figuras (8-12) muestran representaciones de los objetos, cada vez de una manera más exacta; sin embargo, su interpretación en los k-dendrogramas es al mismo tiempo más compleja, por lo que es recomendable recurrir en lugar de éstas a las listas de familias de grupos para cada nivel de asociación.

Algunos autores[6] han definido ciertas medidas para apreciar las deformaciones sufridas por tablas de coeficientes de disimilaridad dadas, al aplicárseles métodos como los presentados en este capítulo, así como algunas medidas de aislamiento o cohesión entre grupos para evaluar de acuerdo a criterios establecidos las uniones entre objetos o grupos de objetos.

[6] Véase "A Model for Taxonomy", de N. Jardine y R. sibson, 1968.

CAPITULO 3

3.1 Coeficientes de Disimilaridad

Como se señala en el capítulo 1, el paso básico en los procesos de clasificación es la estimación del parecido entre los objetos en consideración.

Definición. Un *Coeficiente de Disimilaridad* en un conjunto P de objetos (con $|P| < \infty$), es una función

$$d: P \times P \to \mathbb{R},$$

que satisface las siguientes propiedades:

i) $d(A,A) = 0$ para todo $A \in P$,

ii) $d(A,B) \geq 0$ para todos $A, B \in P$,

iii) $d(A,B) = d(B,A)$ para todos $A, B \in P$.

Los coeficientes de disimilaridad pueden pensarse informalmente como distancias, aunque no necesariamente constituyen una distancia en P, en el sentido usual.

Definición. Sea CD(P) el conjunto de coeficientes de disimilaridad en P; sean d_1 y $d_2 \in CD(P)$. Se dice que d_1 domina a d_2 si:

$$d_1(A,B) \geq d_2(A,B) \text{ para todos } A, B \in P.$$

La manera de calcular estos coeficientes es muy diversa y depende del tipo de tabla de información original con que se cuente[1].

[1] Véase el apéndice A, sección A.1.

3.2 Dendrogramas Generalizados

Definición. Un *Dendrograma Generalizado*[2] es una función

$$c: I \to RS(P),$$

donde $I = A \cup \{0\}$, con $A \subseteq \mathbb{R}^+$ e $|I| < \infty$; $RS(P)$ es el conjunto de relaciones reflexivas y simétricas en P; que satisface las siguientes condiciones:

i) Existe $h \in I$ tal que $c(h) = PxP$,

ii) $h < h' \Rightarrow c(h) \subseteq c(h')$ con $c(h) \neq c(h')$, h y $h' \in I$.

Al conjunto de dendrogramas generalizados en P se le denota por $DG(P)$.

Lo que se busca es una identificación entre los coeficientes de disimilaridad definidos en el conjunto P y los dendrogramas generalizados; de tal manera que dada una tabla cualquiera sea posible asociarle un solo dendrograma generalizado que la represente, y viceversa. Con este propósito se presentan las siguientes definiciones:

Definición. Dado $d \in CD(P)$, se define la función

$$T_d: I \to R(P),$$

donde $R(P)$ es el conjunto de relaciones en P, e I se define como anteriormente; de tal manera que

$$T_d(h) = \{(A,B) \in PxP \mid d(A,B) \leq h\},$$

para $h \in I$.

[2] Esta definición no la introducen N. Jardine y R. Sibson; sin embargo en el presente trabajo se utiliza para lograr una mejor identificación entre los conceptos que manejan estos autores.

Definición. Sea $S(P) = \{f \mid f: I \to R(P)\}$. Se define la función

$$T: CD(P) \to S(P),$$

donde $T(d) = T_d$, para $d \in CD(P)$.

Proposición. $T_d \in DG(P)$, para todo $d \in CD(P)$.

Prueba:

i) Primero hay que demostrar que $T_d(h) \in RS(P)$, para todo $h \in I$. Para $A, B \in P$, se tiene que:

a. $d(A,A) \leqslant h$, para todo $h \geqslant 0 \Rightarrow (A,A) \in T_d(h)$, para todo $h \geqslant 0$.

b. Si $d(A,B) \leqslant h$, entonces $d(B,A) \leqslant h$, para todo $h \geqslant 0 \Rightarrow$ si $(A,B) \in T_d(h)$, también $(B,A) \in T_d(h)$, para todo $h \geqslant 0$.

ii) Para $h = \max\{h' \mid h' \in I\}$ se cumple que $T_d(h) = P \times P$.

iii) Dados $h, h' \in I$, si $h \leqslant h'$ entonces $T_d(h) \subseteq T_d(h')$.

$\therefore T(d) = T_d \in DG(P)$, para todo $d \in CD(P)$.

Definición. Dado $c \in DG(P)$, se define la función

$$U_c: P \times P \to \mathbb{R},$$

tal que $U_c(A,B) = \min\{h \in I \mid (A,B) \in c(h)\}$, para todo $(A,B) \in P \times P$.

Definición. Sea $V(P) = \{f \mid f: P \times P \to \mathbb{R}\}$. Se define la función

$$U: DG(P) \to V(P),$$

donde $U(c) = U_c$, para $c \in DG(P)$.

Proposición. $U_c \in CD(P)$, para todo $c \in DG(P)$.

Prueba:

i) Ya que c(h) es una relación reflexiva, se tiene que $(A,A) \in c(h)$ para todo $h \in I$ y para todo $A \in P$. Por lo anterior $U_c(A,A) = \min \{h \in I \mid (A,A) \in c(h)\} = 0$ para todo $A \in P$.

ii) $U_c(A,B) = \min\{ h \in I \mid (A,B) \in c(h)\} \geq 0$ para todo $(A,B) \in P \times P$, puesto que $I \subseteq \mathbb{R}^+ \cup \{0\}$.

iii) $U_c(A,B) = \min\{ h \in I \mid (A,B) \in c(h)\} = \min\{ h \in I \mid (B,A) \in c(h)\} = U_c(B,A)$, ya que c(h) es una relación simétrica.

$\therefore U(c) = U_c \in CD(P)$.

Proposición. La función $T: CD(P) \rightarrow DG(P)$ es biyectiva.

Prueba:

i) Sean d_1 , $d_2 \in CD(P)$, tales que $d_1 \neq d_2$ para alguna pareja de objetos $(A,B) \in P \times P$. Supóngase que $d_1(A,B) < d_2(A,B)$. Sea $d_1(A,B) = h$, entonces:

$$Td_1(h) = \{(X,Y) \in P \times P \mid d_1(X,Y) \leq h\}, \text{ y}$$

$$Td_2(h) = \{(X,Y) \in P \times P \mid d_2(X,Y) \leq h\}.$$

Por lo que $(A,B) \in Td_1(h)$ y $(A,B) \notin Td_2(h)$; entonces $Td_1(h) \neq Td_2(h) \Rightarrow T(d_1) \neq T(d_2)$.

$\therefore$ T es una función inyectiva.

ii) Hay que probar ahora que dado un dendrograma generalizado c, existe un coeficiente de disimilaridad, tal que la función T aplicada a éste, resulta el dendrograma generalizado mismo. Sea ese coeficiente U_c; entonces hay que probar que: $T(U_c) = c$, es decir que

$$T[U_c(h)] = T_{U_c}(h) = \{(A,B) \in P \times P \mid U_c(A,B) \leq h\} =$$

c(h).

a. Sea $(A,B) \in T_{U_c}(h)$, ésto implica que $U_c(A,B) \leq h$. Sea $r = \min \{h' \mid (A,B) \in c(h')\}$; entonces $r \leq h$. Ya que $(A,B) \in c(r)$, y debido a la condición *ii* de la definición de dendrograma generalizado, se tiene que: $c(r) \subseteq c(h)$ y por lo tanto $(A,B) \in c(h)$.

b. Sea $(A,B) \in c(h)$, entonces se tiene que $U_c(A,B) \leq h$ y por lo tanto $(A,B) \in T_{U_c}(h)$.

$\therefore$ T es suprayectiva y U es su inversa.

Dada esta biyección es posible identificar a los coeficientes de disimilaridad en P, con el conjunto de dendrogramas generalizados en P.

3.3 Dendrogramas y Desigualdad Ultramétrica

Es posible observar[3] que la relación $T_d(h)$ en la mayoría de las tablas de disimilaridad no es una relación de equivalencia, ya que falla la transitividad. La siguiente definición corresponde a la dada en el capítulo 1 para dendrogramas, que son los dendrogramas generalizados para los que se obtienen relaciones de equivalencia en cada nivel de asociación. También se establece la condición de ultrametría sobre los coeficientes de disimilaridad; de manera que, los que la cumplen son los que se identifican con los dendrogramas.

[3] Véase la tabla 6 del capítulo 1, sección 1.3.

Definición. Un *Dendrograma*[4] es una función

$$c: I \to E(P),$$

donde $I = A \cup \{0\}$, con $A \subseteq \mathbb{R}^+$ e $|I| < \infty$; $E(P)$ es el conjunto de relaciones de equivalencia en P; que satisface las siguientes condiciones:

i) Existe $h \in I$ tal que $c(h) = PxP$,

ii) $h < h' \Rightarrow c(h) \subseteq c(h')$ con $c(h) \neq c(h')$, h y $h' \in I$.

Al conjunto de dendrogramas en P se le denota por D(P).

Se habla de relaciones de equivalencia en P, porque los métodos Jerárquicos clasifican a loa objetos en grupos que constituyen particiones en P. Para estos métodos el punto de partida son las tablas de coeficientes de disimilaridad, y el punto final son los dendrogramas.

La siguiente definición es la desigualdad ultramétrica, que es una fuerte restricción sobre los coeficientes de disimilaridad. Los diferentes métodos Jerárquicos de clasificación son maneras distintas de transformar coeficientes de disimilaridad dados en coeficientes ultramétricos.

Definición. Sea $d \in CD(P)$. Se dice que d es un *Coeficiente de Disimilaridad Ultramétrico*, si :

$$d(A,B) \leq \max \{d(B,C), d(A,C)\},$$

para todos A, B y $C \in P$.

[4] Esta definición no coincide esencialmente con la dada por N. Jardine y R. Sibson en "Mathematical Taxonomy", 1971. Para observar su equivalencia con la dada en este capítulo, véase el apéndice B, sección B.1.

Las tablas de disimilaridad cuyos coeficientes cumplen con esta propiedad poseen la característica de que si se toman tres objetos cualesquiera de P y se representan como puntos en el plano, separados a una distancia igual a su coeficiente de disimilaridad, entonces los triángulos así formados son isóceles cuyos lados iguales son mayores que el lado desigual; o bien, triángulos equiláteros. Dada esta representación de los objetos, es posible observar que si un coeficiente es ultramétrico en P, entonces también es una métrica en P (*i.e.* $d(A,B) + d(B,C) \geq d(A,C)$). Esta condición no la satisfacen la mayoría de los coeficientes de disimilaridad.

Ya que un dendrograma se define como una función $c: I \rightarrow E(P)$, donde $E(P)$ es el conjunto de relaciones de equivalencia en P, entonces para que $T_d \in D(P)$, se prueba a continuación que los coeficientes sobre los que se aplica la función T, tienen que cumplir la desigualdad ultramétrica.

Proposición. Sea $CU(P)$ el conjunto de coeficientes de disimilaridad ultramétricos en P. Entonces $d \in CU(P)$, si y sólo si $T_d(h)$ es una relación de equivalencia para todo $h \in I$.

Prueba:

i) Sea $d \in CU(P)$, entonces $d(A,B) \leq \max\{d(A,C), d(C,B)\}$ para todos A, B y $C \in P$. $T_d(h) \in RS(P)$ para todo $h \in I$, como se demostró anteriormente. Sea $h \in I$ y A, B y $C \in P$, entonces: (A,C) y $(C,B) \in T_d(h) \Rightarrow d(A,C) \leq h$ y $d(C,B) \leq h \Rightarrow d(A,B) \leq \max\{d(A,C), d(C,B)\} \leq h \Rightarrow (A,B) \in T_d(h)$.

$\therefore T_d(h) \in E(P)$, para todo $h \in I$.

ii) Sea $T_d(h) \in E(P)$, para todo $h \in I$. Dados A, B y C $\in P$, sea $h = \{\max d(A,C), d(C,B)\}$. Entonces: $d(A,C) \leqslant h$ y $d(C,B) \leqslant h$, *i.e.* $(A,C) \in T_d(h)$ y $(C,B) \in T_d(h)$. Por transitividad $(A,B) \in T_d(h) \Rightarrow d(A,B) \leqslant h = \{\max d(A,C), d(C,B)\}$.

$\therefore d \in CU(P)$.

Por lo que, dado un coeficiente de disimilaridad ultramétrico, es posible asociarle un solo dendrograma que lo represente, y viceversa.

3.4 Métodos Jerárquicos

Como se señala anteriormente los métodos Jerárquicos de clasificación se ven como funciones distintas d_x, que transforman tablas de coeficientes de disimilaridad dadas en ultramétricas:

$$d_x: CD(P) \rightarrow CU(P).$$

Método de Conexión Simple. La idea de este método, en términos de teoría de gráficas, es que en cada nivel h los grupos formados son las componentes conexas de la gráfica "G_h"; es decir, se busca para cada nivel h, una partición del conjunto P de objetos, donde cada elemento de la partición está formado por los vértices de una componente conexa de la gráfica G_h inducida por la relación $T_d(h) = \{(A,B) \in P \times P \mid d(A,B) \leqslant h\}$ en RS(P). El algoritmo con el que opera este método es

el siguiente:

Dado $d \in CD(P)$ y $h \in I$,

1. Considérese $h = 0$.
2. Constrúyase la gráfica G_h, en donde los vértices son los elementos de P y las aristas unen aquellos vértices que tienen un coeficiente igual o menor a h.
3. Encuéntrense las componentes conexas de G_h. A cada una de ellas se les denomina C_h.
4. Unanse los vértices de cada C_h. Los grupos a nivel h son los subconjuntos ajenos de P que resultan de esas uniones y los vértices de las componentes conexas que no fueron modificadas.
5. Si todavía no hay un solo grupo, pasar al valor inmediato superior de h y volver al punto 2.

Las disimilaridades escogidas en el punto 1 y después de cada paso por el 5, son las correspondientes a la tabla resultante "d_s" del método de Conexión Simple, que se define de la siguiente manera

Definición.El *Método de Conexión Simple* es una función

$$d_s(d): CD(P) \rightarrow CU(P),$$

tal que $d_s(d)(A,B) = \min \{h \mid (A,B) \in C_h\}$, para $A, B \in P$ y para alguna $C_h \in G_h$.

En general, se puede decir que un método de clasificación es una función

$$d_x: CD(P) \rightarrow Z,$$

donde $Z \subseteq CD(P)$. Para cada subconjunto Z escogido se define un método d_x, tal que

$$d_x(d) = \sup \{ d^* \in Z \mid d^* \leqslant d \}, \quad \text{con } d \in CD(P),$$

es decir, $d_x(d)$ es el elemento más grande de Z que es dominado por la tabla original dada d. Al método así definido se le llama el *Método Z-subdominante* de d. De esta manera, estos métodos reciben el nombre de *Métodos Subdominantes*.

Definición. Sea $Y \subseteq CD(P)$. Se dice que Y es un subconjunto *acotado* si existe $d_o \in CD(P)$ tal que $d \in Y \Rightarrow d \leqslant d_o$.

Definición. Sea $Y \subseteq CD(P)$ acotado. Se define el supremo de Y como:

$$(\sup Y)(A,B) = \sup \{ d(A,B) \mid d \in Y \},$$

para toda pareja $(A,B) \in P \times P$.

Definición. Sea $Z \subseteq CD(P)$ y $d_x(d) = \{ d^* \in Z \mid d^* \leqslant d \} \neq \phi$ tal que $\sup d_x(d) \in Z$, entonces Z define un método subdominante d_x. Un conjunto Z tal, se llama *sup-cerrado*.

Proposición. El método de Conexión Simple ($d_s(d)$) es el método CU-subdominante de d, es decir:

$$d_s(d) = \sup \{ d^* \in CU(P) \mid d^* \leqslant d \}, \text{ para todo } d \in CD(P).$$

Primero hay que probar que CU(P) define un método subdominante, es decir:

i) El conjunto $Y = \{ d^* \in CU(P) \mid d^* \leqslant d \} \neq \phi$, para todo $d \in CD(P)$.

ii) Dado $X \subseteq CU(P)$ acotado, entonce $\sup X \in CU(P)$.

Después se demuestra que:

iii) $d_s(d) \in CU(P)$, para todo $d \in CD(P)$.

iv) $d_s(d) \leqslant d$, para todo $d \in CD(P)$.

v) Dado $d' \in CU(P)$ tal que $d' \leqslant d$, entonces $d' \leqslant d_s(d)$.

Y por lo tanto $d_s(d) = \sup \{d^* \in CU(P) \mid d^* \leqslant d\}$.

Prueba:

i) El conjunto $Y = \{d^* \in CU(P) \mid d^* \leqslant d\} \neq \phi$, ya que al menos la tabla que tiene solamente ceros en sus entradas (d_0), es uno de sus elementos por ser ultramétrica y porque $d_0 \leqslant d$, para todo $d \in CD(P)$.

ii) Sean A y $B \in P$, sea $X \subseteq CU(P)$ acotado. Ya que X es acotado, entonces sup X existe. Sea $(\sup X)(K,L) = d_r(K,L)$ para todo $(K,L) \in P \times P$; entonces dado $C \in P$:

$$d_r(A,B) = \sup_{d \in X} \{d(A,B)\} \leqslant \sup_{d \in X} [\max \{d(A,C), d(B,C)\}] =$$

$$= \max \{\sup_{d \in X} [d(A,C)], \sup_{d \in X} [d(B,C)]\} =$$

$$= \max \{d_r(A,C), d_r(B,C)\} .$$

$\therefore d_r \in CU(P)$.

Entonces, de *i* e *ii*, CU(P) define un método subdominante.

iii) Para demostrar que $d_s(d) \in CU(P)$, para todo $d \in CD(P)$, como se pidió en la definición, basta observar que dados tres elementos cualesquiera de P, el valor mayor de $d_s(d)$ entre ellos, aparece al menos dos veces. Sean A, B y $C \in P$. Sea $\alpha = d_s(d)(A,B) = \max \{d_s(d)(A,C), d_s(d)(C,B), d_s(d)(A,B)\}$. Sea $\beta = \max \{d_s(d)(A,C), d_s(d)(C,B)\}$. Entonces, por construcción de C_h, para

el nivel de asociación β, existe una componente conexa C_β tal que $(A,B) \in C_\beta$, por lo que: $\alpha = d_s(d)(A,B) \leq \beta \leq \alpha$, $\therefore \alpha = \beta$.

iv) Por construcción de C_h, ya que en todo nivel sólo se aumentan aristas, si $d(A,B) = \alpha$, entonces $(A,B) \in C_\alpha \Rightarrow d_s(d)(A,B) = \min \{h \mid (A,B) \in C_h\} \leq \alpha = d(A,B)$.

v) Sea $d' \in CU(P)$, tal que $d' \leq d$. Supóngase que $d' > d_s(d)$, entonces existe $(A,B) \in PxP$ tal que $d'(A,B) > d_s(d)(A,B)$. Sea $\alpha = d_s(d)(A,B) = \min \{h \mid (A,B) \in C_h\}$, entonces existen $A = v_1, v_2, v_3, \ldots, v_q = B \in P$ tales que $d(v_i, v_{i+1}) \leq \alpha$, $i = 1, 2, \ldots, q-1$. Ya que $d' \in CU(P)$, se tiene que

$$d'(A,B) \leq \max \{d'(A,v_2), d'(v_2,v_3), \ldots, d'(v_{q-1},B)\}$$
$$\leq \max \{d(A,v_2), d(v_2,v_3), \ldots, d(v_{q-1},B)\}$$
$$\leq \alpha = d_s(d)(A,B), \text{ lo que es una contradicción.}$$

$\therefore d' \leq d_s(d)$.

$\therefore d_s(d)(A,B) = \min \{h \mid (A,B) \in C_h\} = \sup \{d^\star \in CU(P) \mid d^\star \leq d\}$ para todo $(A,B) \in PxP$.

En teoría de gráficas, un *Arbol Generador* de una gráfica G, es una subgráfica conexa en la que no se forman ciclos y que contiene a todos los puntos de G. Si a cada arista se le asigna un valor numérico, entonces la *longitud* de un árbol es la suma de los valores de sus aristas. Se define un *Arbol de Mínimo Alcance* de una gráfica G, como un árbol generador cuya longitud es mínima con respecto a las longitudes de to-

dos los posibles árboles generadores de G. Existen diversos algoritmos para encontrar el árbol de mínimo alcance de una determinada gráfica (que no tiene porqué ser único). Entre éstos se encuentra, por ejemplo, el siguiente:

1. Escoger la arista cuyo valor asociado sea el más pequeño, para formar parte del árbol de mínimo alcance. Esta arista no debe formar ciclos con las anteriormente escogidas, de lo contrario no se selecciona.
2. Repetir el paso anterior N-1 veces, donde N es el número de vértices de G.

Se ha observado que existe una relación entre el concepto de árbol de mínimo alcance y el método de Conexión Simple. Es posible demostrar que la familia de grupos determinada por un cierto nivel X en el dendrograma obtenido al aplicar el método de Conexión Simple a una tabla de disimilaridades, corresponde a las componentes conexas que resultan al quitar del árbol de mínimo alcance todas las aristas cuyo valor es mayor que X. Para ilustrar esta idea se presenta el siguiente ejemplo, considerando la tabla que se muestra en el capítulo 1 para aplicar el método de Conexión Simple.

	A	B	C	D	E	F	G	H
A	0	3	4	5	10	7	15	20
B		0	3	2	6	8	13	20
C			0	1	5	5	6	8
D				0	4	6	10	12
E					0	3	6	9
F						0	5	4
G							0	3
H								0

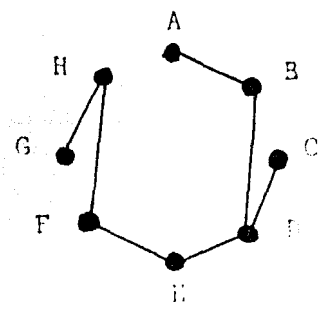


Figura 1

El anterior es un árbol de mínimo alcance para la gráfica cuyos vértices son los objetos A, B, C,...,H y cuyas aristas tienen como longitud a los valores de los coeficientes de disimilaridad correspondientes.

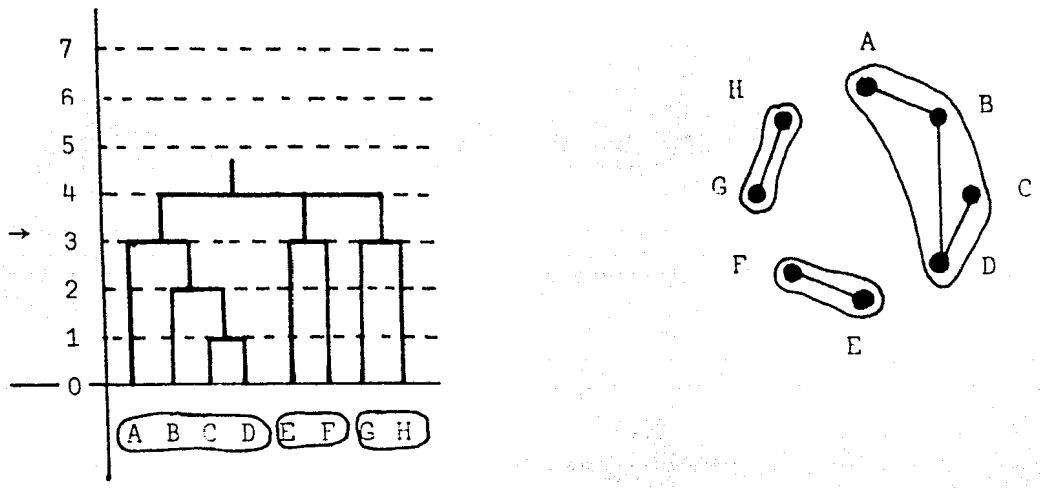


figura 2

Para el nivel de asociación 3, la figura anterior muestra la familia de grupos correspondiente a las componentes conexas del árbol de mínimo alcance.

Algunos comentarios y demostraciones de que éste es un método continuo pueden verse en "Fuzzy Relations and Dendrograms", de A. López y G. Espinosa, y en "Mathematical Taxonomy", de N. Jardine y R. Sibson.

Método de Conexión Completa. El algoritmo con el que trabaja este método ($d_c(d)$), es el siguiente:

1. Se selecciona el coeficiente de disimilaridad más pequeño, uniéndose los objetos correspondientes en un grupo (si hay varios iguales, entonces el orden en que se consideran es determinante para los re-

sultados que se obtienen, como se señala anteriormente).

2. Se recalcula el coeficiente entre el nuevo grupo y el resto de los objetos de la siguiente manera:

$$d(X,Y) = \max \{d(x,y) \mid x \in X,\ y \in Y\},$$

donde X y Y son objetos o grupos de objetos.

3. Si todavía no hay un solo grupo, volver al paso 1.

Los coeficientes seleccionados en el paso 1 son los resultantes $d_c(d)$ en la tabla final.

En términos de gráficas, los grupos en las familias obtenidas en cada nivel de asociación diferente, son subgráficas maximales completas de la gráfica inducida por $T_d(h)$. A pesar de que éste es un método Jerárquico, en el que las familias resultantes están constituídas por grupos cuya intersección es vacía, puede suceder que en un momento dado ocurra un traslape en vértices comunes a las subgráficas maximales completas para un determinado nivel de asociación. Considérese el siguiente ejemplo:

	J	K	L	M	N
J	0	2	6	3	3
K		0	5	8	10
L			0	8	12
M				0	2
N					0

G_3:

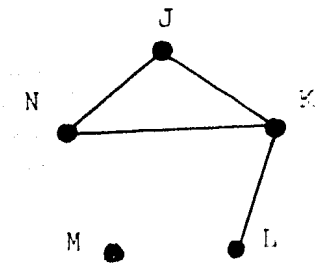


figura 3

A un nivel de asociación 3, las subgráficas maximales comple-

tas de G_3 son las determinadas por los siguientes conjuntos: {J}, {K}, {L}, {M}, {N}, {J,N,M}, {J,K}. Estas subgráficas representan a los grupos en la clasificación, y no son ajenas. Este problema surge como consecuencia de una mala definición del método; sin embargo ésto pasa inadvertido ya que dependiendo del orden de entrada de los coeficientes de disimilaridad, se considera como grupo en el dendrograma a la primera subgráfica maximal completa que aparece.

Métodos Promedio[5]. Como se señala en el capítulo 1, en estos métodos los valores de salida no son funciones continuas de los datos de entrada; de igual manera sucede con el método de Conexión Completa. N. Jardine y R. Sibson consideran que estos métodos no están bien definidos, y proponen una idea, que señalan, sigue siendo un método discontinuo; esta idea consiste en unir grupos simultáneamente cuando se encuentran coeficientes iguales, variando sólo la manera de recalcular las disimilaridades al utilizar los distintos métodos. Esto se ilustra con el siguiente ejemplo: considérese la tabla 1 de disimilaridades, donde ε es una constante pequeña, ya sea positiva, igual a cero, o negativa; a esta tabla se le aplican diferentes métodos, entre ellos, un método Promedio.

	A	B	C
A	0	$1+\varepsilon$	2
B		0	$1-\varepsilon$
C			0

tabla 1

[5] Véase el apéndice A, sección A.1 para los diferentes métodos.
[6] Ejemplo tomado de "Mathematical Taxonomy" de N. Jardine y R. Sibson, 1971.

El siguiente cuadro muestra los dendrogramas y tablas de disimilaridad para los diferentes valores de ε.

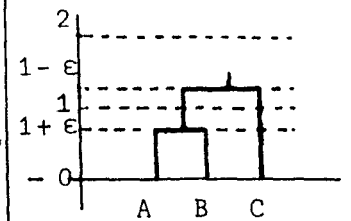
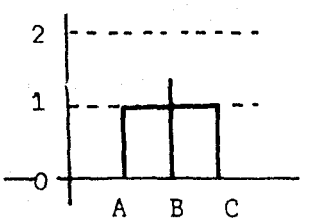
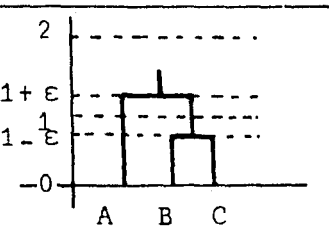
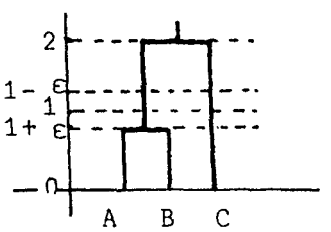
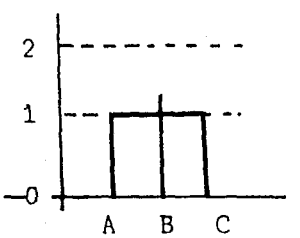
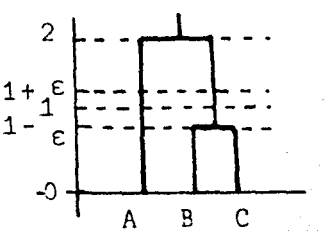
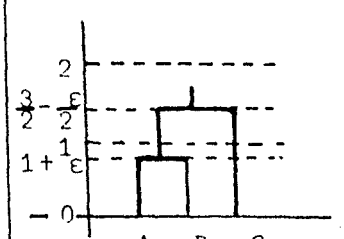
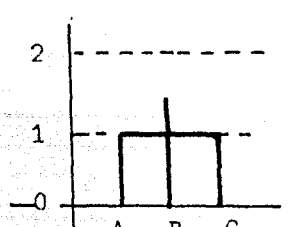
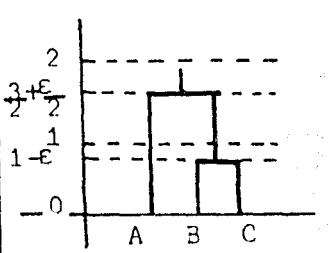
	$\varepsilon < 0$	$\varepsilon = 0$	$\varepsilon > 0$
Conexión Simple			
	A | B | C A | 0 | 1+ε | 1-ε B | | 0 | 1-ε C | | | 0	A | B | C A | 0 | 1 | 1 B | | 0 | 1 C | | | 0	A | B | C A | 0 | 1+ε | 1+ε B | | 0 | 1-ε C | | | 0
Conexión Completa			
	A | B | C A | 0 | 1+ε | 2 B | | 0 | 2 C | | | 0	A | B | C A | 0 | 1 | 1 B | | 0 | 1 C | | | 0	A | B | C A | 0 | 2 | 2 B | | 0 | 1-ε C | | | 0
Promedio entre grupos			
	A | B | C A | 0 | 1+ε | 3/2 - ε/2 B | | 0 | 3/2 - ε/2 C | | | 0	A | B | C A | 0 | 1 | 1 B | | 0 | 1 C | | | 0	A | B | C A | 0 | 3/2+ε/2 | 3/2+ε/2 B | | 0 | 1-ε C | | | 0

tabla 2

Es posible observar que en el caso del método de Conexión Completa y en el de los métodos Promedio, que proceden en forma similar a la del método Promedio entre grupos, al hacer variar de poca manera los datos de la matriz original, se producen cambios considerables en las tablas resultantes. Solamente en el caso del método de Conexión Simple se trata de una función continua en este ejemplo, aunque, como se dijo anteriormente, es posible demostrar que ésto pasa para todos los casos.

CAPITULO 4

4.1 Conjuntos Maximales Completos y k-transitividad

En el capítulo anterior se expone la construcción de un sistema Jerárquico de clasificación, en donde los grupos obtenidos por los diferentes métodos corresponden a las clases de equivalencia inducidas por las particiones resultantes en cada nivel de asociación. En la generalización a un sistema Numéricamente Estratificado se extiende el concepto de relación de equivalencia al de "k-equivalencia", de manera que los conjuntos obtenidos de esta relación al aplicar los métodos, dan lugar a grupos que cumplen con la propiedad de ser conjuntos maximales completos.

Definición. Sea $R \in RS(P)$. Un *Conjunto Maximal Completo* para R, es un subconjunto $S \subseteq P$ que satisface las siguientes condiciones:

$i)$ $S \times S \subseteq R$,

$ii)$ $A \notin S \Rightarrow$ existe $B \in S$ tal que $(A,B) \notin R$.

Esta definición equivale al concepto de subgráfica maximal completa de teoría de gráficas; es decir, a una subgráfica S de la gráfica G asociada a la relación R, para la que existen todos los posibles enlaces para el conjunto de vértices que la forman, y que no está propiamente contenida en otra subgráfica completa.

Definición. Sea $R \in R(P)$. Se dice que R es una relación *k-transitiva* si satisface la siguiente condición:

$$\{(\{A\} \times S) \cup (S \times S) \cup (S \times \{B\})\} \subseteq R \Rightarrow (A,B) \in R,$$ donde $A, B \in P$, $S \subseteq P$ tal que $|S| = k$.

Cuando R es una relación reflexiva, esta condición equivale a la usual de transitividad cuando $k = 1$ y se sustituye al subconjunto S por un tercer elemento "C". En el caso general, este elemento individual se extiende a S, que es un subconjunto completo de P.

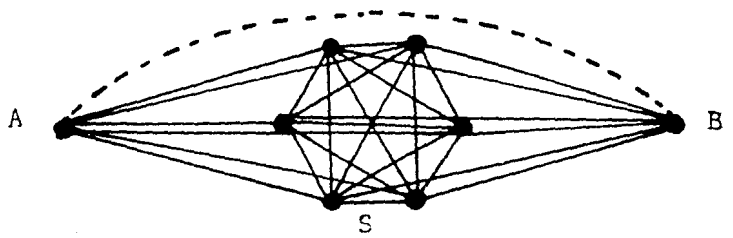


figura 1

Definición. Sea $R \in R(P)$. R es una relación de *k-equivalencia*, si $R \in RS(P)$ y R cumple con la k-transitividad.

4.2 k-Dendrogramas y Desigualdad k-Ultramétrica

La relación $T_d(h)$ no cumple con la transitividad en la mayoría de las tablas. De la misma forma esta relación difícilmente es k-transitiva. Para obtener los dendrogramas generalizados para los que se consiguen relaciones de k-equivalencia en cada nivel de asociación, se presenta en esta sección la definición de k-dendrograma, que corresponde a la dada en el capítulo 2. De igual manera que se generaliza el concepto de dendrograma al de k-dendrograma, se extiende la condición de ultrametría a

la de k-ultrametría, para así demostrar que los coeficientes k-ultramétricos se identifican bajo la función T con los k-dendrogramas.

Definición. Un *k-Dendrograma*[1] es una función

$$c: I \rightarrow F_k(P),$$

donde $I = A \cup \{0\}$, con $A \subseteq \mathbb{R}^+$ e $|I| < \infty$; $E_k(P)$ es el conjunto de relaciones de k-equivalencia en P; que satisface las siguientes condiciones:

i) Existe $h \in I$ tal que $c(h) = PxP$,

ii) $h < h' \Rightarrow c(h) \subseteq c(h')$ con $c(h) \neq c(h')$, h y $h' \in I$.

Al conjunto de k-dendrogramas en P se le denota por $D_k(P)$.

Definición. Sea $S \subseteq P$ y $d \in CD(P)$, entonces el *diámetro* de S, dado d, se define de la siguiente manera:

$$diam(d,S) = \max\{ d(X,Y) \mid X, Y \in S \}.$$

Los coeficientes de disimilaridad que cumplen con la siguiente condición son los coeficientes k-ultramétricos. La aplicación de distintos métodos Numéricamente Estratificados, al hacer variar la k, dá lugar a formas diferentes de transformar tablas de disimilaridad dadas, en k-ultramétricas.

Definición. Sea $S \subseteq P$, tal que $|S| = k$. Sea $d \in CD(P)$ y, A y B

[1] Igual que en el capítulo 3, esta definición no coincide con la de N. Jardine y R. Sibson en la misma publicación. Véase el apéndice B, sección B.2.

$\in P$. Se dice que d es un coeficiente de disimilaridad *k-Ultramétrico* si cumple con la siguiente condición:

$$d(A,B) \leq \max \{ \text{diam}(d, S \cup \{A\}), \text{diam}(d, S \cup \{B\}) \},$$

o bien:

$$d(A,B) \leq \max \{ d(X,Y) \mid X \in S \cup \{A,B\}, Y \in S \}.$$

De la misma manera que en el capítulo anterior, ya que un k-dendrograma se define como una función $c: I \rightarrow E_k(P)$, donde $E_k(P)$ es el conjunto de relaciones de k-equivalencia en P, entonces para que $T_d(h) \in D_k(P)$, se procede a probar que los coeficientes sobre los que se aplica la función T tienen que ser k-ultramétricos.

Proposición. Sea $CU_k(P)$ el conjunto de coeficientes de disimilaridad k-ultramétricos en P. Entonces $d \in CU_k(P)$, si y sólo si $T_d(h) \in E_k(P)$, para todo $h \in I$.

Prueba:

i) Sea $d \in CU_k(P)$. $T_d(h) \in RS(P)$, para todo $h \in I$, como se demostró anteriormente. Sea $h \in I$, A y B $\in P$, $S \subseteq P$ tal que $|S| = k$, entonces:

$\{(\{A\} \times S) \cup (S \times S) \cup (S \times \{B\}) \subseteq T_d(h) \Rightarrow d(X,Y) \leq h$, para todo $(X,Y) \in \{(\{A\} \times S) \cup (S \times S) \cup (S \times \{B\})\} \Rightarrow$

$d(A,B) \leq \max \{d(X,Y) \mid X \in S \cup \{A,B\}, Y \in S\} \leq h \Rightarrow$

$(A,B) \in T_d(h)$.

$\therefore T_d(h) \in E_k(P)$, para todo $h \in I$.

ii) Sea $T_d(h) \in E_k(P)$, para todo $h \in I$. Dados A y B $\in P$, $S \subseteq P$ tal que $|S| = k$, sea $h = \max \{d(X,Y) \mid X \in S \cup \{A,B\}$,

$Y \in S\}$. Entonces: $d(X,Y) \leq h$, para todo $(X,Y) \in \{ S \cup \{A,B\}xS\}$, *i.e.* $\{(\{A\}xS) \cup (SxS) \cup (Sx\{B\})\} \subseteq T_d(h)$. Por k-transitividad $(A,B) \in T_d(h) \Rightarrow d(A,B) \leq h = \max \{d(X,Y) \mid X \in S \cup \{A,B\}, Y \in S\}$.

$\therefore d \in CU_k(P)$.

Por lo que, dado un coeficiente de disimilaridad k-ultramétrico, es posible asociarle un solo k-dendrograma que lo represente, y viceversa.

Proposición. Sea $d \in CD(P)$. Si $d \in CU_k(P)$, entonces $d \in CU_{k+1}(P)$.

Prueba:

Sea $d \in CU_k(P)$, $S \subseteq P$ tal que $|S| = k+1$, y $(A,B) \in PxP$.

Sea $C \in S$, y $R = S-\{C\}$.

$d(A,B) \leq \max \{d(X,Y) \mid X \in R \cup \{A,B\}, Y \in R\}$

$\leq \max \{d(X,Y) \mid X \in S \cup \{A,B\}, Y \in S\} \Rightarrow d \in CU_{k+1}(P)$.

Proposición. Sea $d \in CD(P)$. Si $T_d(h) \in E_k(P)$, para todo $h \in I$, entonces $T_d(h) \in E_{k+1}(P)$, para todo $h \in I$.

La figura 2 ilustra la manera en que se contienen los conjuntos definidos a lo largo de este capítulo y del anterior, así como su correspondencia bajo las funciones T y U.

Obsérvese que $CD(P) = CU_{N-1}(P)$, donde $N = |P|$; también se cumple que $RS(P) = E_{N-1}(P)$, ya que toda relación reflexiva

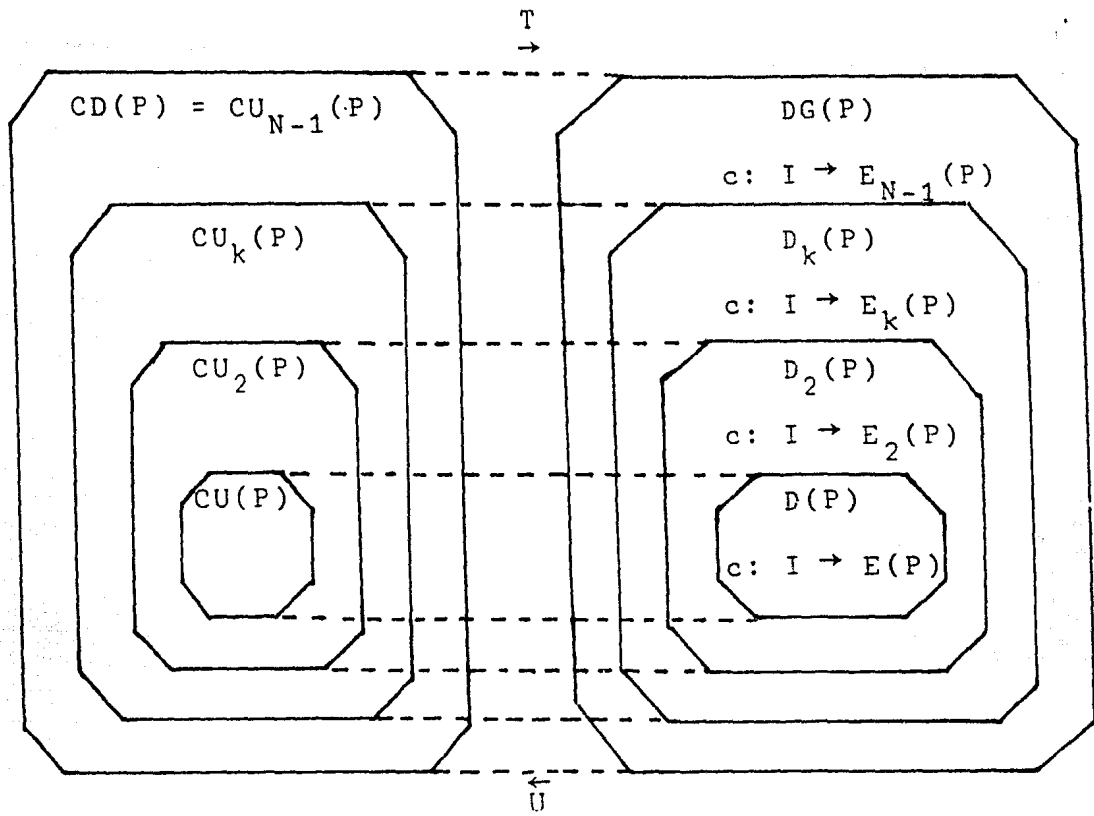


figura 2

y simétrica en un conjunto de |P| puntos, es también (|P-1|)-transitiva.

4.3 Métodos Numéricamente Estratificados

En el capítulo anterior se demostró que el método de Conexión Simple es el método CU-subdominante de d, para d $\in$ CD(P). Esto, en cierto sentido, hace que el resultado obtenido sea el mejor posible bajo condiciones impuestas. Sin embargo, debido al defecto de encadenamiento que posee, produce demasiada simplificación en los datos originales, con un alto grado de deformación. Lo que los métodos Numéricamente Estratificados hacen, es brindar un estado intermedio de simplificación y deformación, construyendo sistemas que se basan en la defini-

ción de métodos subdominantes. Estos métodos, como en el caso Jerárquico, son también maneras diferentes de transformar coeficientes de disimilaridad dados en coeficientes k-ultramétricos , para k = 1, 2, ..., N-1; es decir, son funciones d_k : $CD(P) \rightarrow CU_k(P)$. Los grupos en la clasificación son los conjuntos maximales completos en los datos resultantes para cada nivel de asociación. Si se tomaran en consideración los conjuntos maximales completos inducidos por una relación R, tal que $R \in RS(P)$, entonces éstos podrían formar un conjunto difícil de interpretar, por lo que al dar restricciones en el traslape de ellos, se trata de obtener una mejor descripción de los grupos en la clasificación.

En el capítulo 2 se habla de diferentes tipos de restricciones en el traslape entre grupos, y se exponen más tarde, los métodos B_k que corresponden a restricciones de tipo absoluto. N. Jardine y R. Sibson en "Mathematical Taxonomy" presentan también otro tipo de métodos Numéricamente Estratificados que se basan en restricciones de tipo interno. Estos métodos llamados C_u,no hacen la transformación a coeficientes k-ultramétricos, sino a otro tipo de coeficientes; sin embargo en esta sección no se incluyen, ya que el interés básico de este trabajo radica en los métodos B_k.

Los métodos B_k. Estos métodos están basados en la restricción absoluta de que el traslape entre los diferentes conjuntos maximales completos para un cierto nivel, no contiene más de k-1 elementos de P. La k toma valores 1,2,...,N-1, donde $|P|=$

N. Así, B_1 corresponde al método de Conexión Simple, y a medida que k crece se permite un mayor número de elementos en el traslape, hasta que $k \geq N-1$, con lo que se obtiene la identidad. Para constituir los métodos B_k se construyen conjuntos de coeficientes de disimilaridad que definen métodos subdominantes; es decir, conjuntos sup-cerrados tales que sus elementos corresponden bajo la función T a los dendrogramas generalizados que cumplen con la propiedad de que el traslape entre los grupos resultantes para cada nivel, tiene a lo más k-1 objetos de P. Estos conjuntos sup-cerrados son los coeficientes de disimilaridad k-ultramétricos, y los dendrogramas generalizados con la propiedad mencionada son los k-dendrogramas.

Proposición. Sea $R \in RS(P)$. R es k-transitiva si y sólo si, los conjuntos maximales completos para R se intersectan en a lo más k-1 objetos de P.

Prueba:

i) Sea $R \in RS(P)$, tal que R es k-transitiva. Dados S_1 y S_2 conjuntos maximales completos para R, supóngase que $|S_1 \cap S_2| \geq k$. Sea $A \in S_1$ y $B \in S_2$; debido a la k-transitividad, se tiene que:

$\{(\{A\} \times (S_1 \cap S_2)) \cup ((S_1 \cap S_2) \times (S_1 \cap S_2)) \cup ((S_1 \cap S_2) \times \{B\})\} \subseteq R \Rightarrow (A,B) \in R$. Pero como S_1 y S_2 son maximales completos para R, entonces $A \in S_2$ y $B \in S_1$; es decir: $S_1 = S_2$. P r lo tanto $S_1 \neq S_2 \Rightarrow |S_1 \cap S_2| < k$.

ii) Sea $R \in RS(P)$, tal que dados dos conjuntos maximales

completos para R, S_1 y S_2, si son diferentes, entonces $|S_1 \cap S_2| < k$. Dados A, B $\in$ P, y S $\subseteq$ P tal que $|S| = k$, supóngase que $\{(\{A\}xS) \cup (SxS) \cup (Sx\{B\})\} \subseteq$ R. Entonces $\{A\}xS$ y $Sx\{B\}$ son conjuntos completos para R. Sean S_A y S_B los conjuntos maximales completos que contienen a $\{A\}xS$ y a $Sx\{B\}$ respectivamente. Ya que $S \subseteq S_A \cap S_B$ y $|S| = k$, entonces $|S_A \cap S_B| \geq k \Rightarrow S_A = S_B$. Por lo tanto, $(A,B) \in S_A \Rightarrow (A,B) \in R$.
$\therefore$ R es k-transitiva.

Por lo que, la k-transitividad es la condición necesaria y suficiente para pedir conjuntos maximales completos que se intersecten en a lo más k-1 elementos.

La idea de los métodos B_k, en términos de gráficas, consiste en añadir aristas a la gráfica inducida por la relación $T_d(h)$, de tal forma que se obtienen nuevas gráficas en las que cada pareja de subgráficas maximales completas tiene a lo más k-1 vértices en la intersección. Los vértices de estas subgráficas son los grupos en el nivel correspondiente. El algoritmo con el que proceden estos métodos, es el siguiente:

Dado d $\in$ CD(P) y h $\in$ I, para k = 1,2,...,N-1,

1. Considérese h = 0.
2. Constrúyase la gráfica G_h, en donde los vértices representan a los objetos de P y las aristas unen aquellos vértices cuyos coeficientes de disimilaridad son menores o iguales a h.

3. Encuéntrense las subgráficas maximales completas de G_h.
4. Si dos subgráficas maximales completas se intersectan en k o más elementos, añádanse las aristas necesarias para convertirlas en una sola subgráfica maximal completa.
5. Si se añadieron aristas, vuélvase al punto 3, considerando ahora la nueva gráfica.
6. Los grupos resultantes para este nivel son los conjuntos de vértices correspondientes a las subgráficas maximales completas de la nueva gráfica, a cada una de las cuales se les llama C_k^h.
7. Si aún no hay un solo grupo o una subgráfica maximal completa, increméntese h y vuélvase al punto 2.

Los coeficientes de la tabla d_k, correspondiente a la aplicación del método B_k, se pueden obtener a partir de las C_k^h, de la siguiente manera:

Definición. Los métodos B_k, para $k = 1,2,\ldots,N-1$, donde $N = |P|$, son funciones

$$d_k: CD(P) \rightarrow CU_k(P),$$

tales que:

$$d_k(d)(A,B) = \min \{h \mid (A,B) \in C_k^h\},$$

donde $(A,B) \in PxP$, $h \in I$ y C_k^h es alguna de las subgráficas maximales completas descritas en el punto 6 del algoritmo anterior.

Proposición. Los métodos B_k, con $k = 1,2,\ldots,N-1$, donde $N = |P|$, son los métodos CU_k-subdominantes de d, es decir:

$$d_k(d) = \sup \{d^* \in CU_k(P) \mid d^* \leqslant d\},$$

para todo $d \in CD(P)$.

Primero hay que probar que $CU_k(P)$ define un método subdominante, es decir :

i) El conjunto $Y = \{d^* \in CU_k(P) \mid d^* \leqslant d\} \neq \phi$, para todo $d \in CD(P)$.

ii) Dado $X \subseteq CU_k(P)$ acotado, entonces $\sup X \in CU_k(P)$.

Después se demuestra que:

iii) $d_k(d) \in CU_k(P)$, para todo $d \in CD(P)$.

iv) $d_k(d) \leqslant d$, para todo $d \in CD(P)$.

v) Dado $d' \in CU_k(P)$ tal que $d' \leqslant d$, entonces $d' \leqslant d_k(d)$.

Y por lo tanto $d_k(d) = \sup \{d^* \in CU_k(P) \mid d^* \leqslant d\}$.

Prueba:

i) El conjunto $Y = \{d^* \in CU_k(P) \mid d^* \leqslant d\} \neq \phi$, ya que la tabla que tiene solamente ceros en sus entradas (d_0), es k-ultramétrica y $d_0 \leqslant d$, para todo $d \in CD(P)$.

ii) Sean A y B $\in P$. Sea $Z \subseteq CU_k(P)$ acotado. Ya que Z es acotado, entonces sup Z existe. Sea $(\sup Z)(K,L) = d_z(K,L)$, para todo $(K,L) \in P \times P$; entonces dado $S \subseteq P$, tal que $|S| = k$:

$$d_z(A,B) = \sup_{d \in Z} \{d(A,B)\} \leqslant \sup_{d \in Z} [\max \{d(X,Y) \mid X \in S \cup \{A,B\}, Y \in S\}]$$

$$= \max \{\sup_{d \in Z} \{d(X,Y) \mid X \in S \cup \{A,B\}, Y \in S\}\}$$

$$= \max \{d_z(X,Y) \mid X \in S \cup \{A,B\}, Y \in S\}.$$

$\therefore d_z \in CU_k(P)$.

Entonces, de *i* e *ii*, $CU_k(P)$ define un método subdominante.

iii) Para demostrar que $d_k(d) \in CU_k(P)$, para todo $d \in CD(P)$, basta ver que dados $k+2$ elementos de P, el coeficiente mayor de $d_k(d)$ aparece por lo menos dos veces. Sean $A_0, A_1, A_2, \ldots A_{k+1} \in P$. Sea $\alpha = d_k(d)(A_0, A_{k+1}) = \max \{d_k(d)(A_i, A_j)\}$, i y $j = 0, 1, 2, \ldots, k+1$. Sea $\beta = \max \{d_k(d)(A_i, A_j)\}$ tal que i y j toman los valore anteriores, menos $i = 0$ y $j = k+1$ simultáneamente. Entonces por construcción de C_k^h, para el nivel de asociación β, existe una subgráfica maximal completa C_k^β tal que $(A_0, A_{k+1}) \in C_k^\beta$, por lo que: $\alpha = d_k(d)(A_0, A_{k+1}) \leq \beta \leq \alpha$, $\therefore \alpha = \beta$.

iv) Por construcción de C_k^h, ya que en todo nivel sólo se aumentan aristas, si $d(A,B) = \alpha$, entonces $(A,B) \in C_k^h \Rightarrow d_k(d)(A,B) = \min \{h \mid (A,B) \in C_k^h\} \leq \alpha = d(A,B)$.

v) Antes de demostrar este inciso, se prueba un resultado que es en cierto sentido equivalente, y que sirve para hacer la demostración más sencilla. Se prueba que: $T_{d_k}(h) = \min \{R \in E_k(P) \mid T_d(h) \subseteq R\}$. Ya se demostró que $d \in CU_k(P) \Leftrightarrow T_d(h)$ es k-transitiva para todo h, o lo que es igual, $d \in CU_k(P) \Leftrightarrow T_d(h) \in E_k(P)$ para todo h. También se vió que $d_k(d) \leq d$, y que $d_k(d) \in CU_k(P)$, para todo h. Por lo anterior, y por construcción de T, se sigue que $T_d(h) \subseteq$

$T_{d_k}(h)$ para todo h, y $T_{d_k}(h) \in E_k(P)$ para todo h. Por lo tanto, probar que $T_{d_k}(h) = \{\min R \in E_k(P) \mid T_d(h) \subseteq R\}$, equivale a demostrar que $T_{d_k}(h)$ se construyó añadiendo a $T_d(h)$ el mínimo número de elementos de PxP necesarios para obtener una relación de k-equivalencia. Se desea probar, entonces, que dado $(A,B) \in \{T_{d_k}(h) - T_d(h)\}$, entonces $R = T_{d_k}(h) - \{(A,B),(B,A)\}$ no es una relación de k-equivalencia en P. *Prueba:* con base en la definición del método B_k, y del algoritmo para construir las C_k^h, $(A,B) \in \{T_{d_k}(h) - T_d(h)\} \Rightarrow$ existe $h' \leqslant h$ tal que en la gráfica asociada a $T_{d_k}(h')$ existen S_A y S_B subgráficas maximales completas, tales que $A \in \{S_A - S_B\}$ y $B \in \{S_B - S_A\}$, y $|S_A \cap S_B| \geqslant k$. Como S_A y S_B son subgráficas maximales completas, entonces $(X,Y) \in T_{d_k}(h')$ para todo $X, Y \in S_A$, y $(X,Y) \in T_{d_k}(h')$ para todo $X, Y \in S_B$. Sea $S = \{v_1, v_2, \ldots, v_k\} \subseteq S_A \cap S_B$, entonces $\{A\}xS \subseteq T_{d_k}(h') \subseteq T_{d_k}(h)$ y $\{B\}xS \subseteq T_{d_k}(h') \subseteq T_{d_k}(h)$; además $SxS \subseteq T_{d_k}(h') \subseteq T_{d_k}(h)$. Estas tres últimas afirmaciones implican que $R = T_{d_k}(h) - \{(A,B),(B,A)\}$ no es k-transitiva; $\therefore T_{d_k}(h) = \min \{R \in E_k(P) \mid T_d(h) \subseteq R\}$. En el inciso *iii* se demuestra que el resultado de aplicar el método B_k

es k-ultramétrico; es decir, que el algoritmo añade en cada nivel, aristas suficientes para obtener una relación de k-equivalencia. Ahora se acaba de probar que todas las aristas que añade el algoritmo en un nivel, son necesarias para obtener una relación de k-equivalencia. Con este resultado es inmediato demostrar el inciso v: sea $d' \in CU_k(P)$ tal que $d' \leqslant d$, entonces para todo h se tiene que $T_d(h) \subseteq T_d(h')$ y $T_{d'}(h) \in E_k(P) \Rightarrow T_{d_k}(h) \subseteq T_{d'}(h)$ para todo h $\Rightarrow$ $d' \leqslant d_k(d)$.

$\therefore d_k(d) = \min \{h \mid (A,B) \in C_k^h\} = \sup\{ d^* \in CU_k(P) \mid d^* \leqslant d \}$ para todo $(A,B) \in PxP$.

Proposición. $d_k(d)(A,B) \leqslant d_{k+1}(d)(A,B)$, para todo $(A,B) \in PxP$ y $k = 1,2,\ldots,N-1$, con $|P| = N$.

Prueba:

Anteriormente se demuestra que si $d \in CU_k(P)$, entonces $d \in CU_{k+1}(P)$; ya que $d_{k+1}(d) = \sup\{ d^* \in CU_{k+1}(P) \mid d^* \leqslant d\}$, entonces $d_k(d)(A,B) \leqslant d_{k+1}(d)(A,B)$, para todo $(A,B) \in PxP$.

Dada la forma en que se definieron estos métodos, es posible observar que se trata del método de Conexión Simple cuando k = 1. Así, al ir aumentando el valor de k, se va deformando en menor grado la información contenida en la tabla de disimilaridades original, hasta que $k \geqslant N-1$ se obtiene esta última; es decir $B_k = I$ para $k \geqslant N-1$. Esto sucede, ya que como se señala an

teriormente, toda relación definida en un conjunto de N objetos es N-1 transitiva, y por lo tanto toda tabla de disimilaridades de N objetos es N-1-ultramétrica. Debido a ésto, se tiene que $d_k(d)(A,B) = \min\{ h \mid (A,B) \in C_k^h \} = d(A,B)$, para todo $(A,B) \in PxP$ y $k \geq N-1$; es decir, la tabla original nunca se altera ya que en ningún momento es posible agregar aristas para completar subgráficas maximales completas. En el capítulo 2 se menciona que estos métodos poseen la ventaja de que al aplicarlos para los valores mayores de k, se puede obtener la tabla resultante para valores menores a partir del producto de esta aplicación, y no de la tabla de disimilaridades original. La siguiente proposición establece este hecho.

Proposición. $B_k \circ B_{k+1} = B_k$, para $k = 1,2,\ldots,N-1$.

Prueba:

En las proposiciones anteriores se establece que $d_k(d) \leq d_{k+1}(d) \leq d$, para todo $d \in CD(P)$. Al aplicarles el método B_k a cada uno de los elementos de esta desigualdad, se obtiene: $d_k(d_k(d)) \leq X \leq d_k(d)$. Ya que $d_k(d_k(d)) = \sup \{d^* \in CU_k(P) \mid d^* \leq d_k\} = d_k(d)$, entonces: $d_k(d) \leq X \leq d_k(d)$, por lo que $X = d_k(d)$.

Además de esta ventaja y de la posibilidad de escoger el grado de deformación que el usuario juzgue conveniente que sufran los datos originales[2], se habla también de la continuidad de

[2] Véase la nota 6 del capítulo 2.

teriormente, toda relación definida en un conjunto de N objetos es N-1 transitiva, y por lo tanto toda tabla de disimilaridades de N objetos es N-1-ultramétrica. Debido a ésto, se tiene que $d_k(d)(A,B) = \min\{ h \mid (A,B) \in C_k^h \} = d(A,B)$, para todo $(A,B) \in PxP$ y $k \geq N-1$; es decir, la tabla original nunca se altera ya que en ningún momento es posible agregar aristas para completar subgráficas maximales completas. En el capítulo 2 se menciona que estos métodos poseen la ventaja de que al aplicarlos para los valores mayores de k, se puede obtener la tabla resultante para valores menores a partir del producto de esta aplicación, y no de la tabla de disimilaridades original. La siguiente proposición establece este hecho.

Proposición. $B_k \circ B_{k+1} = B_k$, para $k = 1,2,\ldots,N-1$.

Prueba:

En las proposiciones anteriores se establece que $d_k(d) \leq d_{k+1}(d) \leq d$, para todo $d \in CD(P)$. Al aplicarles el método B_k a cada uno de los elementos de esta desigualdad, se obtiene: $d_k(d_k(d)) \leq X \leq d_k(d)$. Ya que $d_k(d_k(d)) = \sup \{d^* \in CU_k(P) \mid d^* \leq d_k\} = d_k(d)$, entonces: $d_k(d) \leq X \leq d_k(d)$, por lo que $X = d_k(d)$.

Además de esta ventaja y de la posibilidad de escoger el grado de deformación que el usuario juzgue conveniente que sufran los datos originales[2], se habla también de la continuidad de

[2] Véase la nota 6 del capítulo 2.

los métodos B_k. Como en el caso del método de Conexión Simple, los comentarios y demostraciones de esta propiedad no se incluyen en este trabajo, pero pueden verse en "Fuzzy Relations and Dendrograms", de A. López y G. Espinosa, y en "Mathematical Taxonomy", de N. Jardine y R. Sibson.

CAPITULO 5

5.1 Algunas consideraciones sobre la aplicación de Clasificación de grupos

En este capítulo se pretende señalar algunos aspectos importantes para la comprensión y aplicación de la teoría de clasificación. Con este objeto se citan a continuación los puntos que marca M. Anderberg en "Cluster Analysis for Applications":

1. Cualquier conjunto de datos puede ser clasificado de muchas y diversas maneras, todas ellas significativas. Cada clasificación puede resaltar diferentes aspectos de los datos y no es posible hablar de una sola clasificación correcta. Esto es importante ya que puede resaltar aspectos nuevos, desconocidos en la estructura de los datos.
2. La clasificación es un recurso para sugerir hipótesis. La importancia de una clasificación particular radica en el hecho de observar cierta consistencia entre los resultados obtenidos, así como en su estructura y hechos conocidos acerca del problema del que se trata.
3. Ya que un conjunto de grupos resultantes de un método de clasificación no es un resultado definitivo del proceso, sino un dato sugestivo que debe ser estudiado por el usuario, no es recomendable la utilización de algoritmos muy detallados y caros, ya que

muchas veces es posible obtener resultados del mismo tipo, utilizando otros algoritmos.

4. Los métodos de clasificación imponen en los datos una cierta estructura que depende de los criterios utilizados, y revelan otra que originalmente tenían. Esto sucede ya que los métodos prácticos de clasificación trabajan con operaciones sucesivas que pueden ignorar algunos aspectos de los datos y resaltar otros.
5. En relación a los puntos anteriores, puede ser común que el resultado de una clasificación arroje información razonable sólo para una parte de los datos originales. Es posible, entonces, analizar separadamente aquel grupo o grupos, y estudiar la aplicación de otra técnica al resto de los datos.
6. Es posible que sucedan cualesquiera de las dos siguientes situaciones:
   a. Los datos pueden no contener ningún grupo. Esto puede pasar cuando al clasificar variables, éstas sean completamente independientes.
   b. Los datos pueden contener un solo grupo. Esto sucedería cuando al clasificar variables, éstas sean completamente dependientes.

5.2 Necesidad de aplicación de varios métodos en una clasificación

Como se señala anteriormente, cada método de clasificación origina deformaciones diferentes en la matriz de datos originales y de esta manera hace resaltar aspectos distintos en la estructura de los datos. Debido a esta razón es conveniente que el usuario no se base en los resultados obtenidos al aplicar un solo método de clasificación, sino que debe comparar y estudiar los grupos resultantes de diversos métodos; así, se pretende buscar similitudes, consistencias, o propiedades relevantes considerando los criterios impuestos por los diferentes métodos.

Se ha señalado que algunos métodos Jerárquicos y otros Numéricamente Estratificados, reproducen el dendrograma o k-dendrograma original si se cuenta de antemano con tablas ultramétricas o k-ultramétricas. Si se aplicara solamente un método que no cumpliera con esta propiedad, no se sabría si se contaba con una tabla ultramétrica o k-ultramétrica originalmente.

La conveniencia de usar diversos métodos al hacer clasificaciones Jerárquicas, puede verse en "Introducción a los Métodos Jerárquicos de Análisis de Cúmulos", de G. Espinosa y A. López. En cuanto a los métodos Numéricamente Estratificados es recomendable comparar las situaciones resultantes al hacer variar el grado de traslape permitido en los grupos obtenidos.

APENDICE A

A.1 Coeficientes de Disimilaridad (Similaridad)

Sean X y X dos objetos, o dos grupos de objetos, y N el número de características en consideración. Los siguientes son los coeficientes citados en el capítulo 1, sección 1.2:[1]

1. *Coeficiente de Gower*. Es un coeficiente de similaridad que se utiliza para casos de tablas del tipo presencias-ausencias, y para tablas con información cuantitativa en general:

$$s(X,Y) = \sum_{i=1}^{N} \frac{W_{ixy}\, Z_{ixy}}{W_{ixy}} ,$$

donde

$$W_{ixy} = \begin{cases} 1 & \text{si la característica i está definida para los objetos X y Y} \\ 0 & \text{en caso contrario} \end{cases}$$

y Z_{ixy} se puede definir de dos maneras:

a. Para tablas del tipo presencias-ausencias:

$$Z_{ixy} = \begin{cases} 1 & \text{si el estado de la característica i del objeto X, coincide con el estado de la característica i del objeto Y} \\ 0 & \text{en caso contrario} \end{cases}$$

[1] Véase "Cluster Analysis for Applications", de M. Anderberg, 1973.

b. Para tablas de valores cuantitativos en general:

$$Z_{ixy} = \frac{1 - |K_{ix} - K_{ir}|}{R_i}, \text{ donde}$$

K_{ix} = valor de la característica i de X,

K_{iy} = valor de la característica i de Y,

R_i = rango de la característica i = $\max_x K_{ix} - \min_y K_{iy}$.

2. *Métrica de Minkowski*. Es un coeficiente de disimilaridad para tablas de valores cuantitativos en general:

$$d(X,Y) = \left(\sum_{i=1}^{N} |K_{ix} - K_{iy}|^r \right)^{1/r},$$

donde $r \in \mathbb{R}^+$ y es fijado por el usuario. Cuando r = 1 se obtiene la *Distancia de Manhattan* (3) o *City Block*, y cuando r = 2, se obtiene la *Distancia Euclidiana* (4).

Los siguientes coeficientes son todos de similaridad, y son aplicables para el caso de presencias-ausencias. Para definirlos se utiliza el siguiente código:

X y Y son objetos o grupos de objetos,

p = número de características presentes en ambos objetos,

a = número de características ausentes en ambos objetos,

x = número de características presentes en el objeto X y no en el Y,

y = número de características presentes en el objeto Y y no en el X,

N = número total de características en consideración (p+a+x+y).

5. *Coeficiente de Jaccard*.

$$s(X,Y) = \frac{p}{p + x + y} .$$

6. *Coeficiente de Dice*.

$$s(X,Y) = \frac{2p}{2p + x + y} .$$

7. *Proporción de Presencias y Ausencias Comunes*.

$$s(X,Y) = \frac{p + a}{N} .$$

8. *Presencias Comunes*.

$$s(X,Y) = p .$$

9. *Presencias y Ausencias Comunes*.

$$s(X,Y) = p + a .$$

A.2 Métodos Jerárquicos Promedio

Los métodos Promedio citados en el capítulo 1, sección 1.4, son los siguientes:[2]

Método Promedio entre grupos. También se le conoce como *Mean Link, Unweighted Average*, o *Average Linkage between merged groups*. El algoritmo con el que trabaja este método es el siguiente:

1. Dada una tabla de disimilaridades, se busca en ella la entrada menor. Se agrupan los objetos correspondientes en un solo grupo. Las disimilaridades menores escogidas de esta forma son las entradas correspondientes en la tabla final.
2. Se recalcula la disimilaridad entre el nuevo grupo y los elementos restantes en consideración, de la siguiente manera:

$$d(X,Y) = \frac{1}{N_X N_Y} \sum_{\substack{x \in X \\ y \in Y}} d(X,Y) \ ,$$

donde X y Y $\in$ P, o son dos grupos de objetos de P;

N_X = número de elementos de X,

N_Y = número de elementos de Y.

3. Se repiten los dos pasos anteriores hasta que haya un solo grupo.

Lo que hace este método es calcular el promedio aritmético de los valores de los coeficientes de disimilaridad entre un objeto y los elementos del grupo al que puede pertenecer, o entre los elementos de dos grupos que se pueden fusionar.

[2] Véase "Cluster Analysis for Applications", de M. Anderberg, 1973.

Método Promedio dentro de grupos. También se le llama *Mean within off diagonal*, o *Average Linkage within the new group*. Este método opera con base en un algoritmo distinto al usado por los métodos vistos anteriormente, que es el mismo en todos ellos con excepción de la manera de recalcular las disimilaridades. El nuevo algoritmo es el siguiente:

1. Dada una tabla de disimilaridades, se construye a partir de ésta, una nueva tabla M de disimilaridades cuyas entradas son una "medida" que se asocia al grupo que se formaría al unir los objetos correspondientes al renglón y a la columna de dichas entradas. Esta medida se calcula de la siguiente manera:

$$M(X) = \frac{1}{N_X(N_X - 1)/2} \sum_{\substack{x,x' \in X \\ x \neq x'}} d(x,x') \ ,$$

   donde X es un grupo de N_X objetos; la diagonal de la matriz M se mantiene igual; y si en los sumandos se incluye a d(x,x'), entonces no se incluye a d(x',x).
2. En esta nueva tabla M se agrupan los objetos o grupos de objetos cuya medida sea la menor.
3. Se repiten los dos pasos anteriores hasta formar un solo grupo.

Si se colocan en una matriz los elementos del grupo X, entonces M(X) se calcula considerando aquellas entradas por arriba

de la diagonal, por lo que se promedia entre $N_X(N_X - 1)/ 2$ elementos. En particular, la matriz M que se obtiene de este método al realizar el primer paso del algoritmo, resulta ser la misma que la original; sin embargo, este algoritmo se presenta así, ya que se utiliza para explicar el funcionamiento de otros métodos en los que si ocurren cambios.

Método Promedio pesado dentro de grupos. También se le conoce como *Weighted Mean within*. Este método opera con el mismo algoritmo que el método expuesto anteriormente, sólo que en el paso 1 la manera de calcular la medida, es la siguiente:

$$M(X) = \frac{1}{N_X} \sum_{x,x' \in X} d(x,x') \ .$$

El "peso" que se le dá a cada grupo, es su propio tamaño.

Método Incremento al Promedio pesado dentro de grupos. Es llamado también *Incremento al Weighted Mean within*. Este método funciona con el agoritmo usual (el del método de Conexión Simple, Conexión Completa, etc.); pero la manera de recalcular las disimilaridades en cada paso es la siguiente:

$$d(X,Y) = M(Z) - M(X) - M(Y) \ ,$$

donde $Z = X \cup Y$; y M es la medida usada en el método Promedio pesado dentro de grupos.

APENDICE B

B.1 Definición de Dendrograma de N. Jardine y R. Sibson. Equivalencia con la definición del capítulo 3

El objeto de este apéndice es presentar la definición de dendrograma y k-dendrograma de N. Jardine y R. Sibson[1], y establecer una equivalencia con las dadas en este trabajo. Estos autores definen un dendrograma como una función

$$c: [0, \infty) \to E(P),$$

donde E(P) es el conjunto de relaciones de equivalencia en P, y que satisface las siguientes condiciones:

i) Existe $h \in [0, \infty)$ tal que $c(h) = PxP$,

ii) $c(h) \subseteq c(h')$ si $h \leq h'$,

iii) Dado h, existe $\delta > 0$ tal que $c(h + \delta) = c(h)$.

Esta tercera condición elimina la posibilidad de que puedan darse, por ejemplo, dendrogramas definidos de la siguiente manera:

$$c(h) = \begin{cases} \{(A,A) \text{ para todo } A \in P\}, \text{ para todo } h \in [0,1], \\ PxP, \text{ para todo } h \in (1, \infty). \end{cases}$$

Este dendrograma tendría la misma tabla de disimilaridades U_c resultante que el siguiente:

$$c(h) = \begin{cases} \{(A,A) \text{ para todo } A \in P\}, \text{ para todo } h \in [0,1), \\ PxP, \text{ para todo } h \in [1, \infty), \end{cases}$$

[1] Véase "Mathematical Taxonomy", de estos autores, 1971.

si se define la función $U_c(A,B) = \inf\{h \mid (A,B) \in c(h)\}$, para todo $(A,B) \in P \times P$.

La condición *iii* asegura que los niveles de asociación en el dendrograma, son números "bien definidos". Un nivel de asociación es un elemento del conjunto $\{h \in [0,\infty) \mid h' < h \Rightarrow c(h) \neq c(h')\}$. Ya que en la práctica, P es un conjunto finito de objetos, las condiciones *i* e *ii* bastan para asegurar que también el número de niveles de asociación es finito. Para la construcción de la definición de dendrograma de este trabajo, se utilizó esta idea haciendo coincidir al conjunto I con el de niveles de asociación en los dendrogramas definidos por Jardine y Sibson. De esta manera, en el capítulo 3 se define un dendrograma como una función

$$l_c: I \to E(P),$$

donde $I = A \cup \{0\}$, con $A \subseteq \mathbb{R}^+$ e $|I| < \infty$; $E(P)$ es el conjunto de relaciones de equivalencia en P; que satisface las siguientes condiciones:

i) Existe $h \in I$ tal que $l_c(h) = P \times P$,

ii) $h < h' \Rightarrow l_c(h) \subseteq l_c(h')$ con $l_c(h) \neq l_c(h')$, h y $h' \in I$.

Lo que se hace a continuación es mostrar una función biyectiva del conjunto de dendrogramas definidos por Jardine y Sibson (JS(P)), al conjunto de dendrogramas definidos en este trabajo (D(P)).

Proposición. Sea $f: JS(P) \to D(P)$, tal que $f(c) = l_c$, con I =

$\{0\} \cup \{h \in [0, \infty) \mid h' < h \Rightarrow c(h') \neq c(h)\}$ y $l_c(h) = c(h)$ para todo $h \in I$. Sea $g: D(P) \to JS(P)$, tal que $g(l) = c_l$ de manera que $c_l(h) = l(h_0)$, donde $h_0 = \max \{h' \in I \mid h' \leq h\}$. Entonces f es una función biyectiva y g es su inversa.

Prueba:

i) Sean c y c' $\in JS(P)$, con $c \neq c'$ y sean I e I' tales que $l_c: I \to E(P)$ y $l_{c'}: I' \to E(P)$.

a) Si $I \neq I'$ entonces $f(c) = l_c \neq l_{c'} = f(c')$.

b) Si $I = I'$ entonces $c \neq c' \Rightarrow$ existe $h \in [0, \infty)$ tal que $c(h) \neq c'(h)$. Ahora bien,

1. Si $h \in I$, entonces $l_c(h) = c(h) \neq c'(h) = l_{c'}(h) \Rightarrow f(c) \neq f(c')$.

2. Si $h \notin I$, sean $h_0 = \max \{K \in I \mid K < h\}$ y $h_1 = \min\{ K \in I \mid K > h\}$, entonces por construcción de I para todo $r \in [h_0, h_1)$ se tiene que $c(r) = c(h_0)$ y $c'(r) = c'(h_0) \Rightarrow c(h) = c(h_0)$ y $c'(h) = c'(h_0) \Rightarrow l_c(h_0) = c(h_0) = c(h)$ y $l_{c'}(h_0) = c'(h_0) = c'(h) \Rightarrow l_c(h_0) \neq l_{c'}(h_0)$ $\Rightarrow f(c) \neq f(c')$.

$\therefore c \neq c' \Rightarrow f(c) \neq f(c')$, y entonces f es inyectiva.

ii) Ahora se demuestra que f es suprayectiva y su inversa es g.

Dado $l \in D(P)$, con $l: I \to E(P)$, sean $g(l) = c_l$ y $f(c_l) = l': I' \to E(P)$.

1. $c_l(0) = l(0) \Rightarrow l'(0) = c_l(0) = l(0)$.

2. Sea $h \in I$ con $h \neq 0$, y sea $h_0 = \max \{K \in I \mid K < 0\}$

entonces por definición de g y por la condición *ii* de la definición de dendrograma de este trabajo, para todo $r \in [h_0, h)$ se tiene que $c_1(r) = c_1(h_0) = l(h_0) \neq l(h) = c_1(h) \Rightarrow h \in I'$ y además $l(h) = c_1(h) = l'(h)$.

De lo visto anteriormente se tiene que $I \subseteq I'$ y que para todo $h \in I$, $l(h) = l'(h)$. Si se prueba que $I' \subseteq I$ se habrá demostrado que $l = l'$, es decir, que $f(g(l)) = l$.

3. Sea $h \in I'$ con $h \neq 0$, entonces $l'(h) = c_1(h)$ y por construcción de I' para todo $r \in [0, h)$ se tiene que $c_1(r) \neq c_1(h)$. Supóngase que $h \notin I$, entonces como $c_1 = g(l)$, $c_1(h) = l(h_0)$ con $h_0 < h$ y $h_0 \in I \Rightarrow c_1(h_0) = l(h_0) = c_1(h)$, que es una contradicción.

   $\therefore h \in I$.

$\therefore$ f es suprayectiva y g es su inversa.

De esta manera es posible asociarle a cada dendrograma definido como lo hacen Jardine y Sibson, un solo dendrograma de los correspondientes a este trabajo, y viceversa.

B.2 Definición de k-dendrograma de N. Jardine y R. Sibson. Equivalencia con la definición del capítulo 4

Estos autores definen un k-dendrograma como una función

$$c: [0, \infty) \rightarrow E_k(P),$$

donde $E_k(P)$ es el conjunto de relaciones de k-equivalencia en

P, y que satisface las siguientes condiciones:

i) Existe $h \in [0, \infty)$ tal que $c(h) = P \times P$,

ii) $c(h) \subseteq c(h')$ si $h \leq h'$,

iii) Dado h, existe $\delta > 0$ tal que $c(h + \delta) = c(h)$.

Con la misma idea que en el caso enterior se procedió a modificar esta definición para fines prácticos. Así, en el capítulo 4 se define un k-dendrograma como una función

$$l_c: I \to E_k(P),$$

donde $I = A \cup \{0\}$, con $A \subseteq \mathbb{R}^+$ e $|I| < \infty$; $E_k(P)$ es el conjunto de relaciones de k-equivalencia en P; que satisface las siguientes condiciones:

i) Existe $h \in I$ tal que $l_c(h) = P \times P$,

ii) $h < h' \Rightarrow l_c(h) \subseteq l_c(h')$ con $l_c(h) \neq l_c(h')$, h y $h' \in I$.

Es posible hacer una demostración análoga a la del caso de los dendrogramas, utilizando las mismas funciones definidas anteriormente.

BIBLIOGRAFIA

1. Anderberg, Michael,
Cluster Analysis for Applications,
Nueva York, Academic Press, 1973.

2. Espinosa, Guillermo y Arturo López,
Introducción a los Métodos Jerárquicos de Análisis de Cúmulos,
México, Universidad Nacional Autónoma de México, IIMAS,
Comunicaciones Técnicas, serie: Notas, Vol. 1, No. 9, 1977.

3. Gower, J. C. y G. J. S. Ross,
Minimum Spanning Trees and Single Linkage Cluster Analysis
Sobretiro de "Applied Statistics", 1969.

4. Hartigan, J. A.,
Clustering Algorithms,
Nueva York, Wiley, 1975.

5. Jardine, Nicholas y Robin Sibson,
Mathematical Taxonomy,
Londres, John Wiley & Sons, 1971.

6. Jardine, Nicholas y Robin Sibson,
At the Construction of Hierarchic and Non-Hierarchic Classifications,
Cambridge, Inglaterra, King's College Press, 1968.

7. Jardine, Nicholas y Robin Sibson
A Model for Taxonomy,
Sobretiro de "Mathematical Biosciences", No. 2, pp 465-482, 1968.

8. López, Arturo y Guillermo Espinosa,
Fuzzy Relations and Dendrograms,
México, Universidad Nacional Autónoma de México, CIMAS,
Comunicaciones Técnicas, serie: Investigación, Vol. 6,
No. 103, 1975.

9. López, Arturo y Guillermo Espinosa,
Escalamiento Multidimensional, Seriación y Taxonomía Numérica,
México, Universidad Nacional Autónoma de México, IIMAS,
Comunicaciones Técnicas, serie: Monografías, Vol. 3, No.
20, 1976.

10. Sierra González, Alejandro,
Algunas Relaciones entre Taxonomía Numérica y Conjuntos Borrosos,
México, Universidad Nacional Autónoma de México, Facultad de Ciencias, tesis profesional, 1979.

11. Sokal, Robert y Peter Sneath,
Principles of Numerical Taxonomy,
San Francisco, W. H. Freeman & Co., 1963.

12. Sokal, Robert y Peter Sneath,
Numerical Taxonomy,
San Francisco, W. H. Freeman & Co., 1973.