

Contribuciones al estudio del problema de la clasificación mediante grafos piramidales

Carles Capdevila i Marquès

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tesisenxarxa.net) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tesisenred.net) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tesisenxarxa.net) service has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized neither its spreading and availability from a site foreign to the TDX service. Introducing its content in a window or frame foreign to the TDX service is not authorized (framing). This rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

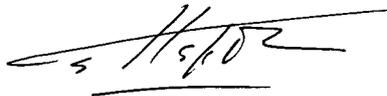
DEPARTAMENT D'ESTADÍSTICA DE LA UNIVERSITAT DE BARCELONA

**PROGRAMA DE DOCTORAT: "PROBABILITATS I ESTADÍSTICA"
BIENI: 1988 - 1990**

**CONTRIBUCIONES AL ESTUDIO DEL PROBLEMA
DE LA CLASIFICACION MEDIANTE
GRAFOS PIRAMIDALES**

Memòria presentada per en Carles Capdevila i Marquès, per optar al títol de Doctor en Matemàtiques per la Universitat de Barcelona.

El Doctorand: Carles Capdevila i Marquès.



El Tutor: Dr. Jordi Ocaña i Rebull.



Vist i Plau



El Director: Dr. Antoni Arcas Pons
Professor Titular d' Estadística i Investigació Operativa.
Departament d'Estadística. Universidad de Barcelona

Barcelona, juny de 1993

"L'information apportée par une classification se situe, en effet, au niveau sémantique: il ne s'agit pas d'atteindre un résultat vrai ou faux, probable ou improbable, mais seulement profitable ou non profitable."

Lance G.N., Williams W.T.(1965)

INDICE

PRÒLEG	5
CAPITULO 1	7
METODOS DE CLASIFICACION Y REPRESENTACION DE DATOS	
1.1.- Introducción.	
1.2.- Algunas Técnicas de Clasificación y Representación.	
1.3.- Métodos Jerárquicos de Clasificación y Representación.	
CAPITULO 2	17
DISIMILARIDADES PIRAMIDALES	
2.1.- Disimilaridades y Distancias.	
2.2.- Disimilaridades y Preordenes Compatibles.	
2.3.- Matrices de Robinson y Disimilaridades Piramidales.	
CAPITULO 3	26
REPRESENTACION MEDIANTE GRAFOS PIRAMIDALES	
3.1.- Introducción.	
3.2.- Axiomática.	
3.3.- Predecesores y Sucesores.	
3.4.- Representación Visual de las Pirámides.	
3.5.- Pirámides y Jerarquías.	
CAPITULO 4	44
PIRAMIDES INDEXADAS Y DISIMILARIDADES PIRAMIDALES	
4.1.- Introducción.	
4.2.- Teoremas de Existencia.	
4.3.- Teoremas de Unicidad.	

CAPITULO 5 **64**
ALGORITMOS PARA LA CONSTRUCCION DE GRAFOS PIRAMIDALES

- 5.1.- Introducción.
- 5.2.- Algoritmo de Clasificación Ascendente Piramidal.
- 5.3.- Depuración de una Pirámide Construida por Alguno de los Algoritmos de Clasificación Piramidal.
- 5.4.- Principales Algoritmos. Propiedades.
 - 5.4.1.- Método del Mínimo.
 - 5.4.2.- Método del Máximo.
 - 5.4.3.- Otros Métodos.
- 5.5.- Programación del Algoritmo CAP.
- 5.6.- Experimento de Simulación.

CAPITULO 6 **120**
ASPECTOS INFERENCIALES SOBRE GRAFOS PIRAMIDALES

- 6.1.- Introducción.
- 6.2.- Las Pruebas de Simulación.
- 6.3.- Test de Significación de la Bondad de Ajuste en una Clasificación Piramidal.
- 6.4.- Potencia-Eficiencia de los Métodos de Clasificación Piramidal.

CONCLUSIONES **144**

ANEXO **148**

- A.1.- Los Programas HSIMULji.SAS y HSIMULA.CLUS.
- A.2.- El Programa PIR.PRO.
- A.3.- Descripción del Programa PIR.PRO.
- A.4.- Relación Entre el Programa PIR.PRO y el Algoritmo CAP.
- A.5.- Ejecución del Programa PIR.PRO.
- A.6.- Salida de Resultados.
- A.7.- Los Programas SIMULU.PRO, SIMULN.PRO Y NSIMUL.PRO.
- A.8.- El Programa ESTADIS.PRO.
- A.9.- Disimilaridades Básicas utilizadas en las Pruebas S3 y S4.

BIBLIOGRAFIA **174**

PRÒLEG

Les primeres consideracions d'aquesta memòria han de ser, forçosament, per expressar la meua gratitud a les moltes i diverses persones que m'han ajudat i han fet possible l'execució d'aquest treball. Ja se que aixó s'acostuma a dir sempre, en ocasions com aquesta, però tots ells saben que, en aquest cas meu, de debó que ho faig ben sincerament.

Gratitut, en primer lloc, al Dr. Antoni Arcas, bon amic i company de promoció, per haver acceptat la direcció del treball que aquí presentem. Ha estat una tasca a la que plegats hem dedicat moltes hores, i que agraeixo particularment les que ell hi ha esmerçat revisant la feina, aportant idees, corregint ..., encoratjant-me en tot moment.

També agraeixo al Dr. Carles M. Cuadras, Cap del Departament d'Estadística de l'Universitat de Barcelona, y a la resta de membres del Departament, la bona acollida que des de bon principi em van dispensar i l'haver-me permés de dur a terme el meu treball en el si del seu Departament, d'aquesta manera la feina se m'ha fet sempre molt mes agradable i en tot moment m'hi he trobat com a casa.

Al Dr. Miquel Salicrú, pels seus assenyats consells i per la col.laboració a la que en tot moment ha estat disposat, fruit de la seua llarga experiència en la línia de recerca.

Al Dr. Tomàs Aluja, de l'Universitat Politècnica de Catalunya, gràcies al qual vaig conèixer els primers articles sobre arbres piramidals, els quals han estat finalment l'objecte d'aquest treball. A ell li dec també la facilitat amb que vaig establir els primers contactes amb el Dr. Bernard Fichet a Marseille.

Als Professors Edwin Diday de l'Universitat de París i de l'Institut National de Recherche en Informatique et en Automatique (INRIA) i Bernard Fichet de l'Universitat d'Aix-Provence, autors dels primers treballs sobre classificacions piramidals, pel seu recolsament i estímul a la nostra tasca i per les seves interpellacions i orientacions que ens han ajudat tothora a concretar el contingut del nostre treball.

Gràcies, al també bon amic i company, Angel Gil de l'Universitat Pompeu Fabra i expert en Prolog, per les seues inestimables aportacions a l'hora de programar l'Algorisme de Clasificació Piramidal.

Agraeixo també a l'amic Pere Solà, professor de filologia francesa de la nostra Universitat, la seua col.laboració en la correcta traducció de tantes cartes i comunicacions.

Y als companys de l'Escola Politècnica d'Informàtica de la UdL, pel seu suport i la paciència que han tingut d'aguantar les moltes hores que m'he passat als seus despatxos, fent treballar els seus ordinadors i compartint els nostres maldecaps.

Finalmet desitjaria que la Susanna, la meua dona, i els meus fills, el Roger i la Blanca no em tinguessin massa en compte el temps, que ja no tornarà, i que he deixat de dedicar-los-hi, els agraeixo de tot cor la seua comprensió, que sempre ha estat molt més generosa del que segurament em mereixia.

Lleida, juny de 1993

CAPITULO 1

METODOS DE CLASIFICACION Y REPRESENTACION DE DATOS

1.1.- INTRODUCCION

En muchos campos de la actividad científica es indispensable, para cualquier estudio o análisis posterior, el disponer de la máxima información posible. Esta información original viene dada, principalmente, por grandes tablas de datos que posteriormente deberán ser analizadas. Por lo general, estas tablas proceden de la observación de diferentes características (variables) sobre un determinado conjunto de individuos, objetos o subpoblaciones de una población Ω , conocida. Estas tablas permiten, por lo tanto, la descripción de un cierto número de objetos a partir de un número reducido de variables.

Si $\Omega = \{\omega_1, \dots, \omega_n\}$ y X_1, \dots, X_r son las variables a través de las cuales obtenemos la tabla de datos, ésta puede expresarse matricialmente de la forma siguiente:

$$X \equiv \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{1r} \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{nr} \end{bmatrix}$$

donde x_{ij} es el valor observado de la variable X_j sobre el individuo ω_i ; es decir, $\forall j=1, \dots, r$:

$$\begin{aligned} X_j : \Omega &\longrightarrow \mathbf{R} \\ \omega_i &\longrightarrow X_j(\omega_i) = x_{ij} \end{aligned}$$

Las tablas de contingencia, cruzando dos particiones de una misma población, o bien las tablas de presencia-ausencia, son otras posibilidades.

"Si fuese necesario definir de alguna forma el Análisis de Datos, podríamos decir que es un conjunto de métodos que tienen como objetivo esencial el poner de manifiesto las relaciones existentes entre los individuos, las variables que los caracterizan y entre los individuos y las variables" (Diday et al. 1982).

Así pues, en el proceso de análisis de los datos proporcionados por la matriz X , parece evidente que una serie de análisis univariantes efectuados por separado, resultarían absolutamente inadecuados, puesto que ignorarían la correlación existente entre las variables. En general, para el tratamiento de grandes tablas de datos, procedentes de observaciones multidimensionales, son las técnicas del Análisis Multivariante las más adecuadas, puesto que permiten analizar, describir e interpretar globalmente dichas tablas. Todo ello ha propiciado su utilización en gran número de estudios estadísticos relacionados con muchos y muy diversos campos de la actividad científica: Economía, Medicina, Agricultura, Biología,....

En este mismo contexto, cabe destacar también los grandes avances que en los últimos años han efectuado los medios informáticos, contribuyendo éstos, en gran manera, al desarrollo creciente de las técnicas del Análisis de Datos Multidimensionales.

1.2.- ALGUNAS TECNICAS DE CLASIFICACION Y REPRESENTACION

De entre los métodos más importantes del Análisis Multivariante, centraremos nuestra atención en aquellos que permiten una clasificación de los individuos de la población y la representación (tanto de los individuos como de los grupos formados en el proceso de clasificación) sobre un determinado espacio geométrico.

En líneas generales, los métodos de clasificación consisten en agrupar los individuos de una determinada población en base a algún criterio previamente establecido con el fin de formar grupos "razonablemente homogéneos". En general distinguiremos entre métodos jerárquicos y no jerárquicos:

Métodos Jerárquicos: Proporcionan una familia de clasificaciones, cada vez menos finas de los individuos de la población. Para ello, los grupos se van fusionando progresivamente hasta obtener Ω , con lo cual, conforme avanza el proceso, los grupos son cada vez menos homogéneos.

Métodos No Jerárquicos: Proporcionan una **única clasificación** de los individuos de la población, sin establecer relaciones entre los grupos formados.

Por lo que a los métodos de representación se refiere, deberemos disponer de una disimilaridad δ , sobre Ω , que cuantifique las diferencias entre los individuos, y de un espacio geométrico E sobre el que efectuaremos la representación y que deberá estar provisto de una distancia d .

Nótese que la disimilaridad definida sobre Ω , puede construirse a partir de la matriz de datos X , con lo cual δ dependerá tanto de los individuos en sí, como de la información recibida sobre los mismos, es decir, de las variables escogidas para su descripción y del método escogido para su cálculo, del cual dependerá, en gran parte, el resultado final del análisis.

El problema entonces de representar los individuos de una población (Ω, δ) sobre un espacio métrico (E, d) , llamado también espacio modelo, consistirá en establecer una correspondencia entre los individuos de Ω y los puntos del espacio E , de manera que la distancia entre dos puntos sea lo más próxima posible a la disimilaridad entre los individuos correspondientes, es decir se tratará de determinar una aplicación

$$\Psi: (\Omega, \delta) \longrightarrow (E, d)$$

de manera que $\forall \omega_i, \omega_j \in \Omega$, $d(\Psi(\omega_i), \Psi(\omega_j))$ sea lo más próxima posible a $\delta(\omega_i, \omega_j)$, según algún criterio de ajuste.

Si es posible encontrar una tal aplicación, diremos que se trata de una representación de los individuos de Ω .

Según el espacio geométrico considerado, pueden distinguirse dos grandes grupos de representaciones: Modelos espaciales o continuos y modelos de redes o discretos.

Los modelos de representación espaciales consisten en representar cada individuo mediante un punto, en un espacio euclídeo, de manera que las distancias entre los mismos reflejen las relaciones de proximidad observadas; mientras que en los modelos discretos, cada individuo viene representado por un nudo en un grafo conexo, de modo que las relaciones entre los nudos del grafo reflejen la disimilaridad observada entre los individuos.

Las principales técnicas relativas a modelos continuos pertenecen al Multidimensional Scaling (MDS), definido por De Leeuw (1982) como aquellos métodos de representación en los que $E = \mathbb{R}^n$. Cuadras (1981), profundizando más en la naturaleza de los datos, considera dentro del MDS aquellas técnicas de representación que tienen por objeto la construcción de una configuración de puntos a partir de una determinada información sobre las disimilaridades entre los individuos.

La Figura 1.2.1 ilustra una representación MDS de un conjunto de 10 ciudades Norte Americanas, según un estudio realizado por Forrest W. Young (1981).

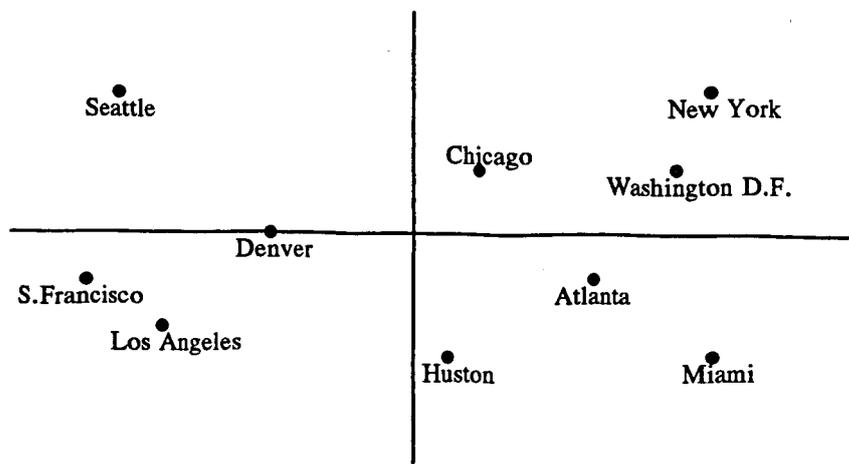


Fig.1.2.1

Por lo que respecta a los modelos discretos, los más usuales son los siguientes:

Arboles Aditivos: E es un grafo conexo, sin ciclos y d una distancia aditiva.

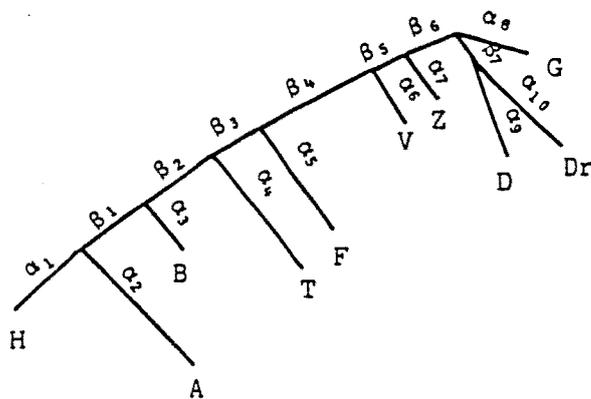
Arboles Ultramétricos: E es un grafo conexo, sin ciclos, con un nudo denominado raíz equidistante de todos los extremos y d una distancia ultramétrica.

Arboles Piramidales: E es un grafo conexo y d una distancia piramidal.

Consideremos algunos ejemplos, a partir de datos reales, de representaciones mediante los modelos citados:

Salicrú (1983) en un estudio sobre la *Drosophila Suboscura* demuestra, utilizando técnicas de taxonomía numérica, que las barreras geográficas son también barreras genéticas, tomando como distancia genética la propuesta por Prevosti (1974), la cual cuantifica la diferencia entre dos poblaciones a partir de las frecuencias cromosómicas producidas por los individuos.

Arcas (1986) partiendo de la misma distancia realiza el estudio mediante árboles aditivos, considerando las poblaciones de Huelva (H), Agadir (A), Barcelona (B), Thessaloniki (T), Fruska Gora (F), Viena (V), Zurich (Z), Drobak (Dr), Dalkeith (D), y Groningen (G) y utilizando el algoritmo ADDTREE, obtiene el árbol aditivo de la Fig.1.2.3.



$\alpha_1 = 0.134$	$\alpha_{10} = 0.150$
$\alpha_2 = 0.293$	$\beta_1 = 0.090$
$\alpha_3 = 0.073$	$\beta_2 = 0.104$
$\alpha_4 = 0.185$	$\beta_3 = 0.050$
$\alpha_5 = 0.152$	$\beta_4 = 0.151$
$\alpha_6 = 0.075$	$\beta_5 = 0.028$
$\alpha_7 = 0.056$	$\beta_6 = 0.042$
$\alpha_8 = 0.076$	$\beta_7 = 0.003$
$\alpha_9 = 0.120$	

Coefficiente correlación entre la distancia original y la distancia aditiva = 0.9965.

Fig.1.2.3

En Francia se realizó un estudio sobre la incidencia que había tenido el accidente nuclear de Tchernobyl en la opinión pública. Para ello se encuestó a una población de 311 personas, cada individuo contestó dos cuestionarios, uno sobre su situación personal y otro en el que se proponía una serie de afirmaciones, relacionadas con el accidente, a las que cada persona encuestada podía responder de una de las cinco maneras posibles:

- | | |
|------------------------------|------------------------|
| 1. Totalmente en desacuerdo. | 4. De acuerdo |
| 2. En desacuerdo | 5. Del todo de acuerdo |
| 3. Podría estar de acuerdo | |

Las afirmaciones fueron las siguientes:

- (1) DIFTF: Las centrales nucleares francesas son muy distintas de Tchernobyl.
- (2) HABI: Durante mucho tiempo no se podrá habitar a menos de 20 Km. de Tchernobyl.
- (3) RADIOF: Después de Tchernobyl, en Francia se está menos de acuerdo con la energía nuclear.
- (4) ACCIDF: En Francia podría suceder un accidente tan grave como el de Tchernobyl.
- (5) INTERT: Lo que ha pasado en Tchernobyl no me interesa.
- (6) DRAMAF: Se han dramatizado las consecuencias sobre Francia del accidente.
- (7) EXPERT: Las explicaciones de los expertos sobre las consecuencias para Francia del accidente han sido incomprensibles para los ciudadanos.
- (8) FERMEF: Después del accidente es necesario cerrar todas las centrales nucleares francesas.

Se consideraron, además dos variables suplementarias:

- (9) ACC: Nivel de acuerdo con las afirmaciones.
- (10) ENG: Nivel de seguridad en las respuestas.

Con los datos de esta encuesta, Durand (1989) pretende realizar una clasificación piramidal de las variables con el fin de poner de manifiesto algunas relaciones entre las mismas. Para ello, a partir de las correlaciones entre las variables, se calcula una matriz de disimilaridades entre las mismas, mediante la transformación $\delta = 1 - |r|$, de manera que dos variables serán próximas si su correlación es alta. Finalmente, a partir de esta disimilaridad, calcula la ultramétrica y la piramidal subdominantes correspondientes y obtiene las representaciones de la Figura 1.2.2

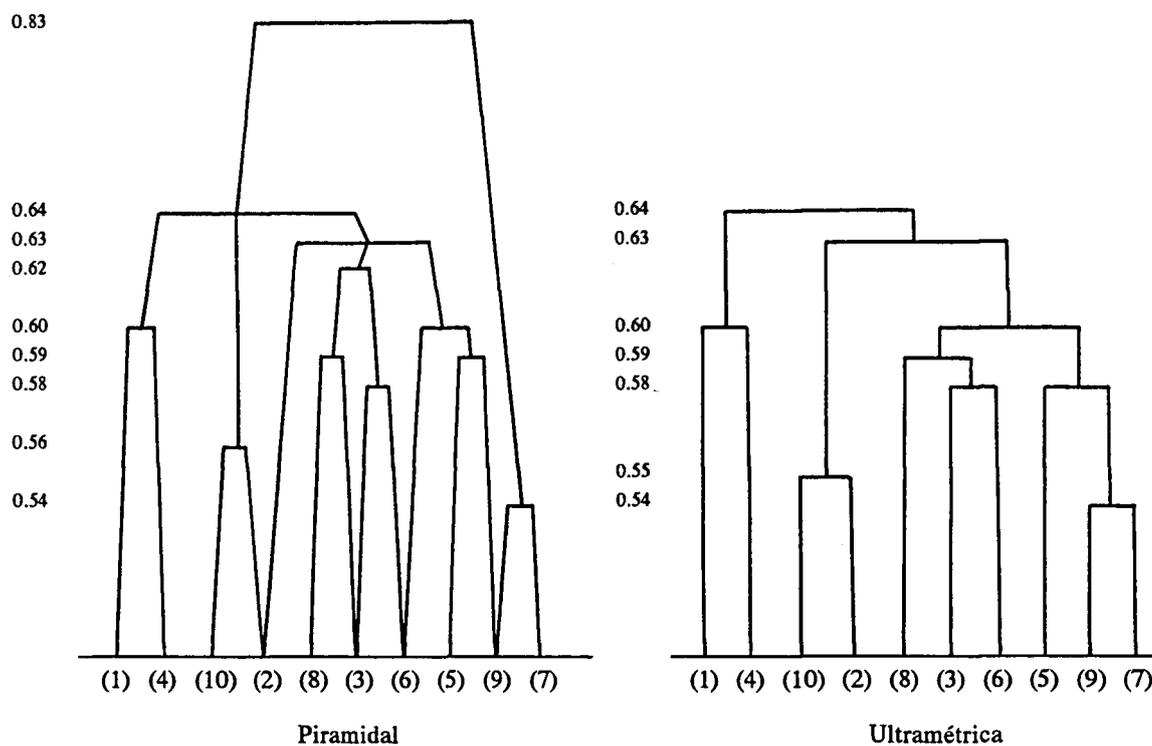


Fig.1.2.2

1.3.- METODOS JERARQUICOS DE CLASIFICACION Y REPRESENTACION

Como es bien conocido, los árboles ultramétricos constituyen el método jerárquico más claro y profusamente utilizado de clasificación y representación de datos, los cuales proporcionan una familia de Clasificaciones (Particiones) de los individuos de una determinada población a diferentes niveles jerárquicos.

Los grafos piramidales, introducidos por E.Diday (1984), pretenden una generalización de los árboles y las clasificaciones ultramétricas mediante estructuras menos restrictivas considerando para ello recubrimientos en lugar de particiones. De

este modo, a cada nivel, los grupos en que queda dividida la población no son disjuntos o encajados, como en el caso ultramétrico, sino que pueden ser también solapados.

En este trabajo pretendemos dar una visión global y rigurosa, desde el punto de vista de la formalización matemática, de las clasificaciones mediante grafos piramidales. Para ello centraremos nuestra atención, fundamentalmente, en tres aspectos:

- * Axiomatización y formalización de las bases matemáticas sobre las que se amparan las Clasificaciones Piramidales (Capítulos 3 y 4).

- * Análisis, puesta a punto y programación informática de algunos algoritmos de clasificación y representación piramidal. En el capítulo 5 se dan algunos resultados interesantes en este sentido.

- * Finalmente, la programación informática de los algoritmos descritos en el capítulo 5, nos permitirá, a través de algunas técnicas de simulación mediante métodos de Montecarlo, abordar algunos problemas inferenciales asociados a los grafos piramidales. En este sentido pues, se muestran en el capítulo 6 algunos resultados relativos a la potencia-eficiencia de los métodos del mínimo y del máximo, así como algún criterio sobre la bondad de una clasificación piramidal, bajo ciertas condiciones que se especificarán.

CAPITULO 2

DISIMILARIDADES DISTANCIAS Y PREORDENES

Dada la importancia de las disimilaridades y las distancias en las técnicas de representación de datos, será conveniente la introducción de algunos resultados al respecto. Prestaremos una especial atención al concepto de disimilaridad piramidal o también llamada de Robinson, puesto que fue este científico quien la introdujo en 1951, para modelizar cronológicamente una serie de datos obtenidos en un estudio efectuado sobre siete yacimientos arqueológicos. Con estas disimilaridades, se intentaba modelizar un problema de seriación cronológica de los datos, a partir de la cual poder ordenar los yacimientos de manera que este orden reflejase, lo mejor posible, la evolución cronológica de los mismos.

La noción de orden será también fundamental en el estudio de las representaciones asociadas a las disimilaridades piramidales. Para ello será conveniente introducir también el concepto de disimilaridad compatible con un cierto preorden.

2.1.- DISIMILARIDADES Y DISTANCIAS

Definición 2.1.1.

Dado un conjunto finito Ω , diremos que una aplicación d , de $\Omega \times \Omega$ en \mathbb{R} es una **disimilaridad** sobre Ω si verifica las siguientes condiciones:

$$\begin{array}{ll} \text{D.1.- } d(x,y) \geq 0 & \forall x,y \in \Omega \\ \text{D.2.- } d(x,x) = 0 & \forall x \in \Omega \\ \text{D.3.- } d(x,y) = d(y,x) & \forall x,y \in \Omega \end{array}$$

Definición 2.1.2.

d es una **disimilaridad métrica** sobre Ω si verifica las condiciones D.1, D.2, D.3 y la desigualdad triangular:

$$\text{D.4.- } d(x,z) \leq d(x,y) + d(y,z) \quad \forall x,y,z \in \Omega$$

Definición 2.1.3.

d es una **disimilaridad definida**, si verifica las condiciones D.1, D.2, D.3 y

$$\text{D.5.- } d(x,y) = 0 \Leftrightarrow x = y \quad \forall x,y \in \Omega$$

Definición 2.1.4.

d es una **distancia** sobre Ω , si es una disimilaridad métrica y definida.

Definición 2.1.5.

Una disimilaridad d , es **ultramétrica**, si cumple las condiciones D.1, D.2, D.3 y la desigualdad ultramétrica

$$\text{D.6.- } d(x,y) \leq \max\{d(x,z), d(y,z)\} \quad \forall x,y,z \in \Omega$$

Definición 2.1.6.

Una disimilaridad d , es **aditiva**, si cumple las condiciones D.1, D.2, D.3 y el axioma del cuarto punto :

$$\text{D.7.- Para cualesquiera } x,y,z,t, \text{ elementos de } \Omega, \\ d(x,y) + d(z,t) \leq \max\{d(x,z) + d(y,t), d(x,t) + d(y,z)\}$$

Definición 2.1.7.

Una disimilaridad d sobre Ω es **Piramidal**, si además de las condiciones D.1, D.2 y D.3, cumple:

$$\text{D.8.- Existe un preorden total } (\leq) \text{ sobre } \Omega, \text{ tal que para todo } \\ x,y,z \in \Omega, \text{ con } x \leq y \leq z, d(x,z) \geq \max\{d(x,y), d(y,z)\}$$

A partir de estas definiciones puede comprobarse que:

- * Toda disimilaridad ultramétrica es también aditiva.
- * Toda disimilaridad aditiva es también métrica.

La demostración puede encontrarse en Arcas y Cuadras (1987).

A continuación veremos como las disimilaridades ultramétricas son un caso particular de las piramidales. Para ello hagamos algunas consideraciones previas.

Si d es una disimilaridad sobre un conjunto finito Ω , cualquier ordenación total (\leq) de sus elementos, dará lugar a una matriz de disimilaridades para d , en la que la i -ésima columna (fila) representa al i -ésimo elemento de Ω (respecto a la ordenación dada). Esta matriz será representada por $M(d, \leq)$.

Recíprocamente, toda matriz $D=(d_{ij})$ correspondiente a una disimilaridad d sobre un conjunto finito $\Omega = \{\omega_1, \dots, \omega_n\}$, en la que $\forall i,j \in \{1,2, \dots, n\}$, $d_{ij} = d(\omega_i, \omega_j)$,

induce un orden total sobre Ω sin más que asociar a cada elemento la fila o columna correspondiente.

Proposición 2.1.1.

Una matriz $(n \times n)$ simétrica de números reales positivos (d_{ij}) , es una matriz asociada a una disimilaridad ultramétrica d , sobre un conjunto $\Omega = \{\omega_1, \dots, \omega_n\}$, si existe una ordenación total de los individuos $(\omega_1 \leq \dots \leq \omega_n)$, de modo que la matriz $M(d, \leq)$ verifique:

a) $d_{ij} \leq d_{ij+1}$ para todo $i \leq j$.

b) Para todo $k \in \{1, \dots, n\}$, si $d_{kk+1} = \dots = d_{kk+s+1} < d_{kk+s+2}$, entonces:

$$d_{k+lj} \leq d_{kj} \quad \text{para } k+1 < j \leq k+s+1$$

$$d_{k+lj} = d_{kj} \quad \text{para } j > k+s+1$$

La demostración puede encontrarse en Lerman (1981).

Es interesante observar la interpretación visual de esta proposición. En esta línea, si d es una disimilaridad ultramétrica sobre Ω , existe una ordenación total de sus elementos \leq , de modo que la matriz $M(d, \leq)$ verifica:

a) A partir de cada término de la diagonal principal, que vale cero, los términos de una misma fila están ordenados de forma creciente (no estrictamente).

b) Considerando solamente la parte triangular superior, los términos de cada columna van creciendo, también a partir de cada término de la diagonal, teniendo en cuenta además que si $d_{ij} < d_{ij+1}$, los términos de la fila $i+1$ son iguales a los de la fila i a partir de la columna $j+1$:

$$\begin{array}{cccccccccccc}
 & \dots & \omega_k & \omega_{k+1} & \omega_{k+2} & \dots & \dots & \dots & \omega_{k+s+1} & \omega_{k+s+2} & \dots & \dots & \dots & \omega_n \\
 \hline
 & \dots \\
 \omega_k & & 0 & d_{kk+1} & = & d_{kk+2} & = & \dots & = & d_{kk+s+1} & < & d_{kk+s+2} & \dots & \dots & d_{kn} \\
 & & & \vee & & \vee & & & & \vee & & \parallel & & & \parallel \\
 \omega_{k+1} & & & 0 & \underbrace{d_{k+1k+2} \dots \dots \dots d_{k+1k+s+1}}_{k+1 < j \leq k+s+1} & & \underbrace{d_{k+1k+s+2} \dots \dots \dots d_{k+1n}}_{j > k+s+1} & & & & & & & &
 \end{array}$$

Concretamente, si d es la disimilaridad ultramétrica, sobre el conjunto

$\Omega = \{\omega_1, \dots, \omega_5\}$, dada por la matriz:

	ω_1	ω_2	ω_3	ω_4	ω_5
	0	3	7	9	2
		0	7	9	3
			0	7	9
				0	9
					0

Existe una ordenación total de los individuos de Ω , tal que la matriz $M(d, \leq)$

satisface las condiciones de la proposición:

	ω_2	ω_1	ω_5	ω_3	ω_4
	0	3	3	7	9
		0	2	7	9
			0	7	9
				0	9
					0

Proposición 2.1.2.

Toda disimilaridad ultramétrica, es también piramidal.

Demostración:

Si d es una disimilaridad ultramétrica, podemos considerar la ordenación de Ω establecida por la proposición 2.1.1.

Sean $\omega_i, \omega_j, \omega_k \in \Omega$ tales que $\omega_i \leq \omega_j \leq \omega_k$. Las condiciones a) y b) de la proposición

anterior implican, respectivamente:

$$d(\omega_i, \omega_j) \leq d(\omega_i, \omega_k) \text{ y } d(\omega_j, \omega_k) \leq d(\omega_i, \omega_k)$$

Por tanto, de ambas desigualdades obtendremos:

$$d(\omega_i, \omega_k) \geq \max\{d(\omega_i, \omega_j), d(\omega_j, \omega_k)\}$$

Así pues d es piramidal.

2.2.-DISIMILARIDADES Y PREORDENES COMPATIBLES

Definición 2.2.1.

Dada una disimilaridad d , y un preorden \leq , definidos sobre un conjunto finito Ω , diremos que d y \leq son **Compatibles**, si y solamente si:

para cualesquiera elementos de Ω , tales que $\omega_i \leq \omega_j \leq \omega_k$,
se verifica: $d(\omega_i, \omega_k) \geq \max\{d(\omega_i, \omega_j), d(\omega_j, \omega_k)\}$

Definición 2.2.2.

Sea Ω un conjunto finito preordenado y $h \in \mathcal{P}(\Omega)$. Diremos que h es una **parte conexa de Ω** respecto al preorden, sii

$$\forall \omega_i, \omega_j \in h, \text{ con } \omega_i \leq \omega_j, \{\omega \in \Omega / \omega_i \leq \omega \leq \omega_j\} \subset h$$

Definición 2.2.3.

Diremos que un preorden definido sobre Ω es **compatible con una familia $A \subset \mathcal{P}(\Omega)$** , sii cualquier $h \in A$, es una parte conexa respecto al preorden.

2.3.- MATRICES DE ROBINSON Y DISIMILARIDADES PIRAMIDALES

En este apartado daremos una caracterización de las matrices asociadas a las disimilaridades piramidales.

Definición 2.3.1.

Diremos que una matriz cuadrada de orden n , $D=(d_{ij})$ es una matriz de **Robinson**, sii para cualesquiera $i,j \in \{1,2,\dots,n\}$: $d_{ij} \geq 0$; $d_{ii}=0$; $d_{ij}=d_{ji}$; y considerando únicamente la parte triangular superior, los términos de las filas y las columnas van creciendo a partir de cada término de la diagonal principal.

Proposición 2.3.1.

Toda matriz ultramétrica es también de Robinson.

La demostración es inmediata teniendo en cuenta la caracterización de las matrices ultramétricas dada en la proposición 2.1.1 y la definición de matriz de Robinson.

El recíproco no es cierto, como puede verse en el ejemplo que sigue:

Consideremos las matrices A y B siguientes:

$$A \equiv \begin{bmatrix} 0 & 3 & 3 & 7 & 9 \\ & 0 & 2 & 5 & 6 \\ & & 0 & 1 & 2 \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \quad B \equiv \begin{bmatrix} 0 & 3 & 3 & 7 & 9 \\ & 0 & 2 & 7 & 9 \\ & & 0 & 7 & 9 \\ & & & 0 & 9 \\ & & & & 0 \end{bmatrix}$$

Observese que la matriz A es de Robinson y sin embargo no es ultramétrica ($a_{13} < a_{14}$ y $a_{24} \neq a_{14}$), en cambio la matriz B es ultramétrica y también de Robinson.

Proposición 2.3.2.

Si (Ω, \leq) es un conjunto finito, totalmente preordenado y d , una disimilaridad sobre Ω , las siguientes propiedades son equivalentes:

- a) d es compatible con \leq .
- b) $M(d, \leq)$ es una matriz de Robinson.
- c) Para todo $\omega, \omega' \in \Omega$ tales que existan $\omega_i, \omega_j \in \Omega$, con $\omega_i \leq \omega \leq \omega' \leq \omega_j$, se cumple $d(\omega, \omega') \leq d(\omega_i, \omega_j)$.

Demostración:

a) \Rightarrow b)

Si suponemos d compatible con \leq , podemos considerar una terna de elementos de Ω tales que $\omega_i \leq \omega_j \leq \omega_k$ y tendremos $d(\omega_i, \omega_k) \geq \max\{d(\omega_i, \omega_j), d(\omega_j, \omega_k)\}$.

En particular, puesto que el preorden sobre Ω es total, para cualquier $\omega_i \in \Omega$, podemos considerar $j=i+1$, $k=i+2$, y tendremos:

$$d(\omega_i, \omega_{i+1}) \leq d(\omega_i, \omega_{i+2}) \quad \text{y} \quad d(\omega_{i+1}, \omega_{i+2}) \leq d(\omega_i, \omega_{i+2})$$

Así pues, si consideramos la matriz $M(d, \leq)$, la primera desigualdad nos indica que los términos de la fila i -ésima van creciendo a partir del término de la diagonal ($d_{i+1} \leq d_{i+2} \leq \dots \leq d_{in}$). Mientras que la segunda desigualdad indicará lo mismo para la columna $(i+2)$ -ésima ($d_{i+1i+2} \leq d_{ii+2} \leq \dots \leq d_{li+2}$).

Así pues, la matriz $M(d, \leq)$ es de Robinson.

b) \Rightarrow c)

Supongamos $\omega_i \leq \omega \leq \omega' \leq \omega_j$, siendo \leq el preorden total sobre Ω inducido por la matriz $M(d, \leq)$, -que será el inicial por construcción de $M(d, \leq)$ -.

Al ser la matriz $M(d, \leq)$ de Robinson, tendremos: $d(\omega_i, \omega') \leq d(\omega_i, \omega_j)$ y $d(\omega_i, \omega') \geq d(\omega, \omega')$.

Combinando ambas desigualdades, resultará: $d(\omega_i, \omega_j) \geq d(\omega, \omega')$, que es lo que queríamos demostrar.

c) \Rightarrow a)

Sean $\omega_r, \omega_s, \omega_t$, elementos cualesquiera de Ω tales que $\omega_r \leq \omega_s \leq \omega_t$.

Aplicando la hipótesis a $\omega_r = \omega_r \leq \omega_s \leq \omega_t$: $d(\omega_r, \omega_t) \geq d(\omega_r, \omega_s)$,

y aplicándola a $\omega_r \leq \omega_s \leq \omega_t = \omega_t$: $d(\omega_r, \omega_t) \geq d(\omega_s, \omega_t)$.

De ambas desigualdades concluimos que si $\omega_r \leq \omega_s \leq \omega_t$, entonces: $d(\omega_r, \omega_t) \geq \max\{d(\omega_r, \omega_s), d(\omega_s, \omega_t)\}$ y por tanto la disimilaridad d es compatible con el preorden.

Proposición 2.3.3.

d es una disimilaridad piramidal sobre Ω , sii existe un preorden total sobre Ω compatible con d .

La proposición es evidente a partir de las definiciones.

Como consecuencia de las proposiciones 2.3.2 y 2.3.3 tendremos:

i) Una disimilaridad es piramidal sii existe un preorden total, \leq compatible con d . Esto equivale a que la matriz $M(d, \leq)$ sea Robinson.

Así pues, las matrices asociadas a las disimilaridades PIRAMIDALES son precisamente las matrices de ROBINSON.

ii) Una disimilaridad es piramidal sii para cualesquiera $\omega, \omega' \in \Omega$ tales que existan $\omega_1, \omega_j \in \Omega$ con $\omega_1 \leq \omega \leq \omega' \leq \omega_j$, se cumple $d(\omega, \omega') \leq d(\omega_1, \omega_j)$.

CAPITULO 3

REPRESENTACION MEDIANTE GRAFOS PIRAMIDALES

3.1.- INTRODUCCION

De entre todas las técnicas de representación discretas, la más conocida y utilizada es, sin lugar a dudas, la representación mediante árboles ultramétricos.

El criterio general para una tal representación es la deformación de una disimilaridad inicial, dada sobre un conjunto finito Ω , hasta obtener una disimilaridad ultramétrica, que ajuste a la primera según un determinado criterio. Estas técnicas se conocen con el nombre de Taxonomía Matemática.

Por otra parte, la obtención de una disimilaridad ultramétrica es equivalente a la de una Clasificación Jerárquica de los individuos de Ω (J.P. Benzecri, 1973).

Finalmente, la representación gráfica de una tal clasificación, se efectúa a través de los denominados árboles ultramétricos o dendrogramas, constituyendo estos la representación visual de las disimilaridades ultramétricas.

Las clasificaciones jerárquicas pues, ya desde Aristóteles, con su "Arbol de la Vida" y C. Lineo con el "Sistema Natural", hasta los trabajos de Sokal y Sneath (1973) de J.P. Benzecri (1973) y I.C. Lerman (1970), pasando por Adanson (1757) con su "algoritmo de construcción jerárquica" para agrupar plantas, han sido un instrumento de gran utilidad para la representación de datos multidimensionales.

Estas clasificaciones, no obstante, conllevan la aceptación de algunas limitaciones importantes, en pro, generalmente, de una mayor simplicidad de la representación.

Los métodos de clasificación tienen, todos ellos, en común, el proporcionar un conjunto de técnicas y de algoritmos que permitan dividir a los individuos de una determinada población en grupos, más o menos homogéneos e interrelacionados, teniendo en cuenta las variables que los caracterizan.

En el caso de las clasificaciones jerárquicas, estos grupos se obtienen a través de particiones sucesivas, organizadas a diferentes niveles jerárquicos. Este hecho peculiar, propio de este tipo de clasificaciones, implica, entre otras cosas, que la distancia ultramétrica entre elementos de clases distintas, a cualquier nivel, es la misma. Así pues, si un individuo de una clase es próximo a los de otra clase distinta, los restantes individuos de la primera, habrán de ser igualmente próximos a los de la segunda, la cual cosa, en algunos casos por lo menos, supone forzar demasiado la realidad.

A pesar de esta fuerte restricción, las clasificaciones jerárquicas continúan teniendo un gran interés en la resolución de problemas prácticos de taxonomía, fundamentalmente por la facilidad con que pueden interpretarse los dendrogramas y porque, la clasificación proporciona, a cada nivel, una partición de la población. A pesar de ello, han habido numerosos intentos para hallar métodos alternativos menos

restrictivos. Los árboles aditivos son quizá uno de los ejemplos más generalizados.

No obstante, en los últimos años, debido al vertiginoso desarrollo de los medios informáticos, numerosos autores se han planteado el problema de sustituir las particiones, que conducen a las clasificaciones jerárquicas, por recubrimientos, es decir han intentado introducir estructuras en las que los grupos no sean, solamente, disjuntos o encajados como en el caso de las jerarquías, sino que puedan existir, además, grupos "solapados". De esta forma, un individuo no será descrito solamente por las características de una única clase, como en el caso ultramétrico, sino por las de todos los grupos a los que pertenezca. Así pues, este nuevo modelo de representación deberá ser más próximo a la realidad que el modelo jerárquico. Para comprender tal afirmación, basta tener en cuenta que la descripción de nuestro entorno natural no genera un método de clasificación que caracterice a cada categoría únicamente por sus propias características, de manera que ninguna de ellas sea común a dos grupos distintos, antes al contrario, siempre existen géneros intermedios con características comunes a dos categorías, por ejemplo los zoofitos, a caballo entre el mundo animal y el vegetal, o bien una determinada característica genética creada por la interacción de dos factores hereditarios que coexisten en un mismo individuo.

En el ejemplo de la Fig. 1.2.2, en el que se efectúa una clasificación ultramétrica y otra piramidal de 10 variables relacionadas con el accidente nuclear de Tchernobyl, si cortamos los grafos a un determinado nivel, obtendremos ciertos grupos de variables. Si deseamos obtener, por ejemplo, tres o cuatro grupos, en el caso ultramétrico obtendremos la partición:

[DITF, ACCIDF] [ENG, HABI] [FERMF, RADIOF, DRAMAF, INTERT, ACC, EXPERT]

Es evidente que las variables de cada grupo están más correlacionadas entre sí, que las de dos grupos distintos, pero en este caso no podemos saber, por ejemplo, si HABI está más correlacionada con ACCIDF que con FERMEF o con EXPERT ...

En el caso piramidal, la clasificación en cuatro grupos, pone de manifiesto las relaciones entre los mismos:

[DIFTF,ACCIDF] [ENG, HABI] [HABI, FERMF,RADIOF,DRAMAF,INTERT, ACC] [ACC, EXPERT]

Observando la clasificación obtenida, podemos decir, por ejemplo, que HABI está más correlacionada con FERMEF que con ACCIDF; que ACC está más correlacionada con EXPERT que con ACCIDF...

Como modelos alternativos a las clasificaciones por particiones, desde el punto de vista clásico de la taxonomía, podemos citar básicamente, los que se han desarrollado:

- a) A partir de la teoría de grafos. Hubert(1974), B. Monjardet (1980).
- b) A partir de la teoría de la probabilidad. Hartigan (1975-1977), Diday et col.(1979).
- c) A partir del M.D.S. por optimización de algún criterio. Shepart et Arabie (1979), Arabie et Carol (1980); o bien por el procedimiento de las nubes dinámicas, Diday (1982).
- d) A partir de la utilización de técnicas que permitan mejorar las jerarquías obtenidas por el método del mínimo, sujetas a efectos de cadena. Jardine et Sibson (1971).

Diday, en 1984, introduce por primera vez el concepto de grafo piramidal, en un intento de dar una representación visual de las disimilaridades de Robinson y establece el concepto de Pirámide Indexada en Sentido Amplio. Paralelamente B. Fichet impone restricciones suplementarias que conducen al concepto de Pirámide Indexada en Sentido Estricto.

Si, tal como ya hemos expuesto en el Capítulo 1, pretendemos dar una visión rigurosa, desde el punto de vista de la formalización matemática, de los métodos de representación piramidal, empezaremos, en el presente capítulo, introduciendo las bases axiomáticas relativas a dichos métodos, a partir de las cuales mostraremos como las jerarquías pueden ser generalizadas mediante otros conjuntos menos restrictivos, dando lugar a otro tipo de representaciones, que llamaremos **piramidales**, más ricas en información y más próximas a los datos iniciales, considerando para ello, **recubrimientos** sucesivos en lugar de particiones de la población Ω .

3.2.- AXIOMATICA

Definición 3.2.1.

Sea Ω un conjunto finito y $\mathcal{P} \subset \mathcal{O}(\Omega)$. Diremos que \mathcal{P} es una **pirámide**, si se verifican las condiciones siguientes:

- P.1.- $\Omega \in \mathcal{P}$.
- P.2.- $\forall \omega \in \Omega, \{\omega\} \in \mathcal{P}$.
- P.3.- $\forall h, h' \in \mathcal{P}, h \cap h' = \emptyset$ ó $h \cap h' \in \mathcal{P}$.
- P.4.- Existe un preorden total sobre Ω , compatible con \mathcal{P} .

Definición 3.2.2.

Sea P una pirámide sobre Ω .

Diremos que una función $i : P \longrightarrow \mathbb{R}^+$ es un índice asociado a P , si verifica las condiciones:

$$I.1.- i(\{\omega\})=0, \quad \forall \omega \in \Omega$$

$$I.2.- \forall h, h' \in P, h \subset h' \Rightarrow i(h) \leq i(h')$$

Si P es una pirámide sobre Ω y i un índice asociado a P , diremos que (P, i) es una Pirámide indexada.

Definición 3.2.3.

Si (P, i) es una pirámide indexada tal que:

$$I.3.- \forall h, h' \in P, h \subsetneq h' \Rightarrow i(h) < i(h')$$

Diremos que (P, i) es una pirámide indexada en sentido estricto.

Si en lugar de la condición I.3 se verifica:

$$I.3'.- \forall h, h' \in P \text{ tales que } h \subsetneq h' \text{ y } i(h) = i(h'), \text{ existen } h_1, h_2 \in P, \text{ distintos de } h, \text{ y tales que } h = h_1 \cap h_2.$$

Entonces diremos que (P, i) es una pirámide indexada en sentido amplio.

De las definiciones anteriores resulta inmediato:

Proposición 3.2.1.

Toda pirámide indexada en sentido estricto, lo es también en sentido amplio.

A partir de ahora, a los elementos de una pirámide los llamaremos "grupos" de la pirámide, y si una pirámide es indexada, al valor del índice sobre cada grupo, lo llamaremos "índice del grupo".

3.3.- PREDECESORES Y SUCESORES

Definición 3.3.1.

Sea $P \subset \mathcal{O}(\Omega)$ una pirámide, y sean p y p' dos grupos de P tales que $p \neq \Omega$ y $p' \neq \{\omega\}$, $\forall \omega \in \Omega$.

Diremos que p' es un **predecesor** de p , sii $p \subset p'$ y no existe ningún $p'' \in P - \{p, p'\}$, tal que $p \subset p'' \subset p'$.

También podemos decir que p es un **sucesor** de p' .

Proposición 3.3.1.

El conjunto de los sucesores de Ω constituye un recubrimiento de Ω .

Demostración:

Si $R = \{p_i \in P / p_i \text{ sucesor de } \Omega\}$, es suficiente demostrar que $\Omega \subset \bigcup_{p_i \in R} p_i$ puesto que la inclusión contraria es evidente.

Para cualquier $\omega \in \Omega$, se verifica: $\{\omega\} \in P$ y $\{\omega\} \subset \Omega$.

Por una parte, si no existe ningún $p \in P$, $p \neq \{\omega\}$ tal que $\{\omega\} \subset p \subset \Omega$, entonces $\{\omega\}$ es un sucesor de Ω , y por tanto $\omega \in \bigcup p_i$.

Si por el contrario existe algún $p \in P$ tal que $\{\omega\} \subset p \subset \Omega$, podemos suponer que p es un grupo maximal (en el sentido de la inclusión) en estas condiciones, así pues p será un sucesor de Ω , es decir $p \in R$ y por tanto $\omega \in \bigcup p_i$.

De forma parecida podríamos demostrar que el conjunto de los sucesores de los grupos de R , es a su vez un nuevo recubrimiento de Ω , y siguiendo este proceso

llegaríamos al recubrimiento formado únicamente por los singletones, ya que en este proceso el cardinal de los grupos va disminuyendo (p sucesor de $\Omega \Rightarrow p \subsetneq \Omega$; p' sucesor de $p \Rightarrow p' \subsetneq p, \dots$).

Proposición 3.3.2.

Cada grupo de una pirámide tiene, a lo sumo, dos predecesores.

Demostración:

En primer lugar observese que cualquier grupo p , de una pirámide P , $p \neq \Omega$, tiene por lo menos un predecesor ($p \subsetneq \Omega$).

Supongamos que p tiene dos predecesores distintos, p_1 y p_2 . Puesto que $p \subsetneq p_i$, para $i=1,2$; tendremos que $p \subset p_1 \cap p_2$, y esta inclusión no puede ser estricta, puesto que si así fuera, y por el hecho de ser p_1 y p_2 distintos, tendríamos: $p \subsetneq p_1 \cap p_2 \subsetneq p_i$, para $i=1,2$, con lo cual p_1 y p_2 no serían predecesores de p .

Así pues, si p_1 y p_2 son dos predecesores distintos de p , será $p = p_1 \cap p_2$.

Supongamos finalmente que p tuviera tres predecesores distintos: p_1, p_2, p_3 , y supongamos que:

$$p_1 = \{\omega_1^1 \leq \dots \leq \omega_{n_1}^1\}, \quad p_2 = \{\omega_1^2 \leq \dots \leq \omega_{n_2}^2\}, \quad p_3 = \{\omega_1^3 \leq \dots \leq \omega_{n_3}^3\}.$$

Puesto que P es una pirámide, en Ω existe un preorden total, supongamos pues que:

$$\omega_1^1 \leq \omega_1^2 \leq \omega_1^3.$$

Por ser p_1, p_2, p_3 predecesores de p , tendremos que : $\omega_{n_1}^1 \leq \omega_{n_2}^2 \leq \omega_{n_3}^3$.

Por otra parte, puesto que p_1, p_2 y p_2, p_3 son pares de predecesores de p , tendremos que:

$$p = p_1 \cap p_2 = \{\omega_1^2 \leq \dots \leq \omega_{n_1}^1\}$$

$$p = p_2 \cap p_3 = \{\omega_1^3 \leq \dots \leq \omega_{n_2}^2\}$$

por lo tanto tendrá que ser: $\omega_1^3 = \omega_1^2$ y $\omega_{n1}^1 = \omega_{n2}^2$, de manera que los grupos p_1 y p_2 , en particular, quedarán de la forma siguiente: $p_1 = \{\omega_1^1 \leq \dots \leq \omega_{n1}^1\}$ $p_2 = \{\omega_1^2 \leq \dots \leq \omega_{n1}^2\}$ con lo cual tendremos que $\omega_1^1 = \omega_1^2$ o bien $\omega_1^1 < \omega_1^2$.

Si fuese $\omega_1^1 = \omega_1^2$, sería $p_1 = p_2$ y por consiguiente p no tendría tres predecesores distintos, y si $\omega_1^1 < \omega_1^2$, $p_2 \subsetneq p_1$ y p_1 no sería predecesor de p , en todo caso pues, llegamos a una contradicción con que p_1, p_2, p_3 sean tres predecesores distintos de p .

En conclusión obtenemos que, en una pirámide, cada grupo tiene, a lo sumo, dos predecesores.

3.4.- REPRESENTACION VISUAL DE LAS PIRAMIDES

Del mismo modo que los dendrogramas proporcionan una representación visual de las Jerarquías Indexadas, es posible construir otro tipo de grafos conexos, a los que llamaremos "grafos piramidales" o "pirámides", que proporcionen una representación visual de las Pirámides Indexadas.

Su construcción consiste en expresar gráficamente las propiedades características de las pirámides:

- * Dos grupos, no encajados, pueden tener intersección no vacía.
- * Cada grupo puede tener hasta dos predecesores.
- * Existencia de un preorden total para el que cada grupo sea conexo.

El grafo piramidal expresará entonces, en un solo diagrama, la imbricación de unos grupos en otros y el índice de la pirámide, de forma que, fijado un determinado nivel,

podamos determinar de forma inmediata, los grupos que forman el recubrimiento correspondiente. Para ello, el proceso a seguir será el siguiente:

- a) Determinar los sucesores o predecesores de cada grupo de la pirámide
- b) Indexar la pirámide.
- c) Tener en cuenta el orden de Ω , compatible con la pirámide.

En Diday (1986-1), se da una condición necesaria y suficiente para que un orden sea compatible con una pirámide, así como un algoritmo para su construcción, utilizando únicamente las propiedades conjuntistas de las pirámides. En la sección 5.2 del Capítulo 5, nosotros proponemos un algoritmo de clasificación piramidal que construye también este orden compatible (uno de los posibles).

Con todo ello, la representación visual de una pirámide se efectuará asociando a cada grupo un segmento horizontal situado, sobre una escala, a una altura igual a la de su índice y unido a sus predecesores mediante unas líneas oblicuas, que llamaremos "aristas", de manera que cada arista una el centro del segmento asociado a un grupo con uno de los extremos de los segmentos asociados a sus predecesores.

Ejemplo 3.4.1.

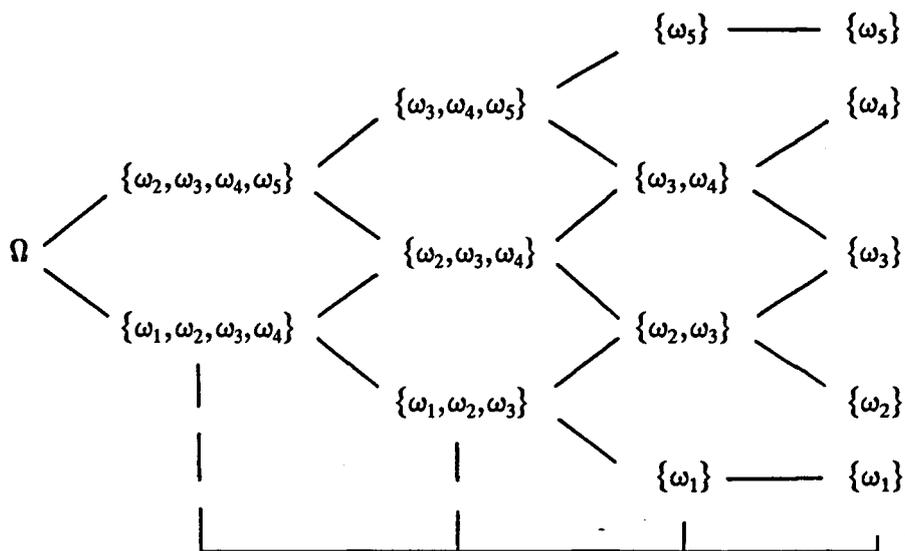
Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$, ordenado por $\omega_1 \leq \omega_2 \leq \omega_3 \leq \omega_4 \leq \omega_5$ y sea

$P = \{\{\omega_1\}, \dots, \{\omega_5\}, \{\omega_2, \omega_3\}, \{\omega_3, \omega_4\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_3, \omega_4, \omega_5\}, \{\omega_2, \omega_3, \omega_4\},$
 $\{\omega_2, \omega_3, \omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3, \omega_4\}, \Omega\}$ una pirámide indexada por:

$$\begin{aligned} i(\{\omega_i\}) &= 0, \text{ para } i=1, \dots, 5 \\ i(\{\omega_2, \omega_3\}) &= 2 \\ i(\{\omega_3, \omega_4\}) &= 2.5 \\ i(\{\omega_1, \omega_2, \omega_3\}) &= 3 \\ i(\{\omega_3, \omega_4, \omega_5\}) &= 4 \\ i(\{\omega_2, \omega_3, \omega_4\}) &= 5 \end{aligned}$$

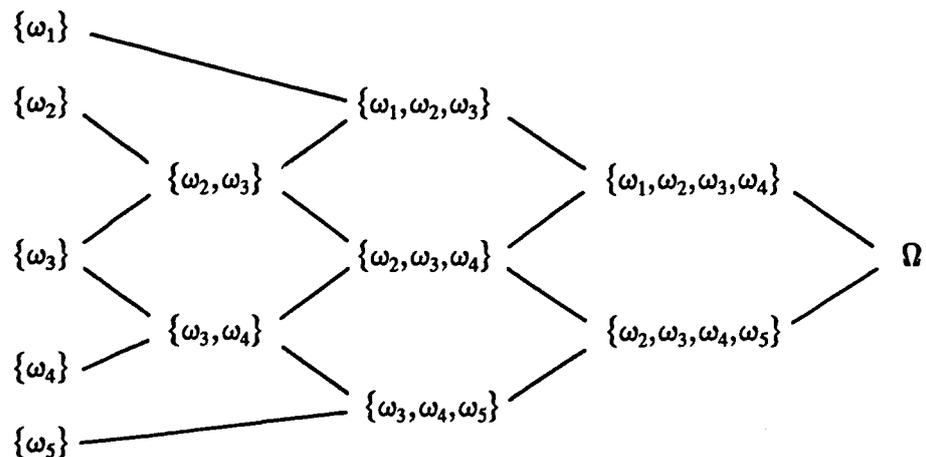
$$\begin{aligned} i(\{\omega_2, \omega_3, \omega_4, \omega_5\}) &= 5 \\ i(\{\omega_1, \omega_2, \omega_3, \omega_4\}) &= 6 \\ i(\Omega) &= 7 \end{aligned}$$

Si partimos de Ω y determinamos los sucesivos sucesores de cada grupo, tendremos el siguiente diagrama:

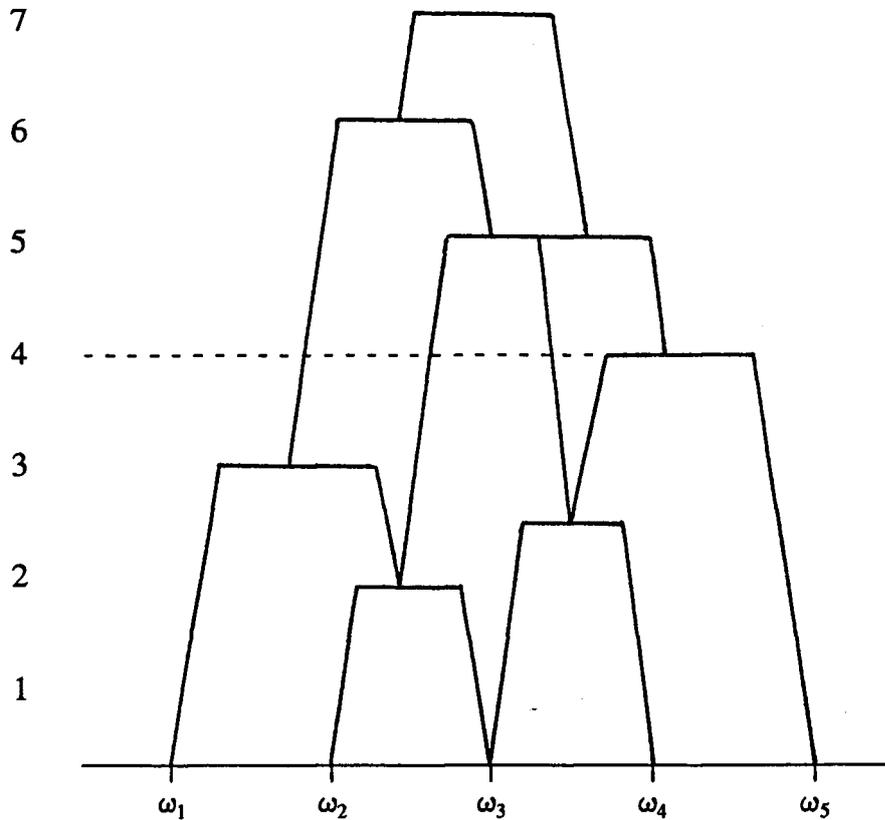


Recubrimientos de Ω . (Prop.3.3.1)

Si partimos de los singletons, $\{\omega_1\}, \dots, \{\omega_5\}$ y determinamos los sucesivos predecesores, obtendremos:



La representación visual de la pirámide (P,i) será:



Observe como cada grupo se une a sus predecesores mediante las aristas de la pirámide.

Observe también como, a cada nivel, la clasificación correspondiente constituye un recubrimiento de la población Ω . A nivel 4, por ejemplo, los grupos en que queda dividida la población son : $\{\omega_1, \omega_2, \omega_3\}$, $\{\omega_2, \omega_3\}$, $\{\omega_3, \omega_4\}$, $\{\omega_3, \omega_4, \omega_5\}$, que obviamente forman un recubrimiento de Ω .

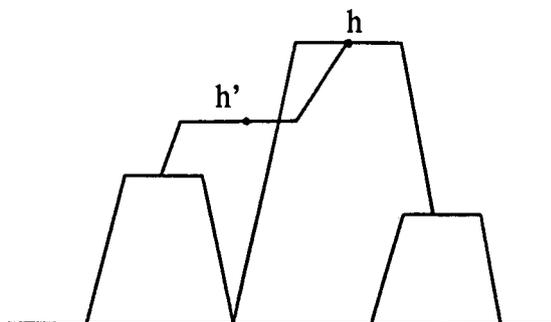
Es evidente que la representación visual de una pirámide es más compleja que la de una jerarquía; será conveniente pues, procurar que los grafos piramidales sean lo

más claros posible. Para ello será muy deseable que **no presenten inversiones ni cruzamientos.**

El concepto de **inversión** en un grafo piramidal se define igual que para el caso ultramétrico, es decir:

Definición 3.4.1.

Un índice i , sobre una pirámide P , presenta **inversión**, si existen $h, h' \in P$ tales que $h \leq h'$ y $i(h) > i(h')$.



* Obsérvese que por la propia definición de índice, no es posible que una pirámide indexada (P, i) presente inversión, no obstante es un concepto que vale la pena mantener, sobretodo de cara a los procesos algorítmicos de construcción piramidal, en los que hay que tener cuidado con que el índice que se va construyendo, a partir de una cierta disimilaridad inicial, lo sea realmente, es decir, no de lugar a inversiones. (ver Ejemplo 5.4.3.)

El concepto de **cruzamiento**, desde el punto de vista de la representación visual, también coincide con dicho concepto en el caso ultramétrico:

Definición 3.4.2.

Un orden sobre Ω da lugar a un **cruzamiento** sobre una pirámide P , si en la representación visual, alguna arista corta a otra arista o a un segmento horizontal.

Para las jerarquías, esta definición es equivalente a decir que: un orden da lugar a un cruzamiento, si existe alguna clase que no sea conexa respecto a dicho orden. Por ejemplo, si $J = \{\{\omega_1\}, \dots, \{\omega_4\}, \{\omega_1, \omega_3\}, \{\omega_2, \omega_4\}, \Omega\}$, el orden $\omega_1 \leq \omega_2 \leq \omega_3 \leq \omega_4$, da lugar a un cruzamiento, mientras que el orden $\omega_1 \leq \omega_3 \leq \omega_2 \leq \omega_4$, no da lugar a cruzamientos para la misma jerarquía (Fig. 3.4.1).

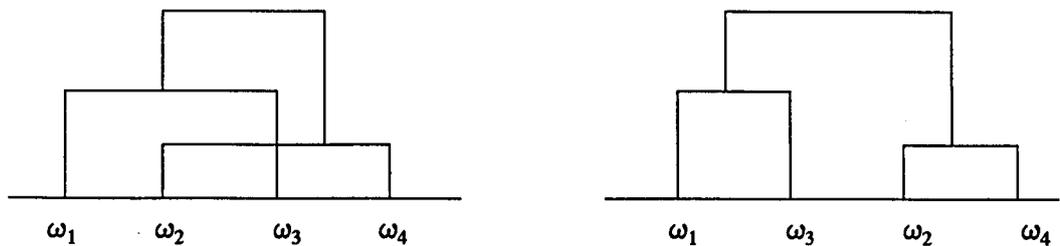


Fig. 3.4.1

Sin embargo en el caso piramidal, el hecho de que cada grupo sea conexo no implica la ausencia de cruzamiento en la pirámide, y esto es debido al hecho de que un grupo pueda tener hasta dos predecesores. En el ejemplo de la Figura 3.4.2, todos los grupos son conexos, y a pesar de ello, existe cruzamiento. Obsérvese que $h \cap h'$ no es vacía y tampoco es un grupo de la pirámide.

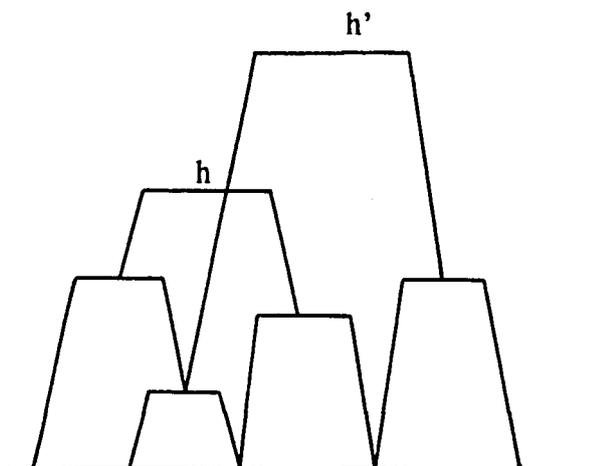


Fig. 3.4.2

Así pues en el caso de una pirámide, la definición 3.4.2, será equivalente a la siguiente:

Definición

Un orden sobre Ω da lugar a un cruzamiento sobre una pirámide \mathbf{P} , si existe algún grupo no conexo respecto a dicho orden o bien si existen $h, h' \in \mathbf{P}$ tales que $h \cap h' \neq \emptyset$ y $h \cap h' \notin \mathbf{P}$.

Por tanto, es evidente también que si \mathbf{P} es una pirámide, en virtud de la propia definición (3.2.1), el grafo correspondiente no puede presentar cruzamientos, puesto que no existirán grupos no conexos ni intersecciones no vacías que no pertenezcan a \mathbf{P} . A pesar de ello, y tal como sucede con la inversión, es un concepto que vale la pena tener en cuenta de cara a los procesos algorítmicos de construcción de grafos piramidales (Teorema 5.2.1).

3.5.- PIRAMIDES Y JERARQUIAS

En este apartado veremos como las pirámides son una generalización natural de las jerarquías.

Proposición 3.5.1.

Toda ultramétrica d sobre Ω , induce un preorden total sobre la población, compatible con la jerarquía total indexada correspondiente a d .

Demostración:

Sea (J, i) la jerarquía total indexada inducida por una disimilaridad ultramétrica d , dada sobre la población $\Omega = \{\omega_1, \dots, \omega_n\}$.

En virtud de la proposición 2.1.1, podemos asegurar que existirá un orden total sobre Ω : $\omega_1 \leq \dots \leq \omega_n$, tal que la matriz de d , (d_{ij}) cumpla las condiciones a) y b) de dicha proposición.

Veamos pues que este orden es compatible con J .

Sea, para ello, h una clase cualquiera de la jerarquía. Hemos de ver que h es conexa respecto al orden correspondiente.

Sean ω_i y ω_j un par de individuos de h tales que exista $\omega_k \in \Omega$ con $\omega_i \leq \omega_k \leq \omega_j$.

Puesto que la disimilaridad d es compatible con el preorden (proposiciones 2.1.2. y 2.3.3.), tendremos que:

$$d(\omega_i, \omega_k) \leq d(\omega_i, \omega_j) \quad \text{y} \quad d(\omega_k, \omega_j) \leq d(\omega_i, \omega_j)$$

Sea h_0 la mínima clase de J , en el sentido de la inclusión, que contiene ω_i y ω_j .

Puesto que d es la ultramétrica inducida por (J, i) , $d(\omega_i, \omega_j) = i(h_0)$; y si (J, i) es la

jerarquía inducida por d , $i(h_0) = \max\{d(\omega, \omega'), \forall \omega, \omega' \in h_0\}$. Si llamamos r_0 al índice de h_0 , tendremos: $d(\omega_i, \omega_k) \leq r_0$ y $d(\omega_k, \omega_j) \leq r_0$.

En estas condiciones vamos a ver que $\omega_k \in h_0$. En efecto, sea ω un elemento cualquiera de h_0 , puesto que $\omega_j \in h_0$, será $d(\omega_j, \omega) \leq r_0$; por otra parte, ya hemos visto que $d(\omega_k, \omega_j) \leq r_0$; y por ser d ultramétrica, $d(\omega_k, \omega) \leq \max\{d(\omega_k, \omega_j), d(\omega_j, \omega)\}$. De estas tres desigualdades, podemos concluir que $\forall \omega \in h_0$, $d(\omega_k, \omega) \leq r_0$.

Así pues, si la distancia de ω_k a cualquier elemento de h_0 es menor o igual que $r_0 = i(h_0)$, ha de ser $\omega_k \in h_0$.

Puesto que $h_0 \subset h$, se verificará $\omega_k \in h$ tal y como queríamos demostrar.

Por tanto, cualquier clase de \mathbf{J} es conexa respecto al preorden de Ω .

Proposición 3.5.2.

Toda Jerarquía Total Indexada es una Pirámide Indexada

Demostración:

Sea (\mathbf{J}, i) una jerarquía total indexada de $\mathcal{P}(\Omega)$. Por ser total, tendremos que $\Omega \in \mathbf{J}$ y $\forall \omega \in \Omega$, $\{\omega\} \in \mathbf{J}$.

Por otra parte, dados cualesquiera h y h' de \mathbf{J} , $h \cap h' \in \{\emptyset, h, h'\}$, en todo caso $h \cap h' = \emptyset$ ó $h \cap h' \in \mathbf{J}$.

Finalmente, \mathbf{J} induce una ultramétrica sobre Ω , la cual, según la proposición 3.5.1. induce un preorden total sobre Ω que es compatible con \mathbf{J} . Por tanto, existe un preorden total sobre Ω compatible con \mathbf{J} .

Así pues, \mathbf{J} es una pirámide de $\mathcal{P}(\Omega)$.

Si además, \mathbf{J} es indexada como jerarquía, considerada como pirámide, también lo será

y con el mismo índice.

Nota 3.5.1.

En una jerarquía cada clase tiene **un único** predecesor. La demostración es inmediata:

Sea $h \in J$, $h \neq \Omega$. h por lo menos tendrá un predecesor puesto que $h \subset \Omega$.

Supongamos que h tuviese dos predecesores, h_1 y h_2 , distintos.

Puesto que $h \subset h_1$ y $h \subset h_2$, tendremos $h_1 \cap h_2 \neq \emptyset$ y por ser J una jerarquía tendrá que ser $h_1 \subset h_2$ o bien $h_2 \subset h_1$ con lo cual uno de los dos no sería predecesor de h .

CAPITULO 4

PIRAMIDES INDEXADAS Y DISIMILARIDADES PIRAMIDALES

4.1.- INTRODUCCION

Teniendo en cuenta que la base matemática de las representaciones jerárquicas se fundamenta en la relación existente entre la obtención de una disimilaridad ultramétrica sobre una población Ω y una jerarquía indexada de $\mathcal{P}(\Omega)$, en el presente capítulo trataremos de demostrar la existencia de una biyección entre las pirámides indexadas y las disimilaridades piramidales, de modo que los grafos piramidales puedan interpretarse como la representación visual de dichas disimilaridades.

Intentaremos abordar este problema, en primer lugar, siguiendo, mientras ello sea posible, un cierto paralelismo con el caso de las jerarquías y las ultramétricas antes mencionado.

Sea Ω un conjunto finito y d una **disimilaridad piramidal** sobre Ω . Nuestro objetivo será pues el de asociar a la disimilaridad d una pirámide indexada de $\mathcal{P}(\Omega)$.

Sea $A = \{\alpha \in \mathbb{R} / \exists \omega_i, \omega_j \in \Omega \text{ y } d(\omega_i, \omega_j) = \alpha\}$

Para cualquier $\alpha \in A$, definimos la siguiente relación sobre Ω :

$$\forall \omega, \omega' \in \Omega \quad \omega R_\alpha \omega' \Leftrightarrow d(\omega, \omega') \leq \alpha$$

En el caso ultramétrico, esta relación es de equivalencia. En este caso las propiedades reflexiva i simétrica se satisfacen de forma inmediata al ser d piramidal, sin embargo este hecho no es suficiente para asegurar la transitividad de la relación, por lo que las relaciones R_α no serán de equivalencia, y por tanto no darán lugar, en general, a particiones de Ω , sino a unos ciertos **recubrimientos**.

Así pues, para cada $\alpha \in A$ podemos considerar los subconjuntos de Ω formados por individuos relacionados entre sí (a distancia menor o igual que α). Si d es piramidal, para cada α , estos subconjuntos formarán un recubrimiento de Ω , al que denominaremos P_α . A pesar de todo, podríamos pensar, siguiendo una analogía con el razonamiento que se sigue en el caso de las disimilaridades ultramétricas, que la unión de todos estos recubrimientos, $P' = \bigcup_{\alpha \in A} P_\alpha$ es la pirámide inducida por la disimilaridad d. Pues bien, esto no siempre es cierto, puesto que la intersección de dos grupos de un tal P' , no siempre es de P' o vacía, tal como puede verse en el Ejemplo 4.1.1.

Ejemplo 4.1.1.

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ y d una disimilaridad piramidal sobre Ω , dada por la matriz:

$$\begin{bmatrix} 0 & 1 & 1 & 3 \\ & 0 & 1 & 2 \\ & & 0 & 2 \\ & & & 0 \end{bmatrix}$$

En este caso, $A = \{0, 1, 2, 3\}$, y construyendo los distintos recubrimientos de Ω , para

los distintos valores de α , tendremos:

$$\alpha=0 \quad P_0 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}\}$$

$$\alpha=1 \quad P_1 = \{\{\omega_1, \omega_2, \omega_3\}, \{\omega_4\}\}$$

$$\alpha=2 \quad P_2 = \{\{\omega_1, \omega_2, \omega_3\}, \{\omega_2, \omega_3, \omega_4\}\}$$

$$\alpha=3 \quad P_3 = \{\Omega\}$$

y $P' = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_2, \omega_3, \omega_4\}, \Omega\}$.

Evidentemente P' no es una pirámide puesto que:

$$\{\omega_1, \omega_2, \omega_3\} \cap \{\omega_2, \omega_3, \omega_4\} = \{\omega_2, \omega_3\} \notin P'$$

Así pues deberemos reconsiderar el método de construcción de una pirámide a partir de una disimilaridad piramidal. En este empeño, enunciamos la siguiente proposición:

Proposición 4.1.1.

Sea d una disimilaridad piramidal sobre Ω , y

$A = \{\alpha \in \mathbb{R} / \exists \omega_i, \omega_j \in \Omega \text{ y } d(\omega_i, \omega_j) = \alpha\}$. Fijado $\alpha \in A$, se verifica:

$$\begin{aligned} h = \{x \in \Omega / \exists \omega_i, \omega_j \in \Omega; d(\omega_i, \omega_j) = \alpha \text{ y } \omega_i \leq x \leq \omega_j\} \text{ maximal} &\Leftrightarrow \\ \Leftrightarrow h = \{\text{individuos de } \Omega \text{ con interdistancias } \leq \alpha\}. \end{aligned}$$

Demostración:

\Rightarrow

Sea $\alpha \in A$ y sean $\omega_i, \omega_j \in \Omega$ tales que $d(\omega_i, \omega_j) = \alpha$, siendo éstos los individuos extremos, en términos del preorden total asociado a d , a distancia α .

Si $\omega, \omega' \in h$, por las proposiciones 2.3.1 y 2.3.3 tendremos que $d(\omega, \omega') \leq d(\omega_i, \omega_j) = \alpha$, con lo cual, cualquier elemento de h pertenecerá al conjunto de

individuos de Ω con interdistancias $\leq \alpha$.

⇐

Sea ω un individuo de h , y $\alpha \in A$. $\forall z \in h$, $d(\omega, z) \leq \alpha$.

Al ser $\alpha \in A$, existirán $\omega_i, \omega_j \in \Omega$ con $d(\omega_i, \omega_j) = \alpha$. Sean éstos los individuos de Ω extremos a distancia α ; en estas condiciones, forzosamente habrá de ser $\omega_i \leq \omega \leq \omega_j$, puesto que si fuese $\omega < \omega_i \leq \omega_j$, por ser d piramidal tendríamos $d(\omega, \omega_j) \geq \alpha$. Por otra parte, puesto que $d(\omega_i, \omega_j) = \alpha$, ω_i y ω_j serán de h , y puesto que ω también es de h , será $d(\omega, \omega_j) \leq \alpha$. Por tanto tendríamos $d(\omega, \omega_j) = \alpha$ y $\omega < \omega_i \leq \omega_j$, con lo cual ω_i y ω_j no serían los individuos extremos a distancia α .

Así pues, si $\omega \in h$, existirán ω_i y ω_j de Ω tales que $d(\omega_i, \omega_j) = \alpha$ y $\omega_i \leq \omega \leq \omega_j$ y además, al ser ω_i y ω_j los individuos extremos a distancia α , el conjunto de los individuos entre ω_i y ω_j será maximal, por construcción, que es lo que queríamos demostrar.

Si hubiésemos supuesto $\omega_i \leq \omega_j < \omega$, habríamos llegado al mismo resultado.

4.2.- TEOREMAS DE EXISTENCIA

En primer lugar justificaremos la definición, que en el teorema 4.2.1. daremos, de pirámide obtenida a partir de una disimilaridad piramidal.

Sea d una disimilaridad piramidal sobre Ω , y sea A el conjunto de valores **distintos** que toma d sobre los pares de individuos de Ω .

Para cualquier $\alpha \in A$, definimos la siguiente familia de $\mathcal{P}(\Omega)$:

$$C_\alpha = \{h \in \mathcal{P}(\Omega) / h = \{\omega \in \Omega / \exists \omega_i, \omega_j \in \Omega; d(\omega_i, \omega_j) = \alpha \text{ y } \omega_i \leq \omega \leq \omega_j\}\}$$

Consideremos, para cualquier $\alpha \in A$, una nueva familia:

$$C_\alpha^* = \{h \in C_\alpha / h \text{ es maximal, en } C_\alpha, \text{ en el sentido de la inclusión}\}$$

Por comodidad, llamaremos a los conjuntos de C_α^* : $h_\alpha^1, \dots, h_\alpha^{n_\alpha}$.

A partir de aquí, podríamos intentar definir la pirámide inducida por d como:

$$P^* = \bigcup_{\alpha \in A} C_\alpha^*$$

Ahora bien, después de la proposición 4.1.1. este nuevo conjunto P^* , coincide con el P' definido en la introducción. Así pues, P^* tampoco será una pirámide.

Teorema 4.2.1.

Toda disimilaridad piramidal d sobre Ω , define una pirámide indexada en sentido amplio de $\mathcal{P}(\Omega)$.

Demostración:

Dada una disimilaridad piramidal d sobre Ω , definimos el siguiente subconjunto de $\mathcal{P}(\Omega)$:

$$P = P^* \cup \{\text{intersecciones no vacías de dos grupos de } P^*\} \quad (1)$$

Con ello, tendremos que los conjuntos de P o bien son grupos de P^* , o bien son intersecciones de dos grupos de P^* , que son precisamente los que le faltan a ésta para que pueda ser pirámide.

Veamos, en primer lugar, que P es una pirámide.

⁽¹⁾ $h \in P$ sii $\left| \begin{array}{l} h \in P^*: h = [\omega_i, \omega_j]_\alpha \text{ maximal (h intervalo maximal en el sentido de} \\ \text{que } \forall \omega' \notin h, \exists \omega \in h \text{ y } d(\omega', \omega) > \alpha). \\ \text{ó} \\ h \text{ es intersección de intervalos maximales.} \end{array} \right.$

P.1.- $\Omega \in \mathbf{P}$

En efecto, $\Omega \in \mathbf{P}^*$ puesto que será $\Omega = h_\alpha$, con $\alpha = \max(A)$.

P.2.- $\forall \omega \in \Omega; \{\omega\} \in \mathbf{P}$

$\forall \omega \in \Omega, \{\omega\} \in \mathbf{P}^*$ puesto que $\{\omega\} = h_0$ y $0 \in A$.

P.3.- $\forall h, h' \in \mathbf{P}, h \cap h' = \emptyset$ o $h \cap h' \in \mathbf{P}$

Sean h, h' dos grupos de \mathbf{P} tales que $h \cap h' \neq \emptyset$

Si h y h' fuesen de \mathbf{P}^* , entonces por definición, $h \cap h' \in \mathbf{P}$.

Si h y h' no fuesen de \mathbf{P}^* , entonces existirían $h_i \in \mathbf{P}^*, i \in \{1, 2, 3, 4\}$, tales que :

$h = h_1 \cap h_2$ y $h' = h_3 \cap h_4$.

Puesto que d es piramidal, tiene sentido suponer la existencia de un preorden total sobre Ω , consideremos, para cualquier $i \in \{1, 2, 3, 4\}$, ω_i y ω'_i los individuos extremos de h_i ($\omega_i \leq \omega'_i$).

Si consideramos $\omega_j = \max\{\omega_i\}_{i=1,2,3,4}$

$$\omega'_k = \min\{\omega'_i\}_{i=1,2,3,4}$$

resulta inmediato, a partir de la consideración de que el preorden sobre Ω es total, que: $h \cap h' = h_j \cap h_k$, con $h_j, h_k \in \mathbf{P}^*$. (Ver Fig. 4.2.1)

Así pues, tendremos que $h \cap h' \in \mathbf{P}$.

Obsérvese que si h ó h' fuese de \mathbf{P}^* , llegaríamos a la misma conclusión.

$h \cap h' = (h_1 \cap h_2) \cap (h_3 \cap h_4) = h_1 \cap h_4 :$

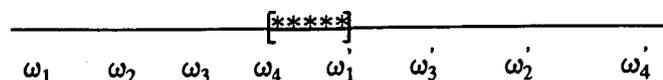


Fig. 4.2.1

P.4.- Existe un preorden total sobre Ω , compatible con \mathbf{P} .

Puesto que d es piramidal, existe un preorden total sobre Ω compatible con d , y este mismo preorden es compatible con \mathbf{P} , puesto que cualquier grupo de \mathbf{P}^* es conexo respecto a este preorden por construcción, y la intersección no vacía de dos grupos conexos es otro grupo conexo (puesto que el preorden es total). Por tanto cualquier grupo de \mathbf{P} será conexo respecto al preorden.

Finalmente vamos a ver que esta pirámide puede ser indexada en sentido amplio.

Para ello, definamos la siguiente aplicación:

$$\begin{aligned} i: \mathbf{P} &\longrightarrow \mathbf{R}^+ \\ h &\longrightarrow i(h) = \max\{d(\omega, \omega'); \forall \omega, \omega' \in h\} \end{aligned}$$

Entonces se satisface:

I.1.- $\forall \omega \in \Omega, i(\{\omega\}) = 0$

Inmediato a partir de la definición.

I.2.- $\forall h, h' \in \mathbf{P}$, si $h \subset h'$, entonces $i(h) \leq i(h')$

Si $h \subset h'$ es evidente que $\max\{d(x, y), \forall x, y \in h\} = i(h)$ es menor o igual que $\max\{d(x, y), \forall x, y \in h'\} = i(h')$, con lo cual $i(h) \leq i(h')$.

I.3.- $\forall h, h' \in \mathbf{P}$ con $h \subsetneq h'$ y $i(h) = i(h')$, existirán $h_1, h_2 \in \mathbf{P}$, distintos de h , y tales que $h = h_1 \cap h_2$.

Si $h \subsetneq h'$ y son del mismo índice, entonces $h \notin \mathbf{P}^*$, puesto que h no es maximal para $\alpha = i(h)$.

Por tanto, si $h \in \mathbf{P}$ y $h \notin \mathbf{P}^*$, existirán h_1 y h_2 de $\mathbf{P}^* \subset \mathbf{P}$, distintos de h , y tales que $h = h_1 \cap h_2$.

Así pues, (\mathbf{P}, i) es una Pirámide indexada en sentido amplio.

Ejemplo 4.2.1.

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ y d la disimilaridad piramidal dada por la siguiente matriz de Robinson

$$\begin{bmatrix} 0 & 3 & 3 & 3 & 6 \\ & 0 & 2 & 2 & 5 \\ & & 0 & 2 & 4 \\ & & & 0 & 4 \\ & & & & 0 \end{bmatrix}$$

Siguiendo el método dado en el teorema anterior, vamos a construir la pirámide indexada en sentido amplio correspondiente a esta disimilaridad piramidal:

En primer lugar, $A = \{0, 2, 3, 4, 5, 6\}$.

Para cada $\alpha \in A$, construimos la correspondiente familia C_α de $\mathcal{P}(\Omega)$.

$$C_0 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}\}$$

$$C_2 = \{\{\omega_2, \omega_3\}, \{\omega_2, \omega_3, \omega_4\}, \{\omega_3, \omega_4\}\}$$

$$C_3 = \{\{\omega_1, \omega_2\}, \{\omega_1, \omega_2, \omega_3\}\}$$

$$C_4 = \{\{\omega_3, \omega_4, \omega_5\}, \{\omega_4, \omega_5\}\}$$

$$C_5 = \{\{\omega_1, \omega_2, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_4, \omega_5\}\}$$

$$C_6 = \{\Omega\}$$

Si de cada familia consideramos únicamente los grupos maximales, tendremos:

$$C_0^* = \{h_0^k = \{\omega_k\}\}_{k=1, \dots, 5}$$

$$C_2^* = \{h_2 = \{\omega_2, \omega_3, \omega_4\}\}$$

$$C_3^* = \{h_3 = \{\omega_1, \omega_2, \omega_3\}\}$$

$$C_4^* = \{h_4 = \{\omega_3, \omega_4, \omega_5\}\}$$

$$C_5^* = \{h_5^1 = \{\omega_1, \omega_2, \omega_3, \omega_4\}, h_5^2 = \{\omega_2, \omega_3, \omega_4, \omega_5\}\}$$

$$C_6^* = \{h_6 = \Omega\}$$

Así pues, $P^* = \{h_0^1, h_0^2, h_0^3, h_0^4, h_0^5, h_2, h_3, h_4, h_5^1, h_5^2, h_6\}$.

Para obtener la pirámide P , tendremos que añadir a P^* , las intersecciones no vacías correspondientes. En este caso serán:

$$h_2 \cap h_3 = \{\omega_2, \omega_3\} \notin P^*$$

$$h_2 \cap h_4 = \{\omega_3, \omega_4\} \notin P^*$$

Por tanto, $P = P^* \cup \{\{\omega_2, \omega_3\}, \{\omega_3, \omega_4\}\}$

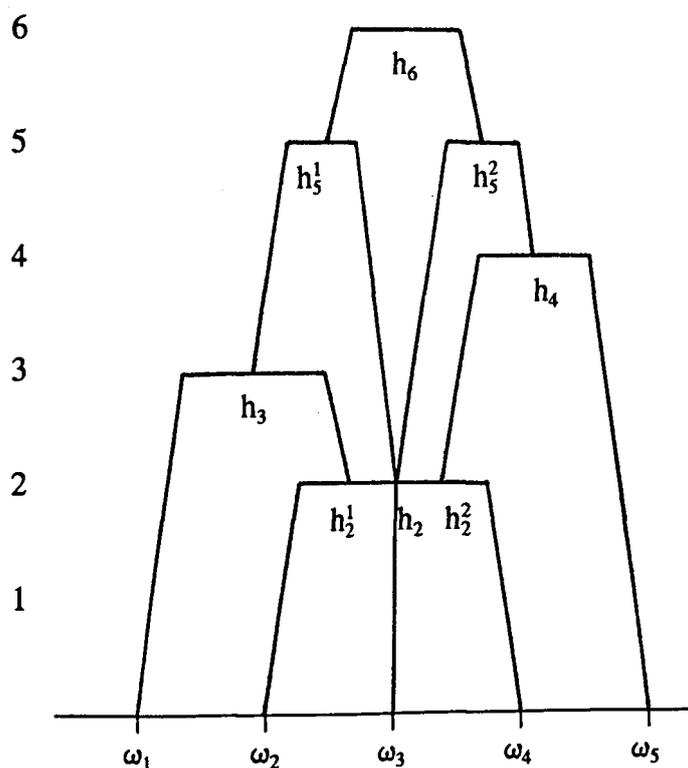
Finalmente construimos el índice sobre P :

$$i(h) = \max\{d(\omega_i, \omega_j), \forall \omega_i, \omega_j \in h\}; \forall h \in P$$

Con lo cual:

$$\begin{array}{ll} i(h_0^k) = 0, \forall k = 1, 2, 3, 4, 5 & i(h_5^1) = i(h_5^2) = 5 \\ i(h_2) = 2 & i(h_6) = 6 \\ i(h_3) = 3 & i(\{\omega_2, \omega_3\}) = d(\omega_2, \omega_3) = 2 \\ i(h_4) = 4 & i(\{\omega_3, \omega_4\}) = 2 \end{array}$$

Si llamamos h_2^1 a $\{\omega_2, \omega_3\}$ y h_2^2 a $\{\omega_3, \omega_4\}$, la visualización de la pirámide será la siguiente:



Teorema 4.2.2.

Toda pirámide indexada (P, i) de $\mathcal{O}(\Omega)$, induce una disimilaridad piramidal sobre Ω .

Demostración:

Dada una pirámide indexada (P, i) , para cada par de individuos ω y ω' de Ω , definimos:

$$d(\omega, \omega') = i(h_0)$$

siendo h_0 , el mínimo grupo de P , en el sentido de la inclusión, que contiene ω y ω' .

Esta definición es correcta puesto que si existe un único grupo de P que contiene ω y ω' , entonces éste es el mínimo, y si existen varios, su intersección, que también es de P , es el mínimo, con lo cual tiene sentido considerar el mínimo grupo de P que contiene ω y ω' .

Veamos pues que d es una disimilaridad piramidal:

D.1.- $\forall \omega, \omega' \in \Omega, d(\omega, \omega') \geq 0$

Inmediato, puesto que para cada $h \in P, i(h) \geq 0$.

D.2.- $\forall \omega \in \Omega, d(\omega, \omega) = 0$

El menor grupo de P que contiene ω , es $\{\omega\}$, y $i(\{\omega\}) = 0$.

D.3.- $\forall \omega, \omega' \in \Omega, d(\omega, \omega') = d(\omega', \omega)$

Inmediato.

D.8.- Existe un preorden total sobre Ω tal que, dados $\omega, \omega', \omega''$ de Ω , tales que $\omega \leq \omega' \leq \omega''$, $d(\omega, \omega'') \geq \max\{d(\omega, \omega'), d(\omega, \omega'')\}$.

Puesto que P es una pirámide, existe en Ω un preorden total compatible con P .

Sean ω, ω' y ω'' elementos de Ω , con $\omega \leq \omega' \leq \omega''$.

Si $h_{\omega\omega''}$ el mínimo grupo de P que contiene ω y ω'' , puesto que $h_{\omega\omega''}$ es conexo respecto al preorden que P induce sobre Ω , tendremos que $\omega' \in h_{\omega\omega''}$.

Si $h_{\omega, \omega'}$ es el mínimo grupo de P que contiene ω y ω' , tendremos que tanto $h_{\omega, \omega''}$ como $h_{\omega', \omega''}$ contienen ω y ω' . Si son grupos conexos y $h_{\omega, \omega'}$ es el mínimo, tendremos que $h_{\omega, \omega'} \subset h_{\omega, \omega''}$, por tanto $i(h_{\omega, \omega'}) \leq i(h_{\omega, \omega''})$ y $d(\omega, \omega') \leq d(\omega, \omega'')$.

Por otra parte, $h_{\omega, \omega''}$ y $h_{\omega', \omega''}$, ambos contienen a ω' y ω'' , por tanto, $h_{\omega', \omega''} \subset h_{\omega, \omega''}$, así pues $i(h_{\omega', \omega''}) \leq i(h_{\omega, \omega''})$ y $d(\omega', \omega'') \leq d(\omega, \omega'')$.

De ambas desigualdades, obtenemos que,

$$d(\omega, \omega'') \geq \max\{d(\omega, \omega'), d(\omega', \omega'')\} \quad \text{c.q.d.}$$

Puesto que el teorema anterior (4.2.2) es válido, en particular, para pirámides indexadas en sentido amplio, estos dos teoremas establecen una correspondencia entre el conjunto de la pirámides indexadas en sentido amplio y el de las disimilaridades piramidales.

En la sección siguiente, demostraremos que esta correspondencia es biunívoca, es decir, cada disimilaridad piramidal induce una única pirámide indexada en sentido amplio y viceversa.

4.3.- TEOREMAS DE UNICIDAD

Proposición 4.3.1.

Sea (P, i) una pirámide indexada de $\mathcal{O}(\Omega)$ y sean ω y ω' dos individuos cualesquiera de Ω . El mínimo grupo de P que contiene a ω y ω' , es también el de menor índice de entre todos los que los contienen.

Demostración:

Sea h_0 el mínimo grupo de P que contiene ω y ω' , y supongamos que exista otro grupo h que también los contenga. En tal caso, tendrá que ser $h_0 \subset h$, puesto que de no ser así, $h_0 \cap h$ sería de P , contendría ω y ω' , y estaría estrictamente contenido en h_0 , con lo cual h_0 no sería el mínimo.

Finalmente, por ser (P, i) pirámide indexada, si $h_0 \subset h$, entonces $i(h_0) \leq i(h)$, de donde se deduce que:

$$i(h_0) = \inf\{i(h), \forall h \in P / \omega, \omega' \in h\}.$$

Proposición 4.3.2.

Sea h un grupo cualquiera de una pirámide P . Si ω y ω' son los individuos extremos de h , entonces h es el mínimo grupo de P que los contiene.

Demostración:

Sea h un grupo cualquiera de P y ω, ω' los individuos extremos de h .

Supongamos que h no sea el mínimo grupo de P que contiene ω y ω' , entonces existirá un $h' \in P$ tal que $h' \subsetneq h$ y $\omega, \omega' \in h'$.

Si $h' \subsetneq h$, existirá $\bar{\omega} \in \Omega / \bar{\omega} \in h$ y $\bar{\omega} \notin h'$, con lo cual tendremos que $\omega, \omega' \in h'$ y

$\bar{\omega} \notin h'$, y puesto que h' es conexo, esto implica que no puede darse la situación $\omega \leq \bar{\omega} \leq \omega'$.

Así pues, tendremos que $\bar{\omega} < \omega \leq \omega'$ o bien $\omega \leq \omega' < \bar{\omega}$ con lo cual ω y ω' no serían los extremos de h , que es lo que habíamos supuesto.

Por tanto h ha de ser el mínimo grupo de \mathbf{P} que contiene ω y ω' .

Proposición 4.3.3.

Sea (\mathbf{P}, i) una pirámide indexada en sentido amplio, y d la disimilaridad piramidal inducida.

Sea $P' = \{h \in \mathcal{P}(\Omega) / \exists \omega_i, \omega_j \in \Omega \text{ y } h = \{\text{individuos de } \Omega \text{ con interdistancias menores o iguales que } d(\omega_i, \omega_j)\}\}$.

Entonces:

$$h \in \mathbf{P} \Leftrightarrow \begin{cases} h \in P' \\ \text{o} \\ \exists h_1, h_2 \in P', \text{ distintos de } h, \text{ y } h = h_1 \cap h_2 \end{cases}$$

Demostración:

\Rightarrow

Sea h un grupo cualquiera de \mathbf{P} . Puesto que \mathbf{P} es una pirámide, tiene sentido considerar los individuos extremos de h respecto al preorden total correspondiente sobre Ω , sean éstos, ω_i y ω_j . En estas condiciones, la proposición 4.3.2. nos asegura que h es el mínimo grupo de \mathbf{P} que contiene ω_i y ω_j .

Supongamos que $h \notin P'$, entonces existirá $\omega_0 \in \Omega$, tal que:

$$\omega_0 \notin h \text{ y } \forall x \in h, d(\omega_0, x) \leq d(\omega_i, \omega_j) \quad (1)$$

Si ω_i, ω_j son los individuos más alejados de h , y $\omega_0 \notin h$, entonces no puede ser $\omega_i \leq \omega_0 \leq \omega_j$, y por tanto tendrá que ser:

$$\omega_0 < \omega_i \leq \omega_j \quad \text{o bien} \quad \omega_i \leq \omega_j < \omega_0$$

Supongamos $\omega_0 < \omega_i \leq \omega_j$. Puesto que d es piramidal, tendremos:

$$d(\omega_0, \omega_j) \geq d(\omega_i, \omega_j).$$

Por otra parte, para $x = \omega_j$ la desigualdad (1) quedará:

$$d(\omega_0, \omega_j) \leq d(\omega_i, \omega_j).$$

Y combinando ambas desigualdades, tendremos:

$$d(\omega_0, \omega_j) = d(\omega_i, \omega_j) \quad (2)$$

Consideremos, además, h_0 , el mínimo grupo de \mathbf{P} que contiene ω_0 y ω_j .

Puesto que la disimilaridad piramidal d , viene inducida por la pirámide \mathbf{P} , resultará que: $d(\omega_0, \omega_j) = i(h_0)$ y $d(\omega_i, \omega_j) = i(h)$.

De lo cual, y teniendo en cuenta (2), se deduce que $i(h) = i(h_0)$.

Veamos que $h \subseteq h_0$:

Puesto que $\omega_0 < \omega_i \leq \omega_j$ y h_0 es conexo, ω_i también ha de pertenecer a h_0 .

Así pues, h es el mínimo grupo que contiene ω_i y ω_j y h_0 también los contiene, por tanto $h \subseteq h_0$.

Además $\omega_0 \in h_0$ y $\omega_0 \notin h$, con lo cual, $h \subsetneq h_0$.

Resumiendo pues, tendremos que: dado $h \in \mathbf{P}$, si $h \notin \mathbf{P}'$, existirá $h_0 \in \mathbf{P}$ tal que $h \subsetneq h_0$ y $i(h) = i(h_0)$, y puesto que (\mathbf{P}, i) es una pirámide indexada en sentido amplio, existirán $h_1, h_2 \in \mathbf{P}$, distintos de h , tales que $h = h_1 \cap h_2$.

Ahora bien, nuestro objetivo era demostrar que si $h \notin \mathbf{P}'$, entonces es intersección de dos grupos de \mathbf{P}' , y no de \mathbf{P} .

Supongamos pues que por lo menos uno de los dos grupos, h_1 por ejemplo, no es de \mathbf{P}' . Entonces, existirán $h_1', h_1'' \in \mathbf{P}$, distintos de h_1 y tales que $h_1 = h_1' \cap h_1''$, con lo cual $h = h_1' \cap h_1'' \cap h_2$.

Si, a su vez, alguno de estos grupos no es de P' , h_i por ejemplo, existirán \bar{h}_i y \hat{h}_i de P , distintos de h_i , con $h_i = \bar{h}_i \cap \hat{h}_i$, con lo cual $h = \bar{h}_1 \cap \hat{h}_1 \cap h_2$, y así sucesivamente.

En este proceso, observese que $h \subset h_1 \subset h_1' \subset \bar{h}_1 \subset \dots$. Puesto que las inclusiones son estrictas y $\mathcal{P}(\Omega)$ es finito, o bien llegaremos a un grupo de P' , o bien a un grupo \tilde{h}_1 tal que $h \subset h_1 \subset \dots \subset \tilde{h}_1$ y no podrá desdoblarse como intersección de dos grupos de P (esto sucederá, por ejemplo, cuando \tilde{h}_1 contenga uno de los extremos de Ω . Fig. 4.3.1), entonces tendrá que ser $\tilde{h}_1 \in P'$, puesto que de lo contrario, sería intersección de dos grupos estrictamente más grandes con lo cual \tilde{h}_1 no sería el último grupo de la cadena como estamos suponiendo.

" \tilde{h}_1 no puede ser intersección no trivial de dos grupos de P "

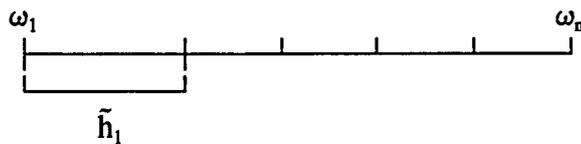


Fig. 4.3.1

Así pues, aplicando este proceso a todos los grupos que sea necesario (a los que no sean de P'), tendremos que h puede expresarse como $h = \tilde{h}_1 \cap \tilde{h}_2 \cap \dots \cap \tilde{h}_r$, con $\tilde{h}_i \in P'$ y $\tilde{h}_i \neq h$, para todo $i \in \{1, 2, \dots, r\}$.

A partir de aquí, por ser el preorden de Ω total, es inmediato demostrar que $h = \tilde{h}_k \cap \tilde{h}_j$, con $k, j \in \{1, \dots, r\}$, que es lo que queríamos.

←

Sean $\omega_i, \omega_j \in \Omega$ y $h \in \mathcal{P}(\Omega)$ el conjunto definido por:

$$h = \{\text{los individuos de } \Omega \text{ con interdistancias } \leq d(\omega_i, \omega_j)\} \in P'$$

En estas condiciones, podemos suponer que ω_i y ω_j son los individuos extremos de h , puesto que si no fuese así, es decir si existiese un $\bar{\omega} \in h$ tal que $\bar{\omega} < \omega_i \leq \omega_j$ (ó bien $\omega_i \leq \omega_j < \bar{\omega}$) tendríamos que: por ser d piramidal : $d(\bar{\omega}, \omega_j) \geq d(\omega_i, \omega_j)$

$$\text{y por ser } \bar{\omega}, \omega_j \in h : d(\bar{\omega}, \omega_j) \leq d(\omega_i, \omega_j)$$

con lo cual $d(\bar{\omega}, \omega_j) = d(\omega_i, \omega_j)$ y por tanto, en la definición de h podríamos sustituir ω_i por $\bar{\omega}$.

Supongamos pues, ω_i, ω_j los extremos de h y sea h_0 el mínimo grupo de \mathbf{P} que los contiene.

Entonces, $\forall \omega \in h$ será $\omega_i \leq \omega \leq \omega_j$ y puesto que h_0 es conexo, $\omega \in h_0$, con lo cual, $h \subset h_0$

Veamos que esta inclusión no puede ser estricta. Para ello tengamos en cuenta que si d es la disimilaridad piramidal inducida por la pirámide (\mathbf{P}, \mathbf{i}) , $d(\omega_i, \omega_j) = \mathbf{i}(h_0)$.

Supongamos pues, $h \subsetneq h_0$. Esto significa que existe por lo menos un $\omega_0 \in h_0$ y $\omega_0 \notin h$.

Si $\omega_0 \in h_0$, puesto que por ser h_0 el mínimo grupo de \mathbf{P} que contiene ω_i, ω_j éstos serán sus extremos (Prop.4.3.2), $\forall x \in h_0$ $d(\omega_0, x) \leq d(\omega_i, \omega_j) = \mathbf{i}(h_0)$. En particular para $x = \omega_j$, $d(\omega_0, \omega_j) \leq \mathbf{i}(h_0)$.

Por otra parte, si $h = \{\text{individuos a interdistancias } \leq \mathbf{i}(h_0)\}$, según la Proposición 4.1.1, será $h = [\omega_i, \omega_j]_{\mathbf{i}(h_0)}$ maximal (ver nota pie de página 48), por tanto si $\omega_0 \notin h$, tendrá que ser $\omega_0 < \omega_i \leq \omega_j$ (ó bien $\omega_i \leq \omega_j < \omega_0$). Por ser d piramidal, $\forall x \in h$ $d(\omega_0, x) \leq d(\omega_0, \omega_j)$, y teniendo en cuenta la desigualdad anterior concluiremos que $d(\omega_0, x) \leq \mathbf{i}(h_0)$. Por tanto, tenemos que $\omega_0 \notin h$ y $\forall x \in h$, $d(\omega_0, x) \leq \mathbf{i}(h_0)$, con lo cual h no sería maximal.

Así pues, la inclusión no puede ser estricta, por tanto $h=h_0$ y en consecuencia, $h \in P$.

Finalmente, si $h=h_1 \cap h_2$, con $h_i \in P'$ y $h_i \neq h$, para $i \in \{1,2\}$, puesto que acabamos de demostrar que $P' \subset P$, resultará que h será intersección de dos grupos de P , por tanto será $h \in P$.

Ejemplo 4.3.1.

Con este ejemplo ilustraremos el resultado establecido en la proposición anterior.

Sea $\Omega = \{1,2,3,4,5,6\}$ y consideremos la siguiente pirámide indexada sobre Ω :

$$P = \{h_0^i = \{i\}_{i \in \Omega}, h_1^1 = \{1,2,3,4\}, h_1^2 = \{3,4,5,6\}, h_2^1 = \{3,4\}, h_2^2 = \{3,4,5\}, h_2^3 = \{1,2,3,4,5\}, h_3 = \Omega\}$$

Si el subíndice de cada grupo indica su índice, la pirámide será indexada en sentido amplio, y la disimilaridad piramidal que induce, según el teorema 4.2.2, es la siguiente:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 3 \\ & 0 & 1 & 1 & 2 & 3 \\ & & 0 & 1 & 1 & 1 \\ & & & 0 & 1 & 1 \\ & & & & 0 & 1 \end{bmatrix}$$

En este caso, $P' = \{h_0^1, h_0^2, h_0^3, h_0^4, h_0^5, h_0^6, h_1^1, h_1^2, h_2, h_3\}$.

Consideremos el grupo h_1^3 , que no pertenece a P' . Como puede verse, existe otro grupo de P , h_1^4 , del mismo índice que h_1^3 y que lo contiene estrictamente. Así pues por ser la pirámide, indexada en sentido amplio, h_1^3 será intersección no trivial de dos grupos de P , por ejemplo: $h_1^3 = h_1^4 \cap h_1^1$ con $h_1^4 \notin P'$, $h_1^1 \in P'$.

Puesto que $h_1^4 \notin P'$, observemos que existe un grupo de su mismo índice, h_1^2 y que lo contiene estrictamente, por tanto h_1^4 será intersección no trivial de dos grupos de P , por

ejemplo: $h_1^4 = h_1^2 \cap h_2$ y éstos son ya grupos de P' .

Finalmente pues, h_1^3 será intersección de dos grupos de P' :

$$h_1^3 = h_1^2 \cap h_2 \cap h_1^1 = h_1^2 \cap h_1^1$$

Teorema 4.3.1.

Dada una pirámide indexada en sentido amplio de $\mathcal{P}(\Omega)$, la disimilaridad piramidal que induce, es única.

Demostración:

Para ello, vamos a ver que si una pirámide indexada en sentido amplio, (P, i) induce una disimilaridad piramidal d' , y a su vez es la pirámide inducida por otra disimilaridad piramidal d , entonces $d = d'$.

Si d' es la disimilaridad piramidal inducida por (P, i) , $\forall x, y \in \Omega$, $d'(x, y) = i(h_0)$, siendo h_0 el mínimo grupo que contiene a x e y .

Si (P, i) es la pirámide inducida por d , entonces $i(h_0) = \max\{d(\omega, \omega'), \forall \omega, \omega' \in h_0\}$.

Por tanto $i(h_0) \geq d(x, y)$.

De donde, $d'(x, y) \geq d(x, y)$.

Vamos a ver que esta desigualdad no puede ser estricta.

Supongamos $d'(x, y) > d(x, y)$. Si $\alpha = d(x, y)$, consideremos el grupo $h_\alpha \in P$, formado por los individuos de Ω situados entre dos a distancia α y maximal, ($h_\alpha \in C_\alpha^*$).

Es claro que $x, y \in h_\alpha$ (no tienen por qué ser los extremos).

Si (P, i) viene inducida por d , $i(h_\alpha) = \alpha < d'(x, y) = i(h_0)$.

Ahora bien, si h_0 es el mínimo grupo de P que contiene x e y , en virtud de la proposición 4.3.1. es también el de menor índice de entre todos los grupos que

contienen x e y , por tanto no puede ser que $i(h_\omega) < i(h_0)$ y la desigualdad $d'(x,y) \geq d(x,y)$ no puede ser estricta, así pues tendrá que ser $d'(x,y) = d(x,y)$ para cualquier par, x e y , de individuos de Ω , y por tanto $d = d'$ c.q.d.

Teorema 4.3.2.

Dada una disimilaridad piramidal d , la pirámide indexada en sentido amplio que induce, es única.

Demostración:

Para ello, veremos que si la pirámide (P, i) induce la disimilaridad piramidal d , y a su vez, ésta induce la pirámide (\tilde{P}, \tilde{i}) , entonces $P = \tilde{P}$ y $i = \tilde{i}$.

Si (\tilde{P}, \tilde{i}) es la pirámide indexada inducida por d , por el teorema 4.2.1. sabemos que:

$$h \in \tilde{P} \Leftrightarrow \begin{cases} h \in \tilde{P}^* \\ \text{o} \\ \exists h_1, h_2 \in \tilde{P}^*, \text{ distintos de } h / h = h_1 \cap h_2 \end{cases}$$

Por otra parte, puesto que gracias a la proposición 4.1.1, $\tilde{P}^* = P'$, resultará :

$$h \in \tilde{P} \Leftrightarrow \begin{cases} h \in P' \\ \text{o} \\ \exists h_1, h_2 \in P', \text{ distintos de } h / h = h_1 \cap h_2 \end{cases}$$

Y por la proposición 4.3.3. esto equivale a que $h \in P$.

Para ver que $i = \tilde{i}$, tendremos en cuenta que $P = \tilde{P}$ y demostraremos que $\forall h \in P$, $i(h) = \tilde{i}(h)$. En efecto,

Si (\tilde{P}, \tilde{i}) es la pirámide indexada inducida por d , para cualquier $h \in P = \tilde{P}$,

$\tilde{i}(h) = \max\{d(\omega, \omega'), \forall \omega, \omega' \in h\}$, por tanto existirán $x, y \in h$ tales que $\tilde{i}(h) = d(x, y)$.

Por otra parte, sean ω_i y ω_j los individuos extremos de h . Puesto que $h \in \mathbf{P}$ y \mathbf{P} es la pirámide que induce d , si ω_i y ω_j son los individuos más alejados de h , éste será el mínimo grupo de \mathbf{P} que los contendrá (Proposición 4.3.2), y por tanto $i(h) = d(\omega_i, \omega_j)$.

Así pues, hasta el momento tenemos que: $x, y \in h \in \mathbf{P}$; ω_i, ω_j son los extremos de h ; y d es la disimilaridad piramidal inducida por (\mathbf{P}, i) ; con lo cual $d(x, y) \leq d(\omega_i, \omega_j)$.

Finalmente puesto que $d(x, y) = \max\{d(\omega, \omega'), \forall \omega, \omega' \in h\}$, tendremos que $d(x, y) \geq d(\omega_i, \omega_j)$.

De ambas desigualdades resultará: $d(x, y) = d(\omega_i, \omega_j)$ y por lo tanto $\forall h \in \mathbf{P}$, $i(h) = \tilde{i}(h)$
c.q.d.

CAPITULO 5

CONSTRUCCION DE GRAFOS PIRAMIDALES. ALGORITMOS

5.1.- INTRODUCCION

En el capítulo 4, el teorema 4.2.1. nos da un método para construir una pirámide a partir de una disimilaridad piramidal. En la realidad, no obstante, la disimilaridad inicial, observada sobre un conjunto de individuos u objetos a clasificar, no será, por lo general, piramidal. En estas condiciones pues, entenderemos un algoritmo de construcción de grafos piramidales, como un método de transformación de una disimilaridad cualquiera en una disimilaridad piramidal.

En este sentido, E. Diday nos indica que sería muy útil dedicar algunos esfuerzos a convertir los métodos de clasificación y representación piramidal, en instrumentos de utilidad práctica, como lo son actualmente los ultramétricos, sugiriendo que, para ello, sería necesario un esfuerzo desde el punto de vista teórico, en el sentido de intentar simplificar las representaciones piramidales. (depuración de los grupos

sobrantes, simplicidad de la representación visual...)

En el presente capítulo, siguiendo las indicaciones mencionadas, describiremos, básicamente, un algoritmo de clasificación piramidal, inspirado en el del propio Diday (1984-1) y convenientemente modificado, en el sentido de propiciar la obtención de una pirámide indexada lo más sencilla posible, desde el punto de vista de la representación visual, con el fin de facilitar sus aplicaciones prácticas.

Este interés por presentar el tema de manera que las ventajas de las clasificaciones piramidales, con respecto a otras más conocidas y utilizadas, puedan ser aprovechadas en la resolución de problemas prácticos, nos ha conducido a la programación informática del algoritmo descrito en la sección 5.2, por los métodos del Mínimo, Máximo y UPGMA.

Una pirámide, construida por alguno de los algoritmos de clasificación, puede contener información redundante. En la sección 5.3, se precisa este último concepto y se dan los criterios para depurar una pirámide.

Finalmente, un sencillo experimento de simulación, nos mostrará, de forma empírica, la mejora que suponen las clasificaciones piramidales respecto a las jerárquicas, en el sentido de que las primeras son, por lo general, mucho más próximas a los datos iniciales que las segundas.

5.2.- ALGORITMO DE CLASIFICACION ASCENDENTE PIRAMIDAL (CAP)

Este algoritmo, inspirado en el de Clasificación Ascendente Jerárquica, debe tener en cuenta las características propias de las pirámides, a saber: cada grupo puede tener, a lo sumo, dos predecesores (por consiguiente, cada grupo podrá ser unido con otro, en sentido ascendente, hasta dos veces) y debe existir un preorden total sobre el conjunto de individuos a clasificar, que sea compatible con la pirámide, es decir, que los grupos formados sean conexos respecto a dicho preorden.

Sea pues $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ un conjunto finito sobre el que tenemos definida una disimilaridad inicial δ_0 .

Fijado un determinado índice de agregación (disimilaridad entre grupos), definimos el siguiente Algoritmo de Clasificación Ascendente Piramidal:

A.1.- Iniciamos el proceso con el recubrimiento $R_0 = \{\{\omega_1\}, \dots, \{\omega_n\}\}$, y la disimilaridad δ_0 .

A.2.- Si $R_{k-1} = \{h_1, \dots, h_p\}$, $k \in \mathbb{N} - \{0\}$, es el recubrimiento correspondiente al paso $k-1$ y δ_{k-1} la disimilaridad sobre los grupos de dicho recubrimiento:

* Unimos los dos grupos de R_{k-1} a distancia mínima, y obtenemos un nuevo recubrimiento R_k .

Si h_i y h_j son grupos de R_{k-1} tales que $\delta_{k-1}(h_i, h_j)$ es mínima, el nuevo recubrimiento será:

$$R_k = \{h_1, \dots, h_p, h_i \cup h_j\}$$

Teniendo en cuenta que, después de unir h_i con h_j , cualquiera de

los dos que ya haya sido unido dos veces, no pertenecerá a R_k .

* Definimos el índice de un grupo $h \in R_k$, como:

$$i(h) = \begin{cases} 0 & \text{si } h = \{\omega_i\}, \forall i = 1, \dots, n \\ \delta_{k-1}(h_i, h_j) & \text{si } h = h_i \cup h_j \end{cases} \quad (2)$$

* Finalmente, definimos una nueva disimilaridad, δ_k , entre los grupos del nuevo recubrimiento R_k , a partir de un cierto criterio (índice de agregación), previamente escogido (ver §5.4).

En este proceso, tendremos en cuenta las siguientes condiciones:

A.2.1.- Debe definirse un preorden total sobre Ω , de manera que cada uno de los grupos formados sea conexo respecto dicho preorden.

A.2.2.- Si $h_r, h_s \in R_{k-1}$, $h_r \subsetneq h_s$, $\delta_{k-1}(h_r, h) = \delta_{k-1}(h_s, h)$, y esta distancia es la mínima entre grupos de R_{k-1} susceptibles de ser unidos⁽³⁾, entonces uniremos h_s con h .

A.2.3.- Si $h_r, h_s \in R_k$; $h_r \subsetneq h_s$ y $i(h_r) = i(h_s)$, entonces eliminamos h_r del recubrimiento R_k .

⁽²⁾La representación visual de una pirámide, es más fácilmente interpretable si el índice de cada grupo, se corresponde con el valor del índice de agregación entre los dos grupos que lo forman, es decir: $i(h_i \cup h_j) = \delta_{k-1}(h_i, h_j)$.

Podríamos preocuparnos por si esta forma de indexar la pirámide, pudiera dar lugar, en general, a inversiones. Sin embargo, en este caso, la forma en que se define el índice de agregación para clasificaciones piramidales, evita este supuesto. (ver Teorema 5.4.1)

⁽³⁾Un grupo es susceptible de ser unido con algún otro, si no ha sido unido dos veces con anterioridad, si contiene alguno de los extremos del mayor grupo que lo contiene ó si no es Ω .

Dos grupos son susceptibles de ser unidos, si cada uno por separado es susceptible de ser unido con algún otro y su unión es conexa.

A.2.4.- Si $h_r \in R_{k-1}$; h_s es el grupo formado en el paso k ; $h_r \subset h_s$ y h_r no contiene ninguno de los extremos de h_s , entonces h_r no interviene en la formación de R_k ($h_r \notin R_k$).⁽⁴⁾ (Fig.5.2.1).

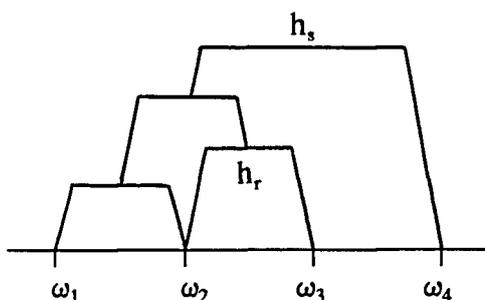


Fig. 5.2.1

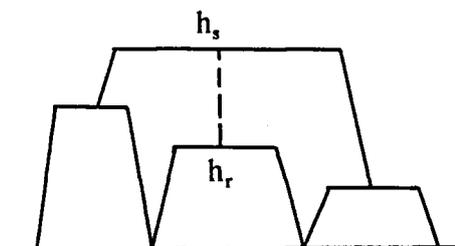


Fig. 5.2.2.

A.3.- Se repite el paso A.2. hasta obtener el recubrimiento $\{\Omega\}$.

Teorema 5.2.1.

El Algoritmo CAP construye una pirámide sin cruzamientos.

Demostración:

Sea P el conjunto de partes de Ω construidas por el algoritmo anterior, es decir:

$$P = \bigcup_{k=0,1,\dots} R_k$$

Veamos que P es una pirámide:

Las propiedades P.1. ($\Omega \in P$) y P.2. ($\{\omega\} \in P$), se satisfacen de forma evidente, puesto que R_0 contiene los singletons y el último recubrimiento es precisamente $\{\Omega\}$.

⁽⁴⁾Si h_s fuese el único predecesor de h_r , en la representación visual de la pirámide, deberemos unir h_r con h_s (Fig. 5.2.2.).

Por otra parte, la condición A.2.1. nos asegura la existencia de un preorden total sobre Ω compatible con P, tal como exige la propiedad P.4.

Verifiquemos pues finalmente la propiedad P.3 :

Sean h y h' dos grupos de P, hemos de ver que: $h \cap h' = \emptyset$ o bien $h \cap h' \in P$.

Supongamos que h y h' son grupos de P, tales que $h \cap h' \neq \emptyset$.

En este caso, existirá algún grupo de P contenido en $h \cap h'$, (por lo menos algún $\{\omega\} \subset h \cap h'$).

Sean pues, h_1, \dots, h_r grupos de P contenidos en $h \cap h'$ y supongamos, además, que sean maximales en el sentido de la inclusión, es decir que no existe ningún $h'_k \in P$, para $k \in \{1, 2, \dots, r\}$, tal que $h_k \subsetneq h'_k \subset h \cap h'$ (en estas condiciones, demostraremos que todos estos grupos se reducen a uno solo).

Supongamos que $h \subsetneq h'$, puesto que si $h \subset h'$, tendríamos $h \cap h' = h \in P$, con lo cual quedaría demostrada la propiedad. (También podríamos suponer $h' \subsetneq h$).

Sean ω'_i y ω'_j los individuos extremos de h' según el orden inducido por la condición A.2.1. del algoritmo. Por otra parte, gracias a la condición A.2.4. para cualquier $k \in \{1, \dots, r\}$, h_k deberá contener ω'_i ó ω'_j .

Si alguno de los h_k contiene a la vez ω'_i y ω'_j , puesto que éstos son los extremos de h' , y h_k es conexo por ser de P, tendremos $h' \subset h_k$, con lo cual $h' = h_k$ y por tanto $h \cap h' = h \cap h_k = h_k = h' \in P$.

Supongamos pues que ningún h_k contiene a la vez ω'_i y ω'_j . Por ejemplo, supongamos que para cualquier $k \in \{1, \dots, r\}$, $\omega'_i \in h_k$ y $\omega'_j \notin h_k$. Consideremos entonces que $\omega' \in \Omega$ sea el individuo más alejado de ω'_i y que pertenezca a alguno de los h_k , para $k \in \{1, \dots, r\}$.

Sea h_0 uno de estos h_k que contienen ω'_i y ω' . Puesto que $h_0 \in P$, la condición A.2.1. del algoritmo nos asegura que h_0 contiene todos los individuos entre ω'_i y ω' , y por tanto, para cualquier $k \in \{1, \dots, r\}$, $h_k \subset h_0$.

Por otra parte, puesto que h_0 es uno de los h_k , $h_0 \subset h \cap h'$. Ahora bien, por hipótesis, h_1, \dots, h_r son los grupos más grandes de P contenidos en $h \cap h'$, con lo cual si $h_k \subset h_0 \subset h \cap h'$, la inclusión $h_k \subset h_0$ no puede ser estricta, así pues, $\forall k \in \{1, \dots, r\}$ $h_k = h_0$.

Por tanto, existirá un único grupo de P , h_0 , contenido en $h \cap h'$ y maximal en el sentido de la inclusión, con lo cual $h \cap h' = h_0 \in P$, que es lo que nos faltaba para demostrar que P era una pirámide.

Finalmente, las condiciones A.2.1. y A.2.4. aseguran la existencia de un preorden total sobre Ω respecto al cual la pirámide no presenta cruzamientos (una arista no corta nunca a otra arista o a un segmento horizontal en la representación visual de la pirámide, definición 3.4.2).

Observese que, en general, la pirámide obtenida a partir del algoritmo CAP, **no es única.**

Proposición 5.2.1.

Si en el paso k -ésimo del algoritmo, unimos los grupos h_1 y h_2 , y en el paso $(k+1)$ -ésimo $h_1 \cup h_2$ es susceptible de ser unido con un $h' \in R_k$, entonces $h' \in R_{k-1}$ y en el paso k -ésimo h' era susceptible de ser unido con h_1 o con h_2 .

Demostración:

Supongamos pues, $h_1, h_2 \in R_{k-1}$; $h = h_1 \cup h_2 \in R_k$; $h \notin R_{k-1}$.

Sea h' un grupo de R_k que en el paso $(k+1)$ -ésimo sea susceptible de ser unido con h .

Puesto que h es el único grupo de R_k que no es de R_{k-1} , ha de ser $h' \in R_{k-1}$. Así pues:

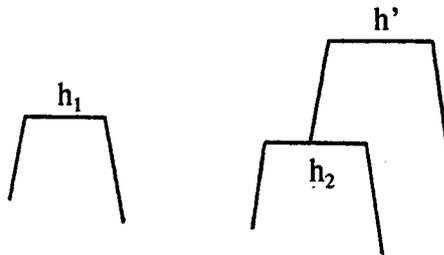
$$R_{k-1} = \{\dots, h', \dots, h_1, \dots, h_2, \dots\}; \quad R_k = \{\dots, h', \dots, h_1 \cup h_2, \dots\}$$

En estas hipótesis vamos a distinguir dos casos:

a) $h_1 \subsetneq h'$ ó $h_2 \subsetneq h'$

Observese, en primer lugar, que no puede darse el caso $h_1 \subsetneq h'$ y $h_2 \subsetneq h'$, puesto que entonces $h_1 \cup h_2 \subset h'$ y no podríamos unir h' con $h_1 \cup h_2$.

Supongamos, por ejemplo, $h_2 \subsetneq h'$ y $h_1 \not\subset h'$. En este caso, después del paso $k-1$, tendríamos la situación siguiente:



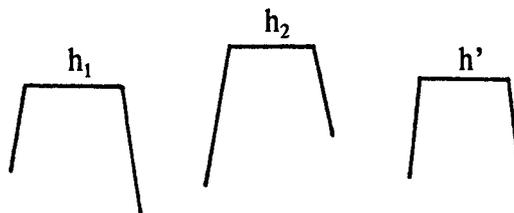
Es evidente que si h_1 y h_2 son susceptibles de ser unidos, h' es susceptible de ser unido con h_1 .

b) $h_1 \not\subset h'$ y $h_2 \not\subset h'$

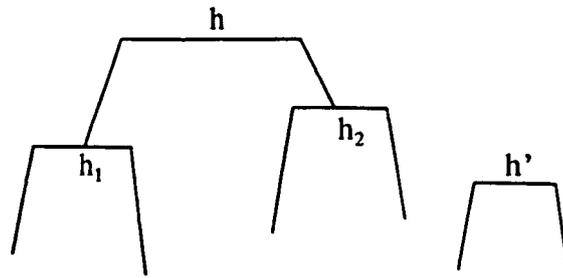
En este caso deben contemplarse las siguientes posibilidades:

b₁) $h_i \cap h' = \emptyset, \forall i \in \{1, 2\}$

Después del paso $k-1$ la situación será:

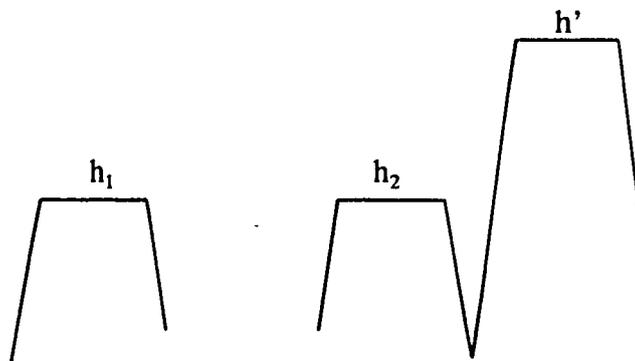


Y después del paso k tendremos:

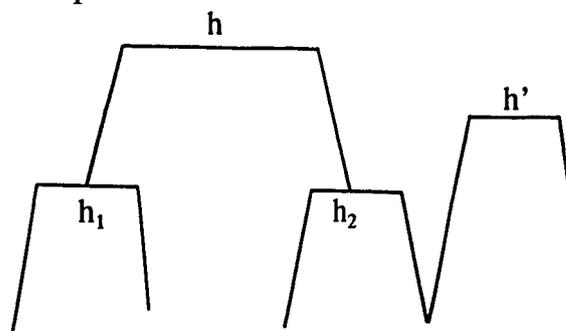


b₂) $h_i \cap h' \neq \emptyset$, para $i=1$ ó $i=2$

Después del paso k-1, si suponemos $h_2 \cap h' \neq \emptyset$, tendremos la siguiente situación:



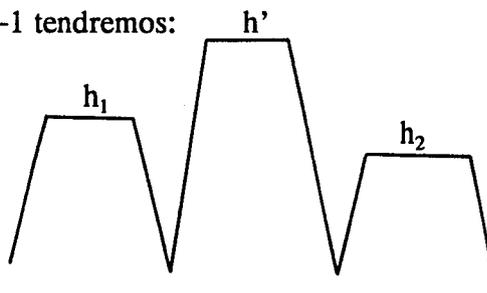
Y después del paso k:



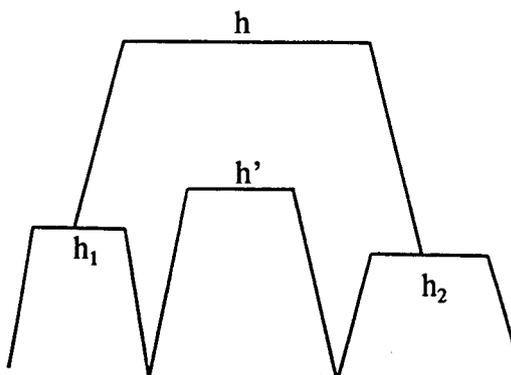
En ambos casos, si h' es susceptible de ser unido con h en el paso $k+1$, entonces en el paso k , h' ya era susceptible de ser unido con h_2 (o bien con h_1 , depende del caso).

b₃) $h_i \cap h' \neq \emptyset, \forall i \in \{1,2\}$

Después del paso k-1 tendremos:



Y después del paso k, una vez unido h_1 con h_2 :



Si en el paso k unimos h_1 con h_2 , tendremos que $h' \subset h$, con lo cual en el paso k+1 no podremos unir h con h' . Así pues, esta situación, en las hipótesis del enunciado, no puede darse.

5.3.- DEPURACION DE UNA PIRAMIDE CONSTRUIDA POR ALGUNO DE LOS ALGORITMOS DE CLASIFICACION PIRAMIDAL

Dada una pirámide indexada (P,i) construida por alguno de los algoritmos de clasificación, podemos preguntarnos si existen grupos "sobrantes", es decir, grupos que proporcionen información redundante sobre la clasificación y por tanto que puedan ser

eliminados de la pirámide sin que varíe la clasificación piramidal.

Para responder a la pregunta con el máximo rigor, será necesario, en primer lugar, precisar el concepto de "grupo sobrante" en una pirámide indexada.

Si (P, i) es una pirámide indexada en sentido amplio, es evidente que ningún grupo puede ser sobrante, puesto que a dicha pirámide le corresponde biunivocamente una disimilaridad piramidal d (§4.2 y 4.3), por tanto si eliminamos algún grupo h de P , a pesar de que $P - \{h\}$ continuase siendo una pirámide indexada en sentido amplio, induciría otra disimilaridad piramidal d' distinta a d y por tanto una clasificación también distinta. Si por el contrario, la pirámide es indexada en sentido estricto, también lo será en sentido amplio (prop. 3.2.1) y por consiguiente tampoco habrá grupos susceptibles de ser eliminados.

Sea pues (P, i) una pirámide indexada pero no en sentido amplio. En este caso, existirán grupos $h, h' \in P$; $h \subsetneq h'$; $i(h) = i(h')$ y h no será intersección no trivial de grupos de P .

Consideremos el conjunto:

$\Sigma_P = \{h \in P / \exists h' \in P, h \subsetneq h', i(h) = i(h') \text{ y } h \text{ no es intersección no trivial de grupos de } P\}$
que será no vacío, por ser la pirámide indexada no en sentido amplio.

En estas condiciones, es evidente que $P - \Sigma_P$ será también una pirámide indexada y además, en sentido amplio.

Si finalmente demostramos que la disimilaridad piramidal inducida por (P, i) es la misma que la inducida por $(P - \Sigma_P, i)$, quedará demostrado que los grupos de Σ_P son realmente los grupos sobrantes de la pirámide P , puesto que P y $P - \Sigma_P$ serán pirámides equivalentes, en el sentido de tener asociada la misma disimilaridad piramidal, y al ser

$(\mathbf{P}-\Sigma_p, i)$ indexada en sentido amplio, no tendrá ningún grupo sobrante, con lo cual podremos asegurar que los grupos de Σ_p serán los únicos grupos sobrantes de \mathbf{P} .

Teorema 5.3.1.

La pirámide indexada sobre Ω (\mathbf{P}, i) , induce la misma disimilaridad piramidal que la pirámide $(\mathbf{P}-\Sigma_p, i)$.

Demostración:

Sabemos que cualquier pirámide indexada de $\mathcal{P}(\Omega)$, sea en sentido amplio o no, induce una disimilaridad piramidal sobre Ω , sin más que definir la distancia entre dos individuos como el índice del menor grupo que los contiene.

Sea pues Ω un conjunto finito y (\mathbf{P}, i) una pirámide indexada. Si lo es en sentido amplio, entonces $\Sigma_p = \emptyset$ y no hay nada que demostrar.

Supongamos pues que (\mathbf{P}, i) es indexada pero no en sentido amplio, con lo cual $\Sigma_p \neq \emptyset$ y $\mathbf{P}-\Sigma_p \subsetneq \mathbf{P}$.

Sean d y d' las disimilaridades piramidales inducidas respectivamente por (\mathbf{P}, i) y $(\mathbf{P}-\Sigma_p, i)$.

Sean x e y dos individuos cualesquiera de Ω , y sea h_0 el mínimo grupo de \mathbf{P} que los contiene. Entonces será $d(x, y) = i(h_0)$.

Si $h_0 \notin \Sigma_p$, entonces $h_0 \in \mathbf{P}-\Sigma_p$, y por tanto $d'(x, y) = i(h_0)$, con lo cual tendremos $d(x, y) = d'(x, y)$, que es lo que queríamos ver.

Si por el contrario, $h_0 \in \Sigma_p$, $h_0 \notin \mathbf{P}-\Sigma_p$, y además existirán otros grupos de \mathbf{P} que contendrán estrictamente h_0 y de su mismo índice. Estos otros grupos lo serán, naturalmente, en número finito, y no todos ellos podrán ser de Σ_p (por lo menos

$\Omega \notin \Sigma_P$).

Sea pues, h'_0 el menor de los grupos de P tales que $h_0 \subseteq h'_0$, $i(h_0) = i(h'_0)$ y $h'_0 \notin \Sigma_P$.

En estas condiciones pues, h'_0 será el menor grupo de $P - \Sigma_P$ que contendrá h_0 y por tanto el menor grupo de $P - \Sigma_P$ que contendrá x e y . Así pues, $d'(x,y) = i(h'_0) = i(h_0)$ y por tanto $d(x,y) = d'(x,y)$ que es lo que queríamos demostrar.

Resumiendo pues, la caracterización de los grupos sobrantes de una pirámide indexada será la siguiente:

* $h \in (P,i)$ es un grupo sobrante sii $h \in \Sigma_P$. Es decir

Si $\exists h' \in (P,i)$ tal que:

$$\left[\begin{array}{l} h \subseteq h' \\ i(h) = i(h') \\ h \text{ no es intersección, no trivial, de grupos de } P \end{array} \right.$$

La **Depuración** de una Pirámide consistirá en **Eliminar los Grupos Sobrantes**.

5.3.1.- Visualización de la Depuración

Teniendo en cuenta la caracterización que acabamos de dar para los grupos sobrantes de una pirámide indexada, y su representación visual, vamos a visualizar también su depuración.

Simplemente observado la representación visual de una pirámide indexada, un grupo h sólo tendrá posibilidades de ser eliminado (sobrante) si, en el mismo segmento

horizontal aparece otro grupo, h' (esto es, si existe otro grupo que lo contiene y de su mismo índice). Pero esto no es suficiente, puesto que además h' no puede ser intersección no trivial de grupos de P . Esto implica que h no puede tener otro predecesor más que h' , puesto que en caso contrario h sería la intersección de sus dos predecesores. Visualmente esto implica que de h no puede salir ninguna arista. Ver Fig.5.3.1.

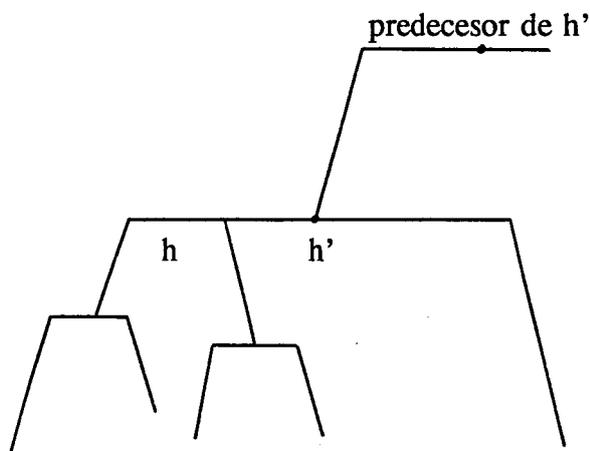


fig.5.3.1

Una vez efectuada la depuración de la pirámide, puede suceder que varias aristas indiquen la inclusión de un mismo grupo en otro, la cual cosa es innecesaria y además complica la representación visual de la pirámide.

En el ejemplo de la figura 5.3.2, h es un grupo sobrante, por tanto una vez depurada la pirámide, la arista a_3 indicará la inclusión de h_1 en h' y las aristas a_1 y a_2 indicarán, respectivamente, $h_1 \subset h_2$ y $h_2 \subset h'$. Después de la depuración pues, la arista a_3 será del todo innecesaria:

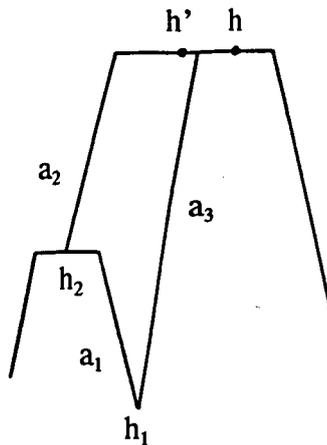


Fig.5.3.2

Así pues, es posible que después de eliminar los grupos sobrantes de una pirámide, desde el punto de vista de la representación visual, existan también **aristas sobrantes**, que deberán ser también eliminadas. Para ello, el criterio a seguir será el siguiente:

Después de la depuración de una pirámide indexada, cualquier arista que no una un grupo con su predecesor es sobrante.

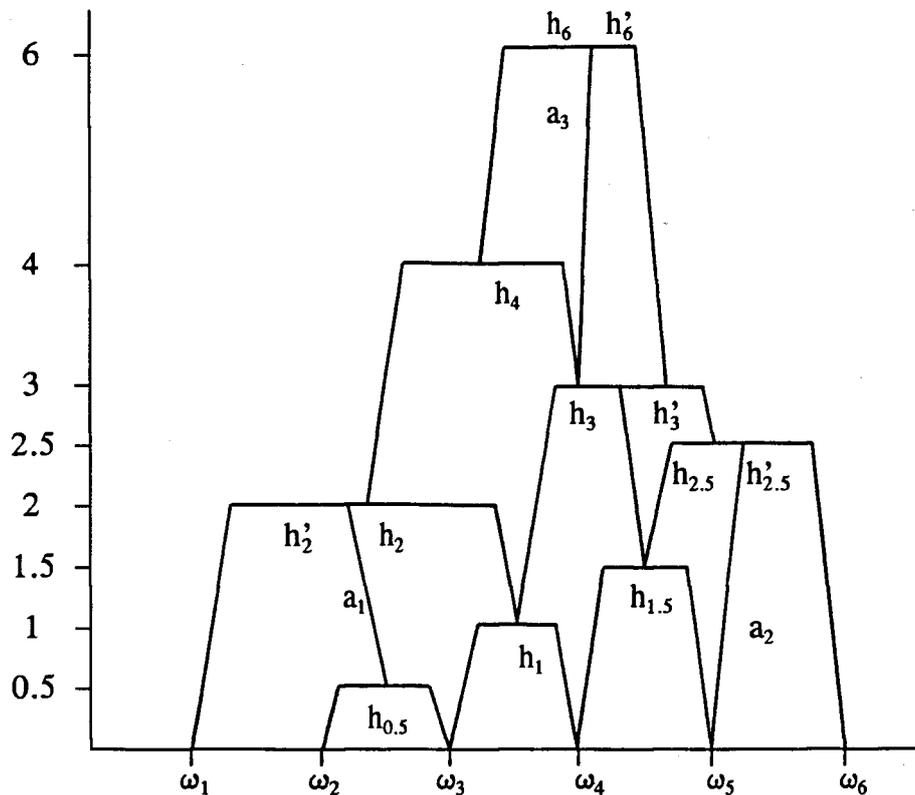
Ejemplo 5.3.1.

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$ y consideremos la pirámide siguiente:

$$\begin{aligned}
 \mathbf{P} = & \{\{\omega_1\}, \dots, \{\omega_6\}, h_{0.5} = \{\omega_2, \omega_3\}, h_1 = \{\omega_3, \omega_4\}, h_2 = \{\omega_1, \omega_2, \omega_3, \omega_4\}, h_2' = \{\omega_1, \omega_2, \omega_3\}, \\
 & h_{1.5} = \{\omega_4, \omega_5\}, h_{2.5}' = \{\omega_5, \omega_6\}, h_{2.5} = \{\omega_4, \omega_5, \omega_6\}, h_3 = \{\omega_3, \omega_4, \omega_5\}, h_3' = \{\omega_3, \omega_4, \omega_5, \omega_6\}, \\
 & h_4 = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, h_6' = \{\omega_3, \omega_4, \omega_5, \omega_6\}, h_6 = \{\Omega\}
 \end{aligned}$$

Indexada por los subíndices indicados en la nomenclatura de cada grupo, excepto los $\{\omega_i\}$ que tienen índice cero.

Su representación visual será la siguiente:



Observese que (P, i) es una pirámide indexada pero ni en sentido amplio ni en sentido estricto.

En este caso los grupos sobrantes, según el criterio establecido anteriormente, serán: $\Sigma_P = \{h'_2, h'_{2.5}, h'_6\}$ y naturalmente, $(P - \Sigma_P, i)$ será una pirámide indexada en sentido amplio.

Una vez eliminados los grupos sobrantes, veamos si alguna de las aristas afectadas puede ser también eliminada:

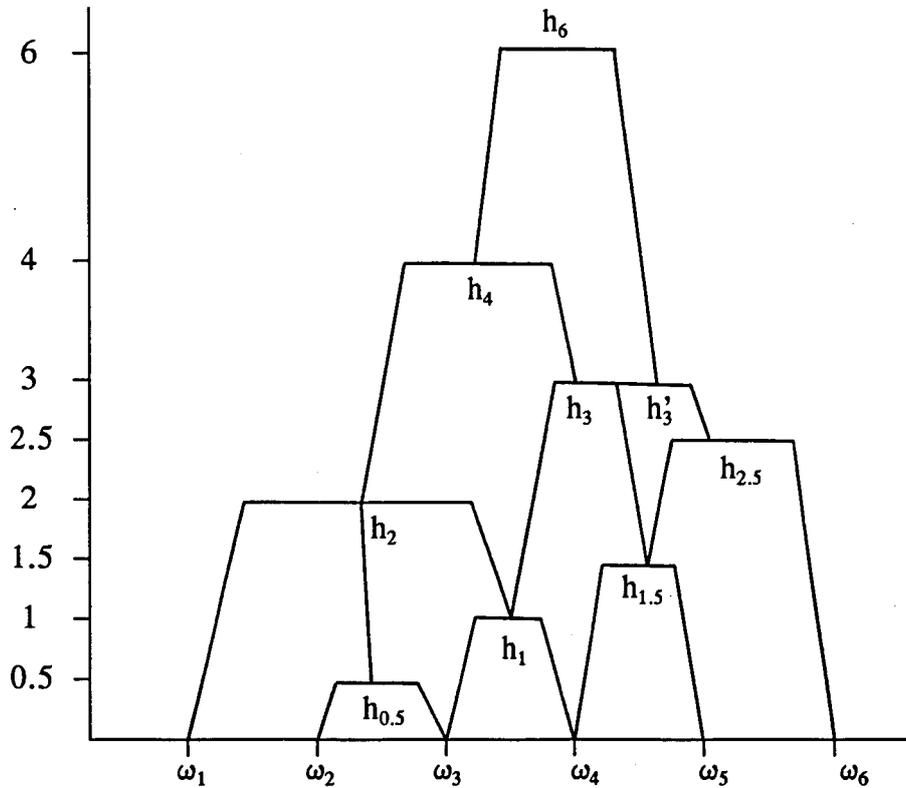
Arista a_1 : No puede ser eliminada puesto que indica la inclusión de h_1 en h_2 que es su predecesor. Además es la única arista que indica esta inclusión.

Arista a_2 : Puede ser suprimida puesto que, una vez eliminado el grupo $h'_{2.5}$, esta arista

indicaría la inclusión de $\{\omega_5\}$ en $h_{2.5}$, que no es su predecesor.

Arista a_3 : También puede ser suprimida, por el mismo motivo que la a_2 .

Después de la depuración, la representación visual de la pirámide será la siguiente:



Observación: La supresión de una arista, en la representación visual de una pirámide, comporta siempre la desaparición de un grupo, pero no recíprocamente.

5.4.- PRINCIPALES ALGORITMOS. PROPIEDADES

Tal y como hemos comentado en el §5.2, el algoritmo CAP necesita de un determinado índice de agregación para poder definir, a cada paso, la disimilaridad entre

los grupos del recubrimiento correspondiente. La elección de este índice de agregación será la que, en definitiva, diferenciará un algoritmo de otro.

Un índice de agregación sobre $\mathcal{P}(\Omega)$, no es más que una aplicación δ , de $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$ en \mathbb{R}^+ tal que para cualesquiera $h, h' \in \mathcal{P}(\Omega)$: $\delta(h, h') \geq 0$ y $\delta(h, h') = \delta(h', h)$.

Recordemos también, que los índices de agregación utilizados en clasificaciones jerárquicas, son de la forma siguiente:

para cualesquiera h y h' clases de una determinada partición P_k ,

$$\delta_k(h, h') = \begin{cases} \delta_{k-1}(h, h') & \text{si } h, h' \in P_{k-1} \\ f(\delta_{k-1}(h, h_1), \delta_{k-1}(h, h_2), \delta_{k-1}(h_1, h_2), \dots) & \text{si } h' = h_1 \cup h_2, \text{ con } h_1 \in P_{k-1} \end{cases}$$

Donde f es la función que determina el índice en cada caso. (Min, Max, ...)

En nuestro caso, puesto que en lugar de particiones tendremos recubrimientos, y por tanto los grupos podrán tener intersección no vacía, será conveniente tener en cuenta esta circunstancia a la hora de formular convenientemente la definición del índice de agregación.

Sea pues, $R_0 = \{\{\omega_1\}, \dots, \{\omega_n\}\}$ el recubrimiento inicial en un proceso de clasificación piramidal y δ_0 la disimilaridad dada entre los grupos del mismo.

Sea R_k el recubrimiento correspondiente al paso k -ésimo del algoritmo. Definimos la nueva disimilaridad δ_k entre los grupos de R_k de la forma siguiente:

Dados cualesquiera $h, h' \in R_k$,

$$\delta_k(h, h') = \left[\begin{array}{ll} 0 & \text{si } h \cap h' \in \{h, h'\} \\ \delta_{k-1}(h, h') & \text{si } h, h' \in R_{k-1} \\ f(\delta_{k-1}(h, h_1), \dots) & \text{si } f(\delta_{k-1}(h, h_1), \dots) \geq i(h') \\ i(h') & \text{en caso contrario} \end{array} \right] \text{ si } h' = h_1 \cup h_2, \text{ con } h_i \in R_{k-1}$$

Teorema 5.4.1.

Toda pirámide indexada $(P-\Sigma_p, i)$ construida por el algoritmo CAP, escogiendo previamente cualquier índice de agregación como el que acabamos de definir, **no presenta inversión y es indexada en sentido estricto.**

Demostración:

Sea (P, i) una pirámide indexada construida por el algoritmo CAP.

Supongamos que en el paso k -ésimo del algoritmo unimos los grupos h_1 y h_2 , pertenecientes al recubrimiento R_{k-1} , para formar un nuevo grupo h' . Recordemos que $i(h') = \delta_{k-1}(h_1, h_2)$.

Para demostrar que la pirámide no presenta inversión, veremos que cada vez que formamos un nuevo grupo, su índice es mayor o igual que el índice de cualquier grupo formado con anterioridad.

En efecto, si tenemos en cuenta la expresión dada para un índice de agregación cualquiera, al calcular la nueva disimilaridad δ_k entre los grupos de R_k , resultará que para cualquier $h \in R_k$ ($h \subset h'$), $\delta_k(h, h') \geq i(h')$. Por lo tanto en el siguiente paso, al formar un nuevo grupo h'' (uniendo los dos más próximos de R_k) su índice será mayor o igual que $i(h')$, puesto que : si h'' es unión de h' con algún otro grupo, la distancia entre ambos será mayor o igual que $i(h')$ como acabamos de ver; y si ninguno de los

dos grupos que forman h'' es h' , querrá decir que estos grupos ya eran de R_{k-1} , con lo cual su interdistancia será mayor o igual que $\delta_{k-1}(h_1, h_2) = i(h')$.

Así pues, dados cualesquiera grupos p y p' construidos en algún paso del algoritmo CAP, si $p \subsetneq p'$, quiere decir que p ha sido formado antes que p' , por tanto $i(p) \leq i(p')$ y por tanto la pirámide no presentará inversión.

Finalmente veamos que, una vez depurada, la pirámide es indexada en sentido estricto.

Para ello, supongamos que existan $p, p' \in P$, tales que $p \subsetneq p'$ y $i(p) = i(p')$. En estas condiciones, el grupo p' habrá sido formado por la unión de p con algún otro grupo. Por otra parte, la condición A.2.3 del algoritmo nos asegura que, el grupo p no puede ser unido con ningún otro grupo. Por tanto p tendrá, a lo sumo, un predecesor y por consiguiente no será intersección no trivial de grupos de P , con lo cual será un grupo sobrante de la pirámide.

Así pues, una vez depurada la pirámide, no existirá ningún grupo contenido estrictamente en otro de su mismo índice, es decir si $p \subsetneq p'$, entonces $i(p) < i(p')$ y por tanto $(P - \Sigma_p, i)$ será una pirámide indexada en sentido estricto. c.q.d.

En este sentido pues, podemos decir que nuestro algoritmo de CAP da lugar siempre a pirámides indexadas en sentido estricto.

Notas:

- * Teniendo en cuenta los resultados de los Teoremas 5.2.1 y 5.4.1, podemos concluir que: **El algoritmo CAP construye pirámides indexadas en sentido estricto y sin cruzamiento.**

* Puesto que la condición A.2.3 del algoritmo es la que nos asegura que la indexación de la pirámide sea en sentido estricto, si suprimiéramos esta condición, la pirámide obtenida sería indexada en sentido amplio. Recuérdese, no obstante, que toda pirámide indexada en sentido estricto, lo es también en sentido amplio (Prop. 3.2.1).

Finalmente estudiaremos algunas particularidades de los métodos del mínimo y del máximo que provienen de una adecuada definición de la función f en la expresión del índice de agregación.

5.4.1.- Método del Mínimo

Definimos $f(\delta_{k-1}(h, h_1), \dots) = \min\{\delta_{k-1}(h, h_1), \delta_{k-1}(h, h_2)\}$

Teorema 5.4.1.1.

El algoritmo CAP por el método del mínimo, siempre da lugar a una jerarquía indexada única.

Demostración:

Ya sabemos que el algoritmo CAP, sea cual sea el índice de agregación escogido, siempre da lugar a una pirámide indexada.

Si queremos ver que el método del mínimo da lugar a una jerarquía, únicamente es necesario demostrar que cada grupo de la pirámide obtenida por este método, tiene

un único predecesor, (Ver Nota 3.5.1).

Para ello, supongamos que en un determinado paso del algoritmo, el k -ésimo por ejemplo, unimos los grupos h_1 y h_2 . Esto significa que h_1 y h_2 son grupos de R_{k-1} y que $h_1 \cup h_2 \in R_k - R_{k-1}$.

En estas condiciones, vamos a ver que ni h_1 ni h_2 pueden volver a unirse con ningún otro grupo.

En efecto, supongamos que en un paso posterior (en el siguiente para no complicar la notación), h_1 , por ejemplo, pueda ser unido con algún otro grupo $h \in R_k$ y que $\delta_k(h, h_1)$ sea mínima entre los pares de grupos de R_k susceptibles de ser unidos, ($h \in R_k$, $h_1 \in R_{k-1}$).

En estas condiciones:

$$\delta_k(h, h_1 \cup h_2) = \min\{\delta_{k-1}(h, h_1), \delta_{k-1}(h, h_2)\} \leq \delta_{k-1}(h, h_1) = \delta_k(h, h_1), \quad \text{con lo cual}$$

$$\delta_k(h, h_1 \cup h_2) \leq \delta_k(h, h_1).$$

Por otra parte, si h_1 y h son susceptibles de ser unidos, también lo serán $h_1 \cup h_2$ y h ($h_1, h, h_1 \cup h_2 \in R_k$); y si $\delta_k(h, h_1)$ es la mínima entre los pares de grupos de R_k susceptibles de ser unidos, habrá de ser $\delta_k(h, h_1 \cup h_2) = \delta_k(h, h_1)$, y puesto que $h_1 \subset h_1 \cup h_2$, la condición A.2.2 del algoritmo impone la unión de $h_1 \cup h_2$ con h y no la de h_1 con h .

Así pues, h_1 no tendrá otro predecesor que $h_1 \cup h_2$. Si en lugar de h_1 , hubiésemos supuesto que era h_2 quien podía ser unido con h , habríamos concluido, de la misma manera, es decir, que h_2 tampoco puede tener otro predecesor que $h_1 \cup h_2$.

Consecuentemente pues, ningún grupo de la pirámide construida por el algoritmo del mínimo tendrá dos predecesores, con lo cual la pirámide obtenida será, en realidad,

una jerarquía indexada, y por tanto, de la misma forma que sucede con el algoritmo del mínimo en clasificación jerárquica, la pirámide, jerarquía en este caso, obtenida es única.c.q.d.

Como consecuencia inmediata de este teorema, podemos decir que si deseamos efectuar una clasificación piramidal de los individuos de una población finita, el método del mínimo no será probablemente el más adecuado, puesto que nos proporcionaría, con toda seguridad, una clasificación jerárquica.

Ejemplo 5.4.1.

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ y δ_0 una disimilaridad sobre Ω , dada por la matriz:

$$\begin{bmatrix} 0 & 5 & 1 & 4 & 1 \\ & 0 & 2 & 6 & 4 \\ & & 0 & 5 & 3 \\ & & & 0 & 3.5 \\ & & & & 0 \end{bmatrix}$$

Vamos a efectuar una clasificación piramidal de los individuos de Ω , aplicando el algoritmo CAP por el método del mínimo.

Para ello, partimos del recubrimiento $R_0 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}\}$ y de la disimilaridad δ_0 entre los grupos de dicho recubrimiento.

Para unificar la notación llamemos h_i a los grupos $\{\omega_i\}$, para todo $i \in \{1, 2, 3, 4, 5\}$.

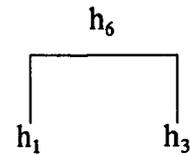
Empecemos pues, seleccionando los pares de grupos de R_0 a distancia mínima. En este caso son: $[h_1, h_3]$ y $[h_1, h_5]$.

Elegimos cualquiera de las dos posibilidades, $[h_1, h_3]$ por ejemplo, y formamos el primer nuevo grupo: $h_6 = h_1 \cup h_3 = \{\omega_1, \omega_3\}$

Puesto que la condición A.2.1. del algoritmo nos impone la construcción de un preorden total sobre Ω de manera que los grupos formados sean conexos, empezamos ordenando los dos individuos que forman h_6 . Definimos pues, $\omega_1 \leq \omega_3$.

Finalmente definimos el índice del nuevo grupo:

$$i(h_6) = \delta_0(h_1, h_3) = 1$$



Con todo ello, podemos formar un nuevo recubrimiento R_1 y definir una nueva disimilaridad sobre los grupos del mismo, δ_1 . (En la definición de δ_1 es precisamente donde interviene el índice de agregación correspondiente).

$$R_1 = \{h_1, h_6, h_3, h_2, h_4, h_5\}$$

0	0	1	5	4	1
	0	0	2	4	<u>1</u>
		0	2	5	3
			0	6	4
				0	3.5
					0

Los pares de grupos de R_1 susceptibles de ser unidos y a distancia mínima, son: $[h_1, h_3]$ y $[h_6, h_5]$ -distancias en negrilla-

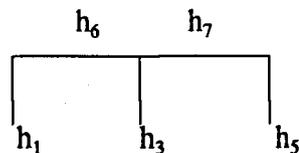
En este caso, la condición A.2.2. del algoritmo, nos impone unir h_6 con h_5 puesto que $h_1 \subset h_6$ y están a la misma distancia de h_5 , por tanto el nuevo grupo será:

$$h_7 = h_6 \cup h_5 = \{\omega_1, \omega_3, \omega_5\}$$

Una vez fijado $\omega_1 \leq \omega_3$, para que h_7 sea conexo deberá ser $\omega_1 \leq \omega_3 \leq \omega_5$ o bien

$\omega_5 \leq \omega_1 \leq \omega_3$. Escogemos la primera de las dos opciones.

Finalmente $i(h_7) = 1$



Con lo cual, el nuevo recubrimiento R_2 y la nueva disimilaridad δ_2 serán:

$$R_2 = \{ h_1, h_7, h_5, h_2, h_4 \}$$

0	0	1	5	4
	0	0	<u>2</u>	3.5
		0	4	3.5
			0	6
				0

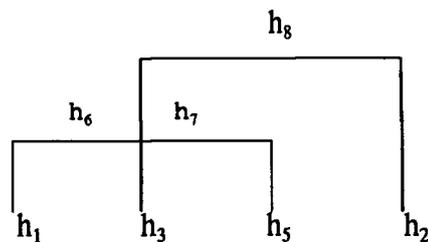
Observese que en el nuevo recubrimiento no aparece h_6 , puesto que $h_6 \subset h_7$ y son del mismo índice, por tanto, según la condición A.2.3, h_6 no debe formar parte del nuevo recubrimiento.

En este caso, el único par de grupos de R_2 susceptible de ser unidos y a distancia mínima, es $[h_7, h_2]$, por tanto el nuevo grupo será:

$$h_8 = h_7 \cup h_2 = \{ \omega_1, \omega_3, \omega_5, \omega_2 \}$$

Y definimos: $\omega_5 \leq \omega_2$

$i(h_8) = 2$



R_3 y δ_3 serán:

$$R_3 = \{ h_1, h_7, h_8, h_2, h_4 \}$$

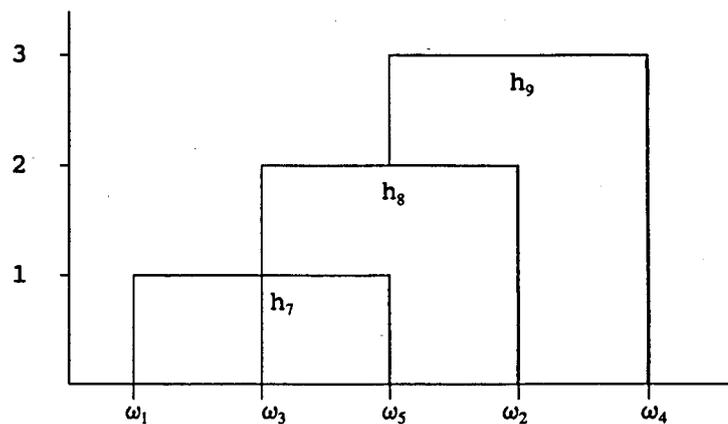
0	0	0	5	4
	0	0	2	3.5
		0	0	3.5
			0	6
				0

En este caso los pares de grupos de R_3 susceptibles de ser unidos y a distancia mínima son $[h_7, h_4]$ y $[h_8, h_4]$.

Puesto que $h_7 \subseteq h_8$, la condición A.2.2. nos indica que debemos unir h_8 con h_4 para formar el nuevo grupo: $h_9 = h_8 \cup h_4 = \Omega$

Y definimos: $i(h_9) = 3.5$ y $\omega_2 \leq \omega_4$, con lo cual queda completado el preorden total sobre Ω , respecto al cual los grupos serán conexos, éste será pues:

$$\omega_1 \leq \omega_3 \leq \omega_5 \leq \omega_2 \leq \omega_4$$



El nuevo recubrimiento y la disimilaridad correspondiente serán: $R_4 = \{\Omega\}$ $\delta_4 \equiv [0]$

Con lo cual finaliza el algoritmo, y la Pirámide obtenida será:

$$P = \{ \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}, h_6, h_7, h_8, h_9 \}$$

Que, una vez depurada, (h_6 es un grupo sobrante) resultará:

$$P = \{ \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}, h_7, h_8, h_9 \}$$

Que, evidentemente, es una Jerarquía indexada.

Finalmente, la disimilaridad piramidal, ultramétrica en este caso, equivalente es:

ω_1	ω_3	ω_5	ω_2	ω_4
0	1	1	2	3.5
	0	1	2	3.5
		0	2	3.5
			0	3.5
				0

5.4.2.- Método del Máximo

En este caso, la definición del índice de agregación puede reducirse a la forma siguiente:

$$\delta_k(h, h') = \begin{cases} 0 & \text{si } h \cap h' \in \{h, h'\} \\ \delta_{k-1}(h, h') & \text{si } h, h' \in R_{k-1} \\ \max\{\delta_{k-1}(h, h_1), \delta_{k-1}(h, h_2)\} & \text{si } h' = h_1 \cup h_2, h_i \in R_{k-1} \end{cases}$$

La peculiaridad de este método radica precisamente en el hecho que simplificando la definición de δ_k , la pirámide obtenida, continua siendo **sin inversión**, la cual cosa permite simplificar la programación del algoritmo (§5.5).

Para demostrarlo basta ver que el índice de un grupo construido por este método, siempre es menor o igual que el de cualquier otro grupo construido en un paso posterior.

Supongamos pues que en el paso k-ésimo del algoritmo unimos los grupos h_1 y h_2 para obtener un nuevo grupo h' ($h_i \in R_{k-1}$; $h' \in R_k - R_{k-1}$).

En el paso siguiente formaremos un nuevo grupo, h'' uniendo los dos más próximos de R_k .

Si ninguno de los dos grupos que forman h'' es h' , querrá decir que estos grupos eran de R_{k-1} , por tanto su interdistancia, $i(h'')$, será mayor o igual que $\delta_{k-1}(h_1, h_2) = i(h')$ es decir, $i(h') \leq i(h'')$.

Si por el contrario fuese $h'' = h' \cup \bar{h}$ (con $\bar{h} \subset h'$), entonces:

$$i(h'') = \delta_k(h', \bar{h}) = \max\{\delta_{k-1}(\bar{h}, h_1), \delta_{k-1}(\bar{h}, h_2)\}$$

Si \bar{h} puede ser unido con $h' = h_1 \cup h_2$, en el paso $k+1$ -ésimo, entonces $\bar{h} \in R_{k-1}$ y, en el paso k -ésimo, \bar{h} era susceptible de ser unido con h_1 o con h_2 (Propos. 5.2.1.), con lo cual $\delta_{k-1}(\bar{h}, h_1)$ ó $\delta_{k-1}(\bar{h}, h_2)$ ha de ser mayor o igual que $\delta_{k-1}(h_1, h_2)$, de lo contrario, en el paso k -ésimo, habríamos unido h_1 ó h_2 con \bar{h} en lugar de unir h_1 y h_2 . Por tanto:

$$i(h'') = \max\{\delta_{k-1}(\bar{h}, h_1), \delta_{k-1}(\bar{h}, h_2)\} \geq \delta_{k-1}(h_1, h_2) = i(h')$$

c.q.d.

Ejemplo 5.4.2.

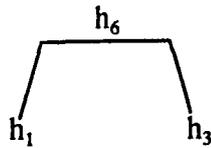
Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ y la matriz de disimilaridades iniciales:

$$\delta_0 \equiv \begin{bmatrix} 0 & 5 & \underline{1} & 4 & 1 \\ & 0 & 2 & 6 & 4 \\ & & 0 & 5 & 3 \\ & & & 0 & \underline{3.5} \\ & & & & 0 \end{bmatrix}$$

Vamos a efectuar una clasificación piramidal por el método del máximo. Para ello, partimos, como siempre, del recubrimiento $R_0 = \{h_1, h_2, h_3, h_4, h_5\}$ y de la disimilaridad δ_0 entre los grupos de dicho recubrimiento.

Los pares de grupos de R_0 a distancia mínima son $[h_1, h_3]$ y $[h_1, h_5]$. Elegimos como nuevo grupo: $h_6 = h_1 \cup h_3 = \{\omega_1, \omega_3\}$

Y definimos: $\omega_1 \leq \omega_3$ (podría ser al revés) y $i(h_6) = 1$



Con lo cual el nuevo recubrimiento y la nueva disimilaridad δ_1 , serán:

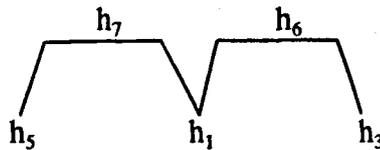
$$R_1 = \{ h_1, h_6, h_3, h_2, h_4, h_5 \}$$

0	0	1	5	4	$\frac{1}{3}$
	0	0	5	5	$\frac{3}{3}$
		0	2	5	$\frac{3}{3}$
			0	6	$\frac{4}{3}$
				0	$\frac{3.5}{3}$
					0

El único par de grupos de R_1 , susceptibles de ser unidos y a distancia mínima, es $[h_1, h_5]$, por tanto el nuevo grupo será: $h_7 = h_1 \cup h_5 = \{\omega_5, \omega_1\}$

Y puesto que en el paso anterior hemos definido $\omega_1 \leq \omega_3$, ahora, para que h_7 sea conexo, debemos definir $\omega_5 \leq \omega_1$, con lo cual el orden parcial construido hasta el momento es: $\omega_5 \leq \omega_1 \leq \omega_3$.

Además $i(h_7) = 1$.



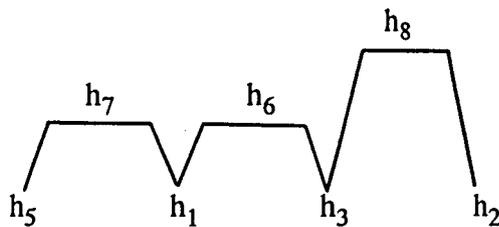
El nuevo recubrimiento R_2 y la nueva disimilaridad δ_2 , serán:

$$R_2 = \{ h_5, h_7, h_6, h_3, h_2, h_4 \}$$

0	0	3	3	4	$\frac{3.5}{3}$
	0	3	3	5	$\frac{4}{3}$
		0	0	5	$\frac{5}{3}$
			0	$\frac{2}{3}$	$\frac{5}{3}$
				0	$\frac{6}{3}$
					0

El nuevo grupo será $h_8 = h_3 \cup h_2 = \{\omega_3, \omega_2\}$, y para que sea conexo, el orden debe ser $\omega_3 \leq \omega_2$.

Además $i(h_8)=2$

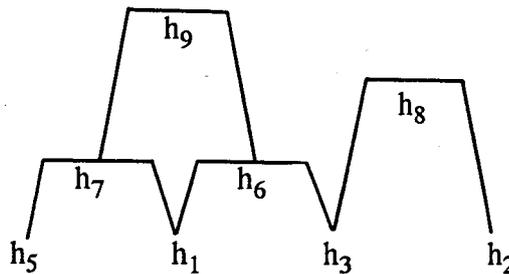


R_3 y δ_3 , serán respectivamente:

$$R_3 = \{ h_5, h_7, h_6, h_8, h_2, h_4 \}$$

0	0	3	4	4	3.5
	0	<u>3</u>	5	5	4
		0	5	5	5
			0	0	6
				0	6
					0

En este caso, la condición A.2.2. nos impone unir h_7 con h_6 , puesto que $h_5 \subseteq h_7$, $\delta_3(h_5, h_6) = \delta_3(h_7, h_6)$ y ésta es la mínima, con lo cual el nuevo grupo será: $h_9 = h_7 \cup h_6 = \{ \omega_5, \omega_1, \omega_3 \}$, y su índice: $i(h_9) = 3$.



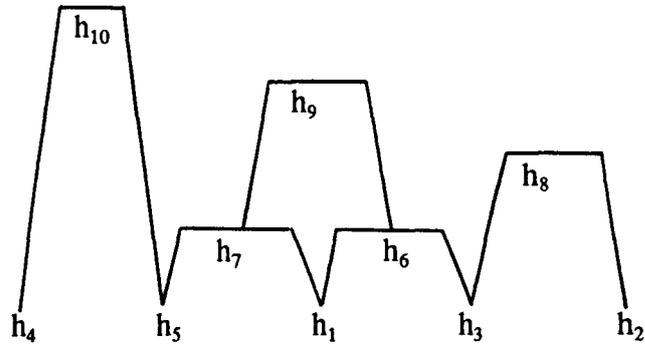
R_4 y δ_4 serán:

$$R_4 = \{ h_5, h_7, h_9, h_6, h_8, h_2, h_4 \}$$

0	0	0	3	4	4	<u>3.5</u>
	0	0	3	5	5	4
		0	0	5	5	5
			0	5	5	5
				0	0	6
					0	6
						0

Unimos h_5 con h_4 a nivel 3.5, así pues tendremos: $h_{10} = h_5 \cup h_4 = \{\omega_4, \omega_5\}$

Para que este grupo sea conexo debemos definir $\omega_4 \leq \omega_5$, con lo cual queda ya definido el preorden total sobre Ω : $\omega_4 \leq \omega_5 \leq \omega_1 \leq \omega_3 \leq \omega_2$ y $i(h_{10}) = 3.5$



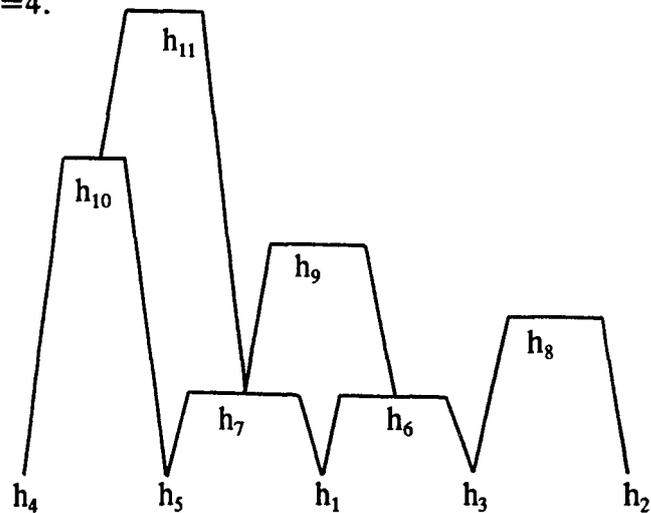
R_5 y δ_5 serán:

$$R_5 = \{h_{10}, h_7, h_9, h_6, h_8\}$$

0	$\frac{4}{0}$	5	5	6
	0	0	3	5
		0	0	5
			0	5
				0

Unimos h_{10} con h_7 , a nivel 4, con lo cual: $h_{11} = h_{10} \cup h_7 = \{\omega_4, \omega_5, \omega_1\}$

Y su índice, $i(h_{11}) = 4$.



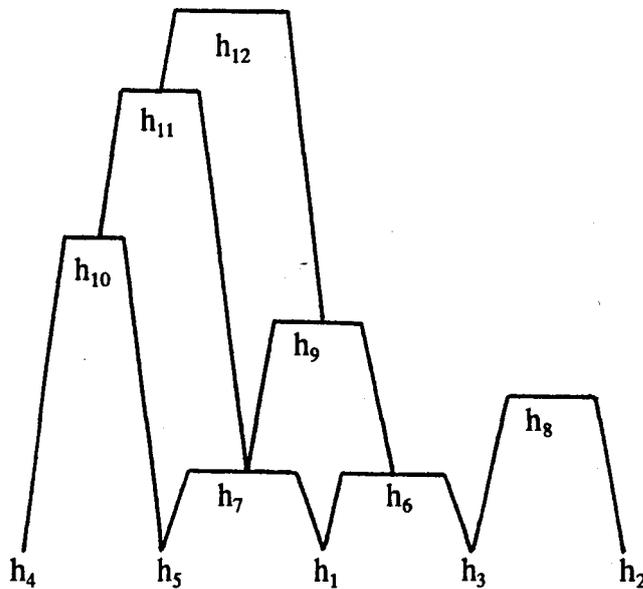
R_6 y δ_6 , serán:

$$R_6 = \{ h_{11}, h_9, h_6, h_8 \}$$

0	<u>5</u>	5	6
	0	0	<u>5</u>
		0	5
			0

Por la condición A.2.2. del algoritmo, las cuatro posibilidades que tenemos a la hora de formar un nuevo grupo, quedan reducidas a dos: h_{11} con h_9 o bien h_8 con h_9 .

Sea pues, $h_{12} = h_{11} \cup h_9 = \{\omega_4, \omega_5, \omega_1, \omega_3\}$, y su índice $i(h_{12}) = 5$.



R_7 y δ_7 serán:

$$R_7 = \{ h_{12}, h_9, h_6, h_8 \}$$

0	0	0	6
	0	0	<u>5</u>
		0	5
			0

$h_{13} = h_9 \cup h_8$ y $i(h_{13}) = 5$

R_8 y δ_8 serán:

$$R_8 = \{h_{12}, h_{13}\}$$

$$\begin{array}{|c|c|} \hline 0 & \underline{6} \\ \hline & 0 \\ \hline \end{array}$$

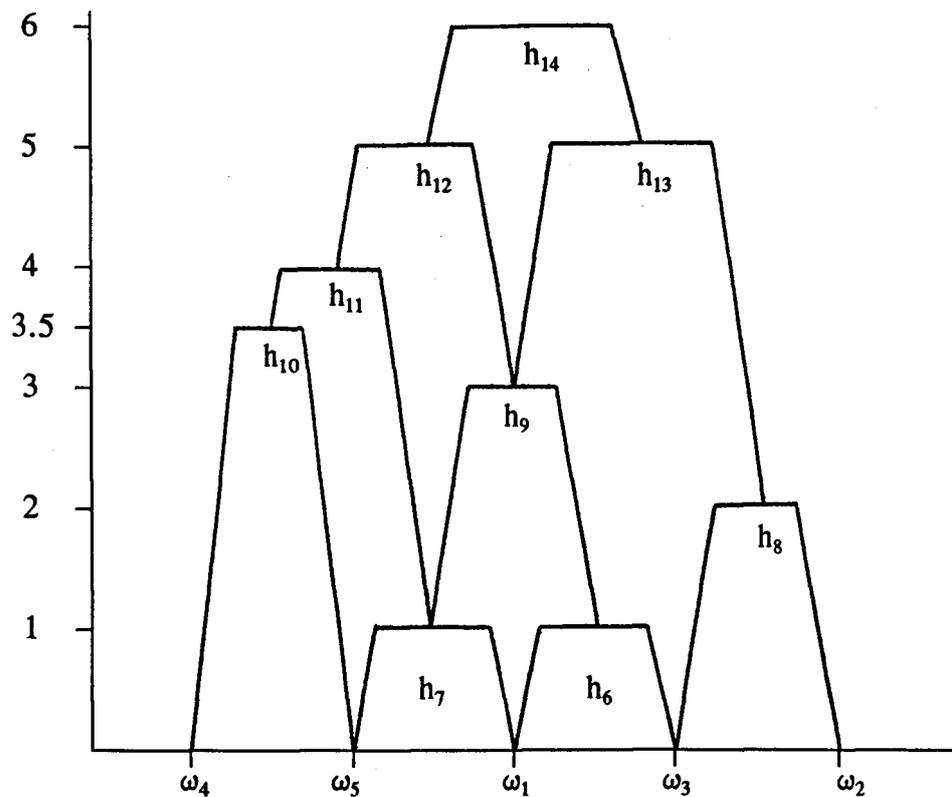
$$h_{14} = h_{12} \cup h_{13} \quad i(h_{14}) = 6$$

Y finalmente $R_9 = \{\Omega\}$

Con lo cual, la pirámide obtenida será:

$$P = \{ \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}, h_6, h_7, h_8, h_9, h_{10}, h_{11}, h_{12}, h_{13}, h_{14} \}$$

El grafo piramidal correspondiente:



Y la disimilaridad piramidal equivalente:

ω_4	ω_5	ω_1	ω_3	ω_2
0	3.5	4	5	6
	0	1	3	5
		0	1	5
			0	2
				0

5.4.3.- Otros Métodos

Cualquiera de los índices de agregación conocidos en clasificación jerárquica, y otros que pudieran darse, pueden ser utilizados en clasificación piramidal adaptándolos convenientemente.

Ejemplo 5.4.3.

En este ejemplo se pone de manifiesto la importancia que en general tiene, para cualquier método que no sea el del máximo, el hecho que $\delta_k(h, h_1 \cup h_2) \geq i(h_1 \cup h_2)$. De no imponer esta condición en la definición general del índice de agregación, el algoritmo podría quedar truncado, como puede verse en el presente ejemplo, por el método UPGMA.

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$ y la matriz de disimilaridades iniciales:

$$\delta_0 \equiv \begin{bmatrix} 0 & \frac{1}{0} & 1 & 1 & 1 & 1 \\ & & 2 & 2 & 2 & 2 \\ & & & 0 & 2 & 3 & 3 \\ & & & & 0 & 3 & 3 \\ & & & & & 0 & 3 \\ & & & & & & 0 \end{bmatrix}$$

Supongamos que el índice de agregación viniese definido por

$$\delta_k(h, h') = \begin{cases} 0 & \text{si } h \cap h' \in \{h, h'\} \\ \delta_{k-1}(h, h') & \text{si } h, h' \in R_{k-1} \\ \frac{n_1 \delta_{k-1}(h, h_1) + n_2 \delta_{k-1}(h, h_2)}{n_1 + n_2} & \text{si } h = h_1 \cup h_2 \end{cases}$$

igual que en el caso de las clasificaciones ultramétricas. El proceso sería el siguiente:

$$h_7 = h_1 \cup h_2 = \{\omega_1, \omega_2\}$$

$$\omega_1 \leq \omega_2$$

$$i(h_7) = 1$$

$$R_1 = \{h_1, h_7, h_2, h_3, h_4, h_5, h_6\}$$

0	0	1	1	1	1	1
	0	0	1.5	1.5	1.5	1.5
		0	2	2	2	2
			0	2	3	3
				0	3	3
					0	3
						0

$$h_8 = h_1 \cup h_3 = \{\omega_3, \omega_1\}$$

$$\omega_3 \leq \omega_1$$

$$i(h_8) = 1$$

$$R_2 = \{h_3, h_8, h_7, h_2, h_4, h_5, h_6\}$$

0	0	1.5	2	2	3	3
	0	0.75	1.5	1.5	2	2
		0	0	1.5	1.5	1.5
			0	2	2	2
				0	3	3
					0	3
						0

Obsérvese que los grupos susceptibles de ser unidos y a distancia mínima son h_7 y h_8 a distancia 0.75, que es menor que el índice de cualquiera de los dos. Por tanto si el nuevo grupo fuese $h_9 = h_7 \cup h_8$, la pirámide presentaría inversión, o dicho de otra forma, $h_8 \subset h_9$ y $i(h_8) > i(h_9)$, con lo cual no estaríamos construyendo una pirámide indexada. Así pues, para cualquier otro método de clasificación que no sea el del mínimo o el del máximo, habremos de prescindir de las distancias estrictamente menores que el índice del último grupo formado. En este caso pues, nos fijaremos en los pares de grupos a distancia 1.5.

$$h_9 = h_8 \cup h_4 = \{\omega_4, \omega_3, \omega_1\}$$

$$\omega_4 \leq \omega_3$$

$$i(h_9) = 1.5$$

$$R_3 = \{h_4, h_9, h_8, h_7, h_2, h_5, h_6\}$$

0	0	1.5	1.5	2	3	3
	0	0	1	1.66	2.33	2.33
		0	0.75	1.5	2	2
			0	0	<u>1.5</u>	1.5
				0	<u>2</u>	2
					0	3
						0

$$h_{10} = h_7 \cup h_5 = \{\omega_1, \omega_2, \omega_5\}$$

$$\omega_2 \leq \omega_5$$

$$i(h_{10}) = 1.5$$

$$R_4 = \{h_4, h_9, h_8, h_7, h_{10}, h_5, h_6\}$$

0	0	1.5	1.5	2	3	3
	0	0	1	1.44	2.33	2.33
		0	0.75	1.16	2	2
			0	0	1.5	1.5
				0	0	<u>2</u>
					0	3
						0

$$h_{11} = h_{10} \cup h_6 = \{\omega_1, \omega_2, \omega_5, \omega_6\}$$

$$\omega_5 \leq \omega_6$$

$$i(h_{11}) = 2$$

$$R_5 = \{h_9, h_8, h_7, h_{10}, h_{11}\}$$

0	0	1	1.44	1.66
	0	0.75	1.16	1.5
		0	0	0
			0	0
				0

El algoritmo queda truncado puesto que todos los valores de δ_5 , son inferiores al índice de h_{11} , que es el último grupo formado.

Vista pues la necesidad de que en una clasificación piramidal, a la hora de definir un índice de agregación, para poder construir una pirámide mediante alguno de los algoritmos de clasificación, es necesario que $\delta_k(h, h_1 \cup h_2) \geq i(h_1 \cup h_2)$, vamos a reiniciar el proceso de clasificación piramidal de los individuos de Ω por el método UPGMA, definiendo correctamente el índice de agregación, que será, en este caso:

$$\delta_k(h, h') = \left[\begin{array}{ll} 0 & \text{si } h \cap h' \in \{h, h'\} \\ \delta_{k-1}(h, h') & \text{si } h, h' \in R_{k-1} \\ \frac{n_1 \delta_{k-1}(h, h_1) + n_2 \delta_{k-1}(h, h_2)}{n_1 + n_2} & \text{si es } \geq \text{ que } i(h') \\ i(h') & \text{en otro caso} \end{array} \right] \quad \left. \vphantom{\delta_k(h, h')} \right\} h = h_1 \cup h_2$$

Partimos, como es natural, del recubrimiento R_0 y de δ_0 .

$$h_7 = h_1 \cup h_2 = \{\omega_1, \omega_2\}$$

$$\omega_1 \leq \omega_2$$

$$i(h_7) = 1$$

$$R_1 = \{h_1, h_7, h_2, h_3, h_4, h_5, h_6\}$$

0	0	1	1	1	1	1
	0	0	1.5	1.5	1.5	1.5
		0	2	2	2	2
			0	2	3	3
				0	3	3
					0	3
						0

$$h_8 = h_1 \cup h_3 = \{\omega_3, \omega_1\}$$

$$\omega_3 \leq \omega_1$$

$$i(h_8) = 1$$

Puesto que h_1 ya ha sido unido dos veces, no aparecerá en el próximo recubrimiento.

$$R_2 = \{h_3, h_8, h_7, h_2, h_4, h_5, h_6\}$$

0	0	1.5	2	2	3	3
	0	1	1.5	1.5	2	2
		0	0	1.5	1.5	1.5
			0	2	2	2
				0	3	3
					0	3
						0

$$h_9 = h_8 \cup h_7 = \{\omega_3, \omega_1, \omega_2\}$$

$$\omega_3 \leq \omega_1 \leq \omega_2$$

$$i(h_9) = 1$$

Obsérvese que h_7 y h_8 están estrictamente contenidos en h_9 y son de su mismo índice, por tanto la condición A.2.3 impone que h_7 y h_8 no deban pertenecer a R_3 .

$$R_3 = \{h_3, h_9, h_2, h_4, h_5, h_6\}$$

0	0	2	2	3	3
	0	0	<u>1.5</u>	1.75	1.75
		0	2	2	2
			0	3	3
				0	3
					0

$$h_{10} = h_9 \cup h_4 = \{\omega_3, \omega_1, \omega_2, \omega_4\}$$

$$\omega_3 \leq \omega_1 \leq \omega_2 \leq \omega_4 \text{ (otra posibilidad sería: } \omega_4 \leq \omega_3 \leq \omega_1 \leq \omega_2 \text{)}$$

$$i(h_{10}) = 1.5$$

$$R_4 = \{h_3, h_9, h_{10}, h_4, h_5, h_6\}$$

0	0	0	2	3	3
	0	0	1.5	<u>1.75</u>	1.75
		0	0	2.063	2.063
			0	3	3
				0	3
					0

$$h_{11} = h_9 \cup h_5 = \{\omega_5, \omega_3, \omega_1, \omega_2\}$$

$$\omega_5 \leq \omega_3 \leq \omega_1 \leq \omega_2 \leq \omega_4$$

$$i(h_{11}) = 1.75$$

$$R_5 = \{h_5, h_{11}, h_{10}, h_4, h_6\}$$

0	0	1.75	3	3
	0	<u>1.75</u>	1.87	2.063
		0	0	2.063
			0	3
				0

Por la condición A.2.2. debemos unir h_{11} con h_{10} a nivel 1.75; y por la condición

A.2.3, h_{11} no pertenecerá al nuevo recubrimiento R_6 . Así pues:

$$h_{12} = h_{11} \cup h_{10} = \{\omega_5, \omega_3, \omega_1, \omega_2, \omega_4\}$$

En este paso no se modifica el preorden definido hasta el momento.

$$i(h_{12})=1.75$$

$$R_6 = \{h_5, h_{12}, h_{10}, h_4, h_6\}$$

0	0	1.75	3	3
	0	0	0	<u>2.063</u>
		0	0	<u>2.063</u>
			0	3
				0

Por A.2.2. debemos unir h_{12} con h_6 y tendremos:

$$h_{13} = h_{12} \cup h_6 = \Omega$$

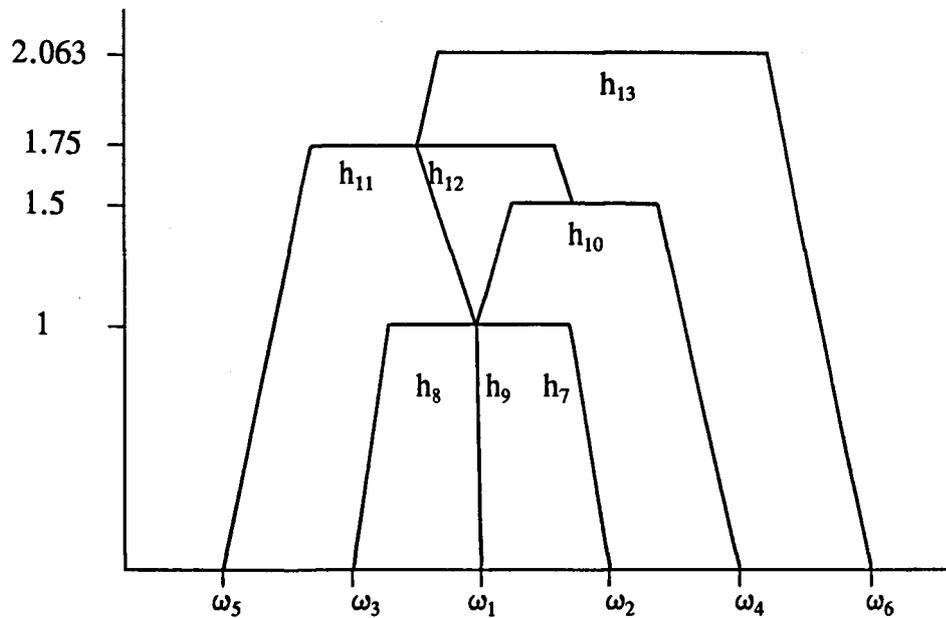
$$\omega_5 \leq \omega_3 \leq \omega_1 \leq \omega_2 \leq \omega_4 \leq \omega_6$$

$$i(h_{13})=2.063$$

$$R_7 = \{\Omega\}$$

[0]

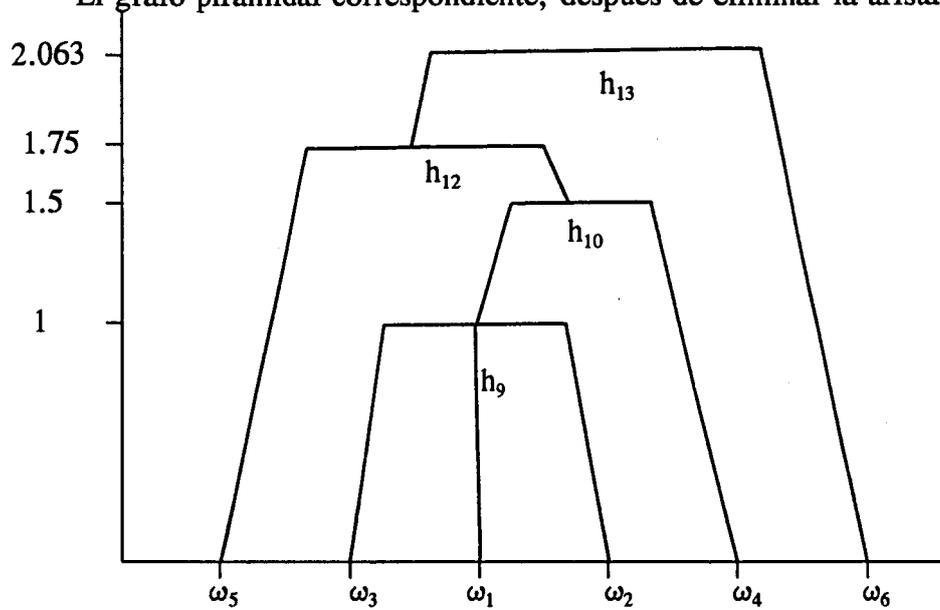
La pirámide resultante será: $P = \{h_1, h_2, \dots, h_{13}\}$



Que una vez depurada, (h_7, h_8, h_{11} , son grupos sobrantes) resultará ser:

$$P = \{h_1, h_2, h_3, h_4, h_5, h_6, h_9, h_{10}, h_{12}, h_{13}\}$$

El grafo piramidal correspondiente, después de eliminar la arista sobrante, será:



Y la disimilaridad piramidal equivalente:

	ω_5	ω_3	ω_1	ω_2	ω_4	ω_6
ω_5	0	1.75	1.75	1.75	1.75	2.063
ω_3		0	1	1	1.5	2.063
ω_1			0	1	1.5	2.063
ω_2				0	1.5	2.063
ω_4					0	2.063
ω_6						0

5.5.- PROGRAMACION DEL ALGORITMO CAP

Si el objetivo principal de este capítulo ha sido el de crear un algoritmo que permita efectuar clasificaciones piramidales y que además éstas puedan tener algún interés, no solamente desde el punto de vista teórico, sino también desde una

vertiente más aplicada a problemas reales de clasificación, es del todo imprescindible la programación informática de dicho algoritmo.

En este empeño pues, ha sido creado un programa, al que hemos llamado **PIR.PRO**, que no debe entenderse como definitivo sino como un primer intento de automatizar las clasificaciones piramidales, suficiente no obstante para mostrar como, de forma automática, puede construirse y visualizarse una pirámide, a partir de una disimilaridad definida sobre los individuos de una población finita, y suficiente también para efectuar los experimentos de simulación planteados en el presente trabajo. Con todo, este primer paso abre el camino a futuros trabajos que, a buen seguro, llevaran a mejores y más eficaces programas de clasificación piramidal.

La programación se ha efectuado en **Turbo Prolog**, lenguaje de inteligencia artificial (creado alrededor de 1970 en la Universidad de Marseille por un grupo de investigación en Inteligencia Artificial, liderado por Alain Colmerauer), diseñado fundamentalmente para manejar problemas lógicos, es decir problemas en los que es necesario tomar decisiones de forma ordenada. El Prolog intenta hacer que la computadora "razone" la forma de encontrar la solución a un problema previamente planteado. Es interesante también la posibilidad de conexión del Prolog con rutinas de otros lenguajes como el Pascal, C, Fortran, etc. Por otra parte, el Prolog dispone de un sistema de detección y control de errores muy eficaz. Finalmente, cabe resaltar la posibilidad de compilación de los programas en Turbo Prolog con lo cual se obtienen programas totalmente independientes que pueden ser ejecutados sin necesidad de conocer el lenguaje. En nuestro caso, el programa compilado es el **PIR.EXE**.

El listado Prolog del programa **PIR.PRO** puede verse en la sección A.2 del Anexo.

Descripción

El programa **PIR.PRO** transforma una disimilaridad inicial, conocida y definida sobre los individuos de una población finita Ω , en una disimilaridad piramidal, equivalente a una pirámide indexada (teoremas 4.2.1 y 4.2.2) y por tanto equivalente a una clasificación piramidal de los individuos de la población.

Este programa, sigue los pasos y satisface las condiciones del algoritmo CAP y permite construir una pirámide por los tres métodos mencionados en el presente capítulo: Mínimo, Máximo y UPGMA.

Además, como medida de adecuación entre la disimilaridad inicial y la piramidal obtenida, el programa calcula el coeficiente de correlación cofenética (Farris J.S. 1969) y el coeficiente gamma de Goodman-Kruskal entre ambas disimilaridades (Goodman, Kruskal 1954 y sección A.6 del Anexo).

Funcionamiento

Una vez han sido introducidos los datos, a saber, el número de individuos que han de ser clasificados -a los que el programa considera, ya de entrada, como grupos con un solo elemento- y las disimilaridades iniciales entre los mismos, **PIR.PRO** asigna a cada par de individuos (grupos del recubrimiento R_0), su disimilaridad.

El primer paso consiste entonces, en construir un nuevo grupo mediante la unión de los dos individuos a distancia mínima. Si existen varias posibilidades, el programa escoge la primera con la que se encuentra.

Al nuevo grupo se le asigna una "estructura" que indica qué dos grupos se ha unido y en qué orden lo han hecho; (en el primer paso este orden siempre es el inicial);

Un "**valor**" que nos indica cuales son los individuos que forman el nuevo grupo. Y un "**índice**".

Después de este primer paso el programa dispondrá de los mismos grupos iniciales más el nuevo, susceptibles todos ellos de ser unidos con algún otro (ver Nota 3 a pie de página 68) a los que identifica como "**disponibles**", y calcula la distancia del nuevo grupo a estos disponibles, teniendo en cuenta el índice de agregación previamente escogido (Mínimo, Máximo,...) .

En un paso cualquiera, después de la formación de un nuevo grupo, el programa calcula su distancia a los restantes grupos disponibles. Busca entonces la menor de las distancias entre grupos disponibles y un par de grupos a dicha distancia.

Una vez localizados, comprueba si son susceptibles de ser unidos (puesto que los dos son disponibles, sólo es necesario que su unión sea conexa -teniendo en cuenta el orden construido hasta el momento- y que no hayan sido unidos en un paso anterior), y en qué orden deben serlo.

En caso afirmativo se crea un nuevo grupo al que se le asigna una "**estructura**" que indica qué dos grupos se han unido y en qué orden lo han hecho; un "**valor**" con los individuos que forman el grupo, ordenados además según el orden que el propio programa va construyendo y respecto al cual, finalmente, los grupos de la pirámide deberán ser conexos; y un "**índice**" que es el valor de la distancia entre los dos grupos. Finalmente el programa se guarda el "**nuevo orden**", que tendrá en cuenta en la formación de los siguientes grupos, y repite el proceso considerando el mismo valor para la distancia, hasta que ya no exista otro par distinto de grupos a dicha distancia, en cuyo caso reiterará el proceso para el valor inmediatamente superior de la misma.

En caso que los grupos a distancia mínima no puedan ser unidos, buscará otro par de grupos a la misma distancia, si existen reiterará el proceso anterior , sino escogerá el valor inmediatamente superior de la distancia y volverá a repetir el proceso.

En todo caso el proceso finaliza cuando el grupo obtenido es Ω .

Limitaciones

El programa **PIR.PRO**, para poblaciones relativamente pequeñas, $n < 25$, no presenta limitaciones importantes. Solamente cuando el tamaño de la población es considerable pueden presentarse problemas derivados de la excesiva recursividad de algunos predicados, no esenciales, por otra parte, en el proceso de construcción piramidal; como por ejemplo el predicado "**member**" que simplemente comprueba si un individuo pertenece a un determinado grupo o no.

Por lo que se refiere a la representación gráfica de la pirámide, las principales limitaciones pueden provenir por una parte, del tipo de pantalla del que disponemos y por otra de su misma configuración.

El programa esta concebido para visualizar el grafo piramidal en un monitor VGA. En caso de no disponer de él, podemos modificar el valor del primer parámetro del predicado "**grafics (6,8,0)**" sustituyéndolo por cualquier otro valor entre 1 y 10 según nuestro tipo de pantalla.

Además hay que tener en cuenta que la pantalla dispone de 30.000 por 30.000 posiciones gráficas, así como 24 filas por 75 columnas de texto, con lo cual y debido a la separación mínima entre los grupos de la base de la pirámide establecida en el programa, podrán visualizarse correctamente pirámides de hasta 28 individuos, y en las

24 filas de texto podrán distribuirse otros tantos índices distintos. Como consecuencia de todo ello, puede suceder que si dos o más índices son muy próximos lleguen a superponerse, con lo cual en la pantalla aparece sólo el mayor de ellos, y lo mismo puede suceder con los segmentos horizontales que representan los grupos de la pirámide. Para darnos cuenta de ello no habrá otra manera que la de comparar el grafo con la lista de grupos formados que nos proporciona el programa.

Por otra parte, aunque el grafo piramidal sea depurado, al menos por lo que a los grupos se refiere, en el listado de grupos formados pueden haber grupos sobrantes, que detectaremos fácilmente observando el grafo correspondiente.

Finalmente, debido a la complejidad en la programación de la representación gráfica de la pirámide, podemos encontrarnos con alguna arista sobrante en el grafo piramidal obtenido.

Ejemplos 5.5.

En los siguientes ejemplos veremos como el programa **PIR.PRO** efectúa una clasificación piramidal de los individuos de un conjunto Ω , a partir de una disimilaridad inicial δ_0 , por los tres métodos que permite dicho programa. Máximo, Mínimo y UPGMA.

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7\}$ y la matriz de disimilaridad inicial δ_0 :

$$\begin{bmatrix} 0 & 3 & 3 & 8 & 8 & 8 & 8 \\ & 0 & 1 & 7 & 7 & 7 & 7 \\ & & 0 & 3 & 3 & 6 & 6 \\ & & & 0 & 2 & 4 & 5 \\ & & & & 0 & 3 & 3 \\ & & & & & 0 & 1 \\ & & & & & & 0 \end{bmatrix}$$

El programa llama h_k al grupo $\{\omega_k\}$, para $k \in \{1, 2, \dots, 7\}$, por tanto el primer

grupo formado será el h8.

METODE DEL MAXIM

MATRIU DE DISSIMILARITAT INICIAL

3	3	8	8	8	8
	1	7	7	7	7
		3	3	6	6
			2	4	5
				3	3
					1

grup h8: **estructura** 2 3;

La "estructura" indica qué dos grupos se han unido para formar el nuevo grupo y en qué orden lo han hecho.

valor [2,3];

El "valor" de un grupo indica cuales son sus elementos, ordenados además según el orden que el programa va construyendo y respecto al cual, finalmente, los grupos de la pirámide deberán ser conexos.

index 1

El "index" indica el índice del nuevo grupo.

nou ordre: [[2,3]]

El "nou ordre" indica, a cada paso, como se va construyendo el orden.

disponibles [1,2,3,4,5,6,7,8]

En "disponibles" el programa nos indica, después de la formación de cada grupo, los que quedan susceptibles de ser unidos con algún otro.

grup h9: **estructura** 6 7; **valor** [6,7]; **index** 1

Observese que, a cada paso, los grupos que se unen están entre los disponibles al final del paso anterior.

nou ordre: [[6,7],[2,3]]
disponibles [1,2,3,4,5,6,7,8,9]

grup h10: **estructura** 4 5; **valor** [4,5]; **index** 2
nou ordre: [[4,5],[6,7],[2,3]]
disponibles [1,2,3,4,5,6,7,8,9,10]

grup h11: **estructura** 3 10; **valor** [3,4,5]; **index** 3
nou ordre: [[2,3,4,5],[6,7]]
disponibles [1,2,5,6,7,8,9,10,11]

grup h12: **estructura** 5 9; **valor** [5,6,7]; **index** 3
nou ordre: [[2,3,4,5,6,7]]
disponibles [1,2,7,8,9,10,11,12]

grup h13: estructura 1 8; valor [1,2,3]; index 3
 nou ordre: [[1,2,3,4,5,6,7]]
 disponibles [1,7,8,9,10,11,12,13]

grup h14: estructura 10 12; valor [4,5,6,7]; index 5
 nou ordre: [[1,2,3,4,5,6,7]]
 disponibles [1,7,8,9,11,12,13,14]

grup h15: estructura 11 14; valor [3,4,5,6,7]; index 6
 nou ordre: [[1,2,3,4,5,6,7]]
 disponibles [1,7,8,9,11,12,13,14,15]

grup h16: estructura 8 15; valor [2,3,4,5,6,7]; index 7
 nou ordre: [[1,2,3,4,5,6,7]]
 disponibles [1,7,9,12,13,14,15,16]

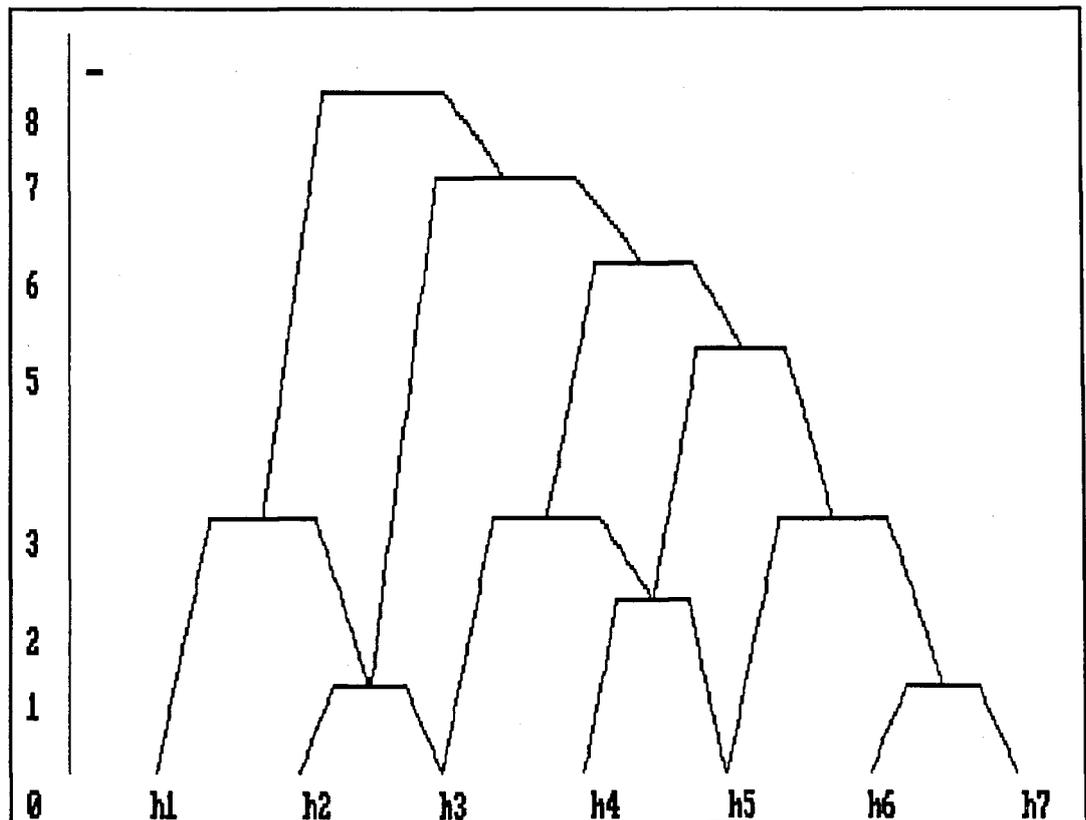
grup h17: estructura 13 16; valor [1,2,3,4,5,6,7]; index 8
 nou ordre: [[1,2,3,4,5,6,7]]

MATRIU DISSIMILARITAT PIRAMIDAL

3	3	8	8	8	8
	1	7	7	7	7
		3	3	6	6
			2	5	5
				3	3
					1

COEFICIENT DE CORRELACIO COFENETICA
 0.99601758695

COEFICIENT DE GOODMAN-KRUSKAL
 1



En este caso, de entre los grupos formados, no existe ninguno sobrante y por consiguiente, la pirámide resultante será: $P=\{h1, \dots, h17\}$

METODE DEL MINIM

MATRIU DE DISSIMILARITAT INICIAL

	3	3	8	8	8	8
		1	7	7	7	7
			3	3	6	6
				2	4	5
					3	3
						1

grup h8: estructura 2 3; valor [2,3]; index 1
nou ordre: [[2,3]]
disponibles [1,2,3,4,5,6,7,8]

grup h9: estructura 6 7; valor [6,7]; index 1
nou ordre: [[6,7],[2,3]]
disponibles [1,2,3,4,5,6,7,8,9]

grup h10: estructura 4 5; valor [4,5]; index 2
nou ordre: [[4,5],[6,7],[2,3]]
disponibles [1,2,3,4,5,6,7,8,9,10]

grup h11: estructura 10 9; valor [4,5,6,7]; index 3
nou ordre: [[4,5,6,7],[2,3]]
disponibles [1,2,3,4,7,8,9,10,11]

grup h12: estructura 11 8; valor [4,5,6,7,2,3]; index 3
nou ordre: [[4,5,6,7,2,3]]
disponibles [1,3,4,8,10,12]

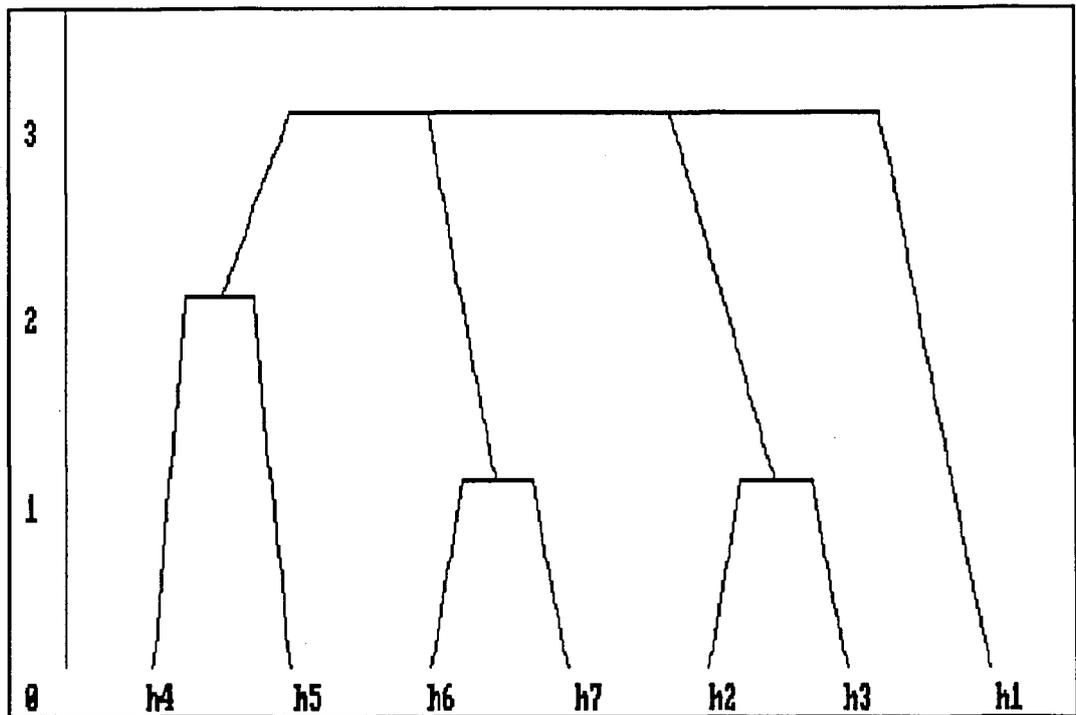
grup h13: estructura 12 1; valor [4,5,6,7,2,3,1]; index 3
nou ordre: [[4,5,6,7,2,3,1]]

MATRIU DISSIMILARITAT PIRAMIDAL

	2	3	3	3	3	3
		3	3	3	3	3
			1	3	3	3
				3	3	3
					1	3
						3

COEFICIENT DE CORRELACIO COFENETICA
0.60558116471

COEFICIENT DE GOODMAN-KRUSKAL
1



En este caso los grupos h11 y h12 son sobrantes, por tanto la pirámide resultante será: $P=\{h1, \dots, h10, h13\}$

METODE UPGMA

MATRIU DE DISSIMILARITAT INICIAL

3	3	8	8	8	8
	1	7	7	7	7
		3	3	6	6
			2	4	5
				3	3
					1

grup h8: estructura 2 3; valor [2,3]; index 1
nou ordre: [[2,3]]
disponibles [1,2,3,4,5,6,7,8]

grup h9: estructura 6 7; valor [6,7]; index 1
nou ordre: [[6,7],[2,3]]
disponibles [1,2,3,4,5,6,7,8,9]

```

grup h10:  estructura 4 5; valor [4,5]; index 2
nou ordre: [[4,5],[6,7],[2,3]]
disponibles [1,2,3,4,5,6,7,8,9,10]

grup h11:  estructura 3 10; valor [3,4,5]; index 3
nou ordre: [[2,3,4,5],[6,7]]
disponibles [1,2,5,6,7,8,9,10,11]

grup h12:  estructura 5 9; valor [5,6,7]; index 3
nou ordre: [[2,3,4,5,6,7]]
disponibles [1,2,7,8,9,10,11,12]

grup h13:  estructura 11 12; valor [3,4,5,6,7]; index 3
nou ordre: [[2,3,4,5,6,7]]
disponibles [1,2,7,8,9,13]

grup h14:  estructura 1 8; valor [1,2,3]; index 3
nou ordre: [[1,2,3,4,5,6,7]]
disponibles [1,7,8,9,13,14]

grup h15:  estructura 8 13; valor [2,3,4,5,6,7]; index 4.66667
nou ordre: [[1,2,3,4,5,6,7]]
disponibles [1,7,9,13,14,15]

grup h16:  estructura 14 15; valor [1,2,3,4,5,6,7]; index 4.66667
nou ordre: [[1,2,3,4,5,6,7]]

```

MATRIU DISSIMILARITAT PIRAMIDAL

3	3	4.66667	4.66667	4.66667	4.66667
	1	4.66667	4.66667	4.66667	4.66667
		3	3	3	3
			2	3	3
				3	3
					1

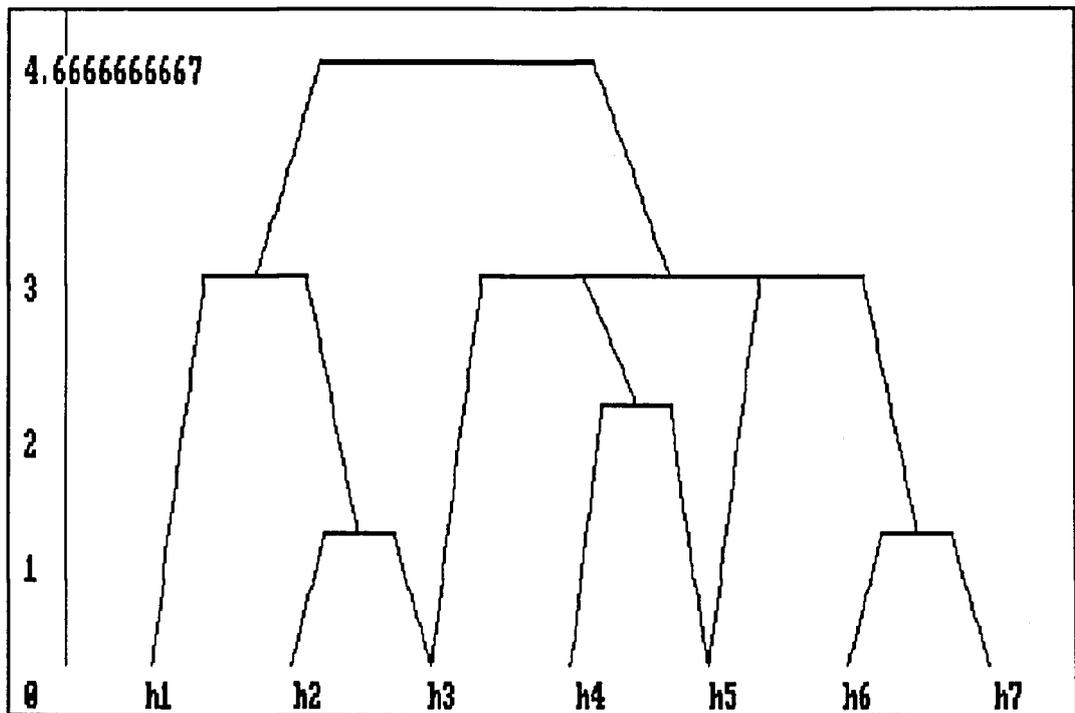
COEFICIENT DE CORRELACIO COFENETICA
0.91864560271

COEFICIENT DE GOODMAN-KRUSKAL
1

Observese que los grupos h11, h12 y h15 son sobrantes, por tanto la pirámide resultará ser:

$$P = \{h1, \dots, h10, h13, h14, h16\}$$

Observese asimismo, en el grafo piramidal correspondiente, que la arista que une h5 con h13 es también sobrante.



5.6.- EXPERIMENTO DE SIMULACION

Por todo cuanto hemos venido exponiendo en el presente trabajo, es razonable suponer que las clasificaciones piramidales son, por lo general, mucho más próximas a la realidad, y por tanto más complejas, que las jerárquicas, en el sentido de que partiendo de una disimilaridad inicial d , definida sobre los individuos de una determinada población finita, la disimilaridad piramidal correspondiente, construida por alguno de los métodos de clasificación piramidal, será más próxima a la inicial que una ultramétrica, construida también por alguno de los métodos de clasificación ascendente

jerárquica.

Esta apreciación no puede enunciarse como si de un teorema se tratase, puesto que es evidente que escogiendo un buen método ultramétrico y un mal algoritmo piramidal, la piramidal correspondiente puede estar más alejada de la inicial que la ultramétrica.

En estas condiciones pues, únicamente nos limitaremos a probar, de una forma empírica, mediante un sencillo experimento de simulación, que por lo general las clasificaciones piramidales son mejores, más próximas a la realidad, que las jerárquicas.

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$, y consideremos dos pirámides indexadas de $\mathcal{P}(\Omega)$, P_1 y P_2 de modo que la primera sea prácticamente una jerarquía y la segunda tenga una estructura estrictamente piramidal, tal como viene indicado en la figura 5.6.1.

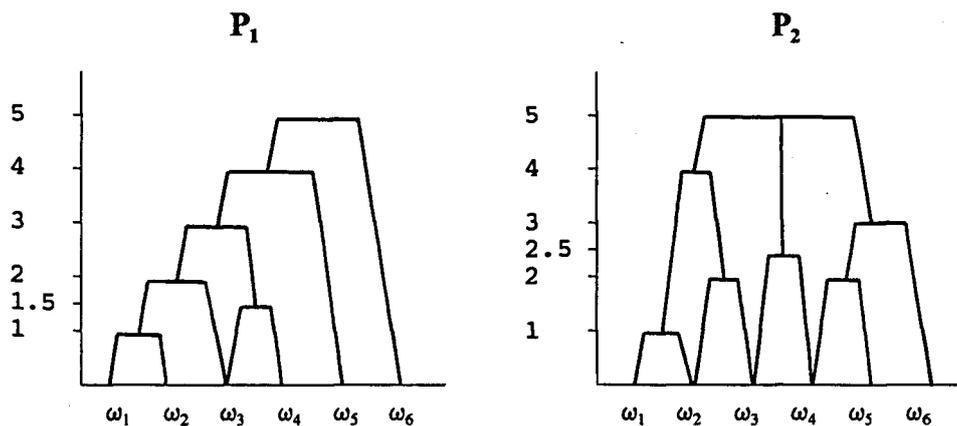


Fig. 5.6.1

Cada una de estas pirámides tiene asociada una única disimilaridad piramidal, sean éstas d_1 y d_2 respectivamente:

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ & 0 & 2 & 3 & 4 & 5 \\ & & 0 & 1.5 & 4 & 5 \\ & & & 0 & 4 & 5 \\ & & & & 0 & 5 \\ & & & & & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 4 & 5 & 5 & 5 \\ & 0 & 2 & 5 & 5 & 5 \\ & & 0 & 2.5 & 5 & 5 \\ & & & 0 & 2 & 3 \\ & & & & 0 & 3 \\ & & & & & 0 \end{bmatrix}$$

El experimento consistirá en perturbar cada una de estas disimilaridades (piramidales) mediante ciertas variables aleatorias, con lo cual obtendremos un cierto número de matrices de disimilaridad "aleatorias". Para cada una de estas nuevas matrices construiremos la ultramétrica y la piramidal correspondiente mediante alguno de los métodos de clasificación y veremos cómo las piramidales son más próximas a la disimilaridad inicial correspondiente que las ultramétricas.

En primer lugar elegimos el tipo de variables aleatorias con las cuales perturbaremos d_1 y d_2 , que serán v.a. Normales. En tal caso, observando las matrices de estas disimilaridades, no es difícil darse cuenta de que las matrices obtenidas después de la perturbación, podrían contener valores negativos, con lo cual no corresponderían a ninguna matriz de disimilaridad. Para evitar este problema basta efectuar una traslación de +10 sobre las matrices d_1 y d_2 .

Para sondear los posibles valores de los parámetros de las Normales que vamos a utilizar, calcularemos en primer lugar las desviaciones típicas de d_1 y d_2 que resultarán ser:

$$S_{d1} = 1.352364185 \quad S_{d2} = 1.38644229^{(5)}$$

Consideremos pues tres variables aleatorias, $\varepsilon_1, \varepsilon_2, \varepsilon_3$ Normales de media cero y desviaciones típicas $\sigma_1, \sigma_2, \sigma_3$ respectivamente, con

$$\begin{aligned} \sigma_1 &= 0.5 & \sigma_2 &= 0.64 \\ \sigma_2 &= 1.0 & \sigma_2 &= 1.30 \\ \sigma_3 &= 1.5 & \sigma_2 &= 1.95 \end{aligned}$$

De esta forma tenemos: $\sigma_1 < \sigma_2 < \sigma_3$.

⁽⁵⁾Observese que las desviaciones típicas son invariantes por traslaciones.

El experimento consistirá finalmente en perturbar d_1 y d_2 ocho veces con cada una de las tres variables, con lo cual obtendremos 48 matrices, correspondientes a otras tantas disimilaridades ⁽⁶⁾: $\{d_{jik}\}_{j=1,2; i=1,2,3; k=1,\dots,8}$

Para ello utilizaremos el generador de variables aleatorias σ^* NORMAL(μ) del paquete estadístico SAS, y mediante los seis programas hSIMULji.SAS descritos en la sección A.1 del Anexo, cada uno de los cuales ejecutaremos 8 veces, obtendremos las 48 matrices de disimilaridad, que serán las disimilaridades "iniciales", a partir de las cuales efectuaremos la clasificación correspondiente.

Finalmente, a partir de cada una de estas matrices efectuaremos una **clasificación jerárquica** por el método UPGMA, utilizando para ello el paquete CLUSTAN, (el programa que efectúa estas clasificaciones viene descrito también en la sección A.1. del Anexo), y una **clasificación piramidal** por el método del Máximo, utilizando el programa PIR.PRO descrito en § 5.5 y especificado en la sección A.2 del Anexo.

En ambos casos compararemos las 48 disimilaridades "iniciales" con las ultramétricas y piramidales correspondientes, mediante el **coeficiente de correlación cofenética**.

Así pues, si d_{jik} es la k -ésima ($k=1,\dots,8$) perturbación de la matriz d_j ($j=1,2$) mediante la variable aleatoria ε_i ($i=1,2,3$), r_j^{ik} indicará el coeficiente de correlación cofenética entre d_{jik} y la ultramétrica o la piramidal correspondiente. En ambos casos calcularemos $\bar{r}_j = 1/8 \sum_k r_j^{ik}$ que es el promedio de los ocho coeficientes de correlación obtenidos para cada i y j fijadas.

⁽⁶⁾Observe que estas 48 matrices serán matrices de datos Normales de media \bar{d}_j , $j \in \{1,2\}$ y desviación típica la σ_i correspondiente, $i \in \{1,2,3\}$.

Los resultados obtenidos son los siguientes:

Clasificación Ultramétrica:

	ε_1	ε_2	ε_3
d_1 :	$\bar{r}_1=0.4801$	$\bar{r}_1=0.4686$	$\bar{r}_1=0.4512$
d_2 :	$\bar{r}_2=0.4903$	$\bar{r}_2=0.4801$	$\bar{r}_2=0.4688$
	$\bar{r}^1=0.4852$	$\bar{r}^2=0.4744$	$\bar{r}^3=0.4601$

Clasificación Piramidal:

	ε_1	ε_2	ε_3
d_1 :	$\bar{r}_1=0.9698$	$\bar{r}_1=0.9139$	$\bar{r}_1=0.8451$
d_2 :	$\bar{r}_2=0.9689$	$\bar{r}_2=0.8834$	$\bar{r}_2=0.8616$
	$\bar{r}^1=0.9694$	$\bar{r}^2=0.8986$	$\bar{r}^3=0.8534$

Observese pues que tanto en el caso ultramétrico como en el piramidal, cuanto mayor es la desviación típica de la variable con la que perturbamos la matriz inicial d_j , menor es la correlación entre la disimilaridad perturbada ("inicial") y la ultramétrica o la piramidal correspondiente.

Por otra parte, la correlación en el caso piramidal es claramente mayor que en el caso ultramétrico.

CAPITULO 6

ASPECTOS INFERENCIALES SOBRE GRAFOS PIRAMIDALES

6.1.- INTRODUCCION

El objetivo principal del Cluster Analysis, es el estudio de algoritmos que permitan detectar una determinada estructura sobre una población dada, a partir de una tabla de datos obtenidos sobre sus individuos.

Los métodos de representación piramidal en particular, pretenden detectar la existencia de una estructura piramidal sobre la población a partir de una matriz de disimilaridades dada sobre la misma. Para ello, el proceso consiste en deformar la disimilaridad inicial hasta transformarla en piramidal. A través de ciertos parámetros se puede evaluar la bondad de ajuste entre la disimilaridad inicial y la piramidal obtenida. Los valores de estos parámetros reflejarán pues, hasta qué punto la clasificación obtenida se adecua a los datos iniciales, o dicho de otra forma, hasta qué punto la estructura inicial de los datos se acomoda a una estructura piramidal.

En nuestro caso, los parámetros utilizados a tal fin son: el coeficiente de correlación cofenética entre ambas disimilaridades, Farris J.S. (1969) y el coeficiente gamma definido por Goodman-Kruskal (1954) y descrito en la sección A.6 del Anexo.

A pesar de disponer de estas medidas de adecuación, en principio será difícil precisar hasta qué punto sus valores, en un caso concreto, son significativos. Por ejemplo, si partimos de una población con 6 individuos, sobre la cual tenemos definida una matriz de distancias d:

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ & 0 & 2 & 1 & 3 & 4 \\ & & 0 & 3 & 2 & 1 \\ & & & 0 & 2 & 3 \\ & & & & 0 & 4 \\ & & & & & 0 \end{bmatrix}$$

y efectuamos una clasificación piramidal por los métodos del mínimo y del máximo, mediante el programa PIR.PRO, y calculamos el valor de los coeficientes, tendremos que para el caso del máximo, $\gamma=0.92$, $\rho=0.89$ y para el caso del mínimo, $\gamma=0.80$ $\rho=0.59$; que como puede observarse son bastante altos, pero nada nos permite asegurar que estos valores sean realmente significativos. Es por ello pues que, en este capítulo, intentaremos dar algún criterio objetivo que nos permita decidir sobre si los valores de dichos parámetros son o no significativos, en el sentido de reflejar el nivel de adecuación de los datos iniciales a la estructura piramidal obtenida.

Por otra parte, sería conveniente también poder decir cual de los métodos considerados es mejor, en el sentido de mejor recuperar una posible estructura piramidal de los datos.

Sobre estas cuestiones es difícil de establecer resultados muy generales, debido fundamentalmente al desconocimiento de la distribución de los parámetros considerados.

Por lo que a ciertos métodos de clasificación jerárquica se refiere, diversos

autores, L.Hubert (1974), F.Baker-L.Hubert (1975), H.H.Bock (1984-85), han realizado algunos estudios y obtenido ciertos resultados inferenciales sobre grafos ultramétricos, que más adelante comentaremos.

En otros aspectos distintos a los que aquí trataremos, pero no tan distantes por lo que a la metodología utilizada se refiere, Pruzansky, Tversky y Carroll (1982), realizan un estudio en el que tratan sobre las relaciones entre modelos continuos y discretos desde una perspectiva empírica, partiendo de una disimilaridad dada sobre los individuos de la población. A partir de estos estudios, Arcas (1986) indica la posibilidad de que la distribución de la disimilaridad observada pueda servir para decidir entre un modelo espacial o un modelo en árbol.

En lo que a los grafos piramidales concierne, debido fundamentalmente a la falta de un algoritmo eficiente de clasificación piramidal, y sobre todo a la inexistencia de software adecuado que permita una ejecución informática de dicho algoritmo, no se conoce, por el momento, ningún trabajo de validación estadística de una clasificación piramidal.

Puesto que disponemos del programa PIR.PRO, descrito en la sección 5.5, es nuestro propósito contribuir a la mejora de los métodos de clasificación mediante grafos piramidales, también desde un punto de vista inferencial.

En el presente capítulo utilizamos algunas técnicas de simulación mediante métodos de Montecarlo, con el fin de estudiar las cuestiones siguientes:

- a) Tabular, a través de la distribución muestral del coeficiente gamma de Goodman-Kruskal, un test de significación de la bondad de ajuste de

una representación piramidal efectuada por los métodos del mínimo y del máximo.

b) Evolución de la distribución de gamma en función del número de individuos a representar.

c) Comparar la eficiencia de los métodos del mínimo y del máximo en función de la distribución de los coeficientes gamma y rho.

6.2.- LAS PRUEBAS DE SIMULACION

Las pruebas de simulación que planteamos, tienen por objetivo fundamental el de dar una respuesta, lo más clara posible, desde un punto de vista empírico, a las cuestiones planteadas en la sección anterior.

Estas pruebas de simulación, basadas en los métodos de Montecarlo, y partiendo de una disimilaridad inicial entre los n individuos de una población finita Ω , consisten en efectuar una clasificación piramidal de dichos individuos, por los métodos del mínimo y del máximo, utilizando el programa PIR.PRO, y en evaluar el grado de adecuación de la representación obtenida a los datos iniciales, comparando la disimilaridad inicial con la piramidal obtenida, mediante el coeficiente gamma y el coeficiente de correlación cofenética.

Para cada uno de los valores de n escogidos para el estudio y para cada uno de los métodos de clasificación utilizados (mínimo y máximo), se repite N veces la prueba. A partir de los N valores obtenidos para cada uno de los coeficientes, podremos

elaborar unas tablas con las medias, desviaciones típicas y cuantiles del coeficiente que deseemos estudiar.

Por lo que a la elección de la disimilaridad inicial concierne, se han contemplado dos posibilidades. Por una parte que sea una disimilaridad totalmente aleatoria, y por otra que, partiendo de una estructura concreta, (a la que nos referiremos como estructura o disimilaridad básica) ésta sea perturbada con una cierta variable aleatoria.

Sobre estos principios, hemos construido tres programas de simulación, **SIMULU.PRO**, **SIMULN.PRO** y **NSIMUL.PRO**, todos ellos a partir del **PIR.PRO**, y que básicamente se diferencian en la forma de generar la disimilaridad inicial (ver sección A.7 del Anexo).

Con ellos se han efectuado cuatro pruebas de simulación que a continuación pasamos a describir.

PRUEBA S1

Consideremos poblaciones con $n = 4, 5, 6, \dots, 17, 18, 20$ y 25 individuos.

Para cada uno de estos valores de n , el programa **SIMULU.PRO** genera $n(n-1)/2$ valores $U(0,1)$ que representan la parte triangular superior de una matriz de disimilaridad entre los n individuos de la población. A partir de esta disimilaridad inicial el programa efectúa una clasificación piramidal por los métodos del mínimo y del máximo y calcula el coeficiente gamma entre la disimilaridad inicial y la piramidal obtenida. Estos resultados se guardan en sendos ficheros, **GUMAXn** y **GUMINn**.

Este proceso se repite N veces para cada uno de los valores de n escogidos. Para $n=25$, $N=200$ y para los restantes valores de n , $N=1000$.

En cada uno de los ficheros de resultados tendremos pues N valores de gamma,

a partir de los cuales, y a través del programa **ESTADIS.C** confeccionado al efecto y descrito en la sección A.8 del Anexo, calcularemos la **media**, la **desviación típica** y los **cuantiles** (de gamma). Estos resultados quedan reflejados en las tablas S1.1, S1.2 y S1.3 descritas en la sección 6.3.

PRUEBA S2

Para efectuar esta segunda prueba utilizaremos el programa **SIMULN.PRO** el cual, para cada uno de los valores de n considerados, (4,5,...18,20,25) genera $n(n-1)/2$ valores $N(0,1)$ a los que suma la constante **10**, para evitar que aparezcan valores negativos en una matriz que representa una disimilaridad. A partir de esta disimilaridad inicial el programa actúa exactamente igual que el programa anterior (**SIMULU.PRO**) y guarda los resultados obtenidos en los ficheros **GNMAXn** y **GNMINn**.

Después de efectuar el estudio estadístico de estos ficheros, los resultados quedan reflejados en las tablas S2.1, S2.2 y S2.3.

PRUEBAS S3 y S4

Consideremos, para estos casos, poblaciones con $n=4,5,6,10,16,20$ individuos. Para cada uno de estos valores de n , consideremos una cierta matriz de disimilaridad a la que llamaremos **Básica**, de la cual conoceremos la desviación típica σ , de sus valores.

Las pruebas, efectuadas con el programa **NSIMUL.PRO**, consistirán entonces en **perturbar** la disimilaridad básica con tres variables aleatorias ε_k , Normales de media cero y varianza σ_k , $k \in \{1,2,3\}$ con $\sigma_1=1/3\sigma$, $\sigma_2=\sigma$, $\sigma_3=3\sigma$, la cual será una medida del error introducido en la disimilaridad básica (bajo, medio y alto) y, a partir de esta

disimilaridad perturbada, efectuar una **clasificación piramidal por los métodos del mínimo y del máximo**. Finalmente se comparará la disimilaridad piramidal obtenida con la básica y con la perturbada, mediante el coeficiente de correlación cofenética y el coeficiente gamma de Goodman-Kruskal. Estos resultados se almacenarán automáticamente en unos ficheros, especificados en la sección A.7 del Anexo, sobre los que actuará el programa ESTADIS.C y se obtendrán las correspondientes tablas de resultados.

Este proceso, para cada n , cada σ_k y cada uno de los dos métodos de clasificación, se repite $N=1000$ veces para $n=4,5,6,10$ y $N=200$ veces para $n=16$ y 20 .

Las pruebas S3 y S4 difieren esencialmente en el tipo de disimilaridad básica considerada.

En la **PRUEBA S3**, la disimilaridad básica considerada, para cada uno de los valores de n , es una disimilaridad piramidal correspondiente a una pirámide que llamaremos "**pseudo-binaria**" puesto que respondería, en cierto modo, a la idea de jerarquía binaria.

En la **PRUEBA S4**, la disimilaridad básica considerada es una disimilaridad ultramétrica, (piramidal) correspondiente a una jerarquía "**encadenada**". Estas disimilaridades básicas, utilizadas en las pruebas, junto con sus grafos y las desviaciones típicas correspondientes, vienen especificadas en la sección A.9 del Anexo.

Después de efectuar el estudio estadístico de los resultados, éstos quedan reflejados en las tablas S3.P, S3.B; S4.P, S4.B.

6.3.- TEST DE SIGNIFICACION DE LA BONDAD DE AJUSTE EN UNA CLASIFICACION PIRAMIDAL

Las pruebas S1 y S2 son las que nos van a permitir tabular, a través de la distribución muestral del coeficiente gamma, un test de significación de la bondad de ajuste de las representaciones piramidales.

A partir de estas pruebas, también podremos decir algo sobre la eficiencia de los métodos del mínimo y del máximo considerados.

**Tabla S1.1 : Relación entre el número de individuos de Ω (n) y la media (M_γ) y desviación típica (S_γ) muestral de gamma, para los métodos del mínimo y del máximo.
Disimilaridad inicial aleatoria U(0,1).**

n	Máximo		Mínimo	
	M_γ	S_γ	M_γ	S_γ
4	0.96	0.07	0.78	0.21
5	0.85	0.10	0.63	0.19
6	0.74	0.10	0.54	0.17
7	0.66	0.09	0.47	0.14
8	0.59	0.09	0.40	0.13
9	0.53	0.08	0.36	0.11
10	0.49	0.07	0.33	0.10
11	0.45	0.07	0.30	0.09
12	0.42	0.07	0.28	0.09
13	0.39	0.06	0.25	0.08
14	0.36	0.06	0.23	0.08
15	0.34	0.06	0.22	0.07
16	0.32	0.05	0.21	0.06
17	0.30	0.05	0.20	0.06
18	0.28	0.05	0.18	0.06
20	0.26	0.05	0.16	0.05
25	0.21	0.04	0.13	0.04

Media y Desv. Típica basadas en una muestra de 200 para n=25 y de 1000 para los restantes casos.

Tabla S1.2 : Relación entre n y los cuantiles del coeficiente γ obtenido por el método del máximo.
Disimilaridad inicial aleatoria $U(0,1)$.

n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.857143	0.857143	0.857143	1.000000	1.000000	1.000000	1.000000
5	0.674419	0.707317	0.783784	0.853659	0.906977	1.000000	1.000000
6	0.562500	0.612245	0.680000	0.752577	0.813953	0.862069	0.900000
7	0.492228	0.526316	0.597938	0.663158	0.724868	0.784314	0.816216
8	0.427632	0.466667	0.525140	0.594771	0.651652	0.709571	0.741379
9	0.393502	0.427007	0.475524	0.533333	0.585586	0.645914	0.669535
10	0.372642	0.393939	0.437573	0.490066	0.535874	0.583653	0.615303
11	0.330827	0.355655	0.401562	0.452055	0.505360	0.553329	0.577586
12	0.289751	0.321485	0.370629	0.421569	0.471944	0.515931	0.540839
13	0.277948	0.303216	0.344127	0.386464	0.431020	0.464392	0.482512
14	0.258931	0.282527	0.321147	0.357998	0.402517	0.438837	0.457928
15	0.239383	0.263137	0.299750	0.339187	0.379609	0.416396	0.436960
16	0.230418	0.247629	0.282639	0.320232	0.356501	0.395044	0.415150
17	0.212063	0.233498	0.263158	0.301455	0.339261	0.372978	0.391679
18	0.206997	0.221208	0.250947	0.284714	0.319969	0.351281	0.367837
20	0.175423	0.193540	0.223776	0.256709	0.288380	0.316252	0.332356
25	0.126055	0.149014	0.182115	0.207617	0.233013	0.255718	0.263527

Tabla S1.3 : Relación entre n y los cuantiles del coeficiente γ obtenido por el método del mínimo.
Disimilaridad inicial aleatoria $U(0,1)$.

n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.454545	0.454545	0.636364	0.818182	1.000000	1.000000	1.000000
5	0.314286	0.371429	0.485714	0.636364	0.771429	0.885714	0.939394
6	0.240506	0.317647	0.432836	0.544304	0.645570	0.759036	0.808219
7	0.236994	0.283237	0.372781	0.462857	0.565714	0.657143	0.717791
8	0.188811	0.232258	0.309211	0.406250	0.487342	0.569620	0.615385
9	0.174419	0.212454	0.283951	0.357692	0.438679	0.510288	0.549587
10	0.162730	0.194245	0.256410	0.325815	0.398082	0.462871	0.503979
11	0.141123	0.171474	0.235690	0.298887	0.360447	0.413978	0.456667
12	0.134752	0.164304	0.219458	0.275806	0.332482	0.384435	0.412801
13	0.121743	0.148280	0.198579	0.253099	0.305715	0.354013	0.381142
14	0.117055	0.139103	0.181234	0.232498	0.286722	0.335585	0.361998
15	0.104725	0.129649	0.171748	0.215774	0.263269	0.311106	0.338557
16	0.099902	0.124112	0.164710	0.205541	0.247488	0.287008	0.312955
17	0.101269	0.121837	0.156131	0.197889	0.238081	0.276234	0.303901
18	0.084565	0.107917	0.142768	0.182978	0.221921	0.256064	0.284830
20	0.082561	0.097704	0.127842	0.162403	0.197669	0.228914	0.248301
25	0.062547	0.076516	0.103286	0.132089	0.162166	0.186245	0.198898

Tabla S2.1 : Relación entre el número de individuos de Ω (n) y la media (M_γ) y desviación típica (S_γ) muestral de gamma, para los métodos del mínimo y del máximo.

Disimilaridad inicial: $N(0,1)+10$.

n	Máximo		Mínimo	
	M_γ	S_γ	M_γ	S_γ
4	0.96	0.07	0.78	0.21
5	0.85	0.10	0.63	0.19
6	0.75	0.11	0.54	0.16
7	0.66	0.10	0.46	0.15
8	0.59	0.09	0.41	0.13
9	0.54	0.09	0.36	0.12
10	0.49	0.08	0.33	0.10
11	0.45	0.07	0.30	0.09
12	0.42	0.07	0.28	0.09
13	0.39	0.07	0.25	0.08
14	0.36	0.06	0.24	0.07
15	0.34	0.06	0.21	0.07
16	0.32	0.06	0.20	0.06
17	0.30	0.05	0.19	0.06
18	0.28	0.05	0.18	0.05
20	0.26	0.05	0.16	0.05
25	0.21	0.04	0.13	0.04

Media y Desv. Típica basadas en una muestra de 200 para $n=25$ y de 1000 para los restantes casos.

Tabla S2.2 : Relación entre n y los cuantiles del coeficiente γ obtenido por el método del máximo.

Disimilaridad inicial: $N(0,1)+10$.

n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.857143	0.857143	0.857143	1.000000	1.000000	1.000000	1.000000
5	0.666667	0.707317	0.789474	0.857143	0.909091	1.000000	1.000000
6	0.562500	0.604167	0.670103	0.752577	0.826087	0.881188	0.911111
7	0.497382	0.531250	0.589744	0.659341	0.730570	0.788889	0.823834
8	0.443478	0.485549	0.536313	0.596439	0.653409	0.712575	0.741573
9	0.396648	0.425703	0.476510	0.535461	0.600707	0.645224	0.678119
10	0.357631	0.388042	0.436385	0.495614	0.546724	0.597315	0.622685
11	0.318731	0.354594	0.408257	0.454145	0.508501	0.549206	0.574572
12	0.301925	0.332280	0.375315	0.422991	0.464930	0.503992	0.528438
13	0.276715	0.301058	0.342545	0.386388	0.432613	0.474747	0.499636
14	0.252436	0.275472	0.317522	0.361478	0.401646	0.446357	0.469355
15	0.244238	0.264328	0.297398	0.335317	0.374361	0.413344	0.432530
16	0.224121	0.247044	0.280323	0.316267	0.355871	0.390244	0.409652
17	0.209308	0.232318	0.264362	0.299587	0.337259	0.370527	0.389637
18	0.198914	0.215161	0.245291	0.282928	0.315273	0.351357	0.370177
20	0.181215	0.195358	0.225310	0.256405	0.288908	0.318573	0.333738
25	0.133563	0.154767	0.182345	0.213535	0.232355	0.254535	0.262130

Tabla S2.3 : Relación entre n y los cuantiles del coeficiente γ obtenido por el método del mínimo.

Disimilaridad inicial: $N(0,1)+10$.

n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.454545	0.454545	0.636364	0.818182	1.000000	1.000000	1.000000
5	0.314286	0.371429	0.485714	0.600000	0.771429	0.885714	0.931034
6	0.270588	0.317647	0.435294	0.549296	0.670588	0.753425	0.811765
7	0.211429	0.268571	0.360000	0.452229	0.558282	0.656051	0.697842
8	0.189873	0.245033	0.329193	0.412587	0.500000	0.576000	0.628378
9	0.171103	0.216117	0.286344	0.365019	0.438017	0.522222	0.562016
10	0.164678	0.199501	0.258294	0.328537	0.396509	0.465882	0.504808
11	0.150519	0.189103	0.239933	0.297297	0.361371	0.420792	0.454685
12	0.140897	0.168035	0.220506	0.272441	0.339880	0.390517	0.416342
13	0.118027	0.147201	0.196941	0.250720	0.306696	0.357882	0.383513
14	0.120685	0.143622	0.184044	0.234653	0.281645	0.335251	0.367704
15	0.104942	0.129388	0.161556	0.209871	0.257779	0.298329	0.323643
16	0.100263	0.122919	0.163416	0.204588	0.246898	0.286209	0.308933
17	0.100689	0.119771	0.153601	0.196364	0.234845	0.272871	0.296653
18	0.088932	0.109512	0.144822	0.181409	0.215341	0.248648	0.266223
20	0.081319	0.098071	0.127628	0.162602	0.199239	0.231434	0.254387
25	0.061431	0.071144	0.102345	0.132342	0.163325	0.183535	0.205345

En primer lugar, observando las tablas anteriores, nos damos cuenta de inmediato, de la gran similitud en los resultados de ambas pruebas, tanto en lo que se refiere a la media muestral del estadístico gamma, como a su desviación típica, como a sus cuantiles, para cualquiera de los valores de n considerados y cualquiera de los dos métodos de clasificación utilizados. Así pues, los resultados obtenidos parecen indicar que la relación entre el número de individuos y el valor esperado del coeficiente gamma, para los métodos del mínimo y del máximo, en el caso de una asignación aleatoria de las disimilaridades iniciales, no depende de la distribución utilizada en la generación de dichas disimilaridades.

Por lo que a la eficiencia de los dos métodos considerados concierne, se observa claramente en las tablas S1.1 y S2.1 que, para cada valor de n considerado, la media muestral de gamma es mayor en el caso del máximo que en el del mínimo.

Así pues, si deseamos comparar ambos métodos, en términos del valor esperado de gamma, podemos decir, a la vista de los resultados, que **el método del máximo es sensiblemente mejor que el del mínimo.**

Este resultado, por otra parte, no es sorprendente si tenemos en cuenta que el Algoritmo Piramidal del Mínimo, da lugar siempre a la jerarquía del mínimo (Teorema 5.4.1.1), que las clasificaciones piramidales son una generalización de las ultramétricas (Proposición 3.5.2) y que, por lo general, tal como se muestra en §5.6, las clasificaciones piramidales reflejan mejor la realidad que las jerárquicas. Por todo ello pues, resulta obvio el hecho de que el algoritmo piramidal del máximo sea mejor que el del mínimo.

Por otra parte, en las tablas S1.1 y S2.1, observamos que las medias muestrales de gamma y sus correspondientes desviaciones típicas, disminuyen a medida que el valor de n aumenta. Además, los resultados obtenidos por el algoritmo piramidal del mínimo se corresponden con los obtenidos en L.Hubert (1974), donde se establece una relación entre n, la media muestral de gamma y su desviación típica correspondiente para el caso de una clasificación jerárquica por el método del mínimo. Concretamente para $n \geq 16$ se observa que las medias muestrales de gamma se pueden aproximar por $1.1 \ln(n)/n$ y las desviaciones típicas correspondientes por $1/n$, con lo cual, para $n \geq 16$,

$$[n\gamma - 1.1 \ln(n)] \sim N(0,1)$$

A pesar de que L.Hubert enuncia esta aproximación de la distribución del estadístico gamma para $n \geq 16$, su tabla de resultados, a partir de la cual extrae el resultado anterior, sólo alcanza hasta $n=25$, y no da ningún argumento que permita validar su resultado para $n > 25$, lo cual nos hace desechar la idea de intentar una aproximación similar para el caso del algoritmo piramidal del máximo.

Puesto que, en general, no parece factible encontrar una expresión manejable para la función de distribución del estadístico γ , intentaremos una aproximación empírica a la misma.

Para ello, partiremos de la base de que, para cada n, las N disimilaridades iniciales generadas aleatoriamente -Uniformes o Normales- son incorrelacionadas, por tanto los **cuantiles** del coeficiente gamma obtenido a partir de dichas disimilaridades, nos proporcionarán una **buena aproximación de la función de distribución**

poblacional de dicho estadístico.

Sea pues una población $\Omega = \{\omega_1, \dots, \omega_n\}$ y sea $\Delta = \{\delta_\Omega\}$ la familia de todas las disimilaridades posibles sobre Ω .

Sea \mathcal{O} un proceso cualquier de transformación de una disimilaridad dada sobre Ω , en una disimilaridad piramidal.

Consideremos entonces la aplicación $\gamma_{\mathcal{O}}$, definida del modo siguiente:

$$\begin{array}{l} \gamma_{\mathcal{O}} : \Delta \longrightarrow \mathbf{R} \\ \delta_\Omega \longrightarrow \gamma_{\mathcal{O}}(\delta_\Omega) = \text{coeficiente de Goodman-Kruskal entre la disimilaridad } \delta_\Omega \\ \text{y la piramidal obtenida por el proceso } \mathcal{O} \end{array}$$

Si δ_Ω es una disimilaridad piramidal, su transformada por cualquiera de los procesos de clasificación piramidal, también lo será y coincidirá con δ_Ω , por consiguiente, lo lógico es esperar que $\gamma_{\mathcal{O}}(\delta_\Omega) = 1$. Si por el contrario δ_Ω es una disimilaridad totalmente aleatoria, esperaremos que $\gamma_{\mathcal{O}}(\delta_\Omega)$ sea un valor próximo a cero.

Considerando pues Δ una población de distancias generadas aleatoriamente, tiene sentido tabular la distribución de γ .

Dada una disimilaridad cualquiera, nos interesará saber si es susceptible de ser representada, lo más fielmente posible, mediante una estructura piramidal. Para ello se considerará la hipótesis nula H_0 de aleatoriedad en la asignación de la disimilaridad inicial δ_Ω , en el sentido de considerar que la población de disimilaridades a la que pertenece es el conjunto de disimilaridades aleatorias generadas mediante algún procedimiento determinado.

En este caso, el recorrido de $\gamma_{\mathcal{O}}$, bajo la hipótesis nula, viene dado por el

recorrido intercuartílico especificado en las tablas S1.2, S2.2 para el método del máximo y en S1.3, S2.3 para el método del mínimo.

Así pues, las tablas de cuantiles obtenidas, nos proporcionarán un test de significación de la bondad de ajuste de una clasificación piramidal (o un método empírico para contrastar una hipótesis nula de aleatoriedad en la asignación de la distancia inicial,) para los métodos el mínimo y del máximo.

En la práctica pues, si después de efectuar una clasificación piramidal, el valor del parámetro gamma supera el valor del cuantil Q_α , podremos rechazar la hipótesis de aleatoriedad de la disimilaridad inicial, a un nivel de significación $1-\alpha$.

6.4.- POTENCIA - EFICIENCIA DE LOS METODOS DE CLASIFICACION PIRAMIDAL

Las pruebas S3 y S4, más que proporcionarnos resultados inferenciales respecto a los métodos de clasificación piramidal, nos permitirán establecer algunos resultados relativos a la potencia-eficiencia de los mismos.

Recordemos que, para cada valor de n fijado, partimos de una disimilaridad, a la que llamamos Básica, que es piramidal en la prueba S3 y ultramétrica encadenada en S4. Las pruebas consisten en aplicar el algoritmo CAP por los métodos del mínimo y del máximo, a la disimilaridad básica perturbada con ciertas variables aleatorias $N(0,\sigma)$, donde σ toma tres valores distintos en cada caso y representa el nivel de perturbación (bajo, medio y alto).

En el análisis de los resultados de estas dos pruebas, tendremos en cuenta el caso en que los coeficientes gamma y rho provienen de comparar la disimilaridad piramidal obtenida con la disimilaridad Básica, así como el caso en que los coeficientes provienen de comparar la disimilaridad piramidal obtenida con la Perturbada. También compararemos los resultados teniendo en cuenta los distintos niveles de error considerados.

En las tablas que siguen, n indicará el número de individuos a clasificar, σ_i el nivel de error considerado, (1) Bajo, (2) Medio, (3) Alto, y M_γ , S_γ , M_ρ , S_ρ las medias y desviaciones típicas muestrales de los estadísticos γ y ρ respectivamente.

Las tablas S3.B y S4.B muestran los resultados obtenidos al comparar la disimilaridad Básica con la Piramidal obtenida.

Las tablas S3.P y S4.P muestran los resultados obtenidos al comparar la disimilaridad Perturbada con la Piramidal.

Tabla S3.B : Disimilaridad Básica: Piramidal.

n	σ_i	M_γ	S_γ	M_p	S_p
4	1	1.000000	0.000000	0.998936	0.000739
	2	0.916727	0.133341	0.904782	0.066595
	3	0.335430	0.354689	0.394067	0.360764
5	1	0.987892	0.026581	0.983291	0.010755
	2	0.789596	0.208644	0.844290	0.152156
	3	0.377528	0.256212	0.432351	0.259531
6	1	0.994795	0.013184	0.986065	0.006565
	2	0.854801	0.137645	0.854233	0.127122
	3	0.376799	0.261546	0.428450	0.262429
10	1	0.992340	0.013115	0.986603	0.005225
	2	0.816205	0.163061	0.880780	0.081670
	3	0.358006	0.206165	0.429746	0.184226
16	1	0.983140	0.015755	0.981912	0.013006
	2	0.857063	0.094006	0.909473	0.054290
	3	0.302856	0.180802	0.397812	0.132669
20	1	0.981385	0.023437	0.986111	0.004733
	2	0.835340	0.145253	0.899045	0.074134
	3	0.260310	0.121246	0.374279	0.117430
<hr/>					
4	1	1.000000	0.000000	0.855520	0.002208
	2	0.816750	0.289796	0.659578	0.207871
	3	0.214179	0.518503	0.229825	0.419401
5	1	0.629228	0.060457	0.681841	0.090533
	2	0.576979	0.177990	0.604922	0.141482
	3	0.313666	0.354496	0.349008	0.319595
6	1	0.459538	0.096304	0.480017	0.051688
	2	0.483029	0.137569	0.501290	0.102330
	3	0.283960	0.242764	0.311387	0.229420
10	1	0.574046	0.179950	0.632349	0.077029
	2	0.551360	0.199615	0.567456	0.112165
	3	0.230515	0.168298	0.249815	0.177826
16	1	0.566487	0.150930	0.628163	0.058011
	2	0.499535	0.152772	0.575599	0.078251
	3	0.175302	0.105273	0.251135	0.096610
20	1	0.542442	0.154461	0.619972	0.074104
	2	0.525712	0.155349	0.592349	0.094592
	3	0.157748	0.105557	0.243384	0.099738

MAXIMO

MINIMO

Tabla S4.B Disimilaridad Básica: Ultramétrica.

n	σ_i	M_γ	S_γ	M_p	S_p
4	1	1.000000	0.000000	0.984635	0.010020
	2	0.917964	0.144077	0.887273	0.105926
	3	0.554909	0.284048	0.530832	0.321110
5	1	1.000000	0.000000	0.983464	0.006871
	2	0.821680	0.159779	0.874868	0.084568
	3	0.396494	0.333871	0.479049	0.326205
6	1	0.997082	0.012901	0.982482	0.006324
	2	0.825469	0.100750	0.868830	0.061678
	3	0.478672	0.294509	0.535545	0.276628
10	1	0.968091	0.021287	0.981037	0.005729
	2	0.689353	0.080508	0.812760	0.070730
	3	0.333045	0.174514	0.407227	0.174060
16	1	0.936232	0.016631	0.979088	0.004848
	2	0.701604	0.108096	0.821757	0.063849
	3	0.266319	0.136567	0.340591	0.102383
20	1	0.923969	0.017198	0.976793	0.004134
	2	0.657375	0.082188	0.772622	0.071975
	3	0.180907	0.102152	0.251431	0.103897
<hr/>					
4	1	1.000000	0.000000	0.994457	0.007137
	2	0.931660	0.194026	0.942193	0.088869
	3	0.543493	0.527343	0.498172	0.509809
5	1	1.000000	0.000000	0.992262	0.008828
	2	0.877499	0.159116	0.908714	0.076388
	3	0.414355	0.461580	0.532239	0.376267
6	1	1.000000	0.000000	0.992293	0.004810
	2	0.899982	0.136944	0.925219	0.053651
	3	0.523246	0.362893	0.565890	0.343331
10	1	0.985614	0.038356	0.989125	0.009284
	2	0.853693	0.104144	0.913439	0.040591
	3	0.418342	0.243163	0.484517	0.254792
16	1	0.971394	0.032003	0.990607	0.004262
	2	0.803565	0.049525	0.923634	0.030909
	3	0.522316	0.121545	0.627650	0.114964
20	1	0.971563	0.021352	0.992287	0.002924
	2	0.851550	0.049203	0.935921	0.027519
	3	0.534067	0.121589	0.614589	0.105640

MAXIMO

MINIMO

Tabla S3.P Disimilaridad Perturbada: Piramidal+N(0, σ)

n	σ_i	M_γ	S_γ	M_p	S_p
4	1	0.960857	0.063716	0.999699	0.000498
	2	0.946549	0.078411	0.967694	0.059864
	3	0.974066	0.064798	0.963288	0.052603
5	1	0.981703	0.031727	0.993848	0.006938
	2	0.884111	0.101127	0.946929	0.055382
	3	0.893018	0.076672	0.918615	0.064048
6	1	0.938890	0.048859	0.989277	0.008523
	2	0.839027	0.093469	0.924298	0.048727
	3	0.756708	0.086255	0.838275	0.078473
10	1	0.842006	0.045966	0.982356	0.004636
	2	0.688543	0.090388	0.877818	0.052313
	3	0.497731	0.074527	0.692679	0.069257
16	1	0.775920	0.036153	0.973189	0.011557
	2	0.555637	0.049813	0.847948	0.023006
	3	0.353982	0.060228	0.583886	0.055093
20	1	0.818239	0.052714	0.975315	0.004756
	2	0.586930	0.088499	0.832361	0.049263
	3	0.280985	0.056276	0.531730	0.049905
<hr/>					
4	1	1.000000	0.000000	0.854957	0.012888
	2	0.889091	0.189989	0.757590	0.165093
	3	0.773091	0.221228	0.760415	0.168082
5	1	0.554787	0.118976	0.684271	0.097457
	2	0.618696	0.199041	0.668283	0.152112
	3	0.736246	0.189963	0.736751	0.167028
6	1	0.427191	0.097476	0.500046	0.070688
	2	0.513236	0.183463	0.599655	0.140979
	3	0.576798	0.146855	0.620867	0.139679
10	1	0.485997	0.182798	0.648936	0.076410
	2	0.443487	0.176222	0.600181	0.130335
	3	0.339196	0.100835	0.502271	0.106861
16	1	0.456448	0.147750	0.638764	0.065941
	2	0.351374	0.092233	0.586052	0.067911
	3	0.231756	0.055034	0.418177	0.103006
20	1	0.464659	0.162299	0.627587	0.069138
	2	0.365687	0.131610	0.571348	0.091115
	3	0.170089	0.050083	0.368809	0.052916

MAXIMO

MINIMO

Tabla S4.P Disimilaridad Perturbada: Ultramétrica+N(0, σ)

n	σ_i	M_γ	S_γ	M_p	S_p
4	1	1.000000	0.000000	0.997792	0.003058
	2	0.992429	0.032005	0.987488	0.020970
	3	0.971429	0.057143	0.975551	0.038555
5	1	0.989472	0.021849	0.994816	0.008523
	2	0.935729	0.070729	0.968107	0.028643
	3	0.862332	0.104843	0.897250	0.090569
6	1	0.967152	0.024619	0.991313	0.005625
	2	0.891089	0.070203	0.944601	0.037306
	3	0.767232	0.100799	0.860349	0.079458
10	1	0.933089	0.024307	0.984793	0.004349
	2	0.729074	0.065622	0.869795	0.055760
	3	0.547176	0.093250	0.709374	0.079951
16	1	0.891822	0.010015	0.975136	0.003482
	2	0.647284	0.074023	0.829115	0.047915
	3	0.359301	0.067465	0.575587	0.068236
20	1	0.876008	0.013435	0.972534	0.004265
	2	0.592427	0.058461	0.775821	0.051197
	3	0.282137	0.061769	0.513443	0.061098
<hr/>					
4	1	1.000000	0.000000	0.984117	0.012676
	2	0.934000	0.142148	0.897918	0.080612
	3	0.767273	0.196296	0.753800	0.138245
5	1	1.000000	0.000000	0.984792	0.008935
	2	0.885771	0.113641	0.901965	0.063220
	3	0.709413	0.169675	0.729640	0.153777
6	1	0.994753	0.011266	0.979959	0.007488
	2	0.832019	0.107759	0.855690	0.055003
	3	0.587117	0.203057	0.687468	0.161329
10	1	0.954268	0.033795	0.975501	0.009286
	2	0.699761	0.064567	0.820026	0.056189
	3	0.401620	0.097835	0.536629	0.121788
16	1	0.910758	0.020405	0.973899	0.004650
	2	0.644933	0.051660	0.816251	0.034280
	3	0.354708	0.061773	0.511280	0.075937
20	1	0.900471	0.012237	0.973774	0.002904
	2	0.636751	0.034260	0.810632	0.032018
	3	0.299549	0.042493	0.448876	0.047136

MAXIMO

MINIMO

Observando las tablas de los resultados obtenidos en las pruebas S3 y S4, ya sea al comparar la disimilaridad piramidal obtenida, con la básica (B) o con la perturbada (P), en general podemos apreciar como, para cada n , el valor de los parámetros γ y ρ disminuye a medida que el nivel de error considerado aumenta. En cambio, las desviaciones típicas aumentan con el nivel de error.

Así pues, la potencia del método disminuye a medida que el nivel de error considerado aumenta, la cual cosa era ya de esperar. No obstante cabe señalar que el descenso de la potencia respecto al nivel de perturbación es más acelerado en el caso de comparar la disimilaridad piramidal obtenida, con la básica (Tablas S3.B y S4.B), que al compararla con la disimilaridad perturbada, sobretodo para valores de n inferiores a 10 (Tablas S3.P y S4.P).

Matizando más este resultado, observemos que para el nivel de error más bajo, los valores obtenidos para γ y ρ al comparar con la disimilaridad básica, son prácticamente iguales a los obtenidos al comparar con la disimilaridad perturbada (Tablas S3.B-S3.P y S4.B-S4.P) y sobretodo en el caso del máximo son valores muy altos (superiores a 0.9), sea cual sea el valor de n . Esta gran similitud es debida, sin duda, a que si la disimilaridad básica es piramidal, la perturbada correspondiente (con un bajo nivel de error) prácticamente lo será también y por tanto al efectuar la clasificación correspondiente la piramidal obtenida será próxima a las dos.

En cambio, para valores de σ elevados los resultados al comparar con la disimilaridad básica son sensiblemente inferiores a los obtenidos al comparar con la perturbada, aunque éstos se van aproximando a medida que aumenta el valor de n . En el caso en que la disimilaridad básica sea ultramétrica y el método de clasificación el

del mínimo, el resultado obtenido es opuesto al resto de los casos, es decir se obtienen mejores resultados al comparar con la disimilaridad básica que con la perturbada. Esta situación es debida al hecho de que la ultramétrica sea encadenada, puesto que la perturbada (con un nivel de error alto) tendrá muy poco que ver con la ultramétrica encadenada, pero al efectuar la clasificación por el método del mínimo éste tenderá a producir de nuevo encadenamientos, con lo cual la disimilaridad piramidal obtenida (ultramétrica) será, sin duda, más próxima a la básica que a la perturbada.

Por otra parte, los valores obtenidos de los parámetros por el método del máximo son superiores a los obtenidos por el mínimo, sea cual sea el valor de n y el nivel de error considerados, exceptuando el caso en que la disimilaridad básica sea ultramétrica (encadenada), en cuyo caso el método del mínimo proporciona mejores resultados que el del máximo, resultado lógico también si tenemos en cuenta que el algoritmo piramidal del mínimo coincide con el algoritmo jerárquico del mínimo, y éste último, como es bien sabido, tiende a producir encadenamientos. Este resultado coincide también con el establecido por Hubert L. (1974) para el caso de las clasificaciones ultramétricas, teniendo en cuenta que Hubert realiza el estudio solamente para una población de 16 individuos.

En general pues, el método piramidal del máximo es más potente (recupera mejor la pirámide inicial) que el del mínimo, exceptuando el caso encadenado que ya hemos comentado.

Desde otro punto de vista, tanto en la prueba S3 como en la S4, podríamos considerar que si el nivel de error con que perturbamos las disimilaridades básicas es

alto, las disimilitudes perturbadas pueden considerarse prácticamente aleatorias. Por tanto puede tener sentido comparar los resultados obtenidos a partir de estas disimilitudes perturbadas con un nivel de error alto (Pruebas S3 y S4), con los obtenidos en el caso aleatorio (Pruebas S1 y S2). De dicha comparación surgen las siguientes tablas:

MAXIMO

n	$M_1(S1)$	$M_1(S2)$	$M_1(S3, \sigma_1)$	$M_1(S4, \sigma_1)$
4	0.96	0.96	0.97	0.97
5	0.85	0.85	0.89	0.86
6	0.74	0.75	0.76	0.77
10	0.49	0.49	0.50	0.55
16	0.32	0.32	0.35	0.36
20	0.26	0.26	0.28	0.28

MINIMO

n	$M_1(S1)$	$M_1(S2)$	$M_1(S3, \sigma_1)$	$M_1(S4, \sigma_1)$
4	0.78	0.78	0.77	0.77
5	0.63	0.63	0.74	0.71
6	0.54	0.54	0.58	0.59
10	0.33	0.33	0.34	0.40
16	0.21	0.20	0.23	0.35
20	0.16	0.16	0.17	0.30

Los resultados son bastante coincidentes, siendo, por lo general, algo mejores los provenientes de las pruebas S3 y S4, la cual cosa es hasta cierto punto natural puesto que las disimilaridades perturbadas provienen de ciertas disimilaridades básicas que son piramidales. Observe, asimismo, que en el caso del mínimo, los resultados de la prueba S4 son mejores que los obtenidos en las restantes pruebas, la cual cosa es también lógica, si tenemos en cuenta que las disimilaridades básicas en S4 son ultramétricas encadenadas y el método del mínimo tiende a producir encadenamientos.

CONCLUSIONES

En esta memoria, se ha desarrollado una metodología, para abordar problemas relativos a la clasificación y representación de un conjunto de objetos o individuos mediante un tipo especial de grafos, llamados grafos piramidales, que vienen a ser una generalización de los grafos ultramétricos en el sentido de admitir, a cada nivel, no solamente grupos disjuntos o encajados, sino también grupos solapados.

En primer lugar se ha situado este método de representación en el marco del Análisis Multivariante, resaltando las ventajas que representa frente a otros modelos de representación discretos, más conocidos y utilizados.

Dado el papel destacado que en las representaciones piramidales tienen las disimilaridades y los preórdenes, se caracterizan las matrices asociadas a las disimilaridades piramidales y se establecen relaciones con los preórdenes compatibles. Asimismo se establece que toda disimilaridad ultramétrica es también piramidal y en consecuencia que toda jerarquía total indexada es una pirámide indexada, lo cual nos permite considerar las clasificaciones piramidales como una generalización de las ultramétricas.

Se ha desarrollado también una nueva formalización de las bases matemáticas sobre las que se amparan las representaciones piramidales, a saber, la axiomática y la equivalencia entre una disimilaridad piramidal y una pirámide indexada (grafo piramidal).

Siguiendo algunas indicaciones del profesor Diday, en el sentido de profundizar en ciertas cuestiones teóricas que puedan conducir a una simplificación de las representaciones piramidales, a fin de que éstas puedan ser utilizadas cómodamente en la realización de trabajos prácticos, como actualmente lo son las representaciones ultramétricas, se ha modificado convenientemente el Algoritmo de Clasificación Ascendente Piramidal, descrito en Diday E. (1984-1), en el sentido de obtener una pirámide indexada lo más sencilla posible desde el punto de vista de la representación visual (pirámide indexada en sentido estricto). Se han caracterizado los grupos sobrantes en una estructura piramidal (aquellos que proporcionan información redundante), con lo cual se obtiene un criterio para simplificar una pirámide, sin que ello suponga pérdida de información. Finalmente se demuestra que las pirámides construidas por el algoritmo CAP descrito en el capítulo 5, escogiendo previamente un índice de agregación adecuado, una vez se han eliminado los grupos sobrantes, son pirámides indexadas en sentido estricto, sin cruzamientos ni inversiones.

Como complemento indispensable al algoritmo, y con el fin de facilitar su utilización en problemas reales de clasificación, se ha creado un programa informático llamado PIR.PRO, que a partir de una disimilaridad inicial dada, la deforma hasta transformarla en piramidal y construye el grafo correspondiente.

Debido a las características particulares del algoritmo estudiado en la memoria,

se establece que el algoritmo piramidal del mínimo es equivalente al algoritmo jerárquico del mínimo.

Un sencillo experimento de simulación nos viene a confirmar la mejora que suponen las representaciones piramidales frente a las ultramétricas, en el sentido de ser más próximas a la realidad, es decir las deformaciones piramidales de una cierta disimilaridad son más próximas a la inicial que las deformaciones ultramétricas.

Finalmente se han efectuado diversas pruebas de simulación por métodos de Montecarlo, a través de las cuales se ha tabulado, a partir de la distribución muestral del estadístico γ de Godman-Kruskal, un test de significación de la bondad de ajuste de una representación piramidal por los métodos del mínimo y del máximo. Se ha estudiado también, desde un punto de vista empírico, la potencia-eficiencia de ambos métodos, observándose que, en general, el método del máximo es bastante más eficiente que el del mínimo.

Los temas desarrollados en esta memoria podrían extenderse en las siguientes líneas:

- a) Del mismo modo que las disimilaridades piramidales están asociadas a las pirámides indexadas en sentido amplio y las ultramétricas a las jerarquías indexadas, podría tener interés intentar caracterizar las disimilaridades unívocamente asociadas a las pirámides indexadas en sentido estricto.
- b) Estudio de nuevos y más eficientes algoritmos de clasificación piramidal.

c) Mejorar el programa PIR.PRO de clasificación piramidal desde el punto de vista de la velocidad de ejecución (por ejemplo confeccionando un programa en C, que efectuase los cálculos de los coeficientes de adecuación, y que pudiese linkarse con el programa PIR.PRO) y de la representación gráfica de la pirámide obtenida.

El programa también podría mejorarse en el sentido de que permitiese construir no solamente una pirámide, sino todas las posibles. Para ello sería necesario que el programa pudiese ir modificando el orden a medida que lo va construyendo, la cual cosa no parece, en principio, fácil de conseguir.

Si esta mejora fuese posible, mediante pruebas de simulación, podría intentarse validar, no solamente la estructura piramidal, como hemos hecho en el Capítulo 6, sino también el orden inducido por esta estructura piramidal, con lo cual podríamos escoger siempre la mejor de las posibles pirámides obtenidas por un determinado método (recordemos que los algoritmos de clasificación no dan lugar a una única pirámide, exceptuando el caso del mínimo).

d) Desarrollar nuevas vías sobre inferencia en grafos piramidales que puedan ayudar a resolver problemas tales como la significación de los grupos a un determinado nivel, comparación con otros modelos etc.

ANEXO

A.1.- LOS PROGRAMAS HSIMUL*j*.SAS Y HSIMULA.CLUS

Para cada $j \in \{1,2\}$ y para cada $i \in \{1,2,3\}$, el programa **HSIMUL*j*.SAS**, utilizando el generador de variables aleatorias σ ***NORMAL**(μ) que posee el paquete estadístico **SAS**, generará una nueva matriz de disimilaridad que provendrá de perturbar la disimilaridad dada d_j , con la variable aleatoria ε_i . Cada uno de estos seis programas será ejecutado ocho veces con lo que obtendremos las 48 matrices citadas en la sección 5.6, que son almacenadas en unos determinados ficheros para más tarde ser utilizadas por los programas **HSIMULA.CLUS** y **PIR.PRO** de clasificación jerárquica y piramidal respectivamente.

El Programa **PIR.PRO** ya ha sido descrito ampliamente en el Capítulo 5 y su programación está detalladamente especificada en la sección A.2 de este Anexo. Los programas **HSIMULA.CLUS**, de clasificación jerárquica y **HSIMUL*j*.SAS**, de generación de variables aleatorias, vienen especificados a continuación.

FILE : HSIMULji SAS

DATA UNO;

INPUT X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15;

X1 = $d_j^{12} + \sigma_1 * \text{NORMAL}(0)$;

.

.

X5 = $d_j^{16} + \sigma_1 * \text{NORMAL}(0)$;

[$d_j^n = d_j(\omega_r, \omega_s)$; valores de la part.
triangular sup. de d_j]

.

.

X15 = $d_j^{56} + \sigma_1 * \text{NORMAL}(0)$;

CARDS;

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

;

* PROC PRINT VAR X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15;

PUT a1 X1 6.3/

a1 X2 6.3/ a8 X6 6.3/

a1 X3 6.3/ a8 X7 6.3/ a16 X10 6.3/

a1 X4 6.3/ a8 X8 6.3/ a16 X11 6.3/ a24 X13 6.3/

a1 X5 6.3/ a8 X9 6.3/ a16 X12 6.3/ a24 X14 6.3/ a32 X15 6.3/

FILE : HSIMULA CLUS

ASSIGN,

FILE = HSIMUL,

SPECIFICATION = 'HSIMULji LISTING'

READ SIMILARITY MATRIX,

CASES = 6,

KLINKAGES = 5,

FORMAT = (5F6.3),

INFILE = HSIMUL

HIERARCHY,

METHOD AVERAGE,

CLUSTERS MIN=2 MAX=2

COMPARE

STOP

A.2.- EL PROGRAMA PIR.PRO

```
code=3000
include "tdoms.pro"
include "tpreds.pro"
include "menu.pro"
domains
a | z=integer
  | r=real*
  | c=z*
  | p=c*
  | pp=p*
  | file=f;fl
database
  | agafat(z)
  | ni(real)
  | pin(real)
  | disfix(r)
  | tr(z,z,z)
  | trl(z,z,z,z,z)
  | pos(z,z,z,real)
  | dis(z,z,real)
  | grup(z,c,c,real)
  | ordre(p)
  | datos(c)
  | disponibles(c)
  | ldis(r)
  | numero(z)
  | elements(z)
  | parametre(z)
  | indicador(z)
predicates
  | controll(z,z,r)
  | leedis(z,z,r)
  | leedis1(z,z,r)
  | escriumetod(z)
dibuix
  | ajudar(integer)
  | continuacio(z,r)
  | finalex(z)
  | gaita
  | infila
  | tri(z,z,z,z,z)
  | col(z,z)
  | darpos(z)
  | generar(c)
  | nar
  | ll(z,c)
  | lll(z,z,c)
  | append(c,c,c)
  | append(p,p,p)
  | append(r,r,r)
  | inclos(c,c)
  | inclos(p,p)
  | member(z,c)
  | member(c,p)
  | delete(z,c,c)
  | delete(c,p,p)
  | delete(real,r,r)
  | sublista(c,c)
  | prefijo(c,c)
  | int(c,c,c)
```

disjunt(c,c)
 parella(c,c,r)
 exder(c,c,c)
 exizq(c,c,c)
 dist(c,p,p,p)
 extendre(c,c,c)
 inici(z,z,c)
 ordenabur(r,r)
 ordenabur(c,c)
 bloc(z,r)
 bdis(z,z,real)
 avant1(z,z,z,z)
 avant(z,z,c)
 gordre(c,p,p,p)
 extendrel(c,p,c)
 writebe(real)
 linea(r)
 marge(z)
 wrmat(z,z,r)
 long(c,z)
 long(r,z)
 piramide(z,r)
 piramidef(string,z,r)
 c(z,c,real)
 darnivell(r,real)
 net(c)
 extrems(c,c)
 eliminar(z,z,c,real,c,c)
 test1(z,c,c)
 test2(z,c,c)
 test3(c,c,c)
 test4(c,c,c,real)
 minim(real,real,real)
 maxi(real,real,real)
 novadis(z,z,z,c)
 dardis(z,z,z,real)
 nou numero(z)
 dispo(c)
 borrar
 unir(z,c,c)
 quitodis(c)
 nougrup(c,c,c)
 final(z,z,c)
 inv(c,c)
 invzap(c,c,c)
 creixent(z,z,c)
 disfinal(c,c,r)
 discom(z,c,r)
 cdis(z,z,real)
 mirod(c,c,z,real)
 sumar1(r,real)
 esperan(r,real)
 sumarlc(r,real)
 restarn(r,real,r)
 mult(r,r,r)
 corr(r,r,real)
 estr(z,c)
 index(z,real)
 piramide1
 interval(z,c,z)
 lloc(z,z)
 lloc1(z,c,z)
 mini(r,real)
 min1(r,real,real)

```

| listadis(z,c,r)
| explora(real,r,r)
| metode
| inicialitzar
| gesmem
| retocar
a | retocardis(z)
| retocargrup(z)
| perfitxer(z,r,z)
| segonsdis(r)
| restatot(r,r)
| restatoto(real,r,r)
| posineg(r,integer,integer)
└─ kruska(r,r)

```

```

goal system("type caratula"),readkey(esc),system(cis),piramide1.
clauses

```

```

└─ append([],X,X).
| append([X|L1],L2,[X|L3]):-append(L1,L2,L3).
| member(_,[]) :-!,fail.
| member(X,[X|_]):-!.
| member(X,[_|Z]):-member(X,Z).
| inclos([],_) :-!.
| inclos([X|Z],Y):-member(X,Y),inclos(Z,Y).
| delete(_,[],[]) :-!.
| delete(X,[X|Y],Z):-delete(X,Y,Z),!.
| delete(X,[A|Z],[A|W]):-delete(X,Z,W).
| sublist([],_) :-!.
| sublist([X|L],[X|M]):-prefijo(L,M),!.
| sublist(L,[_|M]):-sublista(L,M).
| prefijo([],_) :-!.
| prefijo([X|L],[X|M]):-prefijo(L,M).
| int([],[]) :-!.
| int([X|R],Y,[X|Z]):-member(X,Y),!,int(R,Y,Z).
| int([_|R],Y,Z):-int(R,Y,Z).
| disjunt(A,B):-int(A,B,M),M=[],!.
| disjunt(_,_) :-!,fail.
b | exder([A|B],[C|D],[C|X]):-A <> C,exder([A|B],D,X).
| exder([A|B],[A|D],[A|B]):-!,sublista(D,B).
| exder(_,_) :-!,fail.
| exizq([A|B],[C|D],[A|X]):-A <> C,exizq(B,[C|D],X).
| exizq([A|B],[A|D],[A|D]):-!,sublista(B,D).
| exizq(_,_) :-!,fail.

| ordenabur(L,S):-
|   append(X,[A|[B|Y]],L),A=B,!,append(X,[B|Y],M),ordenabur(M,S).
| ordenabur(L,S):-
|   append(X,[A|[B|Y]],L),B < A,!,append(X,[B|[A|Y]],M),ordenabur(M,S).
| ordenabur(L,L).

| bdis(A,A,0):-!.
| bdis(A,B,X):-dis(A,B,X),!.
└─ bdis(A,B,X):-dis(B,A,X),!.

```

```

writebe(X):-writef("%10",X).
linea([A|B]):-writebe(A),linea(B).
linea([]).
marge(0):-!.
c | marge(1):-write("      "),!.
  | marge(N):-write("      "),!,M=N-1,marge(M).
  | wrmat(M,M,[]):-!.
  | wrmat(N,M,L):-P=N-1,marge(P),X=M-N,append(A,B,L),long(A,Z),X=Z,
    | linea(A),nl,N1=N+1,wrmat(N1,M,B).

long([],0):-!.
long(_[B],Z):-long(B,M),Z=M+1.

piramide1:-menu(5,40,7,7,[ajuda,continuar],"pir,pro",0,A),ajudar(A),
nl,nl,menu(5,40,7,7,[per_fitxer,per_pantalla],"Entradades",0,I),
controll(I,N,D),nl,retractall(disfix(_)),asserta(disfix(D)),
d | bloc(N,D),repeat,system(cls),
  | menu(5,40,7,7,[canviar_problema,escollir_metode,sortir],"escollir",0,M),
  | finalex(M),
  | continuacio(N,D).

ajudar(1):-system("more <ajuda"),readkey(esc).
ajudar(_).

controll(1,N,D):-write("fitxer de dades"),NL,readln(Fd),openread(fl,Fd),
readdevice(fl),readint(N),
retractall(elements(_)),asserta(elements(N)),nl,
retractall(grup(_,_,_)),
retractall(dis(_,_)),
leedis(N,N,D),closefile(fl).
e | controll(2,N,D):-write("introdueix n° d'elements"),nl,readint(N),
  | retractall(elements(_)),asserta(elements(N)),nl,
  | retractall(grup(_,_,_)),
  | retractall(dis(_,_)),
  | write("introdueix distancies"),nl,leedis(N,N,D).

bloc(N,D):-inici(1,N,X),parella(X,X,D),asserta(datos(X)).

inici(A,N,[A|X]):-A <= N,!,asserta(grup(A,[A],[A],0)),B=A+1,inici(B,N,X).
inici(_,_):-!.

parella([],[],[]).
parella([A|B],[A|B],D):-!,parella([A|B],B,D).
parella([A|B],[E|F],[D1|D2]):-assertz(dis(A,E,D1)),!,parella([A|B],F,D2).
parella(_[B],[],D):-!,parella(B,B,D).

f | leedis(1,_,[]):-!.
  | leedis(N,M,D):-Z=N-1,leedis1(Z,M,X),leedis(Z,M,D2),append(X,D2,D).
  | leedis1(_,_M):-readterm(r,M).

finalex(2):-!.
finalex(1):-!,piramide1.
finalex(3):-!,exit.

g | continuacio(N,D):- menu(5,40,7,7,[maxim,minim,upgma],"POSSIBLESMETODES",0,In),
  | retractall(indicador(_)),asserta(indicador(In)),
  | metode,
  | menu(5,40,7,7,[fitxer,pantalla],"escriure a",0,Fi),
  | !,perfitxer(N,D,Fi),!.

```

```

metode:-inicialitzar,gesmem,retocar.

inicialitzar:-retractall(ldis(_),disfix(D),asserta(ldis(D)),
               retractall(disponibles(_),datos(X),asserta(disponibles(X)),
               retractall(ordre(_),asserta(ordre(I))),
               retractall(numero(_),elements(N),asserta(numero(N))).

gesmem:-retractall(tr(_,_)),retractall(tr1(_,_,_)),
         retractall(pos(_,_,_)),retractall(parametre(_)).
h |
retocar:-elements(N),retocardis(N),retocargrup(N),!.

retocardis(N):-dis(A,B,_),A>N,retract(dis(A,B,_)),fail.
retocardis(N):-dis(A,B,_),B>N,retract(dis(A,B,_)),fail.
retocardis(_):-!.

retocargrup(N):-grup(A,_,_),A>N,retract(grup(A,_,_)),fail.
retocargrup(_):-!.

perfitxer(N,D,1):-nl,write("introdueix nom fitxer"),NL,readln(F),
                 piramidef(F,N,D),!,dibuix.
i |
perfitxer(N,D,2):-piramide(N,D),!,dibuix.

piramide(N,D):-nl,nl,indicador(Ind),escriumetod(Ind),nl,nl,
               write("MATRIU DE DISSIMILARITAT INICIAL"),nl,
               nl,wrmat(1,N,D),retractall(ni(_),!,repeat,ldis(Dd),darnivell(Dd,Ni),
               c(M,Nv,Ni),final(M,N,Nv),!,datos(Dat),disfinal(Dat,Dat,Dpiramidalinicial),
j |               nl,nl,ordre([Ordre]),disfinal(Ordre,Ordre,Dxx),
               nl,write("MATRIU DISSIMILARITAT PIRAMIDAL"),nl,nl,wrmat(1,N,Dxx),
               nl,write("COEFICIENT DE CORRELACIO COFENETICA"),
               corr(D,Dpiramidalinicial,Rc),nl,
               write(Rc),kruska(D,Dpiramidalinicial).

escriumetod(1):-nl,nl,write(" METODE DEL MAXIM"),nl,nl,!.
escriumetod(2):-nl,nl,write(" METODE DEL MINIM"),nl,nl,!.
escriumetod(3):-nl,nl,write(" METODE UPGMA"),nl,nl,!.

piramidef(F,N,D):-openwrite(f,F),writedevice(f),piramide(N,D),closefile(f).
piramidef(_,_):-closefile(f).

darnivell(_):-numero(K),elements(N),grup(K,_ ,V,_),long(V,N),!.
darnivell(Z,X):-not(ni(_)),mini(Z,X),asserta(ni(X)).
k | darnivell(_ ,X):-agafat(1),ni(X).
darnivell(_ ,X):-agafat(1),ni(X),!.
darnivell(_ ,X):-ni(M),ldis(D),explora(M,D,F),retract(ldis(D)),asserta(ldis(F)),
               mini(F,X),retractall(ni(_)),!,asserta(ni(X)).

c(M,Nv,Ni):-retractall(agafat(_)),dis(A,B,Ni),disponibles(Dispo),
            member(A,Dispo),member(B,Dispo),avantl(A,B,P,S),
            not(grup(_,[P,S],_)),not(grup(_,[S,P],_)),
            avant(P,S,Nv),nounumero(M),asserta(grup(M,[P,S],Nv,Ni)),!,
            eliminar(P,S,Nv,Ni,Dispo,M4),inv(M4,M5),
            novadis(M,P,S,M4),listadis(M,M5,Lup),segonsdis(Lup),
            str_int(Xg,M),frontchar(Gr,'h',Xg),nl,
            write("grup ",Gr,": estructura ",P," ",S,"; valor ",Nv,"; index ",Ni),
            nl,ordre(Ori),write("nou ordre: ",Ori),
            unir(M,M4,Nd),quitodis(Nd),nl,write("disponibles ",Nd),
            asserta(agafat(1)),!,nl.

```

```

— avant1(A,B,A,B):-grup(_,[A],_),grup(_,[B],_),!,fail.
| avant1(A,B,A,B):-grup(_,[A,B],_),!,fail.
| avant1(A,B,A,B):-grup(_,[B,A],_),!,fail.
| avant1(A,B,A,B):-grup(_,[B],_),!,fail.
| avant1(A,B,A,B):-grup(_,[A],_),!,fail.
| avant1(A,B,B,A):-grup(_,[A],_),!,fail.
| avant1(A,B,B,A):-grup(_,[B],_),!,fail.
| avant1(A,B,A,B).
| avant1(A,B,B,A).

| avant(A,B,Nv):-grup(A,_ ,Va,_),grup(B,_ ,Vb,_),nougrup(Va,Vb,Nv),ordre(O),
|   dist(Nv,O,I,D),gordre(Nv,I,D,Of),retract(ordre(_)),
|   asserta(ordre(Of)),!.

| nougrup(A,B,C):-exizq(A,B,C),!.
| nougrup(A,B,C):-disjunt(A,B),append(A,B,C).
| nougrup(_,_):-!,fail.
1 |

| dist(_,[],[],[]).
| /* dist(Nv,[S],[S],[]):-disjunt(Nv,S),!.
| dist(_,[S],[],[S]):-!.*/
| dist(Nv,[S],[S],[]):-sublista(Nv,S),!.
| dist(Nv,[S1|Ss],P,[S1|Ad]):-disjunt(Nv,S1),!,dist(Nv,Ss,P,Ad).
| dist(Nv,[S1|Ss],Ap,D):-sublista(S1,Nv),!,dist(Nv,Ss,Ap,D).
| dist(Nv,[S1|Ss],[S1|Ap],D):-dist(Nv,Ss,Ap,D).

| extendre(Nv,S,R):-exder(Nv,S,R),!.
| extendre(Nv,S,R):-exizq(Nv,S,R),!.
| extendre(_,_):-!,fail.

| gordre(Nv,P,D,Of):-extendrel(Nv,P,R),append([R],D,Of).

| extendrel(Nv,[],Nv).
| extendrel(Nv,[A],A):-sublista(Nv,A),!.
| extendrel(Nv,[A],P):-extendre(Nv,A,P),!,net(P).
| extendrel(Nv,[A,B],P):-extendre(Nv,A,R),extendre(R,B,P),!,net(P).
— extendrel(_,_):-!,fail.

segonsdis(_):-indicador(1),!.
segonsdis(_):-indicador(2),!.
segonsdis(Lup):-indicador(3),!,retract(ldis(Ldi)),append(Lup,Ldi,Noval),
  asserta(ldis(Noval)).

net([]).
net([A|B]):-member(A,B),!,fail.
net(_|B):-net(B).

extrems([A],[A,A]).
extrems([A|B],[A,C]):-append(_,[C],B).

nounumero(N):-retract(numero(X)),N=X+1,asserta(numero(N)),!.

```

```

eliminar(P,S,Nv,D,M,M4):-test1(P,M,M1),!,test2(S,M1,M2),!,test3(M2,M3,Nv),!,
test4(M3,M4,Nv,D),!.

test1(P,M,M1):-grup(,_,P,_,_),!,delete(P,M,M1),!.
test1(,_,M,M):-!.

test2(S,M,M1):-grup(,_,S,_,_),!,delete(S,M,M1),!.
test2(,_,M,M):-!.

test3([A|B],X,Nv):-grup(A,_,V,_,_),inclos(V,Nv),extrems(V,E1),extrems(Nv,E2),
int(E1,E2,D),I=[],!,test3(B,X,Nv).
test3([A|B],[A|X],Nv):-!,test3(B,X,Nv).
test3([],_,_):-!.

m | test4([A|B],X,Nv,D):-grup(A,_,V,D1),D1=D,inclos(V,Nv),D1=D,!,test4(B,X,Nv,D).
test4([A|B],[A|X],Nv,D):-!,test4(B,X,Nv,D).
test4([],_,_):-!.

dispo(M4):-retract(disponibles(_)),asserta(disponibles(M4)),!.

borrar.

unir(Z,X,W):-append(X,[Z],W),!.

quitodis(Nd):-retract(disponibles(_)),asserta(disponibles(Nd)),!.

maxi(A,B,A):-B <= A,!.
maxi(,_,B):-!.

minim(A,B,A):-A <= B,!.
minim(,_,B):-!.

novadis(N,P,S,[M1|M2]):-dardis(P,S,M1,D),!,asserta(dis(N,M1,D)),novadis(N,P,S,M2).
novadis(,_,_,[]):-!.

dardis(P,_,P,0):-!.
dardis(,_,S,0):-!.
dardis(M,_,M,0):-!.
n | dardis(,_,M,M,0):-!.
dardis(,_,M,0):-numero(K),grup(K,_,Nv,_,_),grup(M,_,Mv,_,_),inclos(Mv,Nv),!.
dardis(P,S,M,D):-indicador(1),!,bdis(P,M,D1),bdis(S,M,D2),maxi(D1,D2,D).
dardis(P,S,M,D):-indicador(2),!,bdis(P,M,D1),bdis(S,M,D2),minim(D1,D2,D).
dardis(P,S,M,D):-bdis(P,M,D1),bdis(S,M,D2),grup(P,_,Ep,_,_),grup(S,_,Es,_,_),
long(Ep,Lp),long(Es,Ls),K=((D1*Lp)+(D2*Ls))/(Lp+Ls),numero(Nu),
index(Nu,lnu),lnu < K,!,D=K.
dardis(,_,_,D):-numero(M),index(M,D).

o | final(,_,N,Nv):-long(Nv,N),!.

inv(L1,L2):-invzap(L1,[],L2).
invzap([X|L],L2,L3):-invzap(L,[X|L2],L3).
invzap([],L,L).

disfinal([],[],[]).
disfinal([A|B],[A|B],D):-discom(A,B,D1),disfinal(B,B,D2),append(D1,D2,D).
discom(,_,[],[]).
discom(A,[C|D],[D1|D2]):-cdis(A,C,D1),!,discom(A,D,D2).
p | cdis(A,B,D1):-grup(A,_,Va,_,_),grup(B,_,Vb,_,_),mirod(Va,Vb,B,D1).
creixent(A,N,[A|X]):-A <= N,!,B=A+1,creixent(B,N,X).
creixent(,_,_,[]):-!.
mirod(A,B,C,D):-grup(C,_,Vc,D),inclos(A,Vc),inclos(B,Vc).
mirod(A,B,C,D):-C1=C+1,mirod(A,B,C1,D).

```

```
/* COEFICIENT DE CORRELACIO*/
```

```

sumari([],0).
sumari([X|Y],Z):-sumari(Y,Sy),Z=Sy+X.
esperan(X,Ex):-sumari(X,Sx),long(X,Lx),Ex=Sx/Lx.
sumaric([],0).
sumaric([X|Y],Z):-sumaric(Y,Sy),Z=Sy+X*X.
restarn([],_).
q | restarn([X|Y],C,[R|S]):-R=X-C,restarn(Y,C,S).
| mult([],[],[]).
| mult([A|B],[C|D],[E|F]):-E=A*C,mult(B,D,F).
corr(D,P,Rc):-esperan(D,M),esperan(P,N),sumaric(D,Z),sumaric(P,T),long(D,L)
,A=Z/L,B=T/L,restarn(D,M,V),restarn(P,N,W),mult(V,W,U),esperan(U,C),
Vd=A-(M*M),Dd=sqrt(Vd),Vp=B-(N*N),Dp=sqrt(Vp),not(Dd*Dp=0),!,Rc=C/(Dd*Dp).
corr(_,_):write("CORRELACIO INDETERMINADA").

```

```
/* COEFICIENT DE GOODMAN...*/
```

```

restatot([],[]).
restatot([X|Y],L):-restatoto(X,Y,Z),
restatot(Y,M),append(Z,M,L).
restatoto(_,[],[]).
restatoto(X,[Y|Z],[A|B]):-A=X-Y,restatoto(X,Z,B).
posineg([],0,0).
r | posineg([0|B],P,N):-!,posineg(B,P,N).
| posineg([A|B],P,N):-A>0,!,posineg(B,P1,N),P=P1+1.
| posineg([A|B],P,N):-A<0,!,posineg(B,P,N1),N=N1+1.
| kruska(L1,L2):-restatot(L1,D1),restatot(L2,D2),mult(D1,D2,PD12),
posineg(PD12,P,N),not(P+N=0),!,X=(P-N)/(P+N),nl,nl,
write("COEFICIENT DE GOODMAN-KRUSKAL"),nl,write(X).
kruska(_,_):-nl,write("INDETERMINAT").

```

```
dibuix:-nl,nl,nl,
```

```
write("si desitges visualitzar la piramide fes intro, si no escape"),
!,readkey(cr),graphics(2,6,0),gaita, fail.
```

```

gaita:-numero(N),index(N,P),Par=round(30000/(P+1)),asserta(parametre(Par)),
line(1000,1500,30000,1500,1),
datos(D),long(D,L),Ld=L+1,ordre([O]),interval(Ld,I,N),append(O,I,M),
generar(M),tr1(X,Y,Z,T,W),tri(X,Y,Z,T,W), fail.
gaita:-retractall(pin(_)),numero(Nu),index(Nu,Par),asserta(pin(Par)),!,infila.

```

```
/* imprimeix indexos */
```

```

infila:-grup(_,_),X),pin(Par),Pp=24/(Par+1),Y=24-round(X*Pp),
cursor(Y,1),write(X), fail.
infila:-cursor(1,5),write("_"),readln(_),text,!, fail.

```

```

tri(F1,C1,F2,C2,H):-C2<C1,!,tri(F2,C2,F1,C1,H).
tri(F1,C1,F1,C2,F1):-!,line(F1,C1,F1,C2,1).
tri(F1,C1,F2,C2,F2):-!,D=C2-C1,Pas=D/4,C11=C1+Pas,
line(F1,C1,F2,C11,1),line(F2,C11,F2,C2,1).
tri(F1,C1,F2,C2,F1):-!,D=C2-C1,Pas=D/4,C21=C2-Pas,
line(F1,C1,F1,C21,1),line(F1,C21,F2,C2,1).
tri(F1,C1,F2,C2,H):-D=C2-C1,Pas=D/4,C11=C1+Pas,C21=C2-Pas,
line(F1,C1,H,C11,1),line(H,C11,H,C21,1),
line(H,C21,F2,C2,1).

```

```
col(X,Y):-pos(X,_Y,_).
```

```

darpos(X):-grup(X,[X],_),!,lloc(X,Lx),elements(Num),C=Lx*round(30000/(Num+1)),
asserta(pos(X,30000,C,0)),
Z=round(Lx*75/(Num+1)),cursor(24,Z),str_int(M,X),frontchar(Y,'h',M),write(Y).

```

```

darpos(X):-grup(X,[Y,Z],_,_),!,col(Y,C1),col(Z,C2),grup(X,_,_,D),
parametre(Par),F=round(30000-I*Par),C=C1+(C2-C1)/2,!,asserta(pos(X,F,C,D)).

generar().
generar([X|Y]):-grup(X,[X],_,_),!,darpos(X),generar(Y).
generar([H|W]):-estr(H1,[Ho,H]),index(H1,I),index(H,D),estr(Ho,[L,Hb]),
estr(H,[Hb,Hbp]),!,darpos(H),pos(H,F,C,D),pos(Hbp,F1,C1,_),Pas=(C1-C)/2,C11=C+Pas,
line(F,C,F,C11,1),line(F,C11,F1,C1,1),!,generar(W).
generar([H|W]):-estr(H1,[H,Ho]),index(H1,I),index(H,D),estr(Ho,[Hb,_]),
estr(H,[Hbp,Hb]),!,darpos(H),pos(H,F,C,D),pos(Hbp,F1,C1,_),Pas=(C-C1)/2,C11=C-Pas,
line(F1,C1,F,C11,1),line(F,C11,F,C,1),!,generar(W).
generar([H|W]):-estr(H,[H1,H2]),estr(Ho,[Hb,Hc]),estr(H1,[L,Hb]),estr(H2,[Hc,_]),
!,darpos(H),pos(Ho,F,C,_),pos(H,F1,C1,_),line(F,C,F1,C1,1),pos(H2,F2,C2,_),
index(H,I),parametre(Par),Fi=round(30000-Ih*Par),pos(H1,F3,C3,_),
asserta(tr1(F3,C3,F2,C2,Fi)),!,generar(W).
generar([X|W]):-darpos(X),estr(X,[Y,Z]),
pos(Y,F1,C1,_),pos(Z,F2,C2,_),index(X,D),parametre(Par),F=round(30000-I*Par),
asserta(tr1(F1,C1,F2,C2,F)),!,generar(W).

nar:-write("introduiu nom fitxer"),nl,readln(X),consult(X).

ll(X,Y):-lll(X,1,Y).

lll(X,X,[X]):-!.
lll(X,Y,[Y|Z]):-S=Y+1,lll(X,S,Z).

estr(X,Y):-grup(X,Y,_,_).

index(X,Y):-grup(X,_,_,Y).

interval(X,[X],X):-!.
interval(X,[X|Z],F):-Y=X+1,interval(Y,Z,F).

lloc(X,Y):-ordre([O]),lloc1(X,O,Y).

lloc1(X,[X|_],1):-!.
lloc1(X,[_|Z],M):-lloc1(X,Z,K),M=K+1.

mini([X|C],M):-min1(C,X,M).

min1([],X,X).
min1([C1|C2],X,M):-C1<X,!,min1(C2,C1,M).
min1([_|C2],X,M):-min1(C2,X,M).

listadis(_,[],[]).
listadis(X,[A|B],[C|D]):-dis(X,A,C),listadis(X,B,D).

explora(_,[],[]).
explora(X,[A|B],C):-A<=X,!,explora(X,B,C).
explora(X,[A|B],[A|C]):-explora(X,B,C).

```

A.3.- Descripción del Programa PIR.PRO

Ya sabemos que a partir de una disimilaridad inicial, definida entre los individuos de una población finita, y de un índice de agregación fijado, el programa **PIR.PRO** construye una pirámide siguiendo los pasos y las condiciones del Algoritmo de Clasificación Ascendente Piramidal (§5.2).

En la presente sección, efectuaremos una breve descripción de las partes más importantes del programa:

- a:** Lista de los nombres de los predicados utilizados por el programa.
- b:** Predicados de uso general.
- c:** Escribe las matrices de disimilaridad.
- d:** Inicia la ejecución del programa.
- e:** Controla la entrada de datos.
- f:** Lee las distancias iniciales.
- g:** Permite elegir el método de clasificación.
- h:** Permite cambiar de método sin necesidad de volver a introducir los datos.
- i:** Permite guardar los resultados de la clasificación en un fichero.
- j:** Es el núcleo central del programa y es quien rige la construcción de la pirámide.
- k:** Asigna a cada grupo su nivel.
- l:** Controla la formación de grupos y la construcción del orden piramidal.
Si la unión de dos grupos es compatible con el orden construido hasta el

momento, forma un nuevo grupo y modifica convenientemente el orden, en caso contrario rechaza la unión de dichos grupos.

- m:** Elimina grupos sobrantes del recubrimiento correspondiente.
- n:** Calcula la distancia del nuevo grupo a los restantes grupos disponibles.
- o:** Controla el final del proceso.
- p:** Escribe la matriz de disimilaridad piramidal construida, distribuyendo a los individuos de la población según el orden piramidal construido.
- q:** Calcula el coeficiente de correlación cofenética.
- r:** Calcula el coeficiente de Goodman-Kruskal.

Finalmente el programa efectúa la representación gráfica de la pirámide construida.

A.4.- RELACION ENTRE EL PROGRAMA PIR.PRO Y EL ALGORITMO CAP

En este párrafo vamos a poner de manifiesto que el programa **PIR.ROP** sigue los pasos y satisface las condiciones del Algoritmo de Clasificación Ascendente Piramidal.

En primer lugar, el programa, al igual que el algoritmo, se inicia con los individuos a clasificar y las distancias definidas entre los mismos.

En un paso cualquiera, el algoritmo une los dos grupos a distancia mínima y define el índice del nuevo grupo como la distancia entre los dos grupos que lo forman. El programa, también une los grupos a distancia mínima y define de la misma forma el

índice del nuevo grupo.

Las condiciones del algoritmo son también satisfechas por el programa:

* Un grupo no puede ser unido con otro más de dos veces. Esta condición viene garantizada por los predicados "test1" y "test2" apartado (m).

* Mediante el predicado "avant" , apartado (l) del programa, éste va construyendo un orden total sobre los individuos a medida que va formando los grupos, de manera que cada uno de ellos sea conexo respecto a dicho orden (Condición A.2.1.)

* En cada paso el programa siempre intenta unir en primer lugar el último grupo formado, si ello no es posible, lo intenta con el penúltimo, i así sucesivamente. Esto asegura el que si existen dos grupos que pueden ser unidos con un tercero y uno de los dos está contenido en el otro, tendrá prioridad el mayor de los dos puesto que habrá sido formado con posterioridad, (Condición A.2.2.).

* La condición A.2.3. del algoritmo es satisfecha por el predicado "test4", apartado (m) del programa.

* La condición A.2.4. se cumple gracias al predicado "test3" del programa.

Finalmente, el programa va reiterando el proceso, igual que el algoritmo, hasta obtener el grupo total Ω .

A.5.- EJECUCION DEL PROGRAMA PIR.PRO

En primer lugar se llama al programa con la instrucción "PIR". Aparecerá entonces una carátula con el nombre del programa y el año de su creación:

```
*****
*****      **      *****          *****      *****      *****
*      *      *      *      *      *      *      *      *
*****      **      *****          *****          *****      *      *
*      **      *      *      *      *      *      *      *
*      **      *      *      *      *      *      *      *
*****
1993
*****
Fes ESCAPE
```

Después de hacer "ESCAPE", aparecerá un menú con dos opciones:

```
┌ Pir.Pro
│ Ajuda
│ Continuar
└
```

Con "Ajuda" entramos en un fichero en el que se explica esquemáticamente el funcionamiento del programa.

Si deseamos "Continuar", aparecerá otro menú, referido a la forma en que deseamos entrar los datos:

```
┌ Entrada de dades
│ Per Fitxer
│ Per Pantalla
└
```

Si deseamos entrar los datos desde un fichero externo, "Per Fitxer", bastará con especificar el nombre del fichero en el que se encuentran los datos. Este fichero deberá ser un fichero DOS i, suponiendo que el número de individuos a clasificar sea n, deberá tener el siguiente formato:

```

n
[d(1,2),d(1,3),.....,d(1,n)]
[d(2,3),d(2,4),.....,d(2,n)]
[d(3,4),d(3,5),....,d(3,n)]
.
.
.
[d(n-1,n)]

```

Si deseamos introducirlos a través de la terminal, "Per Pantalla", deberemos indicar igualmente el número de individuos a clasificar y sus interdistancias, en el mismo formato anterior.

Una vez introducidos los datos aparecerá un nuevo menú:

```

—Escollir—
canviar problema
escollir mètode
sortir

```

La primera opción, "canviar problema", nos permite cancelar la ejecución del programa e iniciarlo de nuevo con otros datos. La última, "sortir", es para salir de él.

Si seleccionamos "escollir mètode", aparecerá un nuevo menú que nos permite escoger el método por el cual deseamos efectuar la clasificación:

```

—Possibles Mètodes—
Màxim
Mínim
UPGMA

```

Después de seleccionar el método, un nuevo menú nos permitirá decidir si queremos guardar el resultado en un fichero externo, o bien visualizarlo por pantalla.

```

—Escriu a—
fitxer
pantalla

```

En el primer caso el programa nos pedirá el nombre del fichero.

En cualquier caso, una vez efectuada la clasificación, el programa nos pedirá si queremos visualizar el grafo piramidal correspondiente, en caso afirmativo bastará con hacer "intro".

Cabe señalar que el grafo no quedará guardado en el fichero de resultados, caso de haber elegido dicha opción, si nos interesa conservarlo, deberemos capturarlo con algún programa de captura de gráficos del procesador de textos del que dispongamos, por ejemplo el programa **GRAB** del **WP51**.

A.6.- SALIDA DE RESULTADOS

La salida de resultados es encabezada por el nombre del método utilizado para efectuar la clasificación, seguido de la parte triangular superior de la matriz de disimilaridades iniciales, con los individuos ordenados de 1 a n.

A continuación se especifican, de forma ordenada, los nuevos grupos que se van formando, h_{n+1} , h_{n+2} , ... indicando en cada caso su **estructura**, su **valor** y su **índice**.

A cada paso, la especificación **nou ordre** [[...][...] ...] indica de qué forma el programa va construyendo el nuevo orden sobre la población, respecto al cual cada grupo será conexo. Y en la especificación **disponibles** se indican los grupos que, después de la formación del nuevo grupo, son susceptibles de ser unidos con algún otro.

Una vez finalizada la formación de grupos aparece la matriz de la disimilaridad piramidal equivalente a la pirámide construida, con los individuos ordenados según el orden piramidal indicado después de la formación del último grupo.

Como medidas de adecuación entre la disimilaridad inicial y la piramidal obtenida, el programa proporciona dos coeficientes:

El coeficiente de correlación cofenética entre las dos matrices, definido por

$$\rho = \text{Cov}(d, \delta) / \sigma(d) \cdot \sigma(\delta)$$

Donde d es la disimilaridad inicial y δ la piramidal obtenida.

Y el coeficiente de Goodman-Kruskal definido por

$$\gamma = (S_+ - S_-) / (S_+ + S_-)$$

donde

S_+ = Número de "Pares Concordantes", esto es: número de pares de pares de individuos $\{i,j\}$, $\{k,l\}$ tales que $(d_{ij} - d_{kl}) \cdot (\delta_{ij} - \delta_{kl}) > 0$.

S_- = Número de "Pares Discordantes", esto es: número de pares de pares de individuos $\{i,j\}$, $\{k,l\}$ tales que $(d_{ij} - d_{kl}) \cdot (\delta_{ij} - \delta_{kl}) < 0$.

Si el denominador de alguno de los dos coeficientes es cero, no puede efectuarse el cálculo i el programa da el mensaje "INDETERMINAT".

Finalmente, si lo deseamos, el programa nos proporciona el grafo piramidal.

A.7.- LOS PROGRAMAS DE SIMULACION SIMULU.PRO, SIMULN.PRO Y NSIMUL.PRO

Estos programas, escritos en Turbo Prolog, permiten la generación automática de matrices de disimilaridad y su transformación piramidal por los métodos del mínimo y del máximo. Asimismo proporcionan, como coeficientes de adecuación entre la disimilaridad inicial la piramidal obtenida, el coeficiente de correlación cofenética y el

de Goodman-Kruskal, que son almacenados en ficheros DOS. Permiten, finalmente, la reiteración automática del proceso, tantas veces como el usuario desee.

Para cada uno de los programas, el subprograma que efectúa la clasificación piramidal es el **PIR.PRO** (ver § A.2). La diferencia básica pues entre los mismos, estriba en la generación de la disimilaridad inicial. Vamos a describir brevemente este último aspecto en cada uno de los tres programas.

SIMULU.PRO

Si el usuario indica el número de individuos a clasificar y el número de veces que desea reiterar el proceso, el programa genera, en cada caso, una matriz de disimilaridad aleatoria con valores $U(0,1)$, efectúa una clasificación piramidal por los métodos del mínimo y del máximo y calcula el coeficiente gamma de Goodman-Kruskal. Si n es el número de individuos a clasificar, estos resultados son almacenados en los ficheros GUMAX n y GUMIN n .

El generador de valores uniformes utilizado es el que lleva incorporado el TBP.

SIMULN.PRO

En este caso, fijado también el número de individuos a clasificar, el programa genera una matriz de disimilaridad con valores $N(0,1)$ a los cuales suma la constante 10 para evitar la presencia de valores negativos, efectúa también una clasificación piramidal por los métodos del Mínimo y del Máximo y calcula el coeficiente gamma. Estos resultados son almacenados en los ficheros GNMAX n y GNMIN n .

El método elegido para la generación de valores Normales es el de Box-Muller,

descrito en Sanchez-Ocaña (1990), que consiste en generar, no una sino dos variables Normales a partir de sendas variables $U(0,1)$ del modo siguiente:

$$\begin{aligned} &\text{Sean } M \text{ y } N \text{ dos valores } U(0,1). \\ &A=2\cdot M-1, B=2\cdot N-1, R=A^2+B^2, R < 1, \\ &F=\sigma\cdot\text{sqrt}(-2\text{ Ln}(R)/R), \\ &Y=A\cdot F, Y'=B\cdot F \end{aligned}$$

Y, Y' serán entonces dos variables $N(0,\sigma)$.

La programación TBP de este generador es la siguiente:

```
llistar(X,[Y|Z]):-repeat, random(M), random(N),
  A=(M*2)-1, B=(N*2)-1, R=(A*A)+(B*B), R < 1,
  Fac=sqrt(-2*log(R)/R)*K, Y=(A*Fac),
  X1=X-1,!,llistar(X1,Z,K).
```

Teniendo en cuenta que el programa SIMULN.PRO debe generar valores $N(0,1)+10$, basta modificar el listado anterior con: $K=1, Y=(A*Fac)+10, Y > 0$, `llistar(X1,Z)`.

NSIMUL.PRO

Este programa, a diferencia de los anteriores, permite la introducción de una matriz de disimilaridad, a la que llamaremos **Básica**, la cual es perturbada con una variable $N(0,\sigma)$ generada automáticamente por el método Box-Muller. A partir de esta disimilaridad perturbada, el programa efectúa, como en los casos anteriores, una clasificación piramidal por los métodos del mínimo y del máximo y calcula los coeficientes de Goodman-Kruskal y de correlación correspondientes entre la disimilaridad piramidal obtenida y la básica y entre la piramidal y la perturbada.

Más concretamente, los parámetros calculados y los nombres de los ficheros en que se almacenan son los siguientes:

* Coeficiente de Correlación entre la disimilaridad Perturbada y la Piramidal

obtenida. Estos resultados se almacenan en los ficheros RMAXnk y RMINnk, donde n es el número de individuos y k es un número entero relacionado con el nivel de error con que perturbamos la disimilaridad básica: 1-Bajo, 2-Medio, 3-Alto.

* Coeficiente de Correlación entre la disimilaridad Básica y la Piramidal obtenida. Ficheros de resultados, RBMAXnk y RBMINnk.

* Gamma entre la disimilaridad Perturbada y la Piramidal obtenida. Ficheros de resultados, GPMAXnk y GPMINnk.

* Gamma entre la disimilaridad Básica y la Piramidal obtenida. Ficheros de resultados, GBMAXnk y GBMINnk.

Para ejecutar el programa, el usuario deberá introducir el número de individuos a clasificar, la disimilaridad básica y el valor de la desviación típica de la variable normal con la que queremos perturbar dicha disimilaridad básica.

A.8.- EL PROGRAMA ESTADIS.C

Es un programa muy sencillo, escrito en C, el cual, a partir de un conjunto de valores reales, calcula su media, desviación típica y cuantiles del 5,10,25,50,75,90 y 95 por ciento. Guarda los resultados en un fichero DOS con el nombre RESULTAT.DAT. También guarda la lista de valores utilizados para efectuar la estadística ordenados de menor a mayor, en el fichero LLIST.DAT.

Su ejecución es sencilla, basta con crear un fichero DOS en el que figuren los

nombres de los archivos donde se guardan los valores sobre los que debe efectuarse la estadística, indicando, a partir de la columna 13 del texto, el número de valores que hay en cada fichero, por ejemplo:

```
GUMAX4 1000
GUMIN4 1000
GUMAX5 1000
.....
```

El programa se activa con la sentencia "Estadis", y nos pedirá el nombre del fichero en el que figuran los nombres de los archivos con los valores numéricos correspondientes. También deberemos indicar si queremos guardar los resultados o si por el contrario tenemos suficiente con visualizarlos por la pantalla.

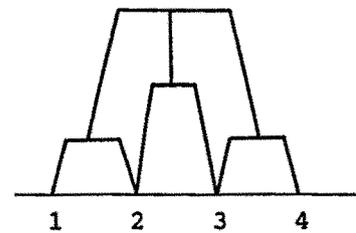
A.9.- DISIMILARIDADES BASICAS PARA LAS PRUEBAS S3 y S4

Para la prueba S3, se han considerado las siguientes pirámides básicas y sus disimilaridades asociadas respectivas:

n=4

$$\begin{bmatrix} 0 & 11 & 14 & 14 \\ & 0 & 12 & 14 \\ & & 0 & 11 \\ & & & 0 \end{bmatrix}$$

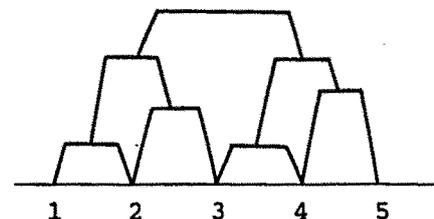
$\sigma=1.37$



n=5

$$\begin{bmatrix} 0 & 11 & 14 & 15 & 15 \\ & 0 & 12 & 15 & 15 \\ & & 0 & 11 & 14 \\ & & & 0 & 13 \\ & & & & 0 \end{bmatrix}$$

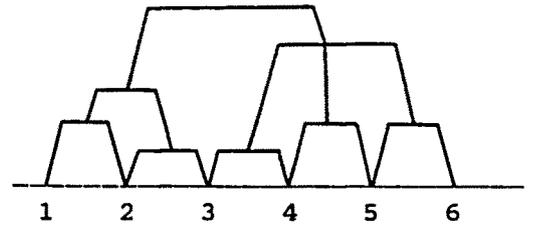
$\sigma=1.56$



n=6

0	12	13	15	15	15
	0	11	15	15	15
		0	11	14	14
			0	12	14
				0	12
					0

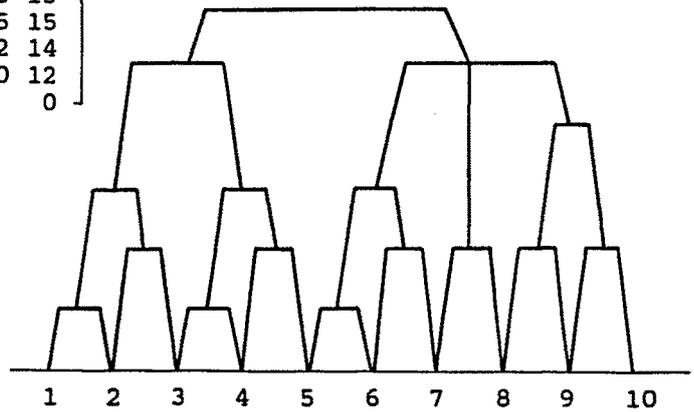
$\sigma=1.5$



n=10

0	11	13	15	15	16	16	16	16	16
	0	12	15	15	16	16	16	16	16
		0	11	13	16	16	16	16	16
			0	12	16	16	16	16	16
				0	11	13	15	15	15
					0	12	15	15	15
						0	12	15	15
							0	12	14
								0	12
									0

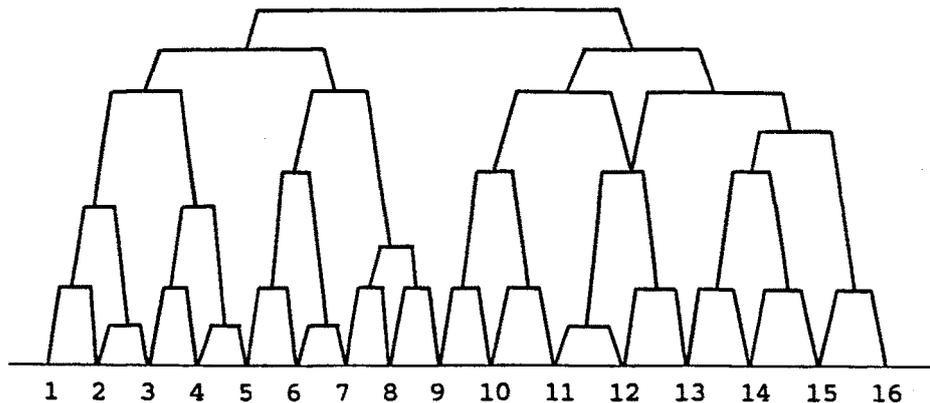
$\sigma=1.7$



n=16

0	12	14	17	17	18	18	18	18	19	19	19	19	19	19	19
	0	11	17	17	18	.	.	18	19	19
		0	12	14	18	.	.	18	19	19
			0	11	18	.	.	18	19	19
				0	12	15	17	17	19	19
					0	11	17	17	19	19
						0	12	13	19	19
							0	12	19	19
								0	12	15	17	17	18	18	18
									0	12	17	17	18	18	18
										0	11	15	17	17	17
											0	12	17	17	17
												0	12	15	16
													0	12	16
														0	12
															0

$\sigma=2.43$



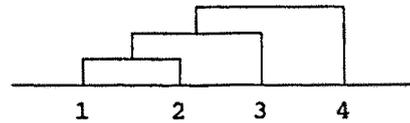
Para la prueba S4, se han considerado las siguientes jerarquías (pirámides)

encadenadas básicas y sus disimilitudes correspondientes:

$n=4$

$$\begin{bmatrix} 0 & 11 & 12 & 13 \\ & 0 & 12 & 13 \\ & & 0 & 13 \\ & & & 0 \end{bmatrix}$$

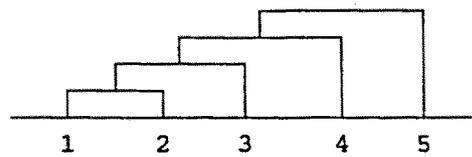
$\sigma=0.74$



$n=5$

$$\begin{bmatrix} 0 & 11 & 12 & 13 & 14 \\ & 0 & 12 & 13 & 14 \\ & & 0 & 13 & 14 \\ & & & 0 & 14 \\ & & & & 0 \end{bmatrix}$$

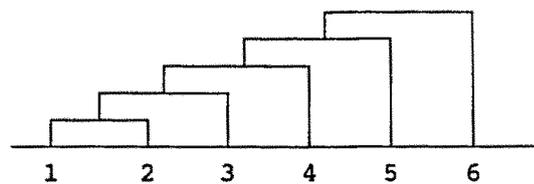
$\sigma=1$



$n=6$

$$\begin{bmatrix} 0 & 11 & 12 & 13 & 14 & 15 \\ & 0 & 12 & 13 & 14 & 15 \\ & & 0 & 13 & 14 & 15 \\ & & & 0 & 14 & 15 \\ & & & & 0 & 15 \\ & & & & & 0 \end{bmatrix}$$

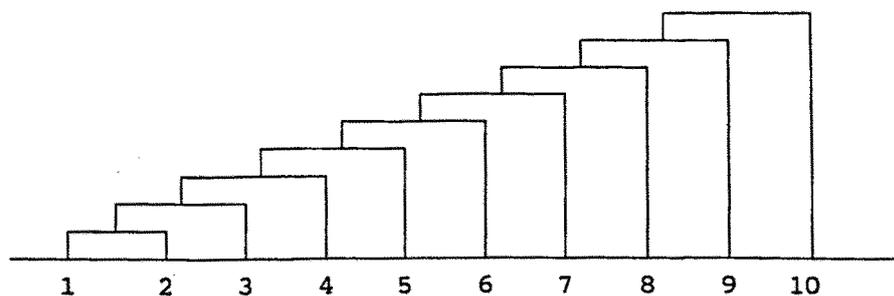
$\sigma=1.25$



$n=10$

$$\begin{bmatrix} 0 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 \\ & 0 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 \\ & & 0 & 13 & 14 & 15 & 16 & 17 & 18 & 19 \\ & & & 0 & 14 & 15 & 16 & 17 & 18 & 19 \\ & & & & 0 & 15 & 16 & 17 & 18 & 19 \\ & & & & & 0 & 16 & 17 & 18 & 19 \\ & & & & & & 0 & 17 & 18 & 19 \\ & & & & & & & 0 & 18 & 19 \\ & & & & & & & & 0 & 19 \\ & & & & & & & & & 0 \end{bmatrix}$$

$\sigma=2.21$



BIBLIOGRAFIA

ADANSON M. (1757)

"Histoire Naturelle du Sénégal"
Banche. Paris.

ARABIE P.; CARROL J.D. (1980)

"MAPCLUS: a mathematical programing approach to fitting the
ADCLUS model"

ARCAS A. (1983)

"Contribuciones a la Construcción de Clasificaciones Estratificadas"
Tesina. Facultad de Matemáticas, Universidad de Barcelona.

ARCAS A. (1984)

"Sobre la no Unicidad de Clasificaciones Jerárquicas Asociadas a un
Método de Clasificación Taxonómico"
Actas del XIV Congreso Nacional de Estadística e Investigación Operativa.

ARCAS A. (1986)

"Contribuciones a la Representación de Datos Multidimensionales
Mediante Arboles Aditivos"
Tesis Doctoral. Fac. Matemáticas Univeridad de Barcelona.

ARCAS A.; CUADRAS C.M. (1987)

"Métodos Geométricos de Representación Mediante Modelos en Arbol".
Publicaciones de Bioestadística y Biomatemática U.B.

ARCAS A.; SALICRU M. (1984)

"Sobre la No Unicidad de la Clasificación Jerárquica Asociada a una
Disimilaridad por los Métodos del Máximo y UPGMA".
Qüestió, Vol.8, nº3, pp.113-120.

BAKER F.B. (1974)

"Stability of two Hierarchical Grouping Techniques. Case I: Sensitivity to Data
Errors"
Journal of the American Statistical Association. Vol.69, nº346.

BAKER F.B.; HUBERT L.J. (1975)

"Measuring the Power of Hierarchical Cluster Analysis"
Journal of the American Statistical Association. Vol.70, nº349.

BENZECRI J.P. (1973)

"L'Analyse des Données: La Taxonomie" Tome 1.
Dunod.

BENZECRI J.P. et col. (1985)

"Introduction a la Classification Ascendente Hiérarchique d'Après un Exemple
des Données Economiques".
Journal de la Societé Statistique de Paris, N°1.

BERTRAND P.; DIDAY E. (1985)

"A visual Representation of the Compatibility Between an order and a
dissimilarity index: The Pyramids."
Computational Statistics Quartely, V. 2 Issue, 1, 1985, 31-40.

BOCK H.H. (1984)

"Statistical Testing and Evaluation Methods in Cluster analysis"
International Conference on Statistics: Applications and New Directions.

BOCK H.H. (1985)

"On Some Significance Tests in Cluster Analysis"
Journal of Classification, 2: 77-108.

BOCK H.H. (1986)

"Multidimensional Skaling in the Framework of Cluster Analysis"
Annual Meeting of the Geseleschaft für Klassifikation.
P.O. Degens, H.Hermes, O.Opitz (Eds.).

BOCK H.H. (1987)

"On the Interface Between Cluster Analysis, Principal Componets Analysis, and
Multidimensional Skaling"
Multivariate Statistical Modeling and Data Analysis, 17-34.
H.Bozdogan, A.K.Gupta (Eds.).

BOCK H.H. (1989)

"Probabilistic Aspects in Cluster Analysis"
Conceptual and Numerical Analysis of Data.
Conference of the Gesellschaft für Klassifikation. Ausburg 1989.

CAILLET F. PAGES J.P. (1976)

"Introduction a l'Analyse des Données"
S.M.A.S.H.

CAPDEVILA C.; ARCAS A. (1992)

"Sobre un Algoritmo de Clasificación Mediante Grafos Piramidales"
Comunicación I Congreso Iberoamericano y XX Reunión Nacional de
Estadística e Investigación Operativa. Cáceres, 1992.

CAPDEVILA C.; ARCAS A. (1993)

"Some Aspects About Statistical Inference In Piramidal Trees"
Comunicación nº41 a la 4ème. Conférence de la Fédération
Internationale des Sociétés de Classification. París, 1993. (Aceptado).

CRITCHLEY F.

"On Exchangeability-Based Equivalence Relations Induced by Strongly Robinson
and, in Particular, by Quadripolar Robinson Dissimilarity Matrices".

CRITCHLEY F.; VAN CUTSEN B. (1989)

"Predissimilarities, Prefilters and Ultrametrics on an Arbitrary Set".
Research Raport of Both the Department of Statistics. Warwick University,
Coventry.

CUADRAS C.M. (1980)

"Curso de Análisis de la Varianza"
Publicaciones de Bioestadística y Biomatemática, N°1. Barcelona.

CUADRAS C.M. (1980)

"Métodes de Representació de Dadeses i la seva Aplicació en Biologia"
Col. Soc. Catalana de Biologia, 13, 95-113.

CUADRAS C.M.

"Técnicas del Análisis Multivariante"
P.P.U.B.

CUADRAS C.M. (1983)

"Análisis Algebraico Sobre Distancias Ultramétricas"
Actas 44 Per. de Sesiones del Instituto Internacional de Estadística,
Madrid, V.2, 554-557.

CUADRAS C.M.; ARCAS A. y otros (1985)

"Métodos Geométricos de la Estadística"
Questió, Vol.9, n°4, pp.219-250.

DIDAY E. et coll. (1979)

"Optimisation en Classification Automatique"
INRIA.

DIDAY E. (1982)

"Croisements, Ordres et Ultrametriques: Application à la Recherche des
Consensus en Classification Automatique".
Rapport de Recerche n°144, INRIA.

DIDAY E. (1983)

"Croisements, Ordres et Ultrametriques"
Mathematiques et Sciences Humaines n°83 p.31-54.

DIDAY E. (1983)

"Problèmes d'Inversion en Classification Hierarchique"
Revue de Statistique Appliquée. Vol.2

DIDAY E. (1986)

"Une Représentation Visuelle des Classes Empietantes: Les Pyramides".
RAIRO. Analyse des données. n°52 p.475-526.

DIDAY E. (1986)

"Ordres and Overlapping Clusters in Pyramids"
Multidimensional Data Analysis.
DSWO Press, Leiden. p.201-234.

DIDAY E. LEMAIRE J. POUGET J. TESTU F. (1982)

"Éléments d'Analyse des Données"
Dunod.

DE LEEUW J. (1982)

"Theory of Multidimensional Scaling".
Handbook of Statistics, V.2. North Holland Publishing Comp. 285-316.

DURAND C. (1986)

"Sur la Représentation Pyramidale en Analyse des Données"
Mémoire de DEA en Mathématiques Appliquées.
Université de Provence. Marseille.

DURAND C. (1988)

"Une Approximation de Robinson Inférieur Maximale"
Rapport de Recherche. Laboratoire de Mathématiques Appliquées et
Informatique. N°88-02.
Université de Provence. Marseille.

DURAND C. FICHET B. (1988)

"One-to-one Correspondences in Pyramidal Representation: A Unified
Approach".
Classification and Related Methods of Data Analysis.
H.H. Bock, Ed. North-Holland. Amsterdam.

DURAND C. (1989)

"Ordres et Graphes Pseudo-Hierarchiques: Theorie et Optimisation
Algorithmique".
Thèse de doctorat en Mathématiques Appliquées.
Universite de Provence. Marseille.

DURAND C. (1989)

"Une Dissimilarité de Robinson Inferieur Maximale".
Actas de l' ISI.

FARRIS J.S. (1969)

"On the Cophenetic Correlation Coefficient"
Syst. Zoology, 18(3), 279-285.

FICHET B. (1981)

"Sur les Approximations d'Indices de Dissimilarités Via les Représentations Euclidiennes et Hierarchiques".
Statistique et Analyse des Données. Vol.6, n°2, pp.1-21.

FICHET B. (1984)

"Sur une Extension de la Notion de Hiérarchie et Son Equivalence Avec Certaines Matrices de Robinson".
Journées de Statistique, Montpellier.

FICHET B. (1988)

" L_p -Spaces in Data Analysis. Classification and Related Methods of Data Analysis".
North-Holland. H.H. Bock Ed. pp.439-444.

FORREST W. YOUNG (1981)

"Encyclopedia of Statistical Sciences".
Campbell B. Read. Ass.Ed.

GONDRAN M. (1976)

"La Structure Algebrique des Classifications Hierarchiques".
Annales de l'INSEE. pp.181-190.

GOODMAN L.A.; KRUSKAL W.H. (1954)

"Mesures of Association for Cross Classifications"
Journal of the American Statistical Association. Vol. 49, Dec. 1954, p.732-764.

GORDON A.D. (1987)

"A Review of Hierarchical Classification"
Royale Statistics Soc. A 150. pp.119-137.

HARTIGAN J.A. (1975)

"Clustering Algorithm"
Wiley.

HARTIGAN J.A. (1977)

"Clustering as modes"
First International Symposium on Data Analysis and Informatics.

HUBERT L. (1974)

"Some Applications on Graph Theory and Related Nonmetrics
Techniques to Problems of Approximate Seriation"
The British Journal of Mathematical and Statistical Psychology.

HUBERT L.J. (1974)

"Approximate evaluation techniques for the single-link and complete-link
hierarchical clustering procedures"
Journal of the American Statistical Association. Vol.69, n°347.

HUBERT L. (1982)

"Inference Procedures for the Evaluation and Comparing of Proximitys
Matrices"
Graduate School of Education. UCLA.

HUBERT L.; ARABIE P. (1985)

"Comparing Partitions"
Fourth European Meeting of the Psychm. Soc. and the Class. Soc.
Cambridge.

HYVER (1973)

"Valeurs Propres des Systèmes de Transformation Représentables par des graphes en arbre"
J. Theoret. Biol., 42. 397-409.

JARDINE N.; SIBSON R. (1971)

"Mathematical Taxonomy"
Wiley, New York.

JUAN J. (1982)

"Le Programme HIVOR de Classification Ascendante Hierarchique Selon les Voisins Réciproques et le Critère de la Variance"
Cahiers d'Analyse des Données. Vol.VII, n°2. pp.173-184.

LEBART L. MORINEAU A. FENELON J.P. (1985)

"Tratamiento Estadístico de Datos".
Marcombo.

LEBART L. MORINEAU A. TABARA N. (1977)

"Techniques de la Description Statistique".
Dunod.

LERMAN I.C. (1970)

"Les Bases de la Classification Automatique".

LERMAN I.C. (1981)

"Classification et Analyse Ordinale des Données".
Dunod.

MARDIA K.V.; KENT J.T.; BIBBY J.M. (1979)

"Multivariate Analysis"
Academic Press. London.

MATHAR R. (1985)

"The Best Euclidian Fit to a Given Distance Matrix in Prescribed Dimension".
Linear Algebra Applications, 67. pp.1-6.

MATULA D.W. (1977)

"In Classification and Clustering"
J.van Ryzing Ed. Academic Press, New York. 95-129.

MONJARDET B. (1950)

"Théories des Graphes et Taxonomie Mathématique in Regards sur la
Théorie des Graphes"
Presses Polytechniques Romandes. Laussane, pp 111-125.

NEI M. et al. (1985)

"Methods for Computing the Standard Errors of Branching Points in an
Evolutionary Tree and Their Application to Molecular Data from Humans and
Apes"
Mol.Biol.Evol. pp.66-85. Univ. of Chicago.

**PANEL ON DISCRIMINANT ANALYSIS, CLASSIFICATION AND
CLUSTERING**

"Discriminant Analysis and Clustering"
Statistical Science, 1989. Vol.4, n°1, pp.34-69.

PERRUCHET CH. (1983)

"Significance Tests for Clusters: Overview and Comments"
Numerical Taxonomy. Springe-Verlag, J. Felsestein Ed.

PREVOSTI A. (1974)

"La Distancia Genética Entre Poblaciones".
Miscellanea Alcobé. 109-118.
Univesdidad de Brcelona.

PRUZANSKI S.; TVERSKY A.; CARROLL J.D. (1982)

"Spatial Versus Tree Representations of Proximity Data".
Psychometrika, Vol.47,1,3-24.

RHAM C. de (1980)

"La Classification Hierarchique Ascendante Selon le Méthode des Voicins Réciproques".
Les Cahiers de l'Analyse des Données. Vol.V. n°2. pp.135-144.

ROBINSON W.S. (1951)

"A Method for Chronological Ordering of Archaeological Deposits".
American Antiquity 16. pp.293-301.

ROBINSON Ph.R. (1987)

"Aplique Turbo Prolog"
Mc. Graw-Hill.

ROLF F.J.; FISHER D.R. (1968)

"Tests for Hierarchical Structure in Random Data Sets"
Systematic Zoology. Vol.17, pp.407-412.

SALICRU M. (1983)

"Consideraciones Sobre Desemejanzas y Clasificaciones Asociadas"
Tesina. Facultad Matemáticas, Universidad de Barcelona.

SAITO T. (1988)

"Cluster Analysis Based on Nonparametric Tests"
Classification and Related Methods of Data Analysis. Ed.H.H.Boock.

SANCHEZ A.; OCAÑA J. (1990)

"Mètodes de Generació de Variables Aleatòries"
Publicacions Dep. Estadística. U.B.

SCHILDT H. (1988)

"Turbo Prolog. Programación Avanzada"
Mc. Graw-Hill.

SHEPART R.; ARABIE P. (1979)

"Additive Clustering: Representation of Similarities as Combinations of
Discrete Overlapping Properties"
Psychological Review. Vol.86, nº2, pp 87-123.

SNEATH P.H. (1966)

"A Comparison of Different Clustering Methods as Applied to Randomly-
spaced Points"
Classif. Soc. Bulletin. Vol.1, pp.2-18.

SNEATH P.H.; SOKAL R. (1973)

"Numerical Taxonomy"
W.H. Freedman and C., San Francisco.

SNEATH P.H. (1980)

"Some Empirical Tests for significance of Clusters. Data Analysis and
Informatics"
E. Diday et al. Eds. North Holland Publishing Company. 491-507.

SOKAL R. (1963)

"Principles of Numerical Taxonomy"
W.H. Freedman and Co. San Francisco.

SOKAL R. (1985)

"Unsolved Problems in Numerical Taxonomy"
Classification and Related Methods of Data Analysis. Ed.H.H.Boock.