UNIVERSIDAD ANAHUAC

ESCUELA DE ACTUARIA

CON ESTUDIOS INCORPORADOS A LA UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

EL ANALISIS DE CUMULOS VISTO COMO UN PROBLEMA DE OPTIMIZACION

T E S I S

QUE PARA OBTENER EL TITULO DE:
A C T U A R I O
P R E S E N T A

SYLVIA TERESITA DEL NIÑO JESUS RAMOS BOLAÑOS

MEXICO, D. F. 1988

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N D I C E

Página

CAPITULO 5

CAPITULO 6

CAPITULO 1

INTRODUCCION

Una de las más comunes y primitivas actividades del hombre ha sido - agrupar, donde, por agrupar se entiende el proceso de organizar elementos de un conjunto dado en subconjuntos homogéneos.

El análisis de cúmulos comprende diversas técnicas de agrupamiento, cuyo objeto es formar grupos o cúmulos, que cumplan con los siguientes - propósitos:

1) Los elementos dentro de cada grupo o cúmulo deben ser de alguna manera "semejantes", es decir; se pretende que los elementos dentro de - un grupo sean homogéneos entre si.

2) Los elementos de los diferentes grupos o cúmulos deben ser "no semejantes", es decir; se desea formar grupos heterogéneos entre si.

Al hablar de semejanza entre individuos o elementos, se presupone que existe una "asociación natural" entre éstos. En muchas ocasiones se distingue con facilidad cuando dos elementos o individuos son semejantes o - no. Por ejemplo, una piedra es diferente a un animal, se puede argumentar que uno es un ser inerte y el otro un ser vivo, y por lo tanto no son semejantes.

Sin embargo, una rana y un gusano al pertenecer al reino animal guar

dan una asociación natural entre ellos. Por otro lado, el gusano pertenece al grupo de los invertebrados y la rana a los vertebrados, visto esto de otra forma se podría decir que no son semejantes.

Del ejemplo anterior, se observa que es necesario definir las características o atributos bajo los cuales se desea agrupar, ya que de esto dependerá que los individuos se consideren semejantes o no.

Hasta este momento sólo se ha mencionado si dos elementos o individuos son semejantes o no; con frecuencia no es suficiente el poder distinguir si dos individuos son o no semejantes, sino es necesario saber - que tanto lo son. Para ello se requiere medir el grado de semejanza entre los elementos.

El grado de semejanza se obtendrá, por medio de alguna medida de asociación entre los elementos, esta a su vez dependerá de las características que definan a dichos elementos.

Aunque el grado de semejanza tome valores en una escala continua, - será necesario definir un criterio respecto al cual el grado de semejanza será "alto" o "bajo", con el objeto de determinar los elementos dentro de un mismo grupo y los que están en diferentes grupos. Tales criterios dependerán del objetivo que se tenga al agrupar. La solución a este problema puede darse, buscando que el grado de semejanza entre los elementos de cada grupo sea mayor o igual a un cierto valor (cota), en este caso el número de grupos no se fijará de antemano, sino se formarán los grupos de -

tal manera que los elementos de cada grupo cumplan con dicha restricción. Otra solución, para el caso en que se fije de antemano el número de grupos, se obtiene al maximizar el grado de semejanza entre los elementos de cada grupo de tal forma que se llegue al número de grupos deseado.

Casi todos los procedimientos usados para descubrir grupos o cúmulos definen como medida de asociación la distancia entre los puntos (métrica) definidos por las características o atributos de los elementos; y por medio de un método iterativo se encuentran los grupos a partir de vecindades definidas en términos de medidas de asociación.

A grandes rasgos, los métodos de agrupación se dividen en "métodos jerárquicos" y "métodos no jerárquicos". Los primeros tienen la propiedad que en cada paso un grupo se obtiene como la unión de grupos obtenidos en pasos previos.

Por otro lado, conviene mencionar que en muchas ocasiones el problema de cúmulos se puede plantear como uno de optimización matemática. - Usando ténicas de programación, ya sea lineal, no lineal, dinámica, entera, etc., dependerá de la naturaleza del problema, el cual define las - restricciones y la función objetivo en el planteamiento de optimización.

En esta tesis se planteará el problema de cúmulos de varias formas, con el objeto de lograr una mayor comprensión de su naturaleza. En especial se hará énfasis en aquellos casos en que el problema puede verse co

mo uno de optimización. Además, se pretende la obtención de un programa para el análisis de cúmulos por métodos jerárquicos, que podrá ser utilizado en microcomputadoras. Lo que parece de gran utilidad ya que este tipo de equipos ha proliferado a últimas fechas y no existe suficiente "sofware" para ellos, sobre todo en lo referente a métodos estadísticos multivariados.

Para lograr lo anterior, la tesis se ha dividido en 8 capítulos. Este es el primer capítulo, que como ya se vió, da una idea general del problema de cúmulos y refiere el resto de la tesis. En el Capítulo 2 se planteará el problema de cúmulos como un problema de partición, se hará un breve análisis de las variables y sus escalas. En los capítulos subsecuentes se tratará el problema para variables continuas y cuantitativas. En el tercer capítulo se definirán las medidas de asociación de mayor relevancia como son las funciones de distancia y de proporción. En el cuarto capítulo se exponen los criterios de optimalidad que servirán para la obtención de la función objetivo del problema de cúmulos, se mencionarán principalmente tres criterios de optimalidad a saber; los que buscan obtener homogenidad dentro de los grupos, los que su objetivo es lograr la heterogenidad entre los grupos y un bicriterio que pretende obtener los dos objetivos anteriores.

En el Capítulo 5 se plantea la solución del problema de cúmulos como uno de optimización y se analiza el uso de métodos de programación matemática para este fin. En particular se describen: la Programación Dinámica , la Entera y el Planteamiento en base a Teoría de Gráficas. En

el mismo capítulo se analizarán métodos propios del análisis de cúmulos, como son los métodos jerárquicos.

El Capítulo 6 describe un programa computacional para resolver el problema de cúmulos por medio de tres de los métodos jerárquicos más importantes como son el de Unión-Simple, Unión-Exhaustiva y el de la Mediana. Una aplicación de estos tres métodos usando dicho programa se hará en el Capítulo 7, en el que se agrupan las entidades federativas de la República Mexicana en base a su rama de actividad económica. Se usarán los tres métodos jerárquicos ya mencionados y se compararán resultados. El último capítulo dará las conclusiones finales de toda la tesis.

CAPITULO 2

PLANTEAMIENTO DEL PROBLEMA DE CUMULOS COMO UNO DE PARTICION

Un problema de cúmulos puede ser planteado como un problema de partición, ya que los grupos o cúmulos que se forman cumplen con la definición de partición.

En este capítulo se dará la definición de partición en términos generales, seguida de la notación que se usará en el resto de la tesis. Se dará la aplicación del problema de partición al de cúmulos y su planteamiento. Finalmente se verán algunas propiedades de la variables que describen a los elementos.

2.1 Definición

Considere un conjunto $I = \{I_1, I_2, \ldots, I_n\}$ y una familia de subconjuntos de I $\pi = \{\pi_1, \pi_2, \ldots, \pi_m\}$ tal que $\pi_k \subseteq I$, $k \in K = \{1,2,\ldots,m\}$; - π se define como una cubierta de I si $\bigcup_{k \in K} \pi_k = I$.

Y si además se tiene

$\pi_k \cap \pi_j = \phi \quad \forall\, j,k \in K \,,\, j \neq k$

π define una partición de I.

Si P es el conjunto de particiones posibles de I, se define la función de costo total $G: P \longrightarrow R$ asociada a cada partición de $P \subseteq R^m$.

El problema de partición consiste en encontrar la partición óptima π^* que minimice el costo total. Es decir, encontrar:

$$G(\pi^*) = \min_{\pi \in P} G(\pi)$$

2.2 Notación

Sea $I = \{I_1, I_2, \ldots, I_n\}$ un conjunto que denota n elementos o individuos de una población.

Suponga que se tiene un conjunto de características o atributos - $C = \{C_1, C_2, \ldots, C_p\}$ observables y poseídas por cada individuo u elemento.

Dicho de otro modo, a cada elemento del conjunto I, I_j le corresponde un vector de atributos X_j, cuyas componentes son las medidas de cada característica asociadas a dicho elemento.

Como cada I_j $(j=1,2,\ldots,n)$ esta descrito por p características, - entonces $X_j=(X_{1j}, X_{2j}, \ldots, X_{pj})$ será un vector de p variables.

Para cada $i=1,2,\ldots,p$ y $j=1,2,\ldots,n$, X_{ij} es la medida de la -- i-ésima característica del j-ésimo elemento.

En consecuencia, el conjunto I de elementos puede ser descrito por una matriz X de p x n que tiene los valores de las p variables para todos los elementos de I.

Por lo tanto X:

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{p1} & X_{p2} & & X_{pn} \end{bmatrix}$$

Conviene hacer notar que en caso de que los elementos que se pretenden agrupar estén definidos por una sola variable unidimensional (en R) - este problema es "sencillo" de resolver, ya que en los Reales (R) existe un orden. Es decir, dado X_i, $X_j \in R$ siempre se puede decir si $X_i > X_j$ o $X_i < X_j$ o $X_i = X_j$; sin embargo, si las variables $X_i, X_j \in R^p$ $p \geq 2$, no se tiene un orden, y el problema de agrupar se vuelve muy "difícil", ya que no se puede decir si $X_i > X_j$ o $X_i < X_j$.

Debido a lo anterior, es necesario utilizar criterios para determinar un orden de un conjunto de vectores $X=(X_1, X_2, \ldots, X_n)$ definidos en - R^p $(p \geq 2)$, estos criterios se determinarán en base a medidas de asociación, las que se tratarán en el Capítulo 3. Dichos criterios no son únicos (dependerán de la medida de asociación que se elija) incluso puede - ocurrir en algunos casos $X_i < X_j$ y en otros $X_i > X_j$. En el siguiente capítulo se propondrán varias medidas de asociación para estas comparaciones.

2.3 El Análisis de Cúmulos como un Problema de Partición

Dada la matriz de observación de los atributos, el problema de cúmu

los pretende determinar m grupos o cúmulos de un conjunto dado de elementos que satisfaga cierto criterio de optimalidad.

Si $\pi = \{\pi_1, \pi_2, \ldots, \pi_m\}$ denota el conjunto de m grupos o cúmulos del conjunto de elementos I, π será entonces una partición de I que cumple con:

$$\bigcup_{k=1}^{m} \pi_k = I$$

y que para $I_j \in I$ sólo pertenecerá a uno y sólo un elemento de π, es decir:

para $I_j \in I$

Si $I_j \in \pi_k \Rightarrow I_j \notin \pi_\ell (\forall \ell \neq k = 1,2,\ldots,m)$

En consecuencia, el problema de cúmulos se reduce a un problema de partición del que se tendrá como objetivo determinar la partición óptima π^* del conjunto de particiones P que satisfaga el criterio de optimalidad, que estará dado en términos de una relación funcional a la que se le dará el nombre de función objetivo.

Los criterios de optimalidad para el problema de cúmulos tenderán a formar grupos heterogéneos entre sí, y a su vez los elementos que forman cada grupo sean lo más homogéneos posibles. Lo que significa que se desea, formar grupos donde la asociación natural de los elementos, que se manifiesta a través de los atributos observados, sea mayor dentro de los grupos, y menor cuando los elementos pertenezcan a diferentes grupos.

2.4 Comportamiento de las Variables

Se considera que se tienen observaciones de cada una de las p variables que describen a los elementos, y que éstas corresponden a un fenómeno empírico, que se caracteriza por la propiedad de que al observarlo bajo determinado conjunto de condiciones, no siempre se obtiene el mismo resultado (de manera que no existe regularidad determinística), sino que los diferentes resultados ocurren con regularidad estadística. Esto quiere decir, que existen números entre 0 y 1, que representan la frecuencia relativa con la que se observan los diferentes resultados en una serie de repeticiones independientes del fenómeno.

Para el análisis de cúmulos puede considerarse dos situaciones que se definen a continuación:

1) Aunque el fenómeno es aleatorio en ocasiones sólo se tiene una observación del fenómeno, se agrupa en base a ésta como si fuese un problema determinístico.

2) En otras ocasiones se cuenta con mayor información acerca del fenómeno (por ejemplo: varias observaciones, su distribución a priori, etc.), por lo que se puede manejar como un problema probabilístico.

Las situaciones antes definidas se plantean independientemente y de acuerdo a la naturaleza del problema. En general, esta tesis estará dedicada al estudio del problema determinístico.

2.5 Las Variables y sus Escalas

En general, las variables observadas se valuan en unidades diferentes y las escalas que se usan para medirlas pueden ser de diferentes tipos.

La diferencia de unidades entre las variables y en especial las diferentes escalas, pueden hacer imposible el uso del análisis de cúmulos.

Una clasificación sistemática de variables, origina una estructura conveniente para identificar diferencias esenciales en los elementos o individuos. Las variables pueden clasificarse de acuerdo a su "recorrido" y a su escala de medida.

Si se entiende por recorrido la cardinalidad del conjunto de valores que la variable puede asumir, el recorrido de una variable se puede clasificar como:

- Finito
- Contable infinito
- No contable infinito

La clasificación de las variables en base a su recorrido debe tomar en cuenta dos factores a saber:

- Su "tipo de recorrido"
- El "tamaño de su recorrido".

En cuanto al tipo de recorrido se hablará más adelante, para ilustrar la importancia del tamaño del recorrido se recurrirá a un ejemplo. Suponga que se desean agrupar empresas usando dos variables que son: ventas (en pesos) y número de empleados. El tipo de recorrido de estas variables es el mismo (contable finito); sin embargo, el tamaño de su recorrido puede ser muy diferente, ya que el recorrido de las ventas se puede suponer de miles de millones de pesos, mientras que el número de empleados a lo más se mide en cientos de pesos.

Basados en estos conceptos, las variables a su vez se clasifican en:

- **Continuas:** pueden tener recorrido finito o no finito, pero no contable. Tales variables pueden asumir cualquier valor dentro de un intervalo o colección de éstos.

- **Discretas:** Tienen un recorrido finito o al menos un recorrido infinito contable.

- **Binarias o Dicotómicas:** son variables discretas que sólo pueden tomar dos valores.

En esta tesis al referirse a variables, sólo se supondrá que éstas son del tipo continuo.

La clasificación de las variables, de acuerdo a su escala de medida se hará en relación al siguiente esquema.

Para la k-ésima variable y los i-ésimo y j-ésimo elementos con los

valores X_{ki}, X_{kj} respectivamente, se dirá que la escala es:

- **Nominal:** si solamente distingue entre clases. Es decir, sólo se puede decir si $X_{ki} = X_{kj}$ o bien si $X_{ki} \neq X_{kj}$.

- **Ordinal:** si puede ordenar elementos. Además de distinguir entre $X_{ki} = X_{kj}$ o $X_{ki} \neq X_{kj}$, se puede decir si $X_{ki} > X_{kj}$ o bien, si $X_{ki} < X_{kj}$.

- **De Intervalo:** si asigna una medida significativa de la diferencia entre dos elementos, es decir se puede afirmar que el i-ésimo elemento es $(X_{ki} - X_{kj})$ unidades diferentes del j-ésimo elemento.

- **De Razón:** si es una escala de intervalo con un cero absoluto si $X_{ki} > X_{kj}$ se puede decir qu el i-ésimo elemento es X_{ki} / X_{kj} veces el j-ésimo elemento.

Se hace notar que el tipo de medida de asociación elegida dependerá del tipo de escala de las variables en cuestión.

Con frecuencia, las variables con escala nominal u ordinal son para variables categóricas o cualitativas; las variables con escalas de intervalo o de razón son variables cuantitativas.

La tesis se limitará a analizar sólo variables cuantitativas y como ya se mencionó, variables continuas.

CAPITULO 3

MEDIDAS DE ASOCIACION

La necesidad de cuantificar el grado de semejanza entre elementos, propició el uso de medidas de asociación entre elementos.

Las medidas de asociación que se utilizan con mayor frecuencia son las de distancia, a las que se hace referencia en este capítulo. Se analizarán también, medidas de asociación que determinan el grado de relación lineal entre los elementos. Finalmente se darán algunas relaciones estadísticas para medir el grado de semejanza entre los elementos.

El uso de una u otra medida de asociación dependerá de la naturaleza del problema en cuestión, es decir, de cómo y qué se desea agrupar y del tipo de limitantes que se tengan.

3.1 Funciones de Distancia

Sea X el conjunto $X=\{X_1, X_2, \ldots, X_n\}$, cuyos elementos se llaman puntos; se dice que X es un espacio métrico, si a cada par de puntos X_i, X_j hay asociado un número real $d(X_i, X_j)$ llamado distancia de X_i a X_j, tal que:

i) $d(X_i,X_j) > 0$ si $i \neq j$; $d(X_i,X_j) = 0$ si $i=j$;

ii) $d(X_i,X_j) = d(X_j,X_i)$;

iii) $d(X_i,X_j) \leq d(X_i,X_k) + d(X_k,X_j)$ $\forall$ $X_k \in X$

La función d se le llama función de distancia o métrica. La métrica de Minkowski, conocida por la norma ℓ_c, se define para X_i, $X_j \in X$ como:

$$d_c(X_i,X_j) = [\sum_{k=1}^{p} | X_{ki} - X_{kj} |^c]^{1/c} \qquad c \geq 1$$

Para diversos valores de c se obtienen diferentes funciones de distancia.

Con c=1, se tiene la norma ℓ_1, definida como:

$$d_1(X_i,X_j) = \sum_{k=1}^{p} | X_{ki} - X_{kj} |$$

La distancia euclidiana o norma ℓ_2 se obtiene cuando c=2

$$d_2(X_i,X_j) = [\sum_{k=1}^{p} (X_{ki} - X_{kj})^2]^{1/2}$$

La métrica de Chebychev se obtiene de:

$$\lim_{c \longrightarrow \infty} d_c(X_i,X_j)$$

que en general se conoce como la norma suprema ℓ_∞, lo que se define:

$$d_\infty (X_i,X_j) = \text{Sup}_{k=1,2,\ldots,p} \{|X_{ki} - X_{kj}|\}$$

Para visualizar el significado geométrico de las funciones de distancia, ya definidas se dará un ejemplo:

Suponga un conjunto de puntos $X_1, X_2, \ldots, X_n$, para los que

$$d_c(X_i, X_j) = 1 \quad \forall \; i \neq j,$$

es decir, un conjunto $U = \{X_i \mid d_c(X_i, X_j) = 1 \quad \forall \; i \neq j\}$ este conjunto se denomina, la esfera unitaria de la métrica.

La representación gráfica de las esferas unitarias de las métricas ℓ_1, ℓ_2, ℓ_∞, en dos dimensiones se tiene en la figura (1).

Las esferas unitarias que las métricas ℓ_c, para $1<c<2$ son curvas conexas, que se encuentran en las esferas unitarias de las métricas ℓ_1 y ℓ_2.

Para el caso en que $2<c<\infty$ la esfera unitaria para una c dada es una curva conexa que se encuentra entre las esferas que las métricas ℓ_2 y ℓ_∞. Tales esferas representan, la superficie formada por la colección de puntos que están a una unidad de distancia del centro.

El siguiente teorema proporciona una ordenación de las funciones de distancia definidas por la norma ℓ_c.

Teorema

La desigualdad

$$d_h(X_i, X_j) \leq d_m(X_i, X_j)$$

REPRESENTACION DE LAS ESFERAS UNITARIAS L_1, L_2 Y L_∞

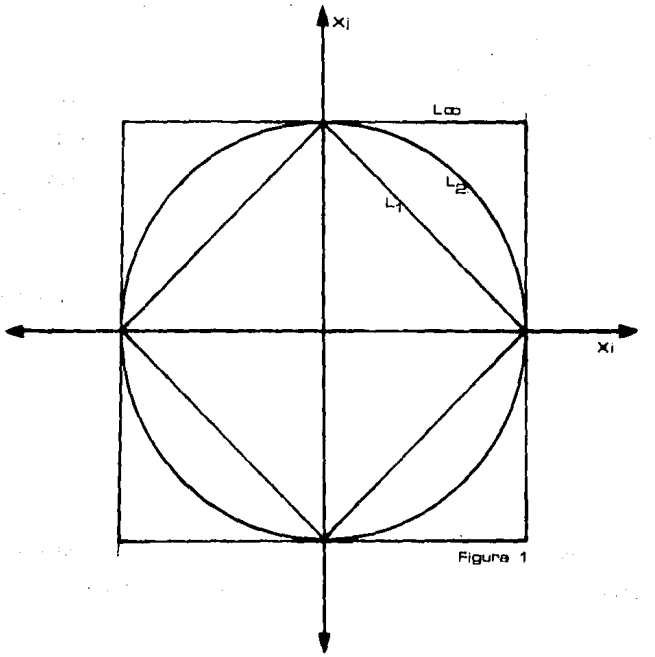


Figura 1

se cumple

$$\nexists \; X_i, X_j \in R^p \iff h \geq m$$

3.2 Comportamiento de las Funciones de Distancia

Cualquier elemento descrito por un conjunto de características, puede ser localizado como un vector o punto en un espacio multivariado. - Usando los atributos del elemento como valores de las coordenadas rectangulares de dicho punto; es decir, X_{ki} es la componente del vector X_i a lo largo del k-ésimo eje coordenado.

Con objeto de facilitar la visualización gráfica y el manejo de ecuaciones, se hará referencia al caso particular de dos dimensiones, es decir, se consideran sólo dos atributos para los vectores X_i, X_j, que se - definen como:

$$X_i = (X_{1i}, X_{2i}) \quad y \quad X_j = (X_{1j}, X_{2j})$$

Se desea analizar el comportamiento de las funciones de distancia - definidas para los vectores X_i, X_j, al rotar dichos vectores.

Como se expuso con anterioridad, la magnitud de la diferencia de los vectores X_i, X_j es la distancia euclidiana entre éstos, representada por:

$$d_2(X_i, X_j) = | X_i - X_j | \quad \text{(donde } || \text{ significa magnitud).}$$

Al rotarse X_i y X_j en los ejes coordenados un ángulo ϕ, para todos los valores de $\phi (0° \leq \phi \leq 360°)$, graficamente X_i y X_j formarán círculos

con radios:

$$|X_i| = \sqrt{X_{1i}^2 + X_{2i}^2} \qquad y \qquad |X_j| = \sqrt{X_{1j}^2 + X_{2j}^2}$$

respectivamente. Lo que se observa en la figura (2). Donde $|X_i|$ es la magnitud de X_i, es decir, la distancia euclidiana de X_i al origen.

Sea R la matriz de rotación definida como:

$$R = \begin{bmatrix} \cos\phi & \operatorname{sen}\phi \\ -\operatorname{sen}\phi & \cos\phi \end{bmatrix}$$

Donde ϕ es el ángulo de rotación.

El determinante de la matriz R de rotación es:

$$|R| = \operatorname{sen}^2\phi + \cos^2\phi = 1$$

Sean X_i' y X_j' los vectores rotados definidos por:

$$X_i' = RX_i \qquad y \qquad X_j' = RX_j$$

Sea C el vector diferencia entre X_i y X_j :

$$C = X_i - X_j = [X_{1i} - X_{1j} \ , \ X_{2i} - X_{2j}],$$

cuya magnitud al cuadrado se define con la distancia euclidiana entre X_i X_j al cuadrado, es decir,

$$|C|^2 = d^2(X_i, X_j) = |X_i|^2 + |X_j|^2 - 2\,|X_i|\,|X_j| \cos\theta$$

donde θ es el ángulo entre X_i y X_j.

ROTACION DE VECTORES

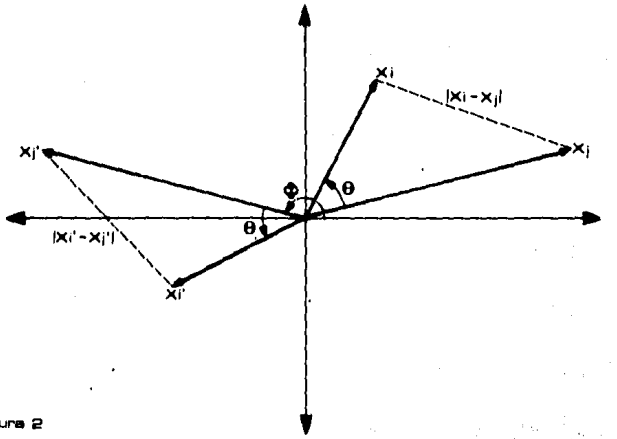


Figura 2

Sea C' el vector diferencia entre X_i' y X_j' $C' = X_i' - X_j'$, cuya magnitud al cuadrado se define:

$$|C'|^2 = |X_i' - X_j'|^2 = |X_i'|^2 + |X_j'|^2 - 2|X_i'||X_j'| \cos\theta$$

donde θ es el ángulo entre X_i' y X_j'.

Como $|X_i'|^2 = |RX_i|^2 = |R|^2 |X_i|^2 = |X_i|^2$

lo que implica que $|C|^2 = |C'|^2$, lo que equivale que:

$$d^2(X_i,X_j) = d^2(X_i',X_j')$$

Por otro lado, $C' = RC$

$$C' = (c_1',c_2') = R[X_{1i} - X_{1j}\ ,\ X_{2i} - X_{2j}]^T$$

$$(c_1',c_2') = \begin{bmatrix} \cos\phi & \text{SEN}\,\phi \\ -\text{SEN}\,\phi & \cos\phi \end{bmatrix} \qquad \begin{bmatrix} X_{1i} - X_{1j} \\ X_{2i} - X_{2j} \end{bmatrix}$$

$$c_1' = (X_{1i} - X_{1j}) \cos\phi + (X_{2i} - X_{2j}) \text{SEN}\,\phi$$

$$c_2' = -(X_{1i} - X_{1j}) \text{SEN}\,\phi + (X_{2i} - X_{2j}) \cos\phi$$

y como $C' = (X_i' - X_j')$, por consiguiente:

$$X_{1i}' - X_{1j}' = (X_{1i} - X_{1j}) \cos\phi + (X_{2i} - X_{2j}) \text{SEN}\,\phi$$

$$X_{2i}' - X_{2j}' = (X_{2i} - X_{2j}) \cos\phi - (X_{1i} - X_{1j}) \text{SEN}\,\phi$$

con lo que se concluye que la norma ℓ_2 comúnmente llamada **distancia euclidiana** entre dos vectores, es invariante de la rotación de ambos, siempre

y cuando el ángulo de separación entre dichos vectores se conserve.

La norma ℓ_1 que se definió con anterioridad como:

$$d_1(X_i,X_j) = \sum_{k=1}^{p} |X_{ki} - X_{kj}|$$

para el caso de dos dimensiones (p=2), se mantiene invariante si los vectores X_i y X_j se rotan un ángulo $\phi = n(90°)$ donde $n \in Z^+$

La norma suprema definida como:

$$\text{Sup}\{|X_{ki} - X_{kj}|\} \quad \text{para} \quad p=2,$$
$$k = 1,2,\ldots,p$$

será la misma para los vectores rotados con $\phi = n(180°)$ $n \in Z^+$

En el anexo A, se generalizará el comportamiento de las funciones de distancia para p dimensiones.

Las funciones de distancia como medidas de asociación entre elementos miden el "grado de disimilitud" de estos, ya que mientras mayor sea la distancia entre sus vectores correspondientes, menos semejantes serán. Se dicen que dos vectores X_i y X_j son iguales si $d(X_i,X_j)=0$, $\forall$ $i \neq j$.

A continuación se da una definición que aclara lo anterior:

Definición:

Para el caso del problema de cúmulos se tiene que un elemento $I_i \in I$

pertenece al grupo π_k de una partición de I <=>

$$d^2(X_i,\bar{X}^k) < d^2(X_i,\bar{X}^\ell) \quad \forall\ \ell \neq k, \quad \ell=1,2,\ldots,m$$

donde $\bar{X}^k = \sum_{i=1}^{nk} X_i/nk$ y $\bar{X}^\ell = \sum_{i=1}^{n\ell} X_i/n\ell$ se conocen como centroides de los grupos π_k y π_ℓ respectivamente; y además, nk y $n\ell$ son el número de elementos de dichos grupos.

En consecuencia, el grado de semejanza entre dos elementos que se mide a través de las funciones de distancia, se hará de manera inversa, es decir, si se desea obtener máximo grado de semejanza entre dos elementos, entonces se minimizará la distancia entre sus vectores.

Lo que se explicará con mayor detalle en las siguientes secciones.

3.3 Funciones de Proporción

Como se mencionó en la sección anterior, el grado de semejanza entre dos elementos puede determinarse en base a la distancia entre los vectores que los definen. Es posible construir otras medidas de asociación basadas en las proporciones que guardan entre sí las componentes de dichos vectores. Una medida de este tipo es el producto punto entre ellos.

Definición:

Sean X_i y X_j vectores de p componentes, el producto punto entre X_i y X_j es:

$$X_i \cdot X_j = \sum_{k=1}^{p} X_{ki}\ X_{kj}$$

el que en términos de magnitud y del ángulo entre los vectores X_i y X_j, se define también como:

$$X_i \cdot X_j = |X_i| \; |X_j| \cos \phi$$

donde ϕ es el ángulo de separación entre X_i y X_j.

El producto punto equivale a lo que en estadística se conoce como - covarianza entre X_i y X_j, la que se define a continuación:

Definición:

La covarianza entre las variables X_i y X_j es

$$COV(X_i, X_j) = E[(X_i - E(X_i)) \; (X_j - E(X_j))]$$

donde $E(X)$ es el valor esperado de X.

Un estimador insesgado de la covarianza entre X_i y X_j es:

$$S(X_i, X_j) = 1/(p-1) \sum_{k=1}^{p} X_{ki} \; X_{kj}$$

cuando $\bar{X}_i = \bar{X}_j = 0$.

Dos elementos representados por los vectores X_i y X_j, son iguales - con referencia a las funciones de distancia si

$$d^2(X_i, X_j) = 0 \qquad i = j$$

sin embargo, para el caso del producto punto, no se puede determinar cuando dos elementos X_i y X_j, son iguales, ya que el producto punto depende

del producto de las magnitudes de los vectores asociados a dichos elemen tos. De lo anterior, se concluye que el producto punto entre dos vecto- res va a depender de la escala de estos.

Puede presentarse el caso que las componentes de los vectores X_i y X_j tengan una misma proporción, es decir,

$$X_{ki} = C\, X_{kj} \quad \forall\ k=1,2,\ldots,p \quad y \quad C \in R;$$

en este caso se puede decir que X_i y X_j son iguales, si su medida de - asociación es el producto punto.

El hecho de que $X_i = C\, X_j$ se le conoce como una relación lineal - entre X_i y X_j.

Basados en la definición del producto punto entre dos vectores X_i y X_j.

$$X_i \cdot X_j = |X_i|\ |X_j| \cos \phi$$

si ϕ que es el ángulo de separación entre dichos vectores, es igual a ce ro (ϕ=0°) entonces los vectores X_i y X_j son semejantes y su producto - punto es:

$$X_i \cdot X_j = |X_i|\ |X_j|$$

el que sólo depende de las magnitudes de los vectores.

Se observa, que si las magnitudes de estos vectores son "pequeñas" su producto punto es "pequeño", mientras que si sus magnitudes son "gran

des" su producto punto es "grande". Lo anterior se debe a que, como ya se mencionó, el producto punto no es independiente de las magnitudes de los vectores. Por consiguiente, se desea una medida de semejanza tal - que distinga los casos en que las componentes de los vectores sean proporcionales de aquellos en que no lo son. Para obtener esto será necesario usar una medida semejante al producto punto, pero que sea independiente de las magnitudes de los vectores. Esto es sencillo de lograr - si se normaliza el producto punto.

Normalizando el producto punto, se puede comparar diferentes pares de vectores sin importar la escala de estos. A esta medida de semejanza se le conoce como el coeficiente de correlación lineal entre dos vectores dados que se define a continuación:

Definición:

El coeficiente de correlación lineal entre dos vectores X_i y X_j es $r_{ij} = X_i \cdot X_j / |X_i| \, |X_j|$.

El coeficiente de correlación lineal entre dos vectores está normalizado y toma valores entre -1 y 1. Lo que viene de dividir X_i entre - su magnitud; el vector que se obtiene es de magnitud uno, es decir,

$$U_i = X_i / |X_i| \quad \text{donde} \quad |U_i|^2 = 1$$

a U_i se le llama vector unitario.

El producto punto entre los vectores unitarios U_i y U_j de X_i y X_j - es:

$$U_i \cdot U_j = X_i \cdot X_j \; / \; |X_i| \; |X_j|$$

y como

$$U_i \cdot U_j = |U_i| \; |U_j| \; \cos \phi \quad \text{donde } \phi \text{ es el ángulo entre } X_i \text{ y } X_j$$

se tiene que

$$\cos \phi = U_i \cdot U_j = X_i \cdot X_j \; / \; |X_i| \; |X_j|$$

El coeficiente de correlación lineal entre X_i y X_j equivale entonces, al coseno del ángulo de separación entre dichos vectores.

Para los vectores

$$X_i = (X_{1i}, X_{2i}, \ldots, X_{pi}) \quad y \quad X_j = (X_{1j}, X_{2j}, \ldots, X_{pj})$$

de p componentes, el coeficiente de correlación lineal entre ellos --- $r(X_i, X_j)$ se define:

$$r(X_i, X_j) = r_{ij} = \sum_{k=1}^{p} X_{ki} \, X_{kj} \; / \; \sqrt{\sum_{k=1}^{p} X_{ki}^2} \; \sqrt{\sum_{k=1}^{p} X_{kj}^2}$$

donde

$$\sum_{k=1}^{p} X_{ki} = \sum_{k=1}^{p} X_{kj} = 0. \quad \text{En consecuencia} \quad -1 \leq r_{ij} \leq 1.$$

El coeficiente de correlación lineal no depende de las magnitudes de los vectores, sino de la proporción que guardan las componentes de dichos vectores.

El siguiente lema muestra que si dos vectores X_i y X_j son propor-

cionales, su coeficiente de correlación lineal $|r_{ij}| = 1$, y por lo tanto X_i y X_j son proporcionales.

Lema:

El coeficiente de correlación lineal $r_{ij} = \pm 1 \iff X_i = K X_j$, - donde $K \in R$.

Prueba:

$$r_{ij} = X_i \cdot X_j / |X_i| |X_j| = X_i K X_i / |X_i| |K| |X_i|$$

$$= K X_i^2 / |K| X_i^2 = K / |K| = \pm 1$$

Por lo tanto, el grado de semejanza entre X_i y X_j es "alto" en - forma positiva cuando r_{ij} tiende a uno; es "alto" en forma negativa cuando r_{ij} tiende a menos uno; y se dice que el grado de semejanza es "bajo" cuando r_{ij} tiende a cero.

Cuando $|r_{ij}| = 1$, X_i y X_j son iguales en base a la proporción entre sus componentes; cuando $r_{ij} = 0$ se dice que X_i y X_j son diferentes, es decir, cuando

$$r_{ij} = 0 \quad \text{el} \quad \cos \phi = 0$$

y para que esto ocurra, se tiene que el ángulo entre X_i y X_j, en este - caso $\phi = \pm 90°$, lo que significa que X_i y X_j son ortogonales.

Para $\bar{X}^k = \sum_{i=1}^{nk} X_i/nk$ y $\bar{X}^\ell = \sum_{i=1}^{n\ell} X_i/n\ell$

en términos de cúmulos, un elemento del conjunto I, I_i pertenece al grupo π_k de una partición π de I, <==>

$$|r(X_i,\bar{X}^k)| > |r(X_i,\bar{X}^\ell)| \quad \text{donde} \quad \ell \neq k \qquad \forall\ \ell=1,2,\ldots,m$$

El hecho de que un elemento I_i pertenezca a un cierto grupo π_k tomando como medida de asociación el coeficiente de correlación lineal, no implica que tomando como medida de asociación alguna función de distancia, dicho elemento I_i perteneciera también al grupo π_k, sino podría pertenecer a otro grupo diferente de π_k, ya que el significado de semejanza cambia en ambos casos.

Esto se explica con mayor detalle en la siguiente sección.

3.4 Comparación entre Funciones de Distancia y Proporción

Hasta ahora se han analizado las funciones de distancia y de proporción como medidas del grado de semejanza entre elementos; sin embargo, cada una de ellas define el concepto de semejanza de manera diferente.

A continuación se ilustrará como se mide el grado de semejanza utilizando las funciones de distancia en comparación con las medidas de semejanza obtenidas de las funciones de proporción.

En la figura (3a) se presentan dos vectores X_i y X_j, cuyo ángulo de separación θ es "pequeño", es decir, que X_i es semejante a X_j usando como medida de semejanza la proporcionalidad. Sin embargo, la magnitud del

vector diferencia de dichos vectores $|X_i - X_j|$ es "grande", por lo tanto el grado de semejanza entre estos vectores, utilizando funciones de distancia es "pequeño".

En la figura (3b) se presenta el caso contrario al anterior, en que θ ángulo entre X_i y X_j es "grande", mientras que la magnitud del vector diferencia entre dichos vectores $|X_i - X_j|$ es "pequeña". En consecuencia, el grado de semejanza medido por funciones de distancia es "grande", mientras que el grado de semejanza que se obtiene al utilizar funciones de proporción es "pequeño".

Un comentario acerca de estas medidas de asociación, es que el coeficiente de correlación lineal está normalizado, ya que toma valores entre -1 y 1, mientras que las funciones de distancia no lo están ya que su valor máximo es arbitrariamente grande, lo anterior complica la normalización de estas funciones.

Otra observación de las funciones de distancia con respecto a las de proporcionalidad, es que toman sus valores de manera inversa, es decir, las funciones de distancia serán más "pequeñas" mientras más semejantes sean los vectores; sin embargo, el valor absoluto de las funciones de proporcionalidad será mayor mientras más semejantes sean dichos vectores.

Ejemplificando el caso descrito por las figuras anteriores se tiene:

SEMEJANZA POR FUNCION DE PROPORCION

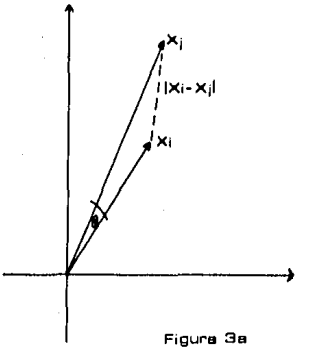


Figura 3a

SEMEJANZA POR FUNCION DE DISTANCIA

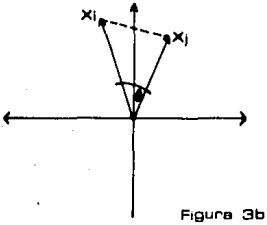


Figura 3b

1. Para la figura (3a).

$X_i = (2,4)$ y $X_j = (3,7)$

$$|X_i - X_j| = 3.1623 \quad r_{ij} = \frac{X_i \cdot X_j}{|X_i|\ |X_j|} = \frac{34}{\sqrt{20}\ \sqrt{58}} = 0.998$$

$\theta = 3°$ de separación entre X_i y X_j.

2. Para la figura (3b).

$X_i = (-1,3)$ y $X_j = (1,4)$

$$|X_i - X_j| = 2.236 \quad r_{ij} = \frac{11}{\sqrt{10}\ \sqrt{17}} = 0.8436$$

$\theta = 32°$ de separación entre X_i y X_j.

En este ejemplo se muestra que para el caso de la figura (3a) X_i y X_j tiene un grado de semejanza "alto" con referencia al coeficiente de correlación lineal, ya que r_{ij} es cercano a 1, mientras que para el caso de la figura (3b), X_i y X_j tiene un grado de semejanza mayor, si se considera de la función de distancia euclidiana para definir semejanza, ya que la distancia entre dichos vectores es menor.

3.5 Métrica de Mahalanobis

En ocasiones resulta necesario asignar diferentes "pesos" a las variables que definen a los elementos. De esta manera se define una función de distancia que escala dichas variables por medio de una transfor-

mación lineal. Esta función de distancia se define como:

$$d_w(X_i,X_j) = [\sum_{k=1}^{p} (W_k\ (X_{ki} - X_{kj}))^2]^{1/2}$$

donde $(W_k,\ k=1,2,\ldots,p)$ es el conjunto de pesos asignado a cada una de las p variables.

$$\text{Sea } X_i' = WX_i = \begin{bmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & W_p \end{bmatrix} \begin{bmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{pi} \end{bmatrix}$$

$$X_j' = WX_j$$

$$d_w(X_i,X_j) = [(W(X_i - X_j))^T\ \ W(X_i - X_j)]^{1/2}$$

$$= [(X_i - X_j)^T\ \ W^T\ W(X_i - X_j)]^{1/2}$$

$$= d_2(X_i',X_j')$$

La función de distancia con pesos d_w en el espacio original es sólo la distancia euclidiana en el espacio transformado.

La transformación anterior (que escala las variables) puede generalizarse a una transformación lineal como la rotación. La matriz de "pesos" para este caso será:

$$W = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1p} \\ W_{21} & W_{22} & \cdots & W_{2p} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ W_{p1} & W_{p2} & \cdots & W_{pp} \end{bmatrix}$$

de lo que resulta una función de distancia generalizada.

$$D_W(X_i,X_j) = [(X_i - X_j)^T \; W^T \; W(X_i - X_j)]^{1/2}$$

Si se define $Q = W^T W$

$$D_W(X_i,X_j) = [\sum_{k=1}^{p} \sum_{\ell=1}^{p} (X_{ki} - X_{kj}) \; q_{k\ell} \; (X_{\ell i} - X_{\ell j})]^{1/2}$$

Para el caso de variables no correlacionadas y de igual varianza, la distancia euclidiana es apropiada para agrupar elementos; sin embargo, en la realidad se presentan casos en que las variables que describen a los elementos están sustancialmente relacionadas (reflejan la misma información). Para ello, se define una generalización de la distancia euclidiana que es la métrica de Mahalanobis, que se relaciona con la función de distancia generalizada $D_W(X_i,X_j)$ arriba definida; ya que para el caso de la métrica de Mahalanobis, Q equivale a la inversa de la matriz de varianza-covarianza (S^{-1}) entre las p variables (en el anexo B se tiene la demostración de que $Q = S^{-1}$).

La métrica de Mahalanobis se define por:

$$D_M(X_i,X_j) = (X_i - X_j)^T \; S^{-1} \; (X_i - X_j)$$

donde S está dada por:

$$S = 1/n \sum_{i=1}^{n} (X_i - \bar{X}) \; (X_i - \bar{X})^T$$

S debe ser positiva definida.

con $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)$ y $\bar{X}_k = \sum_{i=1}^{n} X_{ki}/n$ $\forall\ k=1,2,\ldots,p$

En conclusión, la métrica de Mahalanobis corrige el efecto que se provoca cuando hay gran relación entre las variables, o bien cuando las escalas de éstas difieren en gran medida, lo que da una mejor y más segura estimación del grado de semejanza entre elementos.

Desarrollando la métrica de Mahalanobis D_M que ya se definió se puede observar:

$$D_M(X_i, X_j) = X_i^T S^{-1} X_i + X_j^T S^{-1} X_j - 2X_i^T S^{-1} X_j$$

donde $X_i^T S^{-1} X_j$ es la parte asociada a la proporcionalidad.

Por lo que el índice de correlación lineal se generaliza como se muestra a continuación:

$$R^2(X_i, X_j) = \frac{X_i S^{-1} X_j}{\sqrt{X_i^T S^{-1} X_i}\ \sqrt{X_j^T S^{-1} X_j}}$$

3.6 Similitud entre Perfiles

Una forma gráfica de representar en un espacio bidimensional, vectores con p componentes (características), es en base a perfiles.

Considere una gráfica en el plano cartesiano, donde los valores de

las variables se representan en la ordenada (Y), y el número asignado a cada variable en la abscisa (X): la línea que conecta los puntos, así definidos en el plano (XY) se llama perfil.

De acuerdo a lo anterior, la semejanza entre elementos se reduce a la semejanza entre sus perfiles.

Para $X_i=(X_{1i},X_{2i},\ldots,X_{pi})$ y $X_j=(X_{1j},X_{2j},\ldots,X_{pj})$, se tiene la representación gráfica de sus perfiles en la figura (4).

Para este enfoque se tienen algunas medidas de asociación que determinan el grado de semejanza entre sus elementos.

3.6.1 Coeficiente de DU-MAS

Para medir el grado de semejanza entre elementos, que como ya se dijo, se reduce a medir la semejanza entre sus perfiles. DU-MAS define un coeficiente en términos de la dirección de las pendientes de los segmentos del perfil de cada elemento, comparando con el número total de segmentos que definen los perfiles de dichos elementos.

Un segmento del perfil se define por la línea que va de un punto del valor de la variable al punto que determina el valor de la variable adyacente. Para p variables existe p-1 segmentos de un perfil.

La probabilidad de que un segmento aleatorio resulte con pendiente positiva o negativa es 1/2.

REPRESENTACION GRAFICA DE LOS PERFILES

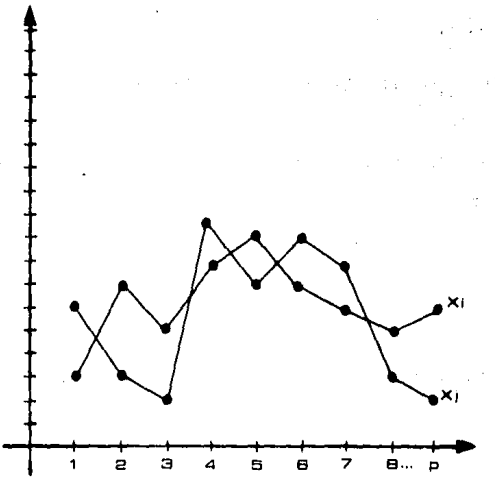


Figura 4

Sea S_{ij} el número de segmentos de los perfiles que definen al i-ésimo y j-ésimo elemento de igual dirección; y sea T_{ij} el número total de segmentos de los perfiles del i-ésimo y j-ésimo elementos, menos los segmentos con pendiente cero.

El valor esperado de S_{ij} será $1/2T_{ij}$, es decir,

$$E(S_{ij}) = 1/2\ T_{ij}$$

donde $T_{ij}=(p-1)$ - [número de segmentos con pendiente cero].

El coeficiente de DU-MAS para X_i y X_j denominado $DM(X_i,X_j)$ se define como:

$$DM(X_i,X_j) = 2(S_{ij}/T_{ij} - 1/2)$$

este coeficiente tiene media 0 y rango (-1,1).

La distribución de S_{ij}/T_{ij} es una distribución Binomial, con parámetros $(T_{ij},1/2)$, la que para T_{ij} "muy grande" puede aproximarse con una - distribución Normal.

Dado que $E(S_{ij})=1/2\ T_{ij}$ se tiene $E(S_{ij}/T_{ij})=1/2$, por otro lado, como

$$VAR\ (S_{ij}) = 1/4\ T_{ij} \quad y \quad VAR\ (S_{ij}/T_{ij}) = 1/T_{ij}^2\ VAR\ (S_{ij})$$

y su desviación estándar es:

$$\sigma_{S_{ij}/T_{ij}} = 1/2\ \sqrt{1/T_{ij}}$$

La varianza del coeficiente de DU-MAS es entonces:

$$VAR\ [DM\ (X_i,X_j)] = VAR(2\ S_{ij}/T_{ij} - 1) = 4\ VAR(S_{ij}/T_{ij})$$

y su desviación estándar es:

$$\sigma_{DM_{ij}} = \sqrt{1/T_{ij}}$$

Una crítica al coeficiente de DU-MAS, es que éste depende de la secuencia en que se ordenan las variables, y que sólo concentra su atención en la dirección de las pendientes de los segmentos y no al tamaño de éstas.

3.6.2 Coeficiente de Cattell

Otro coeficiente que mide el grado de semejanza entre elementos es el coeficiente de Cattell, que se define bajo los supuestos de considerar que las variables son independientes, y se distribuyen Normal (μ_r,σ_r^2) $\forall$ r=1,2,...,p.

Sean X_i y X_j dos elementos definidos por $X_i=(X_{1i},X_{2i},\ldots,X_{pi})$ y $X_j=(X_{1j},X_{2j},\ldots,X_{pj})$; y sean $C_1,C_2,\ldots,C_p$ las p variables (o componentes) que definen las características o atributos de dichos vectores.

Si C_r $\forall$ r=1,2,...,p se distribuyen Normal (μ_r,σ_r^2) se estandarizan determinando nuevas variables Z_r definidas por:

$$Z_r = \frac{C_r - \mu_r}{\sigma_r}$$

que se distribuirán Normal (0,1) ∀ r=1,2,...,p.

La distribución de las diferencias de dichas variables $(Z_{ri} - Z_{rj})$ será Normal con los siguientes parámetros:

$$E(Z_{ri} - Z_{rj}) = E(Z_{ri}) - E(Z_{rj}) = 0$$

$$VAR(Z_{ri} - Z_{rj}) = VAR(Z_{ri}) + (-1)^2 VAR(Z_{rj}) = 2$$

Por lo tanto, $(Z_{ri} - Z_{rj})$ se distribuye Normal (0,2), y en consecuencia $\sum_{r=1}^{p} \frac{(Z_{ri} - Z_{rj})^2}{2}$ se distribuye χ^2 (p).

Con lo que se define el coeficiente de Cattell para X_i y X_j:

$$Ct\ (X_i,X_j) = 1 - \sum_{r=1}^{p} (Z_{ri} - Z_{rj})^2 \ /\ E(\sum_{r=1}^{p} (Z_{ri} - Z_{rj})^2)$$

de donde se tiene:

$$E[\sum_{r=1}^{p} (Z_{ri} - Z_{rj})^2] = \sum_{r=1}^{p} E(Z_{ri} - Z_{rj})^2 = 2p$$

por lo tanto, el coeficiente de Cattell para X_i y X_j es:

$$Ct\ (X_i,X_j) = 1 - \sum_{r=1}^{p} \frac{(Z_{ri} - Z_{rj})^2}{2p}$$

este coeficiente tiene propiedades similares al coeficiente de correlación lineal, que se describe a continuación:

- El coeficiente de Cattell para X_i y X_j toma valores en un intervalo (-1,1), es decir, $-1 \leq Ct\ (X_i,X_j) \leq 1$.

- Si Ct (X_i, X_j) tiende a 1, significa que el grado de semejanza entre X_i y X_j es "alto".

- Si Ct (X_i, X_j) tiende a 0, es que $\sum_{r=1}^{p} (Z_{ri} - Z_{rj})$ tiende a su media (2p).

- Si Ct (X_i, X_j) tiende a -1, significa que el grado de semejanza entre X_i y X_j es "bajo".

Para la práctica el coeficiente de Cattell se basa en dos suposiciones muy fuertes, que un gran número de ocasiones no se cumplen. Estas son, como ya se mencionaron, que las variables son independientes y que su distribución es Normal.

CAPITULO 4

CRITERIOS DE OPTIMALIDAD

Intimamente relacionado con el problema de cúmulos, se tiene el concepto de criterios de optimalidad, el que definirá la función objetivo - del problema. Lo anterior se debe a que, la solución al problema de cúmulos es determinar aquella partición que satisfaga un criterio de optimalidad para una medida de asociación dada. Por ejemplo, se desea obtener la partición de un conjunto de elementos, con un criterio de optimalidad que maximice la distancia entre los grupos que forman dicha partición.

Para el problema de cúmulos se plantearán algunos criterios de optimalidad de los que se obtendrán diferentes funciones objetivo. En general, éstos buscan la homogenidad dentro de cada grupo y/o la heterogenidad entre los grupos.

En consecuencia, dependiendo del criterio de optimalidad que sea - utilizado se llegará a diferentes agrupamientos.

En esta sección se expondrán algunos ejemplos de criterios de optimalidad que estarán dados en términos de una relación funcional, que como ya se mencionó en la sección 2.3, se le da el nombre de función objetivo. La selección más adecuada de ellos dependerá de la naturaleza del problema que se esté analizando.

4.1 Criterios de Optimalidad Cuyo Objetivo Pretende la Homogeneidad

Dado el conjunto de elementos $I=\{I_1, I_2, \ldots, I_n\}$ cuyos vectores de atributos están contenidos en el conjunto $X=\{X_1, X_2, \ldots, X_n\}$, se dice que los elementos I_i, I_j de I pertenecen a un mismo grupo π_k si el grado de semejanza (para una medida de asociación dada) entre X_i y X_j es suficientemente "grande". Por otro lado, se dice que estos pertenecen a diferentes grupos si el grado de semejanza entre X_i y X_j es suficientemente "pequeño". Los criterios de optimalidad que tienen como objetivo buscar la homogeneidad dentro de los grupos, se basan en maximizar el grado de semejanza de los elementos de un mismo grupo.

Si se define un vector $Y_k=(X_1, X_2, \ldots, X_{nk})$ de nk componentes para cada grupo π_k, se tendrá un conjunto $Y=\{Y_1, Y_2, \ldots, Y_m\}$ asociado a cada partición $\pi \in P$, donde P es el conjunto de particiones posibles de I[1].

Es posible definir una función $F: Y \longrightarrow R$ que mida la homogenidad en cada grupo, a partir de la cual se puede evaluar la homogenidad de la partición. Sea $F(Y_k)=y_k$ el valor de dicha función para el k-ésimo grupo; y sea $G: P \longrightarrow R$ donde $P \subset R^m$, la función objetivo que buscará la partición óptima y* del conjunto de particiones posibles de P, que maximice la homogenidad dentro de los grupos o cúmulos de dicha partición; es decir, dado el vector $y=(y_1, y_2, \ldots, y_m)$ $y \in P$ se tiene que la partición deseada será aquella para la cual $G(y^*) = \underset{y \in P}{\text{MIN}}\, G(y)$.

---

[1] El número total de particiones posibles se da por el número de stirling de segunda clase que se verá en el siguiente capítulo.

Cuando G(y) es una función de costos o disimilitud al minimizarse, maximiza el grado de semejanza dentro de los grupos.

Para aclarar lo anterior se tiene el siguiente planteamiento:

PARTICION	FUNCIONES PARCIALES (PARA CADA GRUPO)	FUNCION OBJETIVO (PARA CADA PARTICION)
$\P$	$F(Y_1), F(Y_2), \ldots, F(Y_m)$	$G(y)$

donde $F(Y_k)=y_k \quad \forall \; k=1,2,\ldots,m \quad y \quad y=(y_1,y_2,\ldots,y_m)$, $G(y)$ se define como:

$$G(y) = G(y_1,y_2,\ldots,y_m) = G(F(Y_1), F(Y_2),\ldots,F(Y_m)).$$

A continuación se darán algunos ejemplos de funciones parciales y - funciones objetivo:

FUNCION PARCIAL $F(Y_k) = y_k$ PARA CADA GRUPO	FUNCION OBJETIVO $G(y)$ PARA CADA PARTICION	CRITERIO MAX O MIN
1. $\sum_{i=1}^{n_k} d^2(X_i,\bar{X}^k)$ Varianza del grupo $\P_k$	$G(y) = \sum_{k=1}^{m} y_k$ Suma de las varianzas de los m grupos	MIN $G(y)$ $y\in P$
2. MAX $\{d^2(X_i,X_j)\}$ $i < j = 1,2,\ldots,n_k$ Distancia máxima de - los elementos del gru po $\P_k$	$G(y) = \prod_{k=1}^{m} y_k$ Multiplicación de las máximas distancias de los m grupos	MIN $G(y)$ $y\in P$

FUNCION PARCIAL $F(Y_k) = y_k$ PARA CADA GRUPO	FUNCION OBJETIVO $G(y)$ PARA CADA PARTICION	CRITERIO MAX O MIN
3. $\sum_{i<j=1}^{nk} \lvert r_{ij} \rvert$	$G(y) = \underset{k=1,2,\ldots,m}{\text{MIN}} \{y_k\}$	$\underset{y \in P}{\text{MAX}} \; G(y)$
Suma de la correlación entre los elementos del grupo π_k, en valores - absolutos	Valor mínimo de las sumas de las correlaciones de los m grupos	
4. $1/n_k \sum_{i=1}^{nk} \sum_{j=1}^{nk} d^2(X_i, X_j)_{i<j}$	$G(y) = \sum_{k=1}^{m} y_k$	$\underset{y \in P}{\text{MIN}} \; G(y)$
Suma de distancias de todos los elementos - del grupo π_k	Suma de las distancias de todos los elementos de los m grupos	
5. $\underset{i=1,2,\ldots,n_k}{\text{MIN}} \{d^2 (X_i, \bar{X}^k)\}$	$G(y) = \sum_{k=1}^{m} y_k^2$	$\underset{y \in P}{\text{MIN}} \; G(y)$
Mínima varianza del grupo π_k	Suma de las mínimas varianzas de los m grupos	
6. $\underset{i=1,2,\ldots,n_k}{\text{MAX}} \{d^2 (X_i, \bar{X}^k)\}$	$G(y) = \sum_{k=1}^{m} y_k^2$	$\underset{y \in P}{\text{MIN}} \; G(y)$
Máxima varianza del grupo π_k	Suma de las máximas - varianzas de los m - grupos	
7. $\underset{i<j=1,2,\ldots,n_k}{\text{MAX}} \{d^2 (X_i, X_j)\}$	$G(y) = \underset{k=1,2,\ldots,m}{\text{MAX}} \{y_k\}$	$\underset{y \in P}{\text{MIN}} \; G(y)$
Máxima distancia entre los elementos del grupo π_k	Máxima distancia de entre los elementos de - los m grupos	

donde $\bar{X}^k = \sum_{i=1}^{nk} X_i/n_k$ es la media del grupo π_k,

nk es el número de elementos del grupo π_k,

lo que implica que $n = \sum_{k=1}^{m} n_k$.

Observaciones importantes acerca de los ejemplos anteriores:

En el Ejemplo 1, se pretende encontrar la partición que minimice la suma de las varianzas en los grupos, ya que:

$$\sum_{i=1}^{nk} d^2(X_i, \bar{X}^k) = (X_i - \bar{X}^k)^T \quad (X_i - \bar{X}^k)$$

esta función parcial es igual a la definida en el ejemplo 4, en el que se pretende minimizar la suma de cuadrados de la distancia entre todos los elementos que pertenencen a un mismo grupo. La equivalencia de la función del ejemplo 1 y la del 4, se muestra en el anexo C.

En el Ejemplo 3, el criterio de optimización es maximizar la medida de asociación usada, que es el coeficiente de correlación entre los elementos. (Como se explicó en la sección 3.4, esta medida toma sus valores de manera inversa a las funciones de distancia).

En el Ejemplo 5, se pretende encontrar la partición que minimice la suma cuadrada de las mínimas varianzas dentro de cada grupo, tendiendo a formar grupos extensos y con varianzas desiguales; en comparación con el ejemplo 6 que pretende minimizar la suma cuadrada de las máximas varian-

zas dentro de los grupos, buscando así formar grupos compactos con varian zas similares y distantes entre si.

En el Ejemplo 7, se toma como cota máxima, la máxima distancia entre pares de elementos la que pretenderá ser el máximo diámetro para cada gru po y se tratará de minimizar.

En general, cada uno de los ejemplos anteriores, dará como resultado diferentes tipos de agrupamientos, es decir, la partición óptima será diferente, dependiendo del criterio usado para definir homogenidad y la medida de distancia establecida.

4.2 Criterios de Optimalidad Cuyo Objetivo Pretende la Heterogenidad

En la sección anterior, el objetivo del problema de cúmulos se planteó como el de encontrar la partición óptima, de n elementos dentro de m grupos disjuntos, que satisfaciera un criterio dado de homogenidad dentro de cada grupo. Ahora, se analizará el criterio de heterogenidad entre - los grupos que formarán la partición óptima. Para ello, se busca minimizar el grado de semejanza entre los elementos de diferentes grupos.

Sea F una función que asocia los elementos de un grupo con otro, y sea $Y=\{(Y_k,Y_l)\}$ $k<l$, $\forall$ $l,k=1,2,\ldots,m$ el conjunto de pares ordenados; como ya se definió en la sección anterior $Y_k=(X_1,X_2,\ldots,X_{nk})$ - $\forall$ $k=1,2,\ldots,m$.

$F(Y_k,Y_l)=y_{kl}$ se define como una función que relaciona los elemen-

tos del k-ésimo grupo con los elementos del l-ésimo grupo de una cierta partición. Lo que significa que cada grupo π_k de la dicha partición se asociará con los m-1 grupos restantes.

Por lo tanto, $F: Y \longrightarrow R$, $Y \subset R^2$, genera las componentes para el vector $y=(y_{12},y_{13},\ldots,y_{1m},\ y_{23},y_{24},\ldots,\ y_{2m},\ldots,\ y_{m-1m})$ que tiene $M=m(m-1)/2$ componentes.

La función objetivo del problema $G: P \longrightarrow R$ $P \subset R^M$ será tal que: $G(y^*) = \underset{y\in P}{MAX}\ G(y)$. Esta función representa el beneficio total para cada partición del problema y se maximiza, lo que equivale a minimizar el grado de semejanza entre los grupos.

Para este tipo de criterios se tienen algunos casos en los siguientes ejemplos:

FUNCION PARCIAL $F(Y_k,Y_l) = y_{kl}$ PARA CADA PAR DE GRUPOS	FUNCION OBJETIVO $G(y)$ PARA CADA PARTICION	CRITERIO MAX O MIN
1. $\sum_{j=1}^{n_l}\sum_{i=1}^{n_k} d^2(X_i,X_j)$	$G(y) = \underset{k<l\ =\ 1,2,\ldots,m}{MIN}\ \{y_{kl}\}$	$\underset{y\in P}{MAX}\ G(y)$
Suma de las distancias entre los elementos de los grupos π_k y π_l	Mínima suma de distancias de elementos entre los m grupos	
2. $\underset{\substack{i=1,2,\ldots,n_k\\ j=1,2,\ldots,n_l}}{MIN}\{d^2(X_i,X_j)\}$	$G(y) = \sum_{k=1}^{m}\sum_{l=1}^{m} y_{kl}$	$\underset{y\in P}{MAX}\ G(y)$
Mínima distancia entre los elementos de los grupos π_k y π_l	Suma de mínimas distancias de los m grupos	

FUNCION PARCIAL $F(Y_k,Y_l) = y_{kl}$ PARA CADA PAR DE GRUPOS	FUNCION OBJETIVO $G(y)$ PARA CADA PARTICION	CRITERIO MAX O MIN
3. $\sum_{i=1}^{n_l}\sum_{j=1}^{n_k} d^2(X_i,X_j)$	$G(y) = \underset{k<l=1,2,\ldots,m}{\text{MAX}} \{y_{kl}\}$	$\underset{y\in P}{\text{MAX}}\ G(y)$
Suma de las distancias entre los elementos de los grupos $\P_k$ y $\P_l$	Máxima suma de distancias de elementos entre los m grupos	
4. $\sum_{i=1}^{n_l}\sum_{j=1}^{n_k} d^2(X_i,X_j)/n_k\, n_l$	$G(y) = \underset{k<l=1,2,\ldots,m}{\text{MIN}} \{y_{kl}\}$	$\underset{y\in P}{\text{MAX}}\ G(y)$
Suma promedio de las distancias entre los elementos de los grupos $\P_k$ y $\P_l$	Mínima suma promedio de las distancias entre los m grupos	
5. $\sum_{i=1}^{n_l}\sum_{j=1}^{n_k} r_{ij}^2$	$G(y) = \underset{k<l=1,2,\ldots,m}{\text{MAX}} \{y_{kl}\}$	$\underset{y\in P}{\text{MIN}}\ G(y)$
Suma de la correlación cuadrada entre los elementos de los grupos $\P_k$ y $\P_l$	Máxima suma de correlaciones de elementos entre los m grupos	
6. $\underset{\substack{i=1,2,\ldots,n_k \\ j=1,2,\ldots,n_l}}{\text{MAX}} \{DM(X_i,X_j)\}$	$G(y) = \prod_{k=1}^{m}\prod_{l=1}^{m} y_{kl}^2 \quad k<l$	$\underset{y\in P}{\text{MIN}}\ G(y)$
Máximo coeficiente de DUMAS entre los elementos de los grupos $\P_k$ y $\P_l$	Multiplicación de los máximos coeficientes de DUMAS cuadrados de los m grupos	
7. $n_k\, n_l/(n_k+n_l)d^2(\bar{X}^k\bar{X}^l)$	$G(y) = \sum_{k=1}^{m}\sum_{l=1}^{m} y_{kl} \quad k<l$	$\underset{y\in P}{\text{MAX}}\ G(y)$
Distancia estadística entre los elementos de los grupos $\P_k$ y $\P_l$	Suma de distancias estadísticas de los m grupos	

A continuación se harán algunas observaciones importantes de los - ejemplos anteriores:

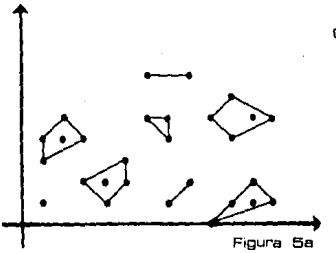
Los Ejemplos 1 y 2 comparten un objetivo parecido al maximizar una función de la mínima distancia entre elementos de diferentes grupos, dando como resultado un gran número de pequeños grupos compactos y distantes entre si. Lo que se ilustra en la figura (5a). Los agrupamientos - óptimos para dichos ejemplos no son necesariamente iguales.

Como un caso "recíproco" al anterior se tiene el Ejemplo 3 donde - se maximiza una función de la máxima distancia cuadrada entre los elementos de los grupos, formándose pocos grupos con un número considerable de elementos y gran dispersión entre los elementos de cada grupo. Esto se ve en la figura (5b).

El Ejemplo 4 representa un caso intermedio de los dos anteriores, - ya que maximiza una función del promedio de la distancia cuadrada entre los elementos de diferentes grupos. Lo que se ilustra en la figura (5c).

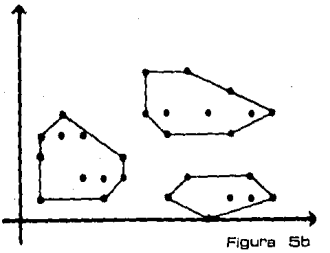
Como se puede observar, existe una infinidad de maneras para plantear un problema de cúmulos, por ello, es de suma importancia que se - analice el tipo de información que se tiene, los objetivos del agrupamiento y las limitaciones para su desarrollo.

Para los Ejemplos 3 y 6 se busca minimizar la función objetivo ya que tanto el coeficiente cuadrado de correlación lineal como el de DUMAS, mientras menor sea su valor menor será el grado de semejanza entre los -



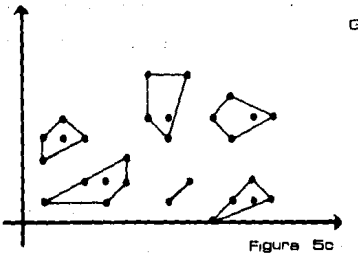
GRUPOS COMPACTOS

Figura 5a



GRUPOS DISPERSOS

Figura 5b



GRUPOS PROMEDIO

Figura 5c

elementos de diferentes grupos.

La función descrita en el Ejemplo 7 se le conoce como la distancia estadística entre el k-ésimo y el l-ésimo grupos, donde $\bar{X}^k = \sum_{i=1}^{nk} X_i/n_k$.

Hasta el momento se han analizado aquellos criterios que pretenden la homogenización de los elementos de un mismo grupo, o bien, la heterogenización de los elementos que pertenecen a diferentes grupos. En la parte siguiente de este capítulo se realizará un breve estudio para los casos en que se plantean ambos criterios a la vez.

4.3 Bicriterio de Optimalidad en el Problema de Cúmulos

En un gran número de aplicaciones tanto el criterio de homogenidad como de heterogenidad de los grupos son de gran importancia, pero con frecuencia se encuentra en conflicto.

La mayoría de los algoritmos para resolver el problema de cúmulos prestan atención a sólo un criterio, en esta sección se planteará el problema optimizando el bicriterio basado en una función de utilidad.

Se empezará por algunas definiciones:

Una partición se le llama eficiente cuando no existe otra con mejor valor para uno de los criterios de optimalidad mencionados, y da al menos un buen valor para el otro criterio. En otras palabras, una parti-

ción eficiente obtiene el mejor valor posible para uno de los criterios, sujeto a la restricción de que el valor del otro criterio no será menor a un valor dado.

Sea $D=\{d_{ij}\}$ una matriz que mide el grado de disimilitud entre los pares de elementos I_i e I_j para $i,j=1,2,\ldots,n$. Como I está asociado a un conjunto de vectores $X=\{X_1,X_2,\ldots,X_n\}$, D puede ser la matriz de distancia entre dichos vectores, ya que de alguna forma ésta mide disimilitud (o bien, D sólo puede tener significado ordinal).

Como π denota una partición de I en m grupos, sea P el conjunto de particiones de I con m grupos no vacios y $P^{\sim}$ el conjunto de particiones de I con a lo más m grupos no vacios.

Recordando que Y_k está asociado al grupo π_k de la partición π y el vector y con la partición π de P.

Se define el diámetro $d(Y_k)$ de un grupo π_k de π como el grado máximo de disimilitud (o bien, máxima distancia) entre los elementos de π_k, es decir:

$$d(Y_k) = \underset{i<j=1,2,\ldots,nk}{MAX} d_{ij}$$

El diámetro d(y) de una partición π(con $d(Y_k) = y_k$ y $y=(y_1,y_2,\ldots,y_m)$

se define como el máximo diámetro de sus grupos o cúmulos, o sea:

$$d(y) = \underset{k=1,2,\ldots,m}{MAX}\ d(Y_k)$$

Una partición de mínimo diámetro π^* de I, en m grupos cumple con:

$$d(y^*) = \underset{\pi \in P}{MIN}\ d(y)$$

o bien:

$$d(y^*) = \underset{\pi \in P}{MIN}\ \underset{\pi_k \in \pi}{MAX}\ d(Y_k) = \underset{\pi \in P}{MIN}\ \{\underset{k=1,2,\ldots,m}{MAX}\ \{d(Y_k)\}\}$$

o bien:

$$d(y^*) = \underset{\pi \in P}{MIN}\ \underset{\pi_k \in \pi}{MAX}\ \underset{I_i, I_j \in \pi_k}{MAX}\ d_{ij} = \underset{\pi \in P}{MIN}\ \{\underset{k=1,2,\ldots,m}{MAX}\ \{\underset{i<j=1,2,\ldots,n_k}{MAX}\ d_{ij}\}\}$$

Se define la separación $S(Y_k, Y_l)$ del grupo π_k de π, como la mínima distancia entre cualquier elemento de π_k y cualquier elemento de otro grupo (π_l) que esté en π, es decir:

$$S(Y_k, Y_l) = \underset{\substack{i=1,2,\ldots,n_k \\ j=1,2,\ldots,n_l}}{MIN}\ d_{ij}$$

La separación S(y) de una partición π (con $S(Y_k, Y_l) = y_{kl}$ y $y = (y_{12}, y_{13}, \ldots, y_{1m}, y_{23}, y_{24}, \ldots, y_{2m}, \ldots, y_{m-1m})$) se define como la mínima separación de sus cúmulos.

Una partición de máxima separación $\hat{\pi}$ de I en m grupos es:

$$S(\hat{y}) = \underset{\pi \in P}{MAX} \; S(y) = \underset{\pi \in P}{MAX} \; \{S(y)\}$$

o bien:

$$S(\hat{y}) = \underset{\pi \in P}{MAX} \; \underset{\pi_k \neq \pi_l \in \pi}{MIN} \; S(Y_k, Y_l) = \underset{\pi \in P}{MAX} \; \{ \underset{1<k=1,2,\ldots,m}{MIN} \; \{ S(Y_k, Y_l) \} \}$$

o bien:

$$S(\hat{y}) = \underset{\pi \in P}{MAX} \; \underset{\pi_k \neq \pi_l \in \pi}{MIN} \; \underset{I_i \in \pi_k, \, I_j \in \pi_l}{MIN} d_{ij} = \underset{\pi \in P}{MAX} \; \{ \underset{1<k=1,2,\ldots,m}{MIN} \; \{ \underset{\substack{i=1,2,\ldots,n_k \\ j=1,2,\ldots,n_l}}{MIN} d_{ij} \} \}$$

Se define una partición π de m grupos como eficiente si y sólo si - no existe una partición $\pi' \in P^{\sim}$ tal que:

$$d(y') < d(y) \quad y \quad S(y') \geq S(y)$$

o tal que:

$$S(y') > S(y) \quad y \quad d(y') \leq d(y)$$

Nótese que en esta definición a π' se le permitirá tener menor o - igual número de grupos o cúmulos no vacios.

Dos particiones eficientes π y $\pi' \in P^{\sim}$ son equivalentes si y sólo si $d(y) = d(y')$ y $S(y) = S(y')$.

Como se definió al principio de esta sección, una partición es eficiente cuando no existe otra con mejor valor para uno de los criterios

de optimalidad mencionados, y da al menos un buen valor para el otro criterio, entonces:

Sea E el conjunto de particiones eficientes de I. Se dice que E es completo si y sólo si cualquier partición eficiente de I pertenece a E ó es equivalente a una partición eficiente de E.

Sea E el conjunto de particiones eficientes de I. Se dice que E es mínimo si y sólo si ningún par de particiones eficientes de E son equivalentes.

Sea U(d,S) una función de utilidad definida para todos los valores del diámetro d y separación S pertenecientes a $D=\{d_{ij}\}$. Se supondrá que U es monótona decreciente en d y monótona creciente en S.

El bicriterio para el problema de cúmulos con una función de utilidad busca determinar $\tilde{\pi}$ tal que:

$$U[d(\tilde{y}), S(\tilde{y})] = \underset{\pi \in P}{MAX} \; U[d(y), S(y)]$$

Como ya se definió E es el conjunto de particiones eficientes de I, la función de utilidad se puede definir entonces:

$$U[d(\tilde{y}), S(\tilde{y})] = \underset{\pi \in E}{MAX} \; U[d(y), S(y)]$$

Un caso particular de la función de utilidad es el siguiente:

$$U[d,S] = \begin{cases} 0 & S < S_{min} \quad 0 \quad d > d_{max} \\ 1 & \text{d.o.f.} \end{cases}$$

donde S_{min} y d_{max} son cotas dadas.

Los elementos de la matriz de disimilitud deberán satisfacer las condiciones de una métrica; esto es:

a) $d_{ij} \geq 0$, b) $d_{ij} = d_{ji}$ y c) $d_{ii} = 0$ para $i \neq j = 1,2,\ldots,n$.

De esta forma, las particiones eficientes determinadas por el algoritmo del bicriterio para el problema de cúmulos, son invariantes a las transformaciones rígidas de $\{d_{ij}\}$. Por que cuando $\{d_{ij}\}$ no tiene un significado métrico, una transformación a la función de utilidad U[d,S] dada, podría dar óptima una partición eficiente diferente a $\bar{\pi}$, en tal caso será necesario formular la función de utilidad de manera diferente.

Por otro lado, es posible definir dos parámetros $Q_h(y)$ y $Q_s(y)$ que indicarán la "calidad" de la homogenidad y de la separación (o heterogenidad) respectivamente, para una partición de π de I.

Sea $Q_h(y)$, la calidad de la homogenidad de una partición π de I, la razón del número de pares de elementos $\{I_i,I_j\}$ con distancia $d_{ij} \geq d(y)$ en diferentes grupos, con el número total de pares de elementos. La calidad de la homogenidad de una partición π de I se define como:

$$Q_h(y) = \{\{I_i,I_j\} \in I \;/\; \exists\, k \neq l \;/\; I_i \in \pi_k,\ I_j \in \pi_l,\ d_{ij} \geq d(y)\} \;/\; n(n-1)$$

La calidad de separación $Q_s(y)$ de una partición π de I se define en forma análoga, como la razón del número de pares de elementos $\{I_i, I_j\}$ - con distancia $d_{ij} \leq S(y)$ del mismo grupo, con el número total de pares de elementos. La calidad de la separación de una partición π de I es:

$$Q_s(y) = \{\{I_i, I_j\} \in I \;/\; \exists\, k \;/\; I_i \in \pi_k,\ I_j \in \pi_k,\ d_{ij} \leq S(y)\} \;/\; n(n-1)$$

El conjunto de particiones eficientes E de I para el diámetro d y separación S, es un conjunto mínimo completo de particiones eficientes para la calidad de homogenidad Q_h y la calidad de la separación Q_s.

Para el caso en que se plantee un bicriterio en el problema de cúmulos, se hará en términos de calidades que será entonces, determinar la - partición $\bar{\pi}'$ de I tal que:

$$U[Q_h(\bar{y}'),\ Q_s(\bar{y}')] = \underset{\pi \in P}{MAX}\ U[Q_h(y),\ Q_s(y)]$$

Con U como función de utilidad definida por Q_h y $Q_s \in [0,1]$ y - $Q_h + Q_s = 1$.

CAPITULO 5

METODOS DE SOLUCION PARA EL PROBLEMA DE CUMULOS

En los capítulos anteriores se mencionó que el problema básico en el análisis de cúmulos es obtener una partición de un conjunto de elementos, de tal manera que se satisfaga un criterio fijado de antemano. Cabe recordar que una partición es una familia de subconjuntos llamados cúmulos o grupos, tal que sus elementos son mutuamente excluyentes y colectivamente exhaustivos. En este contexto, el criterio para obtener la partición consiste en optimizar el grado de homogeneidad de los elementos en cada grupo y/o el grado de heterogeneidad entre los grupos.

Al cuantificar el problema el grado de homogeneidad así como el de heterogeneidad, se expresan en términos de una función que mide el grado de semejanza y/o disimilitud entre los elementos, a la que se le da el nombre de función objetivo. El criterio para hacer la partición es entonces, maximizar o minimizar dicha función objetivo bajo ciertas restricciones que se definen de acuerdo a la naturaleza del problema.

Por lo tanto, el problema de cúmulos se reduce a un problema de optimización. En este capítulo se analizarán algunos algoritmos de programación matemática para su solución.

Se hablará de algoritmos de dos tipos, los que conducen a una solución óptima (de tal forma que puede demostrarse la optimalidad), y los

heurísticos que sólo conducen a una solución cercana al óptimo "Buena" (en algunos casos podría ser la óptima); sin embargo, ésta no se puede demostrar.

En general, los algoritmos que resuelven el problema de cúmulos son heurísticos. Los algoritmos que conducen a una solución exacta están - basados en Programación Entera, Programación Dinámica y Teoría de Gráfi cas. Por otro lado, se tiene la Enumeración Exhaustiva, la cual se limita a problemas muy pequeños, dada la magnitud de cálculos que deben rea lizarse, aunque garantiza una solución óptima.

5.1 Método de Enumeración Exhaustiva

Una forma de resolver el problema de cúmulos es evaluando la función objetivo para cada alternativa de agrupación de un conjunto de elementos I, y así elegir la partición que proporcione el valor óptimo de dicha - función objetivo.

El número de Stirling de segunda clase S(n,m) proporcionará el número de alternativas que se obtienen al agrupar n elementos del conjunto I en m grupos o cúmulos, de tal forma que el orden de los elementos en cada grupo sea irrelevante, y que ningún grupo sea vacio.

El número total de formas de dividir n elementos en m grupos (subconjuntos) no vacios está dado por:

$$S(n,m) = \frac{1}{m!} \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} \; k^n \qquad m \leq n$$

Si el número de grupos o cúmulos m no se especifica, entonces el número de alternativas de agrupación son:

$$\sum_{m=1}^{n} S(n,m) \quad m \leq n$$

Una segunda aproximación del número de Stirling de segunda clase - cuando el número de elementos n es muy grande $(n \rightarrow \infty)$

$$\lim_{n \rightarrow \infty} \frac{S(n,m)}{m^n} = \frac{1}{m!}$$

Asintóticamente, se obtiene:

$$S(n,m) = \frac{m^n}{m!}$$

En la Tabla (1) se tienen los valores de $S(n,m)$ para valores de - $n \leq 8$ y $m \leq 8$ donde $n \geq m$.

NUMERO DE PARTICIONES DE UN CONJUNTO DE
n ELEMENTOS EN m GRUPOS O CUMULOS

n \ m	1	2	3	4	5	6	7	8
1	1							
2	1	1						
3	1	3	1					
4	1	7	6	1				
5	1	15	25	10	1			
6	1	31	90	65	15	1		
7	1	63	301	350	140	21	1	
8	1	127	966	1701	1050	266	28	1

TABLA 1

El algoritmo para resolver el problema de cúmulos por Enumeración - Exhaustiva es:

1) Se elige una función objetivo (como puede ser la suma de varianzas dentro de cada grupo).

2) Se desarrolla una lista con todas las particiones posibles del conjunto de n elementos en m grupos o cúmulos.

3) La función objetivo especificada se evalúa en cada una de las particiones posibles.

4) La alternativa o partición que proporcione el valor óptimo (mínimo o máximo) de la función objetivo se selecciona.

Como se puede ver, este procedimiento es impráctico, aunque n sea muy pequeño. Como dos particiones cualquiera pueden contener algunos grupos o cúmulos iguales, los cálculos se repetirán en el método de Enumeración Exhaustiva.

5.2 Método de Programación Dinámica

El algoritmo de Programación Dinámica aplicado al problema de cúmulos fue desarrollado por Jensen [1].

El propósito del análisis de cúmulos bajo un esquema de Programación Dinámica, es la búsqueda sistemática de agrupamientos que proporcionen el valor mínimo de una función objetivo g(y), eliminando aquellos grupos que no proporcionan valores mínimos de g(y) y aquellos que son redundantes obteniendo así la partición óptima.

En esta sección se darán las características principales de un problema de Programación Dinámica, que servirán como base, para el planteamiento general del algoritmo de Programación Dinámica de Jensen para resolver el problema de cúmulos. Usando este algoritmo se resolverá un ejemplo práctico. Finalmente se compararán los resultados de la solución del ejemplo con los obtenidos usando el método de Enumeración Exhaustiva (ver 5.1).

5.2.1 Características de un Problema de Programación Dinámica

Para resolver un problema de Programación Dinámica no se tiene un algoritmo estándar, sino éste se formula de acuerdo a la situación del problema que se desea resolver. Por esta razón, se darán las características principales de los problemas de Programación Dinámica, para así reconocer una situación que podrá ser planteada como un problema de Programación Dinámica.

Estas características son:

1. El problema se podrá dividir por etapas con una política de decisión para cada etapa.

2. Cada etapa tiene un número de estados asociados.

3. La política de decisión en cada etapa es transformar el estado actual en un estado asociado con la etapa anterior.

4. Dado un estado, la política óptima para las etapas restantes es independiente en la política adoptada en etapas previas.

5. El procedimiento de solución comienza encontrando la política óptima para cada estado de la primera etapa.

6. Una relación recursiva identifica la política óptima en cada estado de la etapa k, dado de la política óptima para cada estado en etapa k-1.

7. Usando esta relación recursiva el procedimiento de solución va etapa por etapa encontrando la política óptima para cada estado hasta encontrar la solución óptima en la etapa final.

5.2.2 Planteamiento del Algoritmo de Jensen

Usualmente un problema de programación dinámica se reduce a una ecuación recursiva que refleja las múltiples decisiones interrelacionadas, - cuyo resultado final es "óptimo".

Jensen [1] da una formulación general para el problema de cúmulos

utilizando Programación Dinámica, en términos de la siguiente ecuación - recursiva:

$$g_k(S,Y_k) = \begin{cases} 0 & \text{para } k = 0 \\ F(Y_k) + g^*_{k-1}(S-Y_k) & \text{para } k=1,2,\ldots,m \end{cases}$$

Para que el problema sea óptimo, se buscará minimizar la función - $g_k(S,Y_k)$ en cada etapa k, es decir, encontrar:

$$g^*_k(S) = \min_{Y_k} \{g_k(S,Y_k)\}$$

En la Figura (8) se encuentra la estructura básica que describe este algoritmo de Programación Dinámica y en la siguiente sección se define la notación utilizada.

5.2.2.1 Definiciones

n ≡ número de elementos a agrupar.

m ≡ número de subconjuntos disjuntos y no vacios en los cuales los n elementos son agrupados.

k ≡ índice o variable de etapa

$$m_o \equiv \begin{cases} m & \text{si } n \geq m \quad y \quad n = 2m \\ n-m & \text{si } n < 2m \end{cases}$$

S ≡ variable de estado que representa al conjunto de elemento que se obtiene en la etapa k.

ESTRUCTURA BASICA DEL ALGORITMO DE PROGRAMACION DINAMICA

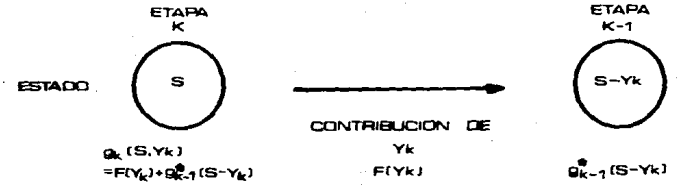


Figura 8

Y_k ≡ variable que representa al grupo π_k de elementos que se adicionan en la etapa k.

$S\text{-}Y_k$ ≡ conjunto de elementos de la etapa k-1.

$F(Y_k)$ ≡ costo de transición de los elementos del grupo o cúmulo π_k. ($F(Y_k)$ se define más adelante).

$g_k(S,Y_k)$ ≡ función objetivo de costos de los elementos del conjunto S divididos en k subconjuntos no vacios, donde Y_k es el k-ésimo grupo.

$g_k^*(S)$ ≡ función de S obtenida del valor mínimo de la función objetivo (en la etapa k) sobre todos los valores de Y_k.

Recordando que π_k es el grupo o cúmulo de n_k elementos, se tiene que el costo transicional $F(Y_k)$ del grupo π_k se define como:

$$F(Y_k) = \frac{1}{2n_k} \sum_{i=1}^{nk} \sum_{j=1}^{nk} d^2(X_i,X_j)$$

que es la suma de varianzas de los n_k elementos del grupo π_k (suponiendo que esta es la medida de asociación que se desea utilizar. Ver capítulo 3).

El número de etapas es $m_o=m$ si $n\geq m$ y $n=2m$; y $m_o=n-m$ si $n<2m$. La razón de esto es que si $n<2m$ habrá $2m-n$ etapas cuyos grupos o cúmulos serán de un sólo elemento. El costo de transición para el grupo de un sólo elemento es 0 (varianza de un elemento), por lo tanto no se altera el valor de la función objetivo, y se puede reducir el número de etapas en -

m_0=n-m. El proceso termina en la etapa m_0 y se asume que los cúmulos - restantes son de un solo elemento.

5.2.2.2 Algoritmo

Para el caso del problema de cúmulos, un algoritmo de Programación Dinámica es el que obtiene agrupamientos óptimos por etapas, de tal manera que en cada etapa la función objetivo se calcula hasta obtener la solución óptima.

Las etapas de este algoritmo son:

Etapa 0.

Se tiene $g_0^*(S) = 0$ por definición.

Etapa 1

Se calcula $g_1(S,Y_1)=F(Y_1)+g_0^*(S-Y_1)$

se tiene $S \equiv Y_1$

entonces $g_1(S,Y_1)=F(Y_1) = F(S)$

y $g_1^*(S) = \min_{Y_1} (F(S)) = F(S)$

Etapa k

Se calcula $g_k(S,Y_k)=F(Y_k)+g_{k-1}^*(S-Y_k)$

donde $S \equiv S + Y_k$

y se determina $g_k^*(S) = \min_{Y_k} (g_k(S,Y_k))$

Etapa m_0

El algoritmo termina al llegar a la etapa m_0 (ver 5.2.2.1),

Se calcula $g_{m_0}(S, Y_{m_0}) = F(Y_{m_0}) + g^*_{m_0-1}(S-Y_{m_0})$

Si $X=(X_1,X_2,\ldots,X_n)$ es el conjunto de valores de los n elementos a agrupar, en esta etapa sólo se tiene un estado que es $S \equiv X$ y así se obtiene:

$$g^*_{m_0}(X_1,X_2,\ldots,X_n) = \min_{Y_{m_0}} \{g_{m_0}(S,Y_{m_0})\}$$

5.2.2.3 Formación de Estados para cada Etapa

El problema principal de un algoritmo de programación dinámica es la formación de estados en cada etapa.

Para el problema de cúmulos se tiene que un estado en la etapa k se define como el conjunto de elementos de k grupos o cúmulos. Como ya se mencionó en la sección 5.2.2.1, S es la variable de estado que se recalcula en cada etapa de acuerdo a los elementos del nuevo grupo que se le adicione.

Por lo tanto, se tiene que S para la etapa k, se forma de los elementos del conjunto S de la etapa k-1 y de los elementos del grupo representado por Y_k, es decir:

$$S \equiv S + Y_k$$

A continuación se describirá la formación de estados para cada etapa:

Etapa 0.

En esta etapa $S = \phi$ por lo tanto sólo se tiene un estado.

Etapa 1.

Si $S \equiv S + Y_1$ y S de la etapa anterior es vacío, entonces $S \equiv Y_1$. - Como Y_1 representa al grupo π_1, la formación de estados se hará de acuerdo a los diferentes grupos o cúmulos π_1. El número máximo de elementos del grupo π_1 estará dado por max(1) = n-m+1, y los grupos restantes serán de un sólo elemento (ya que no pueden ser vacíos).

El número mínimo de elementos del grupo π_1, estará dado por min(1), que se define como sigue:

Si n es múltiplo de m entonces
min (1) = n/m

Si n no es múltiplo de m entonces

$$\min(1) = \begin{cases} [n/m] + 1 & \text{para } 1 \leq n-m[n/m] \\ n-(m-1)\,[n/m] & \text{para } n-m[n/m] < 1 \leq m \end{cases}$$

donde [n/m] es el máximo entero menor o igual a n/m.

Por lo tanto, los estados para esta etapa son los grupos que se obtienen de las combinaciones de los n elementos a agrupar tomados de j en j $\forall$ j = min (1),...,max (1). Es decir, son los grupos de cardinalidad entre min (1) y max (1).

El número total de estados NS(1) en la etapa 1 es entonces:

$$NS(1) = \sum_{j=\min(1)}^{\max(1)} \binom{n}{j}$$

Etapa k.

Se tiene $S \equiv S + Y_k$ lo que significa que los estados de la etapa k están formados por los elementos de k-1 grupos o cúmulos más los elementos del k-ésimo grupo.

El número máximo de elementos en cualquier estado de la etapa k es max (k) = n-m+k, y el número mínimo de elementos para los estados de la etapa k es:

Si n es múltiplo de m $\min(k) = k(n/m)$ o bien, si n no es múltiplo de m, se tiene:

$$\min(k) = \begin{cases} ([n/m]+1)k & \text{para } 1 \leq k \leq n-m\,[n/m] \\ n-(m-k)\,[n/m] & \text{para } n-m[n/m] < k \leq m \end{cases}$$

Por lo tanto, los estados para la etapa k se forman de las combinaciones de los n elementos a agrupar tomados de j en j

$\blacktriangleright$ $j = \min(k),\ldots,\max(k)$.

El número total de estados disponibles en la etapa k es:

$$NS(k) = \begin{cases} 1 & \text{Si } k=0 \\ \sum_{j=\min(k)}^{\max(k)} \binom{n}{j} & k=1,2,\ldots,m_0 \end{cases}$$

se puede observar que esta fórmula es general, para todas las etapas del algoritmo.

Etapa m_0.

En la sección 5.2.2.2 se mencionó que en la etapa final m_0 del algo

ritmo se tiene sólo un estado de n elementos. Esto se probará usando las fórmulas de la etapa anterior.

Para $k = m_0$ si n es múltiplo de m

$$m_0 = m \Rightarrow \max(m_0) = n-m+m$$

$$\Rightarrow \max(m_0) = n$$

el número máximo de elementos de algún estado de la etapa m_0 es n.

$\min(m_0) = m(n/m) = n$,

Lo que significa que el número mínimo y el máximo de elementos en cualquier estado de la etapa m_0 es n.

El número de estados disponibles en la etapa m_0 es entonces:

$$NS(m_0) = \sum_{j=n}^{n} \binom{n}{j} = \binom{n}{n} = 1$$

Por lo tanto, cuando n es múltiplo de m, se tiene un estado de n elementos (los n elementos a agrupar).

Para el caso en que n no es múltiplo de m, se tiene que $m_0 = m$.

Por lo tanto, el número máximo de elementos en algún estado de la etapa m_0 es: $\max(m_0)=n-m+m=n$ y el mínimo es: $\min(m_0)=n-(m-m)[n/m]=n$ y se vuelve a cumplir que en la etapa final del algoritmo se llega a un sólo estado de n elementos.

El número total de estados en este algoritmo, es entonces:

$$\sum_{k=1}^{m_0} NS(k)$$

5.2.2.4 Relación Entre Estados de Diferentes Etapas

A las diferentes maneras de formar un estado en la etapa k se les da el nombre de arcos factibles.

Un arco factible conecta un estado de la etapa k-1 a otro de la etapa k, si los elementos del estado de la etapa k están también contenidos en el estado de la etapa k-1. Esto significa que no puede existir un arco factible entre un estado de la etapa k-1 y otro en la etapa k, si los elementos contenidos en el estado de la etapa k-1 no están contenidos en el estado de la etapa k para $2 \leq k \leq m$.

En el algoritmo de programación Dinámica, el número total de arcos factibles es:

$$NA = NS(1) + \sum_{k=1}^{m_o-1} NA(k)$$

donde NA(k) representa el número total de arcos factibles entre la etapa k y k+1 para $k=1,2,\ldots,m_o$. El valor de NA(k) está dado por:

$$NA(k) = \sum_{i=\min(k)}^{\max(k)} \sum_{j=1}^{\max(k+1)-\min(k)} A(i,j)$$

donde

$$A(i,j) = \begin{cases} \binom{n}{i}\binom{n-i}{j} & \text{si } i \neq j \\ \frac{1}{2}\binom{n}{i}\binom{n-i}{j} & \text{si } i = j \\ 0 & \text{d.o.f.} \end{cases} \quad y \quad \begin{cases} \min(k+1) \leq i+j \leq \max(k+1) \\ (m-k)\, j+i \geq n \\ i \geq j \end{cases}$$

5.2.3 Aplicación al Modelo de Jensen

Planteamiento del Problema:

Se pretende agrupar un conjunto de 6 elementos en 3 subconjuntos - llamados cúmulos, cuando la medida de asociación entre los elementos es la distancia Euclidiana y el criterio de optimalidad es minimizar la suma de varianzas de los elementos de un mismo grupo (criterio definido - en la sección 4.1), que como ya se mencionó es la siguiente:

$$g(y) = \sum_{k=1}^{m} \frac{1}{2n_k} \sum_{i=1}^{nk} \sum_{j=1}^{nk} d^2(X_i, X_j)$$

donde $d^2(X_i, X_j) = (X_i - X_j)^T (X_i - X_j)$

Desarrollo del Problema con el Modelo de Jensen:

En este caso $n=2m$ y como $m=3$ se tendrá que $m_0=3$, lo que significa que el algoritmo consta de 3 etapas.

Se calcula el máximo y el mínimo número de elementos para cada etapa, como sigue:

max (1) = 4
min (1) = 2
max (2) = 5
min (2) = 4
max (3) = 6
min (3) = 6

El número total de estados en las etapas 0, 1, 2 y 3 está dado por:

$NS(0) = 1$

$$NS(1) = \sum_{i=2}^{4} \binom{6}{i} = 50$$

$$NS(2) = \sum_{i=4}^{5} \binom{6}{i} = 21$$

$$NS(3) = \binom{6}{6} = 1$$

El número total de estado es 75 y éstos se listan en la Tabla (2).

El número total de arcos factibles es en cada etapa:

Para la Etapa 1

$$NA(1) = \sum_{i=2}^{4} \sum_{j=1}^{3} \binom{6}{i} \binom{6-i}{j} \qquad \begin{array}{r} 4 \le i + j \le 5 \\ 2j + i \ge 6 \\ i \ge j \end{array}$$

$$= \frac{1}{2} \binom{6}{2} \binom{4}{2} + \binom{6}{3} \binom{3}{2} + \binom{6}{4} \binom{2}{1} = 135$$

$NA(1) = 135$

Para la Etapa 2

$$NA(2) = \sum_{i=4}^{6} \sum_{j=1}^{5} \binom{6}{i} \binom{6-i}{j} \qquad \begin{array}{r} i + j = 6 \\ 2j + i \ge 6 \\ i \ge j \end{array}$$

$$= \binom{6}{4} \binom{2}{2} + \binom{6}{5} \binom{1}{1} = 21$$

$NA(2) = 21$

ESTADOS POR ETAPA PARA UN PROBLEMA DE CUMULOS

(n=6, m=3)

ETAPA 0	ETAPA 1	ETAPA 2	ETAPA 3
1. ( )	1. (1,2,3,4) 2. (1,2,3,5) 3. (1,2,5,4) 4. (1,5,3,4) 5. (5,2,3,4) 6. (1,2,3,6) 7. (1,2,6,4) 8. (1,6,3,4) 9. (6,2,3,4) 10. (1,2,5,6) 11. (1,5,6,4) 12. (5,6,3,4) 13. (5,2,3,4) 14. (1,5,3,6) 15. (5,2,6,4) 16. (1,2,3) 17. (1,2,4) 18. (1,2,5) 19. (1,2,6) 20. (1,3,4) 21. (1,3,5) 22. (1,3,6) 23. (1,4,5) 24. (1,4,6) 25. (1,5,6) 26. (2,3,4) 27. (2,3,5) 28. (2,3,6) 29. (2,4,5) 30. (2,4,6) 31. (2,5,6) 32. (3,4,5) 33. (3,4,6) 34. (3,5,6) 35. (4,5,6) 36. (1,2) 37. (1,3) 38. (1,4) 39. (1,5) 40. (1,6) 41. (2,3) 42. (2,4) 43. (2,5) 44. (2,6) 45. (3,4) 46. (3,5) 47. (3,6) 48. (4,5) 49. (4,6) 50. (5,6)	1. (1,2,3,4,5) 2. (1,2,3,4,6) 3. (1,2,3,5,6) 4. (1,2,4,5,6) 5. (1,3,4,5,6) 6. (2,3,4,5,6) 7. (1,2,3,4) 8. (1,2,3,5) 9. (1,2,3,6) 10. (1,2,4,5) 11. (1,2,4,6) 12. (1,2,5,6) 13. (1,3,4,5) 14. (1,3,4,6) 15. (1,3,5,6) 16. (1,4,5,6) 17. (2,3,4,5) 18. (2,3,4,6) 19. (2,3,5,6) 20. (2,4,5,6) 21. (3,4,5,6)	1. (1,2,3,4,5,6)

TABLA 2

Por lo tanto, el número total de arcos factibles del problema es:

$$NA = NS(1) + \sum_{k=1}^{2} NA(k) = 50 + 135 + 21 = 206$$

$$\Longrightarrow \quad NA = 206$$

Para cada uno de los 206 arcos factibles se tendrá que evaluar la - función de costo transicional $F(Y_k)$ del grupo π_k en la etapa k.

Si se tiene que los 6 elementos se definen en un espacio de atributos o características con p=2 se tiene:

$$X_1 = (1,1) \qquad X_4 = (4,4)$$
$$X_2 = (3,4) \qquad X_5 = (1,2)$$
$$X_3 = (5,5) \qquad X_6 = (5,6)$$

o bien:

$$X = \begin{pmatrix} 1 & 3 & 5 & 4 & 1 & 5 \\ 1 & 4 & 5 & 4 & 2 & 6 \end{pmatrix}$$

la matriz de Distancia cuadrada es:

$$D = \begin{bmatrix} 0 & 13 & 32 & 18 & 1 & 41 \\ & 0 & 5 & 1 & 8 & 8 \\ & & 0 & 2 & 25 & 1 \\ & & & 0 & 13 & 5 \\ & & & & 0 & 32 \end{bmatrix}$$

De acuerdo al algoritmo de Programación Dinámica, se tiene:

Etapa 0: $\quad g^*_0(S) = 0$

Etapa 1: calcular $g_1(S, Y_1) = F(Y_1) + g_0^*(S-Y_1)$

En la etapa 1 $S = Y_1$

entonces $g_1(S, Y_1) = F(Y_1) = F(S)$

$$g_1^*(S) = \min_{Y_1} \{F(S)\} = F(S)$$

Por ejemplo:

$$g_1^*(1,2,3,4) = F(1,2,3,4)$$

$$= \frac{d_{12}^2 + d_{13}^2 + d_{14}^2 + d_{23}^2 + d_{24}^2 + d_{34}^2}{4}$$

$$= 17.75$$

Se calculan los 50 valores de los 50 estados de la etapa 1 (ver tabla (2)).

Esto se representa en la tabla (4).

Etapa 2: calcular $g_2(S, Y_2) = F(Y_2) + g_1^*(S-Y_2)$

para cada conjunto de elementos de la etapa 2. Es decir, para cada estado de la etapa 2 descrito en la tabla (2).

Un ejemplo de esto se tiene en la tabla (3).

TABLA (3)

EVALUACION DE LA FUNCION OBJETIVO (PARA n=6 Y m=3) EN LA ETAPA 2

$$g_2(\{1,2,3,4,5\},\ Y_k) = F(Y_2) + g_1^*(\{1,2,3,4,5\} - Y_2)$$

Para diferentes valores de Y_2, se tiene:

$$g_2(\{1,2,3,4,5\},\ \{5\}\ ) = F(5) + g_1^*(1,2,3,4)$$

$$g_2(\{1,2,3,4,5\},\ \{4\}\ ) = F(4) + g_1^*(1,2,3,5)$$

$$g_2(\{1,2,3,4,5\},\ \{3\}\ ) = F(3) + g_1^*(1,2,4,5)$$

$$g_2(\{1,2,3,4,5\},\ \{2\}\ ) = F(2) + g_1^*(1,3,4,5)$$

$$g_2(\{1,2,3,4,5\},\ \{1\}\ ) = F(1) + g_1^*(2,3,4,5)$$

$$g_2(\{1,2,3,4,5\},\ \{1,2\}\ ) = F(1,2) + g_1^*(3,4,5)$$

$$g_2(\{1,2,3,4,5\},\ \{1,3\}\ ) = F(1,3) + g_1^*(2,4,5)$$

$$g_2(\{1,2,3,4,5\},\ \{1,4\}\ ) = F(1,4) + g_1^*(2,3,5)$$

$$g_2(\{1,2,3,4,5\},\ \{1,5\}\ ) = F(1,5) + g_1^*(2,3,4)$$

$$g_2(\{1,2,3,4,5\},\ \{2,3\}\ ) = F(2,3) + g_1^*(1,4,5)$$

$$g_2(\{1,2,3,4,5\},\ \{2,4\}\ ) = F(2,4) + g_1^*(1,3,5)$$

$$g_2(\{1,2,3,4,5\},\ \{2,5\}\ ) = F(2,5) + g_1^*(1,3,4)$$

$$g_2(\{1,2,3,4,5\},\ \{3,4\}\ ) = F(3,4) + g_1^*(1,2,5)$$

$$g_2(\{1,2,3,4,5\},\ \{3,5\}\ ) = F(3,5) + g_1^*(1,2,4)$$

$$g_2(\{1,2,3,4,5\},\ \{4,5\}\ ) = F(4,5) + g_1^*(1,2,3)$$

TABLA (4)

RESULTADOS DE LA ETAPA 1

S	$g_1^*(S) = F(Y_1)$	Y_1^*
(1,2,3,4)	17.75	(1,2,3,4)
(1,2,3,5)	21.00	(1,2,3,5)
(1,2,3,6)	25.00	(1,2,3,6)
(1,2,4,5)	13.50	(1,2,4,5)
(1,2,4,6)	21.50	(1,2,4,6)
(1,2,5,6)	25.75	(1,2,5,6)
(1,3,4,5)	22.75	(1,3,4,5)
(1,3,4,6)	24.75	(1,3,4,6)
(1,3,5,6)	33.00	(1,3,5,6)
(1,4,5,6)	27.50	(1,4,5,6)
(2,3,4,5)	13.50	(2,3,4,5)
(2,3,4,6)	5.50	(2,3,4,6)
(2,3,5,6)	19.75	(2,3,5,6)
(2,4,5,6)	16.75	(2,4,5,6)
(3,4,5,6)	19.50	(3,4,5,6)
(1,2,3)	16.67	(1,2,3)
(1,2,4)	10.67	(1,2,4)
(1,2,5)	7.33	(1,2,5)
(1,2,6)	20.67	(1,2,6)
(1,3,4)	17.33	(1,3,4)
(1,3,5)	19.33	(1,3,5)
(1,3,6)	24.67	(1,3,6)
(1,4,5)	10.67	(1,4,5)
(1,4,6)	21.33	(1,4,6)
(1,5,6)	24.67	(1,5,6)

...

S	$g_1^*(S) = F(Y_1)$	Y_1^*
(2,3,4)	2.67	(2,3,4)
(2,3,5)	12.67	(2,3,5)
(2,3,6)	4.67	(2,3,6)
(2,4,5)	7.33	(2,4,5)
(2,4,6)	4.67	(2,4,6)
(2,5,6)	16.00	(2,5,6)
(3,4,5)	13.33	(3,4,5)
(3,4,6)	2.67	(3,4,6)
(3,5,6)	19.33	(3,5,6)
(4,5,6)	16.67	(4,5,6)
(1,2)	6.50	(1,2)
(1,3)	16.00	(1,3)
(1,4)	9.00	(1,4)
(1,5)	0.50	(1,5)
(1,6)	20.50	(1,6)
(2,3)	2.50	(2,3)
(2,4)	0.50	(2,4)
(2,5)	4.00	(2,5)
(2,6)	4.00	(2,6)
(3,4)	0.50	(3,4)
(3,5)	12.50	(3,5)
(3,6)	0.50	(3,6)
(4,5)	6.50	(4,5)
(4,6)	2.50	(4,6)
(5,6)	16.00	(5,6)

Entonces se calcula para los 21 estados de la etapa 2.

$$g_2^*(1,2,3,4,5) = \min_{Y_2} \{g_2(\{1,2,3,4,5\}, Y_2)\}$$

Esto se presenta en la Tabla (5).

Etapa 3: calcular $g_3(S,Y_3) = F(Y_3) + g_2^*(S-Y_3)$ y $g_3^*(S) = \min_{Y_3} \{g_3(S,Y_3)\}$ para cada conjunto de elementos (estados de la etapa 3). Para el ejemplo $m_o=m=3$ es la última etapa $S=(1,2,3,4,5,6)$. Hay 21 arcos factibles entre los estados de la etapa 2 y los de la etapa 3. Por lo tanto se eligirá el mínimo de 21 valores

Es decir:

$$g_3^*(1,2,3,4,5,6) = \min_{Y_3} \{g_3(\{1,2,3,4,5,6\}, Y_3)\}$$

$$= \min_{Y_3} \{F(Y_3) + g_2^*(\{1,2,3,4,5,6\} - Y_3)\}$$

En un procedimiento de programación dinámica la forma general de representar la última etapa se da en la Tabla (6).

S \ Y_{m_o}	$g_{m_o}(S,Y_{m_o}) = F(Y_{m_o}) + g^*_{m_o-1}(S-Y_{m_o})$		$Y^*_{m_o}$
	$(X_1)\ldots(X_n)\ldots(X_1,X_2)\ldots(X_1,X_n)\ldots$	$g_{m_o}(S)$	
$(X_1,X_2,\ldots,X_n)$			

Tabla (6) Ultima Etapa (m_o), de un problema de Programación Dinámica.

TABLA (5). RESULTADOS DE LA ETAPA 2

Y_2	$g_2(S, Y_2) = F(Y_2) + g_1^*(S - Y_2)$																					$g_2^*(S)$	Y_2^*
S	(1)	(2)	(3)	(4)	(5)	(6)	(12)	(13)	(14)	(15)	(16)	(23)	(24)	(25)	(26)	(34)	(35)	(36)	(45)	(46)	(56)		
(1,2,3,4,5)	13.50	22.75	13.50	21.00	17.75	-	19.83	23.33	21.67	3.17	-	13.17	19.83	21.33	-	7.83	23.17	-	23.17	-	-	3.17	(15)
(1,2,3,4,6)	5.50	24.75	21.50	25.00	-	17.75	9.17	20.67	13.67	-	23.17	23.83	25.17	-	21.33	21.17	-	11.17	-	19.17	-	5.50	(1)
(1,2,3,5,6)	19.75	33.00	25.75	-	25.00	21.00	25.83	32.00	-	5.17	33.17	27.17	-	28.67	23.33	-	33.17	7.83	-	-	32.67	5.17	(15)
(1,2,4,5,6)	16.75	27.50	-	25.75	21.50	13.50	23.17	-	25.00	5.17	27.83	-	25.17	25.33	14.67	-	-	-	27.17	9.83	26.67	5.17	(15)
(1,3,4,5,6)	19.50	-	27.50	33.00	24.75	22.75	-	32.67	28.33	3.17	33.83	-	-	-	-	25.17	33.83	11.17	31.17	21.83	33.33	3.17	(15)
(2,3,4,5,6)	-	19.50	16.75	19.75	5.50	13.50	-	-	-	-	-	19.17	19.83	6.67	17.33	16.50	17.17	7.83	11.17	15.17	18.67	5.50	(5)
(1,2,3,4)	2.67	17.33	10.67	16.67	-	-	7.00	16.50	11.50	-	-	11.50	16.50	-	-	7.00	-	-	-	-	-	2.67	(1)
(1,2,3,5)	12.67	19.33	7.33	-	16.67	-	19.00	20.00	-	3.00	-	3.00	-	20.00	-	-	19.00	-	-	-	-	3.00	$(^{15}_{23})$
(1,2,3,6)	4.67	24.67	20.67	-	-	16.67	7.00	20.00	-	-	23.00	23.00	-	-	20.00	-	-	7.00	-	-	-	4.67	(1)
(1,2,4,5)	7.33	10.67	-	7.33	10.67	-	13.00	-	13.00	1.00	-	-	1.00	13.00	-	-	-	-	13.00	-	-	1.00	$(^{15}_{24})$
(1,2,4,6)	4.67	21.33	-	20.67	-	10.67	9.00	-	13.00	-	21.00	-	21.00	-	13.00	-	-	-	-	9.00	-	4.67	(1)
(1,2,5,6)	16.00	24.67	-	-	20.67	7.33	22.50	-	-	4.50	24.50	-	-	24.50	4.50	-	-	-	-	-	22.50	4.50	$(^{15}_{26})$
(1,3,4,5)	13.33	-	10.67	19.33	17.33	-	-	22.50	21.50	1.00	-	-	-	-	-	1.00	21.50	-	22.50	-	-	1.00	$(^{15}_{34})$
(1,3,4,6)	2.67	-	21.33	24.67	-	17.33	-	18.50	9.50	-	21.00	-	-	-	-	21.00	-	9.50	-	18.50	-	2.67	(1)
(1,3,5,6)	19.33	-	24.67	-	24.67	19.33	-	32.00	-	1.00	33.00	-	-	-	-	-	33.00	1.00	-	-	32.00	1.00	$(^{15}_{36})$
(1,4,5,6)	16.67	-	-	24.67	21.33	10.67	-	-	25.00	3.00	27.00	-	-	-	-	-	-	-	27.00	3.00	25.00	3.00	$(^{15}_{46})$
(2,3,4,5)	-	13.33	7.33	12.67	2.67	-	-	-	-	-	-	9.00	13.00	4.50	-	4.50	13.00	-	9.00	-	-	4.50	$(^{25}_{34})$
(2,3,4,6)	-	2.67	4.67	4.67	-	2.67	-	-	-	-	-	5.00	1.00	-	4.50	4.50	-	1.00	-	5.00	-	1.00	$(^{24}_{36})$
(2,3,5,6)	-	19.33	16.00	-	4.67	12.67	-	-	-	-	-	18.50	-	4.50	16.50	-	16.50	4.50	-	-	18.50	4.50	$(^{25}_{36})$
(2,4,5,6)	-	16.67	-	16.00	4.67	7.33	-	-	-	-	-	-	16.50	6.50	10.50	-	-	-	10.50	6.50	16.50	4.67	(5)
(3,4,5,6)	-	-	16.67	19.33	2.67	13.33	-	-	-	-	-	-	-	-	-	16.50	15.00	7.00	7.00	15.00	16.50	2.67	(5)

E T A P A 3 (Final)

Y_3 / S	$g_3(S, Y_3) = F(Y_3) + g_2^*(S - Y_3)$																					$g_3^*(S)$	Y_3
	(1)	(2)	(3)	(4)	(5)	(6)	(12)	(13)	(14)	(15)	(16)	(23)	(24)	(25)	(26)	(34)	(35)	(36)	(45)	(46)	(56)		
(1,2,3,4,5,6)	5.50	3.17	5.17	5.17	5.50	3.17	9.17	20.67	13.50	1.50	25.00	5.50	1.50	6.67	5.00	5.00	17.17	1.50	11.17	5.50	18.67	1.50	(15) (24) (36)

TABLA (7)

En la Tabla (7) se tiene el resultado final del ejemplo.

Que en resumen se tiene:

Si $Y_3^*=(X_3,X_6)$ $Y_2^*=(X_2,X_4)$ y $Y_1^*=(X_1,X_5)$ o sea $\P_3=\{I_3,I_6\}$, $\P_2=\{I_2,I_4\}$ y $\P_1=\{I_1,I_5\}$ o bien se puede variar el orden de los grupos pero con el mismo resultado.

5.2.4 Comparación del Modelo de Jensen con el de Enumeración Exhaustiva

Tomando el ejemplo de la sección 5.2.2 en el que se pretende agrupar 6 elementos en 3 cúmulos, se tiene que el número total de particiones o formas de agrupar los 6 elementos en 3 cúmulos (ver 5.1) es:

$$S(6,3) = 1/3! \sum_{k=0}^{3} (-1)^k \binom{3}{k} (3-k)^6 = 90$$

Las 90 alternativas de agrupación o particiones se muestran en las tablas (8), (9) y (10).

Donde, si n_k es el número de elementos del grupo $\P_k$, para k=1,2,3. En la Tabla (8) $n_1=4$, $n_2=1$, $n_3=1$; en la Tabla (9) $n_1=3$, $n_2=2$, $n_3=1$ y en la Tabla (10) $n_1=2$, $n_2=2$ y $n_3=2$. Bajo Enumeración Exhaustiva la función objetivo g(y) necesita ser evaluada para cada una de las 90 alternativas de agrupación, y de éstas se elegirá la alternativa de agrupación (partición) para la que g(y) es mínima.

(1,2,3,4), (5), (6)	(6,2,3,4), (5), (1)
(1,2,3,5), (4), (6)	(1,2,5,6), (3), (4)
(1,2,5,4), (3), (6)	(1,5,6,4), (2), (3)
(1,3,5,4), (2), (6)	(5,6,3,4), (1), (2)
(5,2,3,4), (1), (6)	(5,2,3,6), (1), (4)
(1,2,3,6), (5), (4)	(1,5,3,6), (2), (4)
(1,2,6,4), (5), (3)	(5,2,6,4), (1), (3)
(1,6,3,4), (5), (2)	

Tabla (8). Particiones con $n_1 = 4$, $n_2 = 1$, $n_3 = 1$

(1,2,3), (4,5), (6)	(1,4,5), (2,3), (6)	(3,4,5), (1,6), (2)
(1,2,3), (4,6), (5)	(1,4,5), (2,6), (3)	(3,4,5), (2,6), (1)
(1,2,3), (5,6), (4)	(1,4,5), (3,6), (2)	(3,4,6), (1,2), (5)
(1,2,4), (3,5), (6)	(1,4,6), (2,3), (5)	(3,4,6), (1,5), (2)
(1,2,4), (3,6), (5)	(1,4,6), (2,5), (3)	(3,4,6), (2,5), (1)
(1,2,4), (5,6), (3)	(1,4,6), (3,5), (2)	(3,5,6), (1,2), (4)
(1,2,5), (4,3), (6)	(1,5,6), (2,3), (4)	(2,5,6), (1,4), (2)
(1,2,5), (4,6), (3)	(1,5,6), (2,4), (3)	(3,5,6), (2,4), (1)
(1,2,5), (6,3), (4)	(1,5,6), (3,4), (2)	(4,5,6), (1,2), (3)
(1,2,6), (4,5), (3)	(2,4,5), (1,3), (6)	(4,5,6), (1,3), (2)
(1,2,6), (3,5), (4)	(2,4,5), (1,6), (3)	(4,5,6), (2,3), (1)
(1,2,6), (3,4), (5)	(2,4,5), (3,6), (1)	(1,6,3), (2,5), (4)
(1,4,3), (2,5), (6)	(2,4,6), (1,3), (5)	(4,2,3), (1,5), (6)
(1,4,3), (2,6), (5)	(2,4,6), (1,5), (3)	(4,2,3), (1,6), (5)
(1,4,3), (5,6), (2)	(2,4,6), (3,5), (1)	(4,2,3), (5,6), (1)
(1,5,3), (4,2), (6)	(2,5,6), (1,3), (4)	(5,2,3), (4,1), (6)
(1,5,3), (4,6), (2)	(2,5,6), (1,4), (3)	(5,2,3), (4,6), (1)
(1,5,3), (2,6), (4)	(2,5,6), (3,4), (1)	(5,2,3), (1,6), (4)
(1,6,3), (4,5), (2)	(3,4,5), (1,2), (6)	(6,2,3), (4,5), (1)
(1,6,3), (4,2), (5)		(6,2,3), (4,1), (5)
		(6,2,3), (1,5), (4)

Tabla (9). Particiones con $n_1 = 3$, $n_2 = 2$, $n_3 = 1$

(1,2), (3,4), (5,6)	(1,4), (2,6), (3,5)
(1,2), (3,5), (4,6)	(1,5), (3,4), (2,6)
(1,2), (3,6), (4,5)	(1,5), (3,2), (4,6)
(1,3), (2,4), (5,6)	(1,5), (3,6), (2,4)
(1,3), (2,5), (4,6)	(1,6), (3,4), (5,2)
(1,3), (2,6), (4,5)	(1,6), (3,5), (4,2)
(1,4), (2,3), (5,6)	(1,6), (3,2), (4,5)
(1,4), (2,5), (3,6)	

Tabla (10). Particiones con $n_1 = n_2 = n_3 = 2$

Una observación sobre la lista de alternativas es que por Enumeración Exhaustiva la función objetivo g(y) se calculará más de una vez para alguno de los cúmulos, por ejemplo el cúmulo (1,2,3).

Si, como ya se definió en la sección anterior

$$F(Y_k) = \frac{1}{2n_k} \sum_{i=1}^{nk} \sum_{j=1}^{nk} d^2(X_i, X_j)$$

donde Y_k representa al grupo $\P_k$ de n_k elementos.

$F(Y_k)$ proporciona el cálculo transicional para cada grupo $\P_k$ de una alternativa dada.

El método de Enumeración Exhaustiva hace 3 cálculos transicionales para cada una de las 90 alternativas, resultando un total de 270 cálculos transicionales. El método de Programación Dinámica envuelve 206 - cálculos transicionales (uno para cada arco factible, ver sección anterior), ahorrándose 64 cálculos.

5.3 Problema de Cúmulos Definido como un Problema de Programación Matemática

En el problema de cúmulos es necesario asignar los elementos a grupos o cúmulos en números enteros, por lo que en el problema de cúmulos básicamente maneja variables enteras.

A continuación se plantea el problema general de programación entera como:

$$\max g(y) \quad \text{donde} \quad y \in S \subseteq Z^n \subseteq R^n$$

donde Z^n es el conjunto n-dimensional de vectores enteros y g es una función de vectores enteros definidos en S. El conjunto S es llamado conjunto de restricciones y g es llamada función objetivo.

Cada $y \in S$ se le da el nombre de solución factible del problema. Si existe $y^* \in S$ que satisfaga

$$g(y) \leq g(y^*) < \infty \quad \forall \; y \in S$$

entonces y^* es la solución óptima del problema.

El objetivo en los problemas de programación matemática es establecer si la solución óptima existe, encontrarla, o bien, encontrar todas las soluciones óptimas, si hay más de una. Como se ha visto, el problema de cúmulos puede plantearse de diferentes formas; dependiendo de la función objetivo o criterio y de las restricciones.

En esta sección se estructura el problema de cúmulos en base al modelo de programación lineal entera, para lo que se tienen las siguientes definiciones:

Problema de Programación Lineal Entera

Forma general:

max Cy

y S = {y | Ay = b, y ≥ 0 entero}

Forma estándar:

$$\max \sum_{j=1}^{n} C_j y_j$$

$$\text{s.a.} \quad \sum_{j=1}^{n} \alpha_{ij} y_j = b_i \quad i=1,\ldots,R$$

$$y_j \geq 0 \quad \text{entera} \quad j=1,\ldots,n$$

Donde, A es una matriz de R X n, b es un vector de R componentes, C es un vector de n componentes, y 0 es un vector de n ceros. Se asume que las entradas (o elementos) de A, b y C son siempre enteros.

Un caso especial del modelo de programación lineal entera es el que tiene variables binarias, al que se le añade la restricción de $X_j \leq 1$ - $j=1,\ldots,n$, o bien, en vez de $X \geq 0$ se tiene $X = 0,1$.

Como se explicó en el capítulo 2, el problema de cúmulos es equiva-

lente a un problema de partición, que se puede plantear como un problema de programación lineal entera con variables binarias de la siguiente manera:

Considere el conjunto $I=\{I_1, I_2, \ldots, I_n\}$ y un conjunto $\pi=\{\pi_1, \pi_2, \ldots, \pi_\beta\}$ $(\beta=2^n - 1)$ donde $\pi_j \subseteq I \quad j \in J=\{1,2,\ldots,\beta\}$ un subconjunto $J^* \subseteq J$ define una partición de I si .

$$\bigcup_{j \in J^*} \pi_j = I$$

y $\quad j, k \in J^*, \quad j \neq k \Rightarrow \pi_j \cap \pi_k = \phi$

Sea $C_j > 0$ el costo asociado a cada $j \in J$, el costo total de la partición definida por J^* está dado por $\sum_{j \in J^*} C_j$.

Por lo tanto, el problema de cúmulos puede plantearse como uno de programación lineal entera, de la siguiente manera:

$$\min \sum_{j=1}^{\beta} C_j \, y_j$$

$$\text{s.a.} \sum_{j=1}^{\beta} \alpha_{ij} \, y_j = 1 \quad i=1,2,\ldots,n$$

$$\sum_{j=1}^{\beta} y_j = m$$

$$y_j = 0,1 \qquad j=1,2,\ldots,\beta$$

donde

$$\alpha_{ij} = \begin{cases} 1 & \text{Si } I_i \in \pi_j \qquad i=1,2,\ldots,n \\ 0 & \text{d.o.f.} \qquad j=1,2,\ldots,\beta \end{cases}$$

$$y_j = \begin{cases} 1 & j \in J^* \\ 0 & \text{d.o.f.} \end{cases}$$

$$\beta = \binom{n}{(n-(m-1))} + \binom{n}{(n-m)} + \binom{n}{(n-(m+1))} + \ldots + \binom{n}{(n-(n-1))}$$

β es el número total de subconjuntos de I, considerando que se desea m en cada partición; m es el número de grupos o cúmulos requeridos para cada partición.

El costo para cada grupo $\P_j$, $j=1,2,\ldots,\beta$ puede definirse utilizando una función que mida el grado de disimilitud entre los elementos de dicho grupo, como las vistas en el capítulo 3.

En particular, el problema se puede definir como sigue:

$$C_j = \sum_{i=1}^{n} d^2 (X_i, \bar{X}^j)\ \alpha_{ij}$$

donde X_i es el vector de atributo asociado al elemento I_i $\forall$ $i=1,2,\ldots,n$

$\bar{X}^j$ es la media del grupo $\P_j$ $\forall$ $j=1,2,\ldots,\beta$, que se define:

$$\bar{X}^j = \frac{\sum_{i=1}^{n} X_i\ \alpha_{ij}}{\sum_{i=1}^{n} \alpha_{ij}} = \frac{\sum_{i=1}^{n} X_i\ \alpha_{ij}}{n_j}$$

$n_j = \sum_{i=1}^{n} \alpha_{ij}$ $\forall$ $j=1,2,\ldots,\beta$, es el número total de elementos del grupo $\P_j$.

En la práctica, la manera más fácil de abordar el problema de programación lineal entera es usando el método Simplex (ignorando la restricción de que las variables son enteras) para luego redondear los valores no enteros a enteros en la solución final. Este procedimiento no es siempre adecuado, ya que la solución óptima del problema de programación lineal, no necesariamente es factible después del redondeo. Otra dificultad es que aunque la solución óptima del problema de programación lineal sea factible después del redondeo, no se puede garantizar que sea una solución óptima al problema entero.

Por estas razones es de utilidad analizar métodos que obtengan directamente una solución óptima entera.

Un número considerable de algoritmos se han desarrollado con este propósito.

A continuación se expondrán algunos métodos importantes para resolver el problema de programación lineal entera. Se desarrollará una versión del algoritmo conocido como de Ramificación y Límites (o acotamiento) ver sección 5.3.2 para el problema de cúmulos visto como un problema de programación lineal entera (con variables binarias) planteado anteriormente. Además, se tratará el método conocido como de planos cortantes, ver sección 5.3.1.

5.3.1 Método de Planos Cortantes

El método de planos cortantes busca resolver el problema de progra

mación lineal entera, generando una secuencia de desigualdades lineales que "recortan" parte de la región factible del problema correspondiente de programación lineal dejando intacta la región factible del problema - entero. Cuando se han generado los suficientes hiperplanos cortantes, el problema de programación lineal entera tendrá la misma solución óptima - que su problema correspondiente de programación lineal.

Un ejemplo para ilustrar esto es:

$$\max\ 2X_1 + X_2$$
$$X_1 + X_2 \leq 5$$
$$-X_1 + X_2 \leq 0$$
$$6X_1 + 2X_2 \leq 21$$
$$X_1 , X_2 \geq 0 \qquad \text{enteras}$$

la región factible de su problema correspondiente de programación lineal se muestra en la figura (9) (área dentro de las líneas continuas), los - hiperplanos (líneas punteadas) recortan esa área, para dar lugar a la región factible del problema entero, la cual comparten.

Suponga que el conjunto:

$S = \{X \mid AX = b, \quad X \geq 0 \text{ enteras}\}$ de soluciones factibles, está acotado y que por lo tanto tiene un número finito de puntos.

Se define el conjunto S^+ llamado casco convexo de S como:

$$S^+ = \{y \mid y = \sum \alpha_i X_i, \ \alpha_i \geq 0, \ \sum \alpha_i = 1, \ X_i \in S\}$$

REGION FACTIBLE DEL PROBLEMA DE PROGRAMACION LINEAL ENTERA

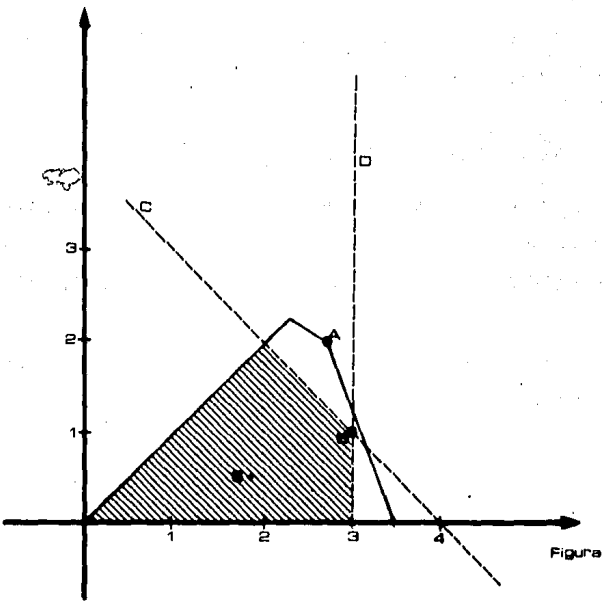


Figura 9

A: Solución óptima por programacion lineal

B: Solución óptima por programacion lineal entera

C: Plano cortante 1

D: Plano cortante 2

donde

$$S \subseteq S^{+} \subseteq T = \{X \mid AX = b,\ X \geq 0\}$$

con T es el conjunto de soluciones factibles del problema correspondiente de programación lineal.

Para el ejemplo anterior, la región compartida es S^{+}. Los dos planos cortantes eliminan $T - S^{+}$ de T.

5.3.2 Método de Ramificación y Límites

Como cualquier problema de programación entera acotado, tiene sólo un número finito de soluciones factibles, es natural considerar el uso de algún método de enumeración para encontrar una solución óptima.

Ramificación y límites es un método de optimización que usa el árbol básico de enumeración.

La idea general del método de ramificación y límites es la siguiente:

Suponga (a ser especificado) que la función objetivo debe ser minimizada. Se considera que el límite superior de la función objetivo está disponible. (Este usualmente es el valor de la función objetivo para la mejor solución factible identificada hasta el momento). El primer paso es dividir el conjunto de todas las soluciones factibles en subconjuntos y obtener el límite inferior de cada uno de ellos, del valor de la función objetivo de las soluciones dentro de cada subconjunto. Aquellos -

subconjuntos cuyos límites inferiores exceden el límite superior del valor actual de la función objetivo, se excluyen de cualquier consideración. (Un subconjunto que se excluye por esta u otra legítima razón se dice que es penetrado). Uno de los subconjuntos restantes, aquel menor límite inferior, se divide entonces en algunos subconjuntos. Se obtienen otra vez sus límites inferiores para excluir algunos de estos de consideración. De todos los subconjuntos restantes se selecciona uno para dividirlo y así sucesivamente. Este proceso se repite una y otra vez hasta que se encuentra una solución factible tal que su valor correspondiente de la función objetivo no sea mayor que el límite inferior de algún subconjunto.

En resumen, ramificación y límites es un método de optimización que usa enumeración en forma de "árbol", y requiere el cálculo de los límites superiores e inferiores de la función objetivo, para acelerar el proceso de penetración y así cortar la enumeración.

5.3.3 Algoritmo General de Ramificación y Límites

Paso Inicial:

Se comienza con un conjunto de soluciones iniciales (incluyendo soluciones no factibles) como el único "subconjunto restante". Sea Z_u el límite superior del valor de la función objetivo.

Paso de Ramificación:

Usar una regla de ramificación para seleccionar uno de los subconjuntos restantes. (Aquellos no penetrados ni divididos) y dividirlo en dos o más subconjuntos nuevos de soluciones.

Paso de Límite:

Para cada nuevo subconjunto se obtiene su límite inferior Z_L del valor de la función objetivo para las soluciones factibles del subconjunto.

Paso de Penetración:

Cada nuevo subconjunto es excluído de otra consideración en base a las siguientes pruebas:

1) $Z_L \geq Z_u$

ó

2) El subconjunto encontrado no tiene soluciones factibles.

ó

3) La mejor solución factible del subconjunto ha sido definida (de manera que Z_L corresponde al valor de la función objetivo); si $Z_L < Z_u$ entonces se reajusta $Z_u = Z_L$, se guarda esta solución como la solución titular y se repite la prueba 1) para todos los subconjuntos restantes.

Regla de Terminación:

El procedimiento se detiene cuando no quedan subconjuntos (sin penetrar). La solución titular actual es óptima. Si no hay solución titular (es decir, si Z_u sigue teniendo el valor inicial), entonces el problema no tiene soluciones factibles. De otra forma, regresa al Paso de Ramificación.

Si el objetivo es maximizar en vez de minimizar la función objetivo, el procedimiento no cambia sólo se intercambian los papeles de los límites superiores e inferiores de la función objetivo. Por lo tanto, Z_u se reemplaza por Z_L y viceversa, y la dirección de las desigualdades se cambia.

Reglas de Ramificación:

Las dos reglas de ramificación más conocidas para seleccionar el subconjunto que se va a dividir es la del mejor límite y la del límite nuevo.

La Regla del mejor Límite selecciona el subconjunto que tiene el límite más favorable (límite inferior en caso de minimización) ya que dicho subconjunto podría ser el más prometedor de contener el óptimo.

La Regla del Límite nuevo selecciona al subconjunto creado más recientemente y que no haya sido penetrado. Para los subconjuntos creados al mismo tiempo se elige el del límite más favorable.

5.3.4 Un Algoritmo de Ramificación y Límites para el Problema de Cúmulos

Planteando el problema de cúmulos como un problema de programación lineal entera de variables binarias (sección 5.3). Se desarrolla un algoritmo para su solución basado en el método de Ramificación y Límites - ya analizado.

La forma ya convenida de plantear este problema es:

$$\min \sum_{j=1}^{\beta} C_j \; y_j$$

$$\text{s.a.} \sum_{j=1}^{\beta} \alpha_{ij} \; y_j = 1 \qquad \text{para} \quad i=1,\ldots,n$$

$$\sum_{j=1}^{\beta} y_j = m$$

$$y_j = 0,1 \quad \text{para} \quad j=1,\ldots,\beta$$

Nota: Las especificaciones sobre α_{ij}, y_j y C_j se dan en la sección - 5.3.

Todas las variables se reordenan de acuerdo a sus costos C_j --- $(j=1,2,\ldots,\beta)$ en forma creciente. Es decir,

$$0 \leq C_1 \leq C_2 \ldots, \leq C_\beta$$

Con el problema de esta forma el algoritmo tenderá a hacer las variables cero tanto como las restricciones se lo permitan, pero dando preferencia a las variables iniciales cuando sea necesario dar el valor de 1 a algunas de las variables.

El algoritmo va definiendo subconjuntos de soluciones, asignando valores a las primeras variables formando así una solución a la que se le da el nombre de solución parcial actual $(y_1,y_2,\ldots,y_\alpha)$ (donde, α es el - número de variables asignadas a subconjuntos en actual consideración), - cuya solución complementaria es $(y_1,y_2,\ldots,y_\alpha, y_{\alpha+1},\ldots,y_\beta)$.

Paso Inicial:

Como se puede verificar a simple vista la solución trivial --- $(y_1=y_2=\ldots=y_\beta=0)$ del problema de cúmulo ya planteado, no es factible (que sería la que diera el mínimo costo). El algoritmo se inicia con una

solución parcial de $y_1=1$ (y todas las demás variables igual a cero), - es decir, con $\alpha=1$. Se tiene como límite superior del valor de la función objetivo $Z_u = \infty$.

Paso de Ramificación: (En el caso inicial se sigue al Paso del Límite).

En base a la solución parcial previa $(y_1,y_2,\ldots,y_{\alpha-1})$, ésta se divide en dos nuevos subconjuntos (de soluciones parciales), uno con $y_\alpha=1$ - y otro con $y_\alpha=0$. Para hacer la selección de la nueva solución parcial actual se usa la Regla de Ramificación del Límite nuevo, que se explica en el siguiente paso.

Paso del Límite:

El límite inferior Z_L del valor de la función objetivo de la solución parcial actual $(y_1,y_2,\ldots,y_\alpha)$ es:

$$Z_L = \begin{cases} \sum_{j=1}^{\alpha} C_j y_j & \text{si } y_\alpha = 1 \\ \sum_{j=1}^{\alpha-1} C_j y_j + C_{\alpha+1} & \text{si } y_\alpha = 0 \end{cases}$$

La razón de añadir $C_{\alpha+1}$ si $y_\alpha=0$ es porque el algoritmo calcula este límite sólo si previamente se encuentra que $(y_1,y_2,\ldots,y_{\alpha-1}y_\alpha=0,\ldots, y_\alpha=0)$ no es factible. Esto se presenta mientras se trata de llevar la solución parcial previa $(y_1,y_2,\ldots,y_M)$ donde $M=\max\{j \mid y_j=1\}$ a la prueba 3 del paso de penetración. Por lo tanto, el límite Z_L siempre es menor (o igual) para $y_\alpha=1$ que para $y_\alpha=0$, así que, usando la Regla del

Límite nuevo siempre se seleccionará primero $y_\alpha=1$ como solución parcial actual si ninguna de las dos ha sido penetrada. El motivo de esto es para que el paso de penetración no se aplique a la solución parcial con $y_\alpha=0$ hasta después de que la solución parcial con $y_\alpha=1$ haya sido penetrada (probablemente penetrando todos sus subconjuntos subsecuentes).

Paso de Penetración:

La penetración de una solución parcial se lleva a cabo bajo las siguientes pruebas:

Prueba 1. $(Z_L \geq Z_u)$ es obvia.

Prueba 2. (El subconjunto encontrado no tiene soluciones factibles). Se hace analizando si alguna de las restricciones no se satisface completando la solución parcial.

Por lo tanto, la solución parcial es penetrada si:

$$\sum_{j=1}^{\alpha} \alpha_{ij}\, y_j + \sum_{j=\alpha+1}^{\beta} \max\,(\alpha_{ij},0) < 1$$

para alguna $i=1,\ldots,n$

ó

$$\sum_{j=1}^{\alpha} \alpha_{ij}\, y_j + \sum_{j=\alpha+1}^{\beta} \min\,(\alpha_{ij},0) > 1$$

para alguna $i=1,\ldots,n$

donde

$$\begin{matrix}\max \\ \min\end{matrix}\,(\alpha_{ij},0) = \begin{matrix}\max \\ \min\end{matrix}\,\{\alpha_{ij}\, y_j \mid y_j = 0,1\}$$

ó

$$\sum_{j=1}^{\alpha} y_j + \sum_{j=\alpha+1}^{\beta} \max \{y_j \mid y_j=0,1\} < m$$

ó

$$\sum_{j=1}^{\alpha} y_j + \sum_{j=\alpha+1}^{\beta} \min \{y_j \mid y_j=0,1\} > m$$

Prueba 3. (Se encuentra la mejor solución factible del subconjunto). Se obtiene analizando si la solución correspondiente al límite inferior Z_L del valor de la función objetivo (sea la solución parcial - más $y_{\alpha+1} = 1-y_\alpha$ y el resto de las variables iguales a cero) actualmente es factible.

Por lo tanto, la tercera forma de que la solución parcial actual - sea penetrada es cuando:

$$\sum_{j=1}^{\alpha} a_{ij} y_j + a_{i\alpha+1}(1 - y_\alpha) = 1 \quad \forall\ i=1,\ldots,n$$

y

$$\sum_{j=1}^{\alpha} y_j + (1 - y_\alpha) = m$$

Si esto ocurre y $Z_L < Z_u$, entonces se hace $Z_u = Z_L$ y se guarda esta solución como la solución titular actual.

Para ilustrar el ordenamiento de los pasos anteriores se tiene el - diagrama de flujo mostrado en la figura (10).

DIAGRAMA DE FLUJO PARA RESOLVER UN PROBLEMA DE DUALIDAD POR EL METODO DE RAMIFICACION Y LIMITES

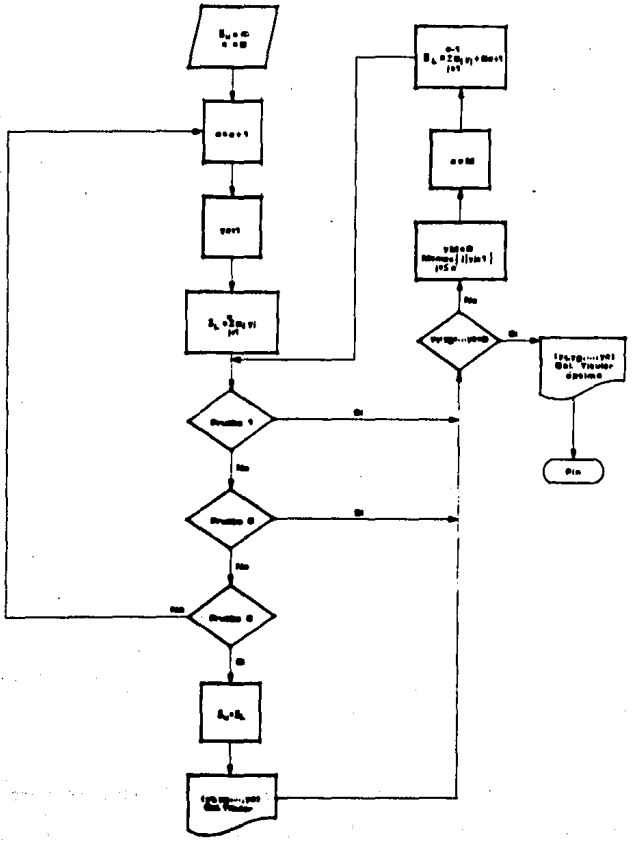


Figura 10

5.3.5 Ejemplificación del Algoritmo de Ramificación y Límites

Con el objeto de ilustrar el algoritmo de Ramificación y Límites, - se usa el ejemplo que se resolvió en la sección anterior, utilizando Programación Dinámica. En dicho ejemplo se tienen n=6 elementos definidos por:

$$X_1 = (1,1) \qquad X_2 = (3,4) \qquad X_3 = (5,5)$$
$$X_4 = (4,4) \qquad X_5 = (1,2) \qquad X_6 = (5,6)$$

El planteamiento de Programación Lineal Entera para resolver este - problema de cúmulos es el siguiente:

$$\min \sum_{j=1}^{\beta} C_j y_j$$

$$\text{s.a.} \sum_{j=1}^{\beta} \alpha_{ij} y_j = 1 \quad \text{para} \quad i=1,\ldots,6$$

$$\sum_{j=1}^{\beta} y_j = 3$$

$$y_j = 0,1 \qquad \text{para} \quad j=1,\ldots,\beta$$

donde $\beta=\binom{6}{4} + \binom{6}{3} + \binom{6}{2} + \binom{6}{1} = 56$ grupos o cúmulos (son los equivalentes a los cúmulos definidos en la Tabla (4) más los 6 grupos de un - sólo elemento).

$$y_j = \begin{cases} 1 & j \in J^* \\ 0 & \text{d.o.f.} \end{cases} \qquad j=1,2,\ldots,56$$

(Si $\P_j$ está dentro de la partición. Ver 5.3).

$$\alpha_{ij} = \begin{cases} 1 & I_i \in \P_j \quad \text{para} \quad i=1,\ldots,6 \\ 0 & \text{d.o.f.} \qquad\qquad\quad j=1,\ldots,56 \end{cases}$$

La matriz $A = \{\alpha_{ij}\}_{6x56}$ para este ejemplo se tiene en la Tabla (11).

Si se define el costo C_j $(j=1,\ldots,56)$ para cada grupo o cúmulo - como:

$$C_j = \sum_{i=1}^{6} d^2 (X_i, \bar{X}^j)\ \alpha_{ij}$$

$$\text{donde } \bar{X}^j = \frac{\sum_{i=1}^{6} X_i\ \alpha_{ij}}{\sum_{i=1}^{6} \alpha_{ij}}$$

se tienen los costos C_j para $j=1,2,\ldots,56$ grupos o cúmulos en la Tabla (12), y de acuerdo a estos se reordenan los grupos o cúmulos en forma creciente, es decir de tal forma que:

$$0 \leq C_1 \leq C_2, \ldots, \leq C_{56}$$

Finalmente, utilizando los valores de α_{ij} y C_j, el planteamiento del problema queda como sigue:

MATRIZ A_{6x56} PARA n=6 ELEMENTOS AGRUPADOS EN m=3 CUMULOS

$$A = \begin{bmatrix}
1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&1&1&1&0&0&0&0&1&0&1&0&0&1&0&1&1&0&1&0&0&1&1&1&1&1&1&1&1&1&1&1&1\\
0&1&0&0&0&0&0&1&0&0&1&0&0&1&1&1&1&1&1&1&0&1&1&0&0&1&0&1&0&1&1&1&0&0&0&1&1&0&1&0&0&0&0&1&0&1&1&0&1&0&0&0&0&1&1&0\\
0&0&1&0&0&0&0&0&1&1&1&0&1&1&0&0&0&1&1&0&0&0&0&0&0&0&0&1&1&1&1&0&0&1&0&0&1&0&1&1&1&1&1&1&1&0&0&1&0&0&1&0&1&1&1&0&0&1\\
0&0&0&1&0&0&0&1&1&0&0&1&1&1&0&0&1&0&1&0&1&0&1&1&1&1&0&0&1&1&1&0&0&0&1&0&1&1&1&0&0&1&0&0&0&0&1&1&1&0&0&1&0&0&1&0\\
0&0&0&0&1&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&1&1&1&0&1&0&1&1&1&1&1&1&0&1&1&0&1&0&0&1&1&1&1&0&0&1&0&0&1&1&0&0&0&1&1&1\\
0&0&0&0&0&1&0&0&0&1&0&1&1&0&0&1&1&1&1&0&0&0&0&0&0&0&0&0&0&1&0&1&1&0&1&0&0&1&0&1&1&1&1&0&1&1&0&1&1&1&1&1&1&1&1&1
\end{bmatrix}$$

TABLA (11)

TABLA (12)

COSTOS C_j PARA CADA CUMULO ORDENADO

GRUPO $\P_j$	COSTO C_j	j
(1,2,3,4)	17.75	39
(1,2,3,5)	21.00	46
(1,2,3,6)	25.00	53
(1,2,4,5)	13.50	31
(1,2,4,6)	21.50	48
(1,2,5,6)	25.75	54
(1,3,4,5)	22.75	49
(1,3,4,6)	24.75	52
(1,3,5,6)	33.00	56
(1,4,5,6)	27.50	55
(2,3,4,5)	13.50	30
(2,3,4,6)	5.50	19
(2,3,5,6)	19.75	43
(2,4,5,6)	16.75	37
(3,4,5,6)	19.50	42
(1,2,3)	16.67	36
(1,2,4)	10.67	26
(1,2,5)	7.33	22
(1,2,6)	20.67	45
(1,3,4)	17.33	38
(1,3,5)	19.33	41
(1,3,6)	24.67	51
(1,4,5)	10.67	25
(1,4,6)	21.33	47
(1,5,6)	24.67	50
(2,3,4)	2.67	14
(2,3,5)	12.67	28

GRUPO $\P_j$	COSTO C_j	j
(2,3,6)	4.67	18
(2,4,5)	7.33	23
(2,4,6)	4.67	17
(2,5,6)	16.00	32
(3,4,5)	13.33	29
(3,4,6)	2.67	13
(3,5,6)	19.33	40
(4,5,6)	16.67	35
(1,2)	6.50	20
(1,3)	16.00	33
(1,4)	9.00	24
(1,5)	0.50	7
(1,6)	20.50	44
(2,3)	2.50	11
(2,4)	0.50	8
(2,5)	4.00	15
(2,6)	4.00	16
(3,4)	0.50	9
(3,5)	12.50	27
(3,6)	0.50	10
(4,5)	6.50	21
(4,6)	2.50	12
(5,6)	16.00	34
(1)	0	1
(2)	0	2
(3)	0	3
(4)	0	4
(5)	0	5
(6)	0	6

$$\min \; 0.5 \sum_{j=7}^{10} y_j + 2.5 \sum_{j=11}^{12} y_j + 2.67 \sum_{j=12}^{14} y_j + 4 \sum_{j=15}^{16} y_j +$$

$$+ \; 4.67 \sum_{j=17}^{18} y_j + 5.5\, y_{19} + 6.5 \sum_{j=20}^{21} y_j + 7.33 \sum_{j=22}^{23} y_j +$$

$$+ \; 9\, y_{24} + 10.67 \sum_{j=25}^{26} y_j + 12.5 \sum_{j=27}^{28} y_j + 13.33\, y_{29} +$$

$$+ 13.5 \sum_{j=30}^{31} y_j + 16 \sum_{j=32}^{34} y_j + 16.67 \sum_{j=35}^{36} y_j + 16.75\, y_{37} + 17.33\, y_{38} +$$

$$+ 17.75\, y_{39} + 19.33 \sum_{j=40}^{41} y_j + 19.5\, y_{42} + 19.75\, y_{43} + 20.5\, y_{44} +$$

$$+ 20.67\, y_{45} + 21\, y_{46} + 21.33\, y_{47} + 21.5\, y_{48} + 22.75\, y_{49} +$$

$$+ 24.67 \sum_{j=50}^{51} y_j + 24.75\, y_{52} + 25\, y_{53} + 25.75\, y_{54} + 27.5\, y_{55} + 33\, y_{56}$$

s.a.

$$y_1 + y_7 + y_{20} + y_{22} + \sum_{j=24}^{26} y_j + y_{31} + y_{33} + y_{36} + y_{38} + y_{39} + y_{41} + \sum_{j=44}^{56} y_j = 1$$

$$y_2 + y_8 + y_{11} + \sum_{j=14}^{20} y_j + y_{22} + y_{23} + y_{26} + y_{28} + \sum_{j=30}^{32} y_j + y_{36} + y_{37} + y_{39} +$$

$$+ \; y_{43} + y_{45} + y_{46} + y_{48} + y_{53} + y_{54} = 1$$

$$y_3 + \sum_{j=9}^{11} y_j + y_{13} + y_{14} + y_{18} + y_{19} + \sum_{j=27}^{30} y_j + y_{33} + y_{36} +$$

$$+ \sum_{j=38}^{43} y_j + y_{46} + y_{49} + \sum_{j=51}^{53} y_j + y_{56} = 1$$

$$y_4 + y_8 + y_9 + \sum_{j=12}^{14} y_j + y_{17} + y_{19} + y_{21} + \sum_{j=23}^{26} y_j + y_{29} +$$

$$+ y_{30} + y_{31} + y_{35} + \sum_{j=37}^{39} y_j + y_{42} + \sum_{j=47}^{49} y_j + y_{52} + y_{55} = 1$$

$$y_5 + y_7 + y_{15} + \sum_{j=21}^{23} y_j + y_{25} + \sum_{j=27}^{32} y_j + y_{34} + y_{35} + y_{37} +$$

$$+ \sum_{j=40}^{43} y_j + y_{46} + y_{49} + y_{50} + \sum_{j=54}^{56} y_j = 1$$

$$y_6 + y_{10} + y_{12} + y_{13} + \sum_{j=16}^{19} y_j + y_{32} + y_{34} + y_{35} + y_{37} + y_{40}$$

$$+ \sum_{j=42}^{45} y_j + y_{47} + y_{48} + \sum_{j=50}^{56} y_j = 1$$

$$\sum_{j=1}^{56} y_j = 3$$

$$y_j = 0,1 \quad \forall\ j=1,\ldots,56$$

Siguiendo el algoritmo de Ramificación y Acotamiento se busca la solución al problema de cúmulos ya planteado como se describió en la sección anterior.

En la Tabla (13) se dan los valores más relevantes para varias iteraciones del algoritmo.

TABLA (13)

ITERACIONES DEL ALGORITMO Y RAMIFICACION Y ACOTAMIENTO

ITERACION	SOLUCION PARCIAL	Z_L	¿PENETRADO?	PRUEBA DE PENETRACION	Z_u	SOLUCION TITULAR
1	(1)	0	NO		∞	-
2	(11)	0	NO		∞	-
3	(111)	0	SI	2	∞	-
4	(110)	0	NO		∞	-
5	(1101)	0	SI	2	∞	-
6	(1100)	0	NO		∞	-
7	(11001)	0	SI	2	∞	-
8	(11000)	0	NO		∞	-
9	(110001)	0	SI	2	∞	-
10	(110000)	0.5	NO		∞	-
11	(1100001)	0.5	SI	2	∞	-
12	(1100000)	0.5	NO		∞	-
	⋮					
-	⋮					
-	⋮					
-	⋮					
	$y_1=1, y_2=1, y_j=0$ $\forall\ j=3,\ldots,41$	19.50	SI	3	19.50	$y_1=1, y_2=1,$ $y_j=0, \forall\ j=3,\ldots,41,$ $43,\ldots,56$ $y_{42}=1$
-	(10)	0	NO		19.50	
-	(101)	0	NO		19.50	
-	(1011)	0	SI	2	19.50	
-	(1010)	0	NO		19.50	
-	(10101)	0	SI	2	19.50	
-	(10100)	0	NO		19.50	
-	(101001)	0	SI	2	19.50	
-	(101000)	0.5	NO		19.50	
-	⋮					
-	⋮					
-	$y_1=1, y_2=0, y_3=1, y_j=0$ $j=4,\ldots,36$	16.75	SI	3	16.75	$y_1=1, y_2=0, y_3=1$ $y_j=0, y_{37}=1$ $\forall\ j=4,\ldots,35$ $38,\ldots,56$
-	(100)	0	NO		16.75	
-	(1001)	0	NO		16.75	
-	(10011)	0	SI	2	16.75	
-	(10010)	0	NO		16.75	
						...

ITERACION	SOLUCION PARCIAL	Z_L	¿PENETRADO?	PRUEBA DE PENETRACION	Z_u	SOLUCION TITULAR
-	(100101)	0	SI	2	16.75	
-	(100100)	0.5	NO		16.75	
-	⋮					
	$y_1=1, y_2=0, y_3=0, y_4=1$ $y_j=0$ $j=5,\ldots,42$	19.75	SI	1	16.75	
-	(1000)	0	NO		16.75	
-	(10001)	0	NO		16.75	
-	(100011)	0	SI	2	16.75	
-	(100010)	0.50	NO		16.75	
-	⋮					
	$y_1=1, y_2=y_3=y_4=0, y_5=1$ $y_j=0$ $j=6,\ldots,18$	5.50	SI	3	5.50	$y_1=1, y_2=y_3=y_4=0$ $y_5=1, y_j=0,$ $y_{19}=1$, $j=6,\ldots,18$ $j=20,\ldots,56$
-	(10000)	0	NO		5.50	
-	(100001)	0	NO		5.50	
-	(1000011)	0.5	SI	2	5.50	
-	(1000010)	0.5	NO		5.50	
-	⋮					
-	$y_1=1$ $y_2=y_3=y_4=y_5=0$ $y_6=1$ $y_j=0$ $j=7,\ldots,29$	13.5	SI	1	5.5	
-	(100000)	0.5	NO		5.50	
-	(1000001)	0.5	SI	2	5.50	
-	(1000000)	0.5	NO		5.5	
-	(10000001)	0.5	NO		5.5	
-	(100000011)	1.0	SI	2	5.5	
-	(100000010)	1.0	NO		5.5	
-	(1000000101)	1.0	SI	2	5.5	
-	(1000000100)	3.0	NO		5.5	
-	(10000001001)	3.0	SI	2	5.5	
-	(10000001000)	3.0	NO		5.5	
-	(100000010001)	3.0	SI	2	5.5	
-	(100000010000)	3.17	NO		5.5	
-	(1000000100001)	3.17	SI	2	5.5	
-	(1000000100000)	3.17	NO		5.5	
-	(10000001000001)	3.17	SI	2	5.5	
-	(10000001000000)	4.5	NO		5.5	
-	(100000010000001)	4.5	SI	2	5.5	
-	(100000010000000)	4.5	NO		5.5	
-	(1000000100000001)	4.5	SI	2	5.5	...

ITERACION	SOLUCION PARCIAL	Z_L	¿PENETRADO?	PRUEBA DE PENETRACION	Z_u	SOLUCION TITULAR
-	(1000000100000000)	5.17	NO		5.5	
-	(10000001000000001)	5.17	SI	2	5.5	
-	(1000000100000000)	5.17	NO		5.5	
-	(10000001000000001)	5.17	SI	2	5.5	
-	(100000010000000000)	6.0	SI	1	5.5	
-	(10000000)	0.5	NO		5.5	
-	(100000001)	0.5	NO		5.5	
-	(1000000011)	1.0	SI	2	5.5	
-	(1000000010)	3.0	NO		5.5	
-	(10000000101)	3.0	SI	2	5.5	
-	(10000000100)	3.0	NO		5.5	
-	(100000001001)	3.0	SI	2	5.5	
-	(100000001000)	3.17	NO		5.5	
-	(1000000010001)	3.17	SI	2	5.5	
-	⋮					
-	(100000001000000000)	6.0	SI	1	5.5	
-	(100000000)	0.5	NO		5.5	
-	⋮					
-	(100000000100000000)	6.0	SI	1	5.5	
-	(1000000000)	2.5	NO		5.5	
-	⋮					
-	(10000000001000)	6.0	SI	1	5.5	
-	(10000000000)	2.5	NO		5.5	
-	⋮					
-	(10000000000100)	6.0	SI	1	5.5	
-	⋮					
-	(100000000000000000)	6.0	SI	1	5.5	
-	(0)	0	NO		5.5	
-	(01)	0	NO		5.5	
-	(011)	0	NO		5.5	
-	(0111)	0	SI	2	5.5	
-	(0110)	0	NO		5.5	
-	⋮					
-	(01100000000000000000)	6.5	SI	1	5.5	
-	(010)	0	NO		5.5	
-	⋮					
-	(000000)	0.5	NO		5.5	
-	(0000001)	0.5	NO		5.5	
-	(00000011)	1.0	NO		5.5	
-	(000000111)	1.5	SI	2	5.5	

...

ITERACION	SOLUCION PARCIAL	Z_L	¿PENETRADO?	PRUEBA DE PENETRACION	Z_u	SOLUCION TITULAR
-	(000000110)	1.5	SI	3	1.5	(0000001101000 00.....000) $y_7=y_8=y_{10}=1$ $y_j=0$ ∀ j=1,...,6,9 j=11,...,56
-	(00000010)	1.0	NO		1.5	
-	(000000101)	1.0	NO		1.5	
-	(0000001011)	1.5	SI	2	1.5	
-	(0000001010)	3.5	SI	1	1.5	
-	(000000100)	1.0	NO		1.5	
-	(0000001001)	1.0	NO		1.5	
-	(00000010011)	3.5	SI	1,2	1.5	
-	(00000010010)	3.5	SI	1	1.5	
-	(0000001000)	3.0	SI	1	1.5	
-	(0000000)	0.5	NO		1.5	
-	(00000001)	0.5	NO		1.5	
-	(000000011)	1.0	SI	2	1.5	
-	(000000010)	3.0	SI	1	1.5	
-	(000000001)	0.5	NO		1.5	
-	(0000000011)	1.0	SI	2	1.5	
-	(0000000010)	3.0	SI	1	1.5	
-	(000000000)	0.5	NO		1.5	
-	(0000000001)	0.5	NO		1.5	
-	(00000000011)	3.0	SI	1,2	1.5	
-	(00000000010)	3.0	SI	1	1.5	
-	(0000000000)	2.5	SI	1	1.5	

Después de haber encontrado la solución titular:

(0000001101000...00)

en las iteraciones subsecuentes se tendría que las soluciones parciales restantes no dan mejor solución factible, por lo tanto, la solución óptima es:

$$y_7 = y_8 = y_{10} = 1, \quad y_j = 0 \quad \forall \; j=1,\ldots,6,9 \;\; y \;\; j=11,\ldots,56$$

donde y_7 representa al grupo $\pi_7 = (1,5)$

y_8 representa al grupo $\pi_8 = (2,4)$

y y_{10} representa al grupo $\pi_{10} = (3,6)$

lo que significa que el conjunto $I = \{I_1, I_2, I_3, I_4, I_5, I_6\}$

tiene la partición óptima:

$\pi = \{\{I_1, I_5\}, \{I_2, I_4\}, \{I_3, I_6\}\}$ que es la misma solución que se obtuvo por Programación Dinámica.

5.4 Teoría de Gráficas y el Análisis de Cúmulos

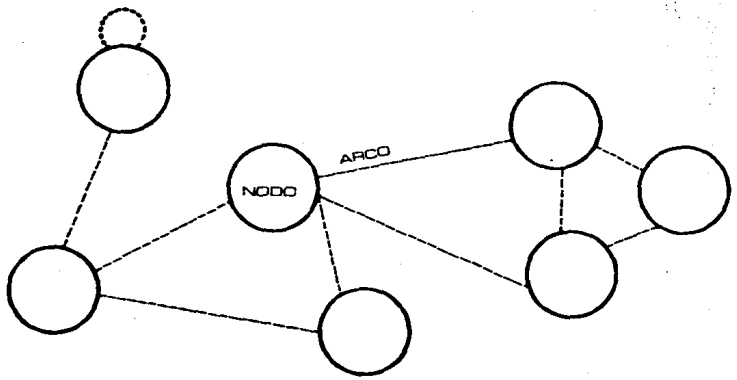
Una gran clase de problemas de Programación Lineal Entera pueden asociarse con el concepto conocido como una gráfica. Una gráfica consiste de un conjunto, cuyos elementos se denominan nodos y de una relación entre parejas de éstos. Si dos nodos están relacionados, se dice que existe un arco entre ellos. Las gráficas tienen la ventaja que pueden representarse pictóricamente, lo que ayuda a comprender y resolver los problemas planteados por medio de éstas.

Una representación pictórica de una gráfica se da en la Figura (11). Los círculos corresponden a los elementos de un conjunto dado (nodos), y las líneas indican la relación entre pares de elementos (arcos).

Definiciones

Sea $X=\{X_i \mid X_i=1,\ldots,n\}$ es un conjunto finito y S es el conjunto de todos los pares desordenados (X_i, X_j) de elementos de X, esto es:

REPRESENTACION PICTORICA DE UNA GRAFICA



NODO

ARCO

Figura II

$$S = \{(X_i, X_j) \mid \forall\, i,j\ ,\ X_i, X_j \in X\}$$

Sea $G=(X,E)$, $E \subseteq S$ una gráfica donde, los elementos de X se le llaman nodos y los elementos de E son los arcos de la gráfica.

Si $e_k=(X_i,X_j)$ el k-ésimo arco se dice que es incidente a los nodos X_i y X_j que son a su vez incidentes al arco k.

La relación de incidencia puede representarse por una matriz. Sea $A=(\alpha_{ik})$, $i=1,\ldots,n$, $k=1,\ldots,q$ (donde n es el número total de nodos y q es el número total de arcos) es la matriz de incidencia que se define en la gráfica como:

$$\alpha_{ik} = \begin{cases} 1 & \text{si } e_k \text{ y } X_i \text{ son incidentes} \\ 0 & \text{d.o.f.} \end{cases}$$

De la matriz A se le da el nombre de matriz de incidencia de la gráfica $G=(X,E)$. AL i-ésimo renglón de A le corresponde el nodo X_i, y sus unos indican todos los arcos incidentes de dicho nodo. A la k-ésima columna de A le corresponde el arco e_k y puede ser un vector unitario o tiene dos elementos iguales a 1. El primer caso ocurre si y sólo si $e_k=(X_i,X_i)$ para alguna i.

Se dice que dos arcos son adyacentes si ambos son incidentes de un mismo nodo. Una secuencia de arcos distintos $(e_{k1},e_{k2},\ldots,e_{kq})$ donde (e_{kt},e_{kt+1}) son adyacentes para $t=1,\ldots,q-1$, se le llama una trayectoria.

Una trayectoria que empieza y termina en el mismo nodo, se le da el nombre de ciclo.

Una gráfica conexa es aquella tal que $\forall X_i, X_j \in X$, $i \neq j$ existe al menos una trayectoria que conecta los nodos X_i y X_j.

Una gráfica $\hat{G}=(\hat{X},\hat{E})$ es subgráfica de $G=(X,E)$, si $\hat{G}$ se puede obtener de G quitándole el subconjunto $E-\hat{E}$ de arcos y algún subconjunto $X-\hat{X}$ de nodos que no son incidentes a algún arco de $\hat{E}$.

Una gráfica conexa se dice que es un árbol si no contiene ciclos. Una subgráfica de $G=(X,E)$ que es árbol y que su conjunto de nodos es X se le da el nombre de árbol de trayectoria de G.

Una forma de representar el problema de cúmulos es usando teoría de gráficas.

El término nodo $X_i \in X$ se puede asociar a un elemento $I_i \in I$ ó a un cúmulo $\Re_i \in \Re$, o bien, a un grupo de cúmulos. Y formar una gráfica exhaustiva $G=(X,E)$ dando peso a los arcos $e_k=(X_i,X_j) \in E \subseteq S$ de la gráfica, de acuerdo al grado de disimilitud $d(X_i,X_j)$ entre sus elementos correspondientes.

Se han definido un gran número de algoritmos que resuelven el problema de cúmulos mediante teoría de gráficas. Probablemente, los algoritmos de teoría de gráficas más conocidos son los Jerárquicos de Unión Simple y Jerárquicos de Unión Exhaustiva que se describen en la siguiente -

sección.

Gower y Ross [2] discuten algunos algoritmos para encontrar el árbol de trayectoria mínima y lo relacionan con el método Jerárquico de Unión Simple. Consideran que la longitud de un árbol es la suma de la longitud de sus arcos que componen el árbol, por lo tanto, el árbol de trayectoria mínima se define como el árbol de longitud mínima.

Si la longitud L_k del arco $e_k=(X_i,X_j)$ $X_i,X_j \in X$ se define como $L_k=d(X_i,X_j)$, el algoritmo va agrupando nodos (los que pueden ser elementos, cúmulos o grupos de cúmulos) cuya longitud de arco (grado de disimilitud) sea mínima, es semejanza al método Jerárquico de Unión Simple.

Los algoritmos Jerárquicos de Unión Simple proporcionan particiones con separación (heterogeneidad entre elementos de distintos grupos) óptima, pero descuidan la homogeneidad entre elementos de un mismo grupo, mientras que los algoritmos Jerárquicos de Unión Exhaustiva proporcionan particiones con buena homogeneidad, pero no toman en cuenta la separación (heterogeneidad).

Usando teoría de gráficas se desarrolló también un algoritmo para resolver el problema de cúmulos con el bicriterio [3] de optimalidad definido en el Capítulo 4. Este algoritmo, primeramente determina las particiones $\hat{\pi}$ de máxima separación (heterogeneidad) por medio de un algoritmo Jerárquico de Unión Simple y la partición π^* de mínimo diámetro (homogeneidad) con un algoritmo no Jerárquico [4] de Unión Exhaustiva.

Y así, trabaja con una secuencia de subgráficas $G_t=(X,E_t)$ de $G=(X,E)$ - donde $E_t=\{e_k=(X_i,X_j) \mid d(X_i,X_j) \geq t\}$, t toma como valor inicial $S(\hat{\P})$ (donde S es la separación máxima), y va tomando valores de $d(X_i,X_j) < S(\hat{\P})$ - en orden decreciente. Si la coloración -p de G se define como la asignación de uno entre los p colores para cada nodo de G, tal que dos nodos - adyacentes no tengan el mismo color. El propósito del algoritmo es que cada subgráfica G_t tenga una coloración óptima (mínimo número de colores).

Se puede demostrar que los grupos de nodos del mismo color en la coloración óptima de la última subgráfica G_t colorable en m colores define la partición de I de mínimo diámetro en m cúmulos; por lo tanto, la partición óptima de m cúmulos que cumple con el Bicriterio para el problema de cúmulos es aquella que se obtiene con $t=d(\P^*)$ (donde d es el mínimo diámetro definido en el Capítulo 4).

5.5 Métodos Jerárquicos

Esta sección está muy relacionada con la parte de teoría de gráficas tratada en la sección anterior, ya que muchos algoritmos para resolver métodos jerárquicos en el análisis de cúmulos han sido desarrollados por medio de gráficas.

Se tiene que $I=\{I_1,I_2,\ldots,I_n\}$ es el conjunto de n elementos a agrupar y $P=\{P_0,P_1,\ldots,P_m\}$ es una sucesión particiones de I tales que:

i) El número de cúmulos de P_i es mayor que el número de cúmulos P_{i+1}.

ii) $P_o=\{\{I_1\},\{I_2\},\ldots,\{I_n\}\}$ es la partición de I en n cúmulos de un sólo elemento.

iii) $P_m=\{I_1,I_2,\ldots,I_n\}$ es la partición de I en un cúmulo de n elementos.

Es decir, $P_o,P_1,\ldots,P_m$ forman una sucesión de particiones de I, - tal que la primera partición (P_o) contiene n grupos cada uno con un sólo elemento y la última partición (P_m) contiene un sólo grupo de n elemen-- tos.

Tomando como medida de asociación entre los elementos i-ésimo y - j-ésimo de la partición P_k de I ($\forall$ k=1,2,...,m), la distancia cuadrada - Euclidiana d^2_{ij}, es posible asociar a cada partición P_k una matriz de distancias cuadradas entre sus elementos, sea esta D_k.

Se tiene que D_o es la matriz de dimensión nxn y D_m es la matriz de dimensión 1 x 1.

A cada partición P_k se le asocia, también, un número f_k llamada fuerza de agrupamiento o punto Rama, que se define como:

$$f_k = \min_{p,q} \{d^2_{pq} \mid \forall\, p \neq q \text{ ésimos elementos de } P_{k-1}\}$$

y que cumple con:

i) $f_o = 0$

ii) $f_k < f_{k+1} \quad \forall\, k=1,2,\ldots,m$

Una representación esquemática de los resultados de los métodos jerárquicos es a través de "árboles" (definidos en la sección de gráficas) o "dendogramas". Un dendograma, es una clase especial de árboles cuyos niveles numéricos están asociados con los puntos Rama (f_k), como se ilustra en la siguiente Figura (12).

La matriz de distancias D_0 para la partición P_0 de I se tiene en la Tabla (14).

	I_1	I_2	I_3	. . .	I_n
I_1	0	d^2_{12}	d^2_{13}	. . .	d^2_{1n}
I_2		0	d^2_{23}	. . .	d^2_{2n}
I_3			0	. . .	d^2_{3n}
$\vdots$					
I_n					0

Tabla (14). Matriz D_0 Para la Partición

$$P_0 = \{\{I_1\},\{I_2\},\{I_3\},\ldots,I_n\})$$

Suponiendo que I_i e I_j son los elementos con mayor grado de semejanza, es decir:

$$d^2_{ij} = \min_{p,q} \{d\,p\,q \mid \text{\#}\, p \neq q \quad \text{ésimos elementos en } P_0\}$$

Entonces I_i e I_j se unen para formar el cúmulo nuevo $\{I_i, I_j\}$. - Ahora, la matriz D_1 de dimensión (n-1) x (n-1) para la partición P_1 se tiene en la Tabla (15).

DENDOGRAMA

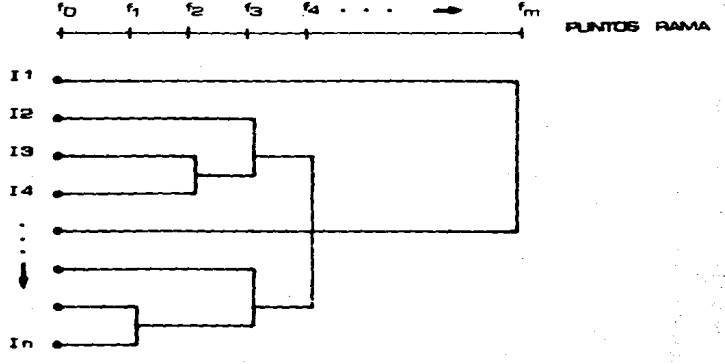


Figura 12

	$\{I_i, I_j\}$	I_1	I_2	$I_3 \ldots I_n$
$\{I_i, I_j\}$	0	d^2_{ij1}	d^2_{ij2}	$d^2_{ij3} \ldots d^2_{ijn}$
I_1		0	d^2_{12}	$d^2_{13} \ldots d^2_{1n}$
I_2			0	$d^2_{23} \ldots d^2_{2n}$
I_3				$0 \ldots d^2_{3n}$
$\vdots$				
I_n				0

$f_1 = d^2_{ij}$

Tabla (15). Matriz D_1 Para la Partición -

$$P_1 = \{\{I_i, I_j\}, I_1, \ldots, I_n\}$$

Se puede observar que (n-2) renglones de la nueva matriz D_1 de distancia cuadradas de P_1 se toman directamente de la matriz anterior D_0, sin recalcularse. Los demás, en este caso el primer renglón se calcula de nuevo. La distancia d^2_{ijk} para $k \neq i \neq j$ $k=1,\ldots,n$, va a estar en - función de la matriz previa, y así la cantidad de cálculos se minimiza.

Lance y Williams [5] dan un esquema recursivo donde el cálculo de la matriz de distancia sólo depende de los elementos de la matriz de - distancia de la etapa previa. A este esquema se le da el nombre de combinatorio, y se define como sigue:

Se tienen dos cúmulos (sean estos $\P_i$ y $\P_j$) de una cierta partición con n_i y n_j elementos respectivamente y cuya distancia es d^2_{ij}. Se considera que d^2_{ij} es la mínima distancia en el sistema, así que $\P_i$ y $\P_j$ se

unen para formar el nuevo grupo $\P_\ell$, es decir, $\P_i \cup \P_j = \P_\ell$ con n_ℓ elementos ($n_\ell = n_i + n_j$).

La determinación de la distancia entre el nuevo grupo $\P_\ell$ y un tercero $\P_h$ se hace en función de los valores de d^2_{hi}, d^2_{hj}, d^2_{ij}, n_i y n_j, que son ya conocidos y que se encuentran en la i-ésima y j-ésima columnas de la matriz de distancia de la etapa previa.

La relación combinatoria para $d^2_{h\ell}$ es entonces:

$$d^2_{h\ell} = \alpha_i\, d^2_{hi} + \alpha_j\, d^2_{hj} + \beta\, d^2_{ij} + \gamma \left| d^2_{hi} - d^2_{hj} \right|$$

donde los parámetros $\alpha_i, \alpha_j, \beta$ y γ se definen de acuerdo a los diferentes métodos jerárquicos como se explican en las siguientes secciones.

5.5.1 Método de Unión Simple

El método de Unión Simple se conoce también con el nombre de Vecino Cercano, ya que los cúmulos se van formando en cada etapa por la mínima y más pequeña unión entre ellos. Es decir, la distancia entre dos cúmulos o grupos se define como la distancia entre sus elementos más cercanos.

De acuerdo a la relación combinatoria, para este método se tiene que $\alpha_i = \alpha_j = \frac{1}{2}$, $\beta = 0$ y $\gamma = -\frac{1}{2}$. Es decir, después de haber unido los grupos $\P_i$ y $\P_j$ para formar un grupo o cúmulo $\P_\ell$, la distancia entre éste y un nuevo grupo $\P_h$ es: $d^2_{h\ell} = \frac{1}{2}\, d^2_{hi} + \frac{1}{2}\, d^2_{hj} - \frac{1}{2} \left| d^2_{hi} - d^2_{hj} \right|$ lo que significa que:

si $d^2_{hi} > d^2_{hj} \Longrightarrow d^2_{h\ell} = d^2_{hj}$

que es la mínima distancia.

O bien, si $d^2_{hj} > d^2_{hi} \Longrightarrow d^2_{h\ell} = d^2_{hi}$

Por lo tanto, para el método de Unión Simple la relación combinatoria se puede expresar como:

$$d^2_{h\ell} = \min \{d^2_{hi}, d^2_{hj}\}$$

Una característica de este método, es que conforme el grupo va creciendo se va acercando a algunos o a todos los elementos restantes. La posibilidad de que un elemento individual se una a un grupo preexistente, en vez de que éste sirva de núcleo a un grupo nuevo, se incrementa. Por lo anterior, este método obtiene grupos o cúmulos en forma de U (no elipsoides), y se dice que el sistema tiene tendencias "encadenadas".

La crítica al método de Unión Simple es que los elementos de los extremos del grupo son marcadamente diferentes.

5.5.2 Método de Unión Exhaustiva

Este método se conoce también como de Vecino Lejano.

En este método la distancia entre dos cúmulos o grupos se define como aquélla entre el par de elementos más lejanos (uno en cada grupo).

Los parámetros de la relación combinatoria son $\alpha_i=\alpha_j=\frac{1}{2}, \beta=0, \gamma=\frac{1}{2}$. Es decir, después de haber unido los grupos π_i y π_j para formar un grupo o cúmulo π_ℓ, la distancia entre éste y un nuevo cúmulo π_h se define como:

$$d^2_{h\ell} = \frac{1}{2} d^2_{hi} + \frac{1}{2} d^2_{hj} + \frac{1}{2} |d^2_{hi} - d^2_{hj}|$$

Lo que significa que:

si $d^2_{hi} > d^2_{hj} \Rightarrow d^2_{h\ell} = d^2_{hi}$ que es la máxima distancia.

O bien, si $d^2_{hj} > d^2_{hi} \Rightarrow d^2_{h\ell} = d_{hj}$

Por lo tanto, para el método de Unión Exhaustiva, la relación combinatoria se puede expresar como:

$$d^2_{h\ell} = \max \{d^2_{hi}, d^2_{hj}\}$$

En este caso, los grupos que se forman conforme van creciendo parecen alejarse; los elementos individuales que aún no están agrupados están más propensos a formar el núcleo de un nuevo grupo.

5.5.3 Método de Unión Promedio Ponderado

Se conoce como el método del centroide ya que considera que un grupo definido en un Espacio Euclidiano éste se reemplaza por las coordenadas de su centroide.

La relación combinatoria en este método se obtiene como sigue:

Si después de haber unido $\P_i$ con $\P_j$ se obtiene $\P_\ell$, se tiene que el centroide del grupo $\P_\ell$ es:

$$\bar{X}^{\ell} = \frac{n_i X_i + n_j X_j}{n_\ell}$$

donde n_i, n_j y n_ℓ son el número de elementos de los grupos $\P_i$, $\P_j$ y $\P_\ell$ respectivamente.

Por definición se tiene que la distancia de un grupo $\P_h$ al grupo $\P_\ell$ es:

$$d^2_{h\ell} = [X_h - (n_i X_i + n_j X_j)/ n_\ell]^2$$

desarrollando se tiene que:

$$d^2_{h\ell} = \frac{n_i}{n_\ell} (X_h - X_i)^2 + \frac{n_j}{n_\ell} (X_h - X_j)^2 - \frac{n_i n_j}{n^2_\ell} (X_i - X_j)^2$$

$$d^2_{h\ell} = \frac{n_i}{n_\ell} d^2_{hi} + \frac{n_j}{n_\ell} d^2_{hj} - \frac{n_i n_j}{n_\ell n_\ell} d^2_{ij}$$

Por lo tanto, los parámetros de la relación combinatoria para el método del centroide son:

$$\alpha_i = n_i/n_\ell \ , \quad \alpha_j = n_j/n_\ell ,$$

$$\beta = -\alpha_i \alpha_j \quad y \quad \gamma = 0$$

5.5.4 Método de la Mediana

Una desventaja del método del centroide es que si n_i y n_j son - "muy diferentes" el centroide del grupo formado por la unión de los grupos $\P_i$ y $\P_j$, ($\P_i \cup \P_j = \P_\ell$) es decir, $\P_\ell$ estará más cercano al grupo con

más elementos, y las características del grupo de menos elementos se perderán.

Para corregir lo anterior se desarrolló el método de la mediana. - Este método es independiente del tamaño del grupo, ya que toma $n_i = n_j$. En este método los parámetros para la relación combinatoria son:

$$\alpha_i = \alpha_j = \frac{1}{2}, \quad \beta = -\frac{1}{4} \quad y \quad \gamma = 0$$

Es decir, la distancia entre el grupo $\P_\ell$ y un nuevo grupo $\P_h$ es:

$$d^2_{h\ell} = \frac{1}{2}\ (d^2_{hi} + d^2_{hj}) - \frac{1}{4}\ d^2_{ij}$$

Lo que en un Espacio Euclidiano significa que usando el método de - la mediana el centro del nuevo grupo $\P_\ell$ está en el punto medio del lado más pequeño del triángulo definido por $\P_i$, $\P_j$ y $\P_h$ y $d^2_{h\ell}$ se mide a lo largo de la mediana de ese triángulo.

5.5.5 Algoritmo de Métodos Jerárquicos

El algoritmo para resolver el problema de cúmulos a través de métodos jerárquicos se describe a continuación.

Paso 1: Sea m=1, calcular la matriz D_0.

Paso 2: Se busca de la partición P_{m-1} los elementos tales que:

$$d^2_{ij} = \min \{d^2_{pq} \mid p \neq q, \ \P_p, \P_q \in P_{m-1}\}$$

Paso 3: Se forma el grupo $\pi_\ell = \pi_i \cup \pi_j$ y conservan los demás grupos - iguales (con esto se define la partición P_m)

Se calculan f_m y $d^2_{h\ell}$ como:

$$f_m = d^2_{ij} \quad y$$

$$d^2_{h\ell} = \alpha_i d^2_{hi} + \alpha_j d^2_{hj} + \beta d^2_{ij} + \gamma \mid d^2_{hi} - d^2_{hj} \mid$$

$\forall\ h \neq i,j$

(donde los parámetros α_i, α_j, β y γ toman valores dependiendo del método jerárquico que se elija).

Paso 4: Si P_m es la partición que contiene un sólo grupo termina, si no se hace m=m+1 y se regresa al paso 2.

En general, los métodos jerárquicos resuelven el problema de cúmulos tratando de optimizar la trayectoria por la que se va obteniendo los cúmulos o grupos. Dando una solución que no necesariamente es óptima.

5.5.6 Aplicación a los Métodos Jerárquicos

Volviendo a utilizar el ejemplo que se usó en Programación Dinámica y en Ramificación y Acotamiento en las secciones anteriores, se resolverá por métodos jerárquicos.

Sea $X_1=(1,1)$, $X_2=(3,4)$, $X_3=(5,5)$, $X_4=(4,4)$, $X_5=(1,2)$ y $X_6=(5,6)$ y se tiene la matriz de distancias cuadradas:

D_0:

	I_1	I_2	I_3	I_4	I_5	I_6
I_1	0	13	32	18	1	41
I_2		0	5	1	8	8
I_3			0	2	25	1
I_4				0	13	5
I_5					0	32
I_6						0

y la partición del conjunto $I=\{I_1,I_2,\ldots,I_6\}$ es

$P_0 = \{\{I_1\}, \{I_2\}, \{I_3\}, \{I_4\}, \{I_5\}, \{I_6\}\}$

Para m=1, se tiene $f_1=1$

y $P_1 = \{\{I_1,I_5\}, \{I_2,I_4\}, \{I_3,I_6\}\}$

usando el método de unión simple (vecino-cercano) se calcula la nueva matriz de distancias cuadradas:

D_1:

	$\{I_1, I_5\}$	$\{I_2, I_4\}$	$\{I_3, I_6\}$
$\{I_1,I_5\}$	0	8	4
$\{I_2,I_4\}$		0	2
$\{I_3,I_6\}$			0

Para m=2, se tiene $f_2=2$

y $P_2 = \{\{I_1,I_5\}, \{I_2,I_4,I_3,I_6\}\}$

D_2:

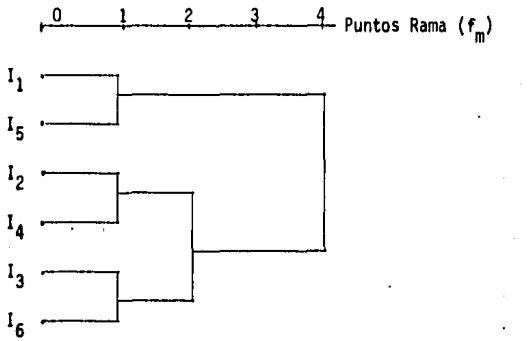
	$\{I_1, I_5\}$	$\{I_2, I_3, I_4, I_6\}$
$\{I_1, I_5\}$	0	4
$\{I_2, I_3, I_4, I_6\}$		0

Para m=3, se tiene $f_3=4$

y $P_3 = \{\{I_1, I_2, I_3, I_4, I_5, I_6\}\}$

$D_3 = 0$

Esto gráficamente se representa por el siguiente dendograma:



El algoritmo permite agrupar los elementos del grupo I en 4 formas diferentes (dada por P_0, P_1, P_2 y P_3) obteniéndose un coeficiente de agrupamiento, f_m para cada una de éstas. Interpretando los resultados obtenidos se tienen los siguientes puntos:

1) Como $f_1 = d^2_{15} = d^2_{24} = d^2_{36}$ se agruparon en la primera iteración del algoritmo. (Se puede observar que este fue el resultado que se obtuvo por Programación Dinámica y por Ramificación y Acotamiento).

2) El grupo $\{I_1, I_5\}$ es el que está más separado del resto ya que es el que se integra a éstos con la mayor fuerza de agrupamiento (Punto Rama) $f_3=4$.

CAPITULO 6

UN PROGRAMA PARA CUMULOS JERARQUICOS

Para poder resolver un problema por medio del análisis de cúmulos es necesario auxiliarse de una computadora, ya que por muy pequeño que sea el problema (como se vió en el capítulo anterior) la información y los cálculos que se necesitan hacer son muchos, y es casi imposible realizarlo sin esta ayuda.

En este capítulo se describirá un programa computacional que se elaboró para resolver el problema de cúmulos por medio de tres de los métodos jerárquicos más relevantes; a saber:

- Método de Unión-Simple
- Método de Unión-Exhaustiva
- Método de la Mediana

El programa original se tomó de Anderbeng [6] se le hicieron cambios y adaptaciones que se explicarán aquí. Además de describir el programa se explica qué información requiere y los resultados que proporciona, obteniéndose así una guía para su operación.

6.1 Descripción

El programa original se tomó de Anderbeng [6] donde se encuentra -

programado en lenguaje Fortran, en esta aplicación se reprogramó en lenguaje Basic y se le hicieron algunos cambios.

El programa que se obtuvo finalmente se encuentra en el anexo D de esta tesis, puede ser utilizado en microcomputadoras IBM/PC o compatible.

Este programa consta de un programa Principal y de 3 subrutinas que se relacionan como se ve en el Diagrama de la Figura (13), y se describen a continuación:

Programa Principal:

Da la capacidad máxima (en palabras) del programa; dimensiona todos los arreglos utilizados en éste; lee las especificaciones al problema de cúmulos que se desea resolver, así como algunas opciones del programa.

El programa principal llama a la Subrutina de Lectura y a la Subrutina de Agrupamiento.

Subrutina de Lectura:

Lee la Matriz de Similitud de una unidad de disco especificada en el Programa Principal, así como su formato de entrada. Este formato puede ser de dos formas:

1) Cada registro es un renglón de la matriz triangular inferior.

2) La matriz triangular inferior se lee por renglones en arreglos de un mismo tamaño, que se especifica en el programa principal.

DIAGRAMA DEL PROGRAMA

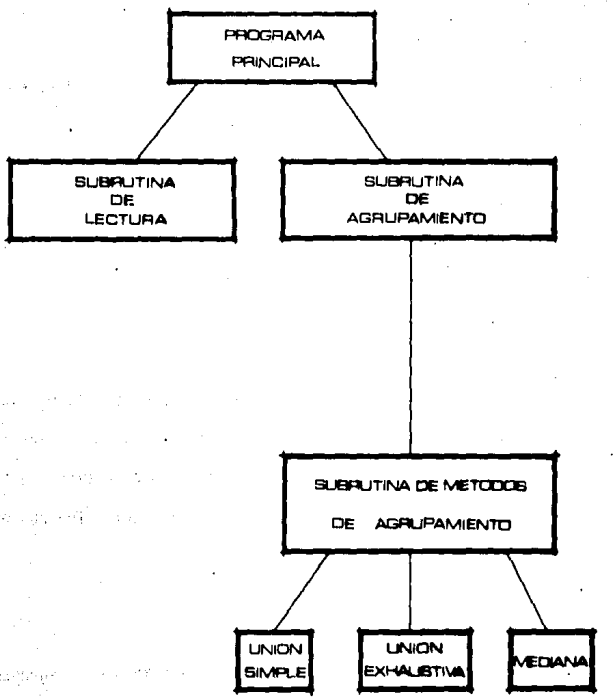


Figura 13

Esta subrutina manda un mensaje de error cuando lee más o menos información de la esperada.

Como se puede ver, es necesario hacer un programa adicional que calcule la matriz de similitud. Este programa debe tener como entrada los datos del problema, como son los elementos a agrupar y sus variables.

En el anexo D, se dará este programa adicional cuya salida es la matriz de similitud en el formato descrito en 1).

Subrutina de Agrupamiento:

Compara los valores de la matriz de similitud obtenida en la etapa previa (primera etapa matriz original) para decidir cuáles serán los grupos que se unan en cada etapa.

Luego, calcula los valores de la nueva matriz de similitud llamando a la subrutina de Métodos de Agrupamiento. Esta subrutina da los resultados finales del programa, en forma de tabla en la que se tiene la información del agrupamiento por etapas con sus correspondientes fuerzas de agrupamiento*.

Subrutina de Métodos de Agrupamiento:

Está formada por tres programas uno para cada Método de Agrupamiento, que como ya se mencionaron son tres:

- Método de Unión-Simple de Johnson [7].

* Ver Sección 5.5.

- Método de Unión-Exhaustiva de Johnson [8].

- Método de la Mediana de Gower [9].

De acuerdo al método elegido para el agrupamiento, se calculan los nuevos valores de la matriz de similitud que dependen de los grupos que se unan en cada etapa. Estos métodos jerárquicos ya han sido explicados detalladamente en el capítulo anterior.

6.2 Entradas y Salidas del Programa

La descripción de entradas y salidas del programa se esquematiza a continuación, la definición de variables y opciones se da en la siguiente sección.

Esquema General:

	Información Requerida (Entradas)	Información Proporcionada (Salidas)
Programa Principal	DATOS: TITLE NE OPCIONES: ISIGN NTSV NTIN INOPT	
Subrutina de Lectura	DATOS: X(I)	MENSAJES DE ERROR: "Se encontró EOF cuando ninguno esperaba" "No hay EOF cuando se esperaba uno"
Subrutina de Agrupamiento	OPCION: MET	RESULTADOS: TITLE I IL(I) II(I) JL(I) JJ(I) MEXT(I) SS(I)

6.3 Descripción de Variables y Opciones

6.3.1 En el Programa Principal

DATOS:

TITLE. Variables alfanumérica de hasta 20 caracteres, que representa el título que se deseee dar al problema de cúmulos que se pretende resolver.

NE. Variable numérica que toma el número de elementos que se desea agrupar. Esta variable determina la capacidad requerida para guardar la matriz de similitud.

Asumiendo que la matriz de similitud es simétrica se tiene $\binom{NE}{2} = \frac{NE(NE-1)}{2}$ pares de elementos, lo que determina el número de datos de la matriz de similitud. Por lo tanto, la capacidad requerida para guardar la matriz de similitud (en relación al número de elementos NE) es entonces:

Elementos	Capacidad Requerida
25	300
50	1225
75	2775
100	4950

En este programa se tiene una capacidad de 2775 para 75 elementos a agrupar. Sin embargo, se pueden ampliar.

OPCIONES:

ISING. Opción de elección de la medida de asociación

Si ISIGN = 1 -Se toma la distancia Euclidiana como medida de asociación que mide el grado de disimilitud entre los elementos.

Si ISIGN = -1 -Toma el coeficiente de correlación lineal como medida de asociación que mide el grado de similitud entre los elementos.

NTSV. Opción de salida de los resultados del agrupamiento.

Si NTSV $\leq$ 0 -Los resultados del agrupamiento no se imprimen.

Si NTSV > 0 -Se imprimen los resultados del agrupamiento.

NTIN. Unidad por la que la matriz de similitud se lee.

NTIN = 1,2,3 -Puede tomar estos tres valores para la microcomputadora que se está usando en este caso.

INOPT. Opción que dá el formato de entrada de la matriz de similitud.

INOPT $\leq$ 0 -La matriz de similitud es una matriz triangular inferior donde cada registro representa un renglón de ésta.

INOPT > 0 -La matriz triangular inferior se lee por renglones en arreglos de un mismo tamaño, es decir, de un tamaño INOPT.

6.3.2 En la Subrutina de Lectura

DATOS:

X(I) Es la variable de entrada de la matriz de similitud, que como ya se explicó está relacionada con el número de elementos NE. En este caso tiene una capacidad máxima de 2775.

MENSAJES DE ERROR:

"Se encontró EOF cuando ninguno se esperaba"

Este mensaje significa que se leyó menos información de la matriz de similitud de la esperada.

"No hay EOF cuando se esperaba uno"

En este mensaje indica que hay más información de la matriz de similitud de la esperada.

6.3.3 En la Subrutina de Agrupamiento

OPCION:

MET. Opción de elección del método que se desea utilizar para hacer la agrupación.

MET = 1 Unión-Simple
MET = 2 Unión-Exhaustiva
MET = 3 Mediana

RESULTADOS: Se tiene una tabla con los resultados de todas las etapas del proceso.

TITLE. Es el título que se da inicialmente.

I. Etapa en la que ocurre la agrupación.

II(I) Cúmulo que inicia el agrupamiento de la etapa I.

JJ(I) Cúmulo que finaliza el agrupamiento de la etapa I.

SS(I) Fuerza de agrupamiento con la que se unieron los cúmulos de la etapa I.

IL(I) Etapa previa en la que el cúmulo inicial II(I) en la etapa I ya ha sido unido.

JL(I) Etapa previa en la que el cúmulo final JJ(I) de la etapa I ya ha sido unido.

MEXT(I) Etapa posterior en la que el cúmulo inicial II(I) de la etapa I, se vuelve a unir.

La subrutina de métodos de agrupamiento no tiene entradas y salidas al usuario, trabaja internamente calculando los nuevos valores para la matriz de similitud, de acuerdo al método elegido y así poder formar los agrupamientos en cada etapa.

6.4 Comentarios

Los métodos jerárquicos como ya se ha mencionado, no dan una solución óptima al problema de cúmulos, sino dan una o varias soluciones aproximadas que se consideran bastante "Buenas" si se analiza con cuidado el proceso de agrupamiento. En particular, esta solución puede elegirse de las etapas del algoritmo de métodos jerárquicos, de acuerdo a la manera en que actúan las fuerzas de agrupamiento, o bien, al número de cúmulos o grupos que se desea obtener.

En el siguiente capítulo se utilizará el programa que se describió aquí, para un problema específico y se analizarán los resultados comparando los tres métodos.

CAPITULO 7

UNA APLICACION PRACTICA DEL ANALISIS DE CUMULOS: LA CONCENTRACION GEOGRAFICA DE ACTIVIDADES EN MEXICO

Una aplicación real del problema de cúmulos es la mejor forma de comprender y analizar los métodos jerárquicos descritos en la sección 5.5 y para los cuales se realizó el programa que se detalla en el capítulo 6.

En este capítulo se agruparán las entidades federativas de la República Mexicana de acuerdo a la importancia relativa de sus ramas de actividad económica. Para ello, se planteará el problema en términos de análisis de cúmulos y se resolverá con los tres métodos jerárquicos programados. Se formarán grupos o cúmulos de entidades federativas cuya distribución de ramas de actividad económica sea semejante. Por último, se compararán los resultados de los tres métodos jerárquicos y se darán conclusiones de interés.

7.1 Planteamiento del Problema

Problema: Agrupar las treinta y dos entidades federativas de la República Mexicana de acuerdo a la distribución porcentual (%) de nueve ramas de actividad económica.

Planteamiento: Sea n el número de elementos a agrupar, en este caso el número de entidades federativas; y sea p el número de atributos o características poseidos por los n elementos, en este caso el número de ramas de actividad económica. Entonces, se tiene n=32 y p=9.

Se define el conjunto de elementos $I=\{I_1,I_2,\ldots,I_{32}\}$ como el conjunto de entidades federativas, y el conjunto de características o atributos sea $C=\{C_1,C_2,\ldots,C_9\}$ como el conjunto de ramas de actividad económica, de la siguiente manera:

I_1 = Aguascalientes
I_2 = Baja California Norte
I_3 = Baja California Sur
I_4 = Campeche
I_5 = Coahuila
I_6 = Colima
I_7 = Chiapas
I_8 = Chihuahua
I_9 = Distrito Federal
I_{10} = Durango
I_{11} = Guanajuato
I_{12} = Guerrero
I_{13} = Hidalgo
I_{14} = Jalisco
I_{15} = México
I_{16} = Michoacán
I_{17} = Morelos
I_{18} = Nayarit
I_{19} = Nuevo León
I_{20} = Oaxaca
I_{21} = Puebla

I_{22} = Querétaro
I_{23} = Quintana Roo
I_{24} = San Luis Potosí
I_{25} = Sinaloa
I_{26} = Sonora
I_{27} = Tabasco
I_{28} = Tamaulipas
I_{29} = Tlaxcala
I_{30} = Veracruz
I_{31} = Yucatán
I_{32} = Zacatecas

Los datos sobre la actividad económica por ramas, se obtuvieron del X Censo General de Población [10] y se pueden ver en la Tabla (16). Las ramas consideradas son las siguientes:

C_1 = Agricultura, Ganadería, Caza, Silvicultura y Pesca.

C_2 = Explotación de Minas y Canteras.

C_3 = Industrias Manufactureras.

C_4 = Electricidad, Gas y Agua.

C_5 = Construcción.

C_6 = Comercio al por mayor, al por menor, Restaurantes y Hoteles.

C_7 = Transporte, Almacenamiento y Comunicaciones.

C_8 = Establecimientos Financieros, Seguros, Bienes Inmuebles, etc.

C_9 = Servicios Comunales, Sociales y Personales.

PERSONAS POR RAMA DE ACTIVIDAD ECONOMICA

	AGRIC. GANAD. ETC.	MINERIA	MANUF.	ELECTR. GAS Y AGUA	CONSTR.	COMERC. TURISM.	TRANSP. ALMAC.	FINANC. SEGUROS INMOB.	SERVICIOS
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
I_1	28615	503	23323	224	10625	16566	7954	2080	20514
I_2	38180	502	54698	1438	25010	55454	16027	9286	67323
I_3	13538	592	5226	292	4876	8289	3362	1264	12143
I_4	42836	292	9925	255	8681	10821	4980	1215	18266
I_5	76343	7532	69841	1956	31698	49163	22047	7830	69510
I_6	30291	922	8155	470	7310	11381	5332	1320	16413
I_7	421561	504	25576	1001	18929	34139	10837	3163	48196
I_8	137909	6405	82286	1594	41285	67457	30294	11551	85323
I_9	202336	333858	407001	72810	321627	134858	37105	162970	222606
I_{10}	110311	3300	27151	891	16763	25920	12474	3679	39126
I_{11}	187495	68886	80307	2953	62693	49464	28200	11882	51065
I_{12}	318424	993	35859	674	22552	49978	15771	3569	97606
I_{13}	187043	3987	42452	663	17939	27197	12307	2428	51945
I_{14}	267824	1938	229277	2580	80092	157843	53741	25225	180655
I_{15}	367888	4115	505855	8718	138731	245000	104705	45736	332344
I_{16}	344325	1478	69745	1165	38135	70661	23603	6722	77073
I_{17}	76303	510	29078	545	22131	29159	10561	3575	43829
I_{18}	84819	351	16241	467	11263	19169	7671	1870	24846
I_{19}	67308	2246	197791	3073	58712	89990	38634	23661	131095
I_{20}	474793	1663	40283	583	18370	34393	11063	2369	67961
I_{21}	447439	2237	120031	1736	39961	82621	27806	9673	109276
I_{22}	65035	1326	39381	377	16296	18171	7962	3014	26589
I_{23}	23136	110	4554	225	4562	9934	3278	1082	12828
I_{24}	181346	4415	47484	630	26191	39957	15559	5182	61032
I_{25}	156542	1225	40197	1238	30211	51912	24474	9688	65999
I_{26}	100765	4330	46493	1530	29206	51286	24344	9761	64843
I_{27}	127459	4678	22266	415	16365	20608	9311	2883	30681
I_{28}	112362	3835	74481	2113	45234	70613	27807	10342	98428
I_{29}	65906	165	25575	181	7599	9740	4913	751	17295
I_{30}	678029	7832	144494	3785	82113	134702	51985	14355	194176
I_{31}	115336	406	35671	929	22433	33621	10763	5038	51499
I_{32}	148474	5901	14427	421	18744	19229	7241	2590	27629

TABLA (16)

7.2 Desarrollo del Problema

A cada entidad federativa del conjunto I, I_i (para i=1,2,...,32) le corresponde un vector de distribución* de ramas de actividad económica - X_i, cuyas componentes son los datos asociados a cada entidad (ver sección 2.2).

Como cada entidad federativa I_i (para i=1,2,...,32) está descrita por 9 ramas de actividad económica) entonces $X_i=(X_{1i},X_{2i},...,X_{9i})$ será un vector de 9 componentes.

Por lo tanto, el conjunto I de entidades federativas se describe por una matriz X de 32x9 elementos que se encuentran en la Tabla (16). En base a la matriz X y usando como medida de asociación entre elementos la - distancia Euclidiana, se calcula el grado de disimilitud entre las entidades federativas, para ello, se utiliza el Programa del Anexo D que calcula la matriz de disimilitud para este problema, la que se encuentra en la Tabla (17).

Esta matriz de disimilitud, servirá de entrada al programa que resuelve métodos Jerárquicos. (Ver Capítulo 6).

7.3 Análisis de Resultados

Los resultados de los agrupamientos por los tres métodos Jerárquicos

---

* Se calcula la distribución porcentual de las ramas de actividad económica de la población económicamente activa para estandarizar la información.

MATRIZ DE DISIMILITUD DE ENTIDADES FEDERATIVAS POR RAMA DE ACTIVIDAD ECONOMICA

.00	.15	.12	.27	.04	.16	.24	.25	.29	.23	.20	.37	.21	.03	.10	.32	.13	.20	.19	.52	.20	.13	.19	.24	.19	.09	.32	.06	.24	.29	.18	.40
.15	.00	.17	.24	.12	.27	.37	.18	.27	.36	.21	.47	.44	.16	.14	.45	.24	.41	.14	.45	.43	.27	.20	.17	.30	.19	.45	.13	.40	.42	.21	.52
.12	.17	.00	.19	.12	.11	.32	.10	.22	.22	.12	.34	.30	.14	.20	.31	.10	.26	.27	.39	.30	.19	.13	.22	.16	.07	.31	.07	.29	.27	.17	.36
.27	.24	.19	.00	.24	.08	.24	.17	.42	.04	.15	.16	.12	.02	.20	.12	.19	.08	.40	.02	.13	.15	.09	.06	.05	.15	.12	.21	.13	.09	.04	.19
.04	.12	.12	.24	.00	.18	.37	.00	.25	.26	.21	.39	.34	.06	.09	.26	.15	.31	.17	.35	.33	.16	.21	.27	.21	.10	.35	.05	.30	.32	.21	.45
.16	.27	.11	.08	.16	.00	.41	.11	.37	.11	.18	.23	.19	.17	.25	.20	.05	.15	.34	.39	.19	.14	.08	.12	.06	.08	.20	.14	.19	.16	.07	.27
.24	.67	.32	.24	.37	.41	.00	.49	.71	.31	.44	.19	.23	.33	.41	.23	.43	.26	.72	.03	.24	.43	.40	.20	.37	.46	.22	.24	.29	.25	.36	.17
.05	.18	.10	.17	.08	.11	.49	.00	.30	.18	.17	.72	.27	.06	.14	.27	.08	.23	.24	.47	.25	.10	.14	.19	.14	.04	.27	.06	.23	.24	.14	.35
.29	.28	.22	.42	.25	.37	.71	.30	.00	.43	.27	.56	.50	.29	.27	.50	.35	.48	.26	.68	.49	.33	.40	.44	.40	.31	.50	.29	.45	.48	.40	.56
.23	.36	.22	.04	.26	.11	.31	.18	.43	.00	.19	.14	.09	.23	.21	.10	.12	.06	.42	.29	.09	.15	.11	.02	.06	.17	.09	.23	.10	.06	.06	.17
.20	.21	.22	.19	.21	.19	.44	.17	.27	.19	.00	.30	.24	.20	.26	.29	.17	.23	.25	.43	.24	.15	.21	.19	.18	.17	.24	.21	.22	.23	.18	.30
.37	.49	.34	.16	.39	.23	.19	.32	.56	.14	.30	.00	.00	.37	.44	.00	.25	.10	.36	.17	.11	.29	.24	.15	.19	.31	.00	.36	.16	.09	.19	.00
.21	.44	.30	.12	.34	.19	.23	.27	.50	.09	.24	.00	.00	.31	.38	.05	.21	.06	.49	.21	.04	.21	.20	.07	.15	.26	.04	.32	.09	.04	.14	.19
.03	.16	.14	.22	.06	.17	.53	.06	.29	.23	.20	.37	.31	.00	.09	.31	.15	.28	.19	.51	.29	.12	.20	.24	.19	.10	.32	.08	.26	.28	.18	.40
.10	.14	.20	.30	.09	.25	.61	.14	.27	.31	.26	.44	.38	.09	.00	.39	.21	.16	.11	.58	.36	.18	.28	.30	.27	.18	.39	.14	.32	.36	.26	.47
.32	.45	.31	.15	.25	.26	.33	.27	.50	.10	.25	.00	.00	.31	.39	.00	.21	.06	.50	.20	.04	.22	.19	.00	.15	.26	.04	.32	.10	.04	.14	.19
.13	.24	.10	.10	.15	.05	.45	.08	.25	.12	.17	.25	.21	.15	.21	.21	.00	.17	.31	.41	.20	.11	.08	.14	.08	.06	.22	.11	.19	.10	.07	.29
.20	.41	.26	.08	.21	.15	.26	.23	.48	.06	.23	.10	.06	.28	.26	.05	.17	.00	.47	.24	.06	.19	.15	.05	.11	.22	.05	.28	.11	.02	.10	.12
.19	.14	.27	.40	.17	.34	.72	.24	.26	.42	.35	.55	.49	.19	.11	.53	.31	.47	.00	.69	.47	.29	.37	.42	.37	.27	.50	.21	.45	.47	.27	.58
.53	.45	.50	.02	.55	.39	.03	.47	.68	.25	.43	.17	.21	.51	.56	.20	.41	.24	.69	.00	.22	.41	.36	.26	.35	.46	.30	.52	.27	.23	.24	.14
.20	.43	.30	.17	.33	.19	.24	.25	.49	.09	.24	.11	.04	.29	.24	.04	.20	.06	.47	.22	.00	.19	.09	.07	.15	.25	.06	.31	.07	.04	.15	.12
.13	.27	.19	.15	.14	.14	.43	.10	.33	.15	.15	.29	.21	.12	.18	.22	.11	.19	.29	.41	.19	.00	.17	.15	.13	.13	.22	.16	.15	.19	.11	.30
.19	.20	.13	.09	.21	.05	.40	.14	.40	.11	.21	.22	.20	.23	.26	.19	.08	.15	.37	.28	.20	.17	.00	.15	.05	.11	.20	.17	.21	.16	.08	.24
.24	.37	.23	.06	.27	.12	.50	.19	.44	.02	.19	.13	.07	.24	.32	.08	.14	.05	.42	.28	.07	.15	.13	.00	.08	.19	.08	.24	.09	.05	.07	.16
.19	.30	.16	.05	.21	.08	.37	.14	.46	.06	.18	.19	.15	.19	.27	.15	.08	.11	.37	.25	.15	.13	.06	.08	.00	.12	.15	.19	.15	.12	.04	.22
.09	.19	.07	.15	.10	.08	.48	.04	.31	.17	.17	.31	.24	.19	.18	.26	.06	.22	.27	.46	.25	.13	.11	.19	.12	.00	.26	.07	.03	.23	.13	.34
.32	.45	.31	.12	.35	.20	.22	.27	.50	.09	.24	.06	.04	.32	.39	.04	.22	.05	.50	.20	.06	.22	.20	.08	.15	.26	.00	.32	.11	.04	.15	.00
.06	.13	.07	.21	.05	.14	.54	.06	.29	.23	.21	.36	.33	.05	.14	.32	.11	.28	.21	.52	.31	.16	.17	.24	.16	.07	.32	.00	.23	.29	.18	.40
.26	.40	.29	.13	.30	.19	.29	.23	.45	.10	.22	.16	.09	.26	.32	.10	.19	.11	.43	.27	.07	.15	.21	.09	.15	.23	.11	.28	.00	.09	.13	.18
.29	.42	.27	.09	.32	.16	.25	.24	.48	.06	.23	.09	.04	.28	.36	.04	.18	.02	.47	.23	.04	.19	.16	.05	.12	.23	.04	.29	.09	.00	.11	.11
.18	.31	.17	.04	.21	.07	.36	.14	.40	.06	.18	.19	.14	.18	.26	.14	.07	.10	.37	.24	.13	.11	.08	.07	.04	.13	.15	.18	.13	.11	.00	.22
.40	.52	.38	.19	.43	.27	.15	.35	.56	.17	.30	.08	.10	.40	.47	.09	.29	.12	.56	.14	.12	.30	.26	.16	.22	.34	.08	.40	.18	.11	.22	.00

TABLA (17)

de Unión-Simple, Unión-Exhaustiva y la Mediana, se encuentran en las tablas (18), (19) y (20) respectivamente.

Cabe mencionar que en la etapa inicial los cúmulos se numeraron de acuerdo a su número de elemento (como el método de Cúmulos Jerárquicos lo indica, ver sección 5.5) es decir: $I_i = \P_i \quad \forall\, i=1,2,\ldots,32$.

En las tablas sólo se da el número de cúmulo y la forma en que se van agrupando etapa a etapa. Al agruparse dos cúmulos el número del cúmulo inicial se mantiene para la unión de estos. Por ejemplo, tomando la Tabla (20) de resultados del método de la Mediana, en la etapa 1 se une $\P_{18}$ con $\P_{30}$ entonces se tiene que:

$$\P_{18} = \P_{18} \cup \P_{30}$$

en la etapa 5 $\quad \P_{16} = \P_{16} \cup \P_{27}$

y en la etapa 6 $\quad \P_{16} = \P_{16} \cup \P_{18}$

lo que significa que:

$$\P_{16} = \P_{16} \cup \P_{27} \cup \P_{18} \cup \P_{30}$$

es decir, el cúmulo $\P_{16}$ en la sexta etapa del método de cúmulos Jerárquicos por la Mediana, está formado por:

$$\P_{16} = \{I_{16}, I_{24}, I_{18}, I_{30}\}$$

Lo que significa que en términos de agrupar entidades federativas, es que Michoacán (I_{16}), Tabasco (I_{27}), Nayarit (I_{18}) y Veracruz (I_{30}) forman un grupo (en la etapa 6) con distribución de ramas de actividad

ecónomica muy* semejante.

Para analizar la manera en que se van uniendo los grupos o cúmulos etapa a etapa, de acuerdo con su grado de disimilitud (fuerza de agrupamiento) se tiene una representación gráfica a la que se llamó dendograma (en la sección 5.5). Los dendogramas que se obtuvieron al resolver el problema ya descrito por los métodos Jerárquicos de Unión-Simple, Unión-Exhaustiva y la Mediana, se tienen en las figuras (14), (15) y (16) respectivamente; y se analizarán en las siguientes secciones. En las etapas en que el rango de las fuerzas de agrupamiento es muy pequeño, se tuvo que reescalar para ampliar esas secciones del dendograma.

7.3.1 Análisis de Resultados del Método de Unión-Simple

Con ayuda de la Tabla (18) y la Figura (14) que contienen la solución al problema, de agrupar las 32 entidades federativas de acuerdo a su rama de actividad económica, por el método jerárquico de Unión-Simple, se determinan los cúmulos o grupos.

Se comienza formando grupos con pares de elementos los que posteriormente se unen a nuevos elementos o a grupos ya formados, hasta que se llega a un sólo grupo compuesto de todos los elementos.

Los grupos que se distinguen marcadamente, antes de que se lleguen a unir todos (etapa 1 a la 19) son:

* Esto habría que analizarlo cuidadosamente en base a la fuerza de agrupamiento con la que se unieron estos grupos, en comparación con el resto.

METODO DE UNION SIMPLE

ETAPA	CUMULO INICIAL	CUMULO FINAL	FZA AGRUPAMIENTO
1	18	30	2.148267E-02
2	10	24	2.240599E-02
3	7	20	3.287578E-02
4	1	14	3.373813E-02
5	16	27	3.593847E-02
6	13	21	3.729735E-02
7	13	18	3.848126E-02
8	13	16	3.907956E-02
9	4	31	3.994072E-02
10	4	10	4.042856E-02
11	4	25	4.147026E-02
12	8	26	4.158026E-02
13	1	5	4.344803E-02
14	6	17	4.552141E-02
15	6	23	4.598829E-02
16	4	13	4.676125E-02
17	4	6	.0512741
18	1	8	5.331921E-02
19	1	28	5.371342E-02
20	1	4	6.417438E-02
21	1	29	6.671036E-02
22	1	3	6.999263E-02
23	1	32	7.695156E-02
24	1	12	.0782564
25	1	15	8.561145E-02
26	1	22	.1006832
27	1	19	.1107824
28	1	2	.1158776
29	1	7	.1364967
30	1	11	.154378
31	1	9	.2542665

TABLA (18)

ETAPA	PRE-ETAPA II	PRE-ETAPA JJ	ETAPA II SIGUE
1	0	0	7
2	0	0	10
3	0	0	29
4	0	0	13
5	0	0	8
6	0	0	7
7	6	1	8
8	7	5	16
9	0	0	10
10	9	2	11
11	10	0	16
12	0	0	18
13	4	0	18
14	0	0	15
15	14	0	17
16	11	8	17
17	16	15	20
18	13	12	19
19	18	0	20
20	19	17	21
21	20	0	22
22	21	0	23
23	22	0	24
24	23	0	25
25	24	0	26
26	25	0	27
27	26	0	28
28	27	0	29
29	28	3	30
30	29	0	31
31	30	0	0

Cont. TABLA (18)

DENDOGRAMA DEL METODO DE UNION SIMPLE

(Etapas K=1, 2,...19, para K>19, grupos fuera de rango)

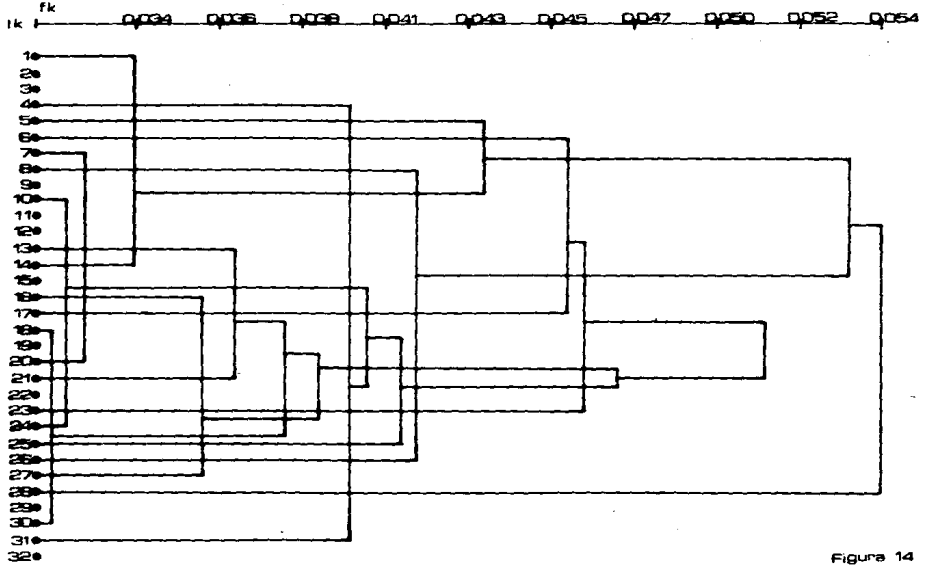


Figura 14

1) $\P_7 = \{I_7, I_{20}\}$ este grupo se considera aislado de todos los demás, ya que se forma en la etapa 3 y se vuelve a unir hasta la etapa 29.

2) $\P_{13} = \{I_{13}, I_{21}, I_{18}, I_{30}, I_{16}, I_{27}\}$ este grupo se obtuvo de la unión de dos grupos $\P_{18} = \{I_{18}, I_{30}\}$ y $\P_{16} = \{I_{16}, I_{27}\}$ ya formados en etapas previas (etapa 1 y etapa 5 respectivamente) y dos nuevos elementos I_{13} e I_{21} a partir de la sexta etapa.

3) $\P_4 = \{I_4, I_{31}, I_{10}, I_{24}, I_{25}\}$ este grupo se comienza a formar a partir de la etapa 2 con $\{I_{10}, I_{24}\}$ al que se le une en la etapa 10 $\{I_4, I_{31}\}$ y finalmente el elemento I_{25} en la etapa 11.

4) $\P_6 = \{I_6, I_{17}, I_{23}\}$ este grupo se obtiene de la unión del grupo -- $\P_6 = \{I_6, I_{17}\}$ (en la etapa 14) con el elemento I_{23} (en la etapa 15). Este grupo puede considerarse como parte de $\P_4$ ya que se une a éste en la etapa 17.

5) $\P_1 = \{I_1, I_{14}, I_5, I_8, I_{26}, I_{28}\}$ este grupo se comienza a formar en la etapa 4 como $\P_1 = \{I_1, I_{14}\}$ a este se le une I_5 en la etapa 13, posteriormente se le une el grupo obtenido en la etapa 12 $\P_8 = \{I_8, I_{26}\}$ y finalmente en la etapa 19 se le une I_{28}.

Los 10 grupos restantes son de un sólo elemento, a saber:

$\P_{29}=\{I_{29}\}$, $\P_3=\{I_3\}$, $\P_{32}=\{I_{32}\}$, $\P_{12}=\{I_{12}\}$, $\P_{15}=\{I_{15}\}$, $\P_{22}=\{I_{22}\}$, $\P_{19}=\{I_{19}\}$, $\P_2=\{I_2\}$, $\P_{11}=\{I_{11}\}$ y $\P_9=\{I_9\}$.

Estos grupos se van uniendo al grupo $\P_1$ a partir de la etapa 20, de ahí la tendencia encadenada del método jerárquico de Unión-Simple (ver - 5.5.1).

Otra forma de determinar la solución es tomando en cuenta que en la etapa 16 y 17 se forma un grupo que une a los grupos $\P_4$, $\P_{13}$, y $\P_6$ ya definidos, al que se nombra $\P_4$ como sigue:

$$\P_4 = \P_4 \cup \P_{13} \cup \P_6$$

Este grupo puede considerarse como "bueno" ya que el grado de disimilitud (fuerza de agrupamiento) con el que se lleva a cabo la unión es "pequeña" dentro del rango que se tiene en el problema, es decir:

Para $\P_4=\P_4 \cup \P_{13} \cup \P_6$ se tiene un grado de disimilitud de 0.054 - dentro de un rango de (0.021, 0.254); por lo tanto, otra solución a este problema es con los grupos:

$$\P_1 = \{I_1, I_{14}, I_5, I_8, I_{26}, I_{28}\} ,$$
$$\P_4 = \{I_4, I_{31}, I_{10}, I_{24}, I_{25}, I_{13}, I_{21}, I_{18}, I_{30}, I_{16}, I_{27}, I_6, I_{17}, I_{23}\}$$

y los 10 grupos de un sólo elemento.

Se puede observar, que conforme el grupo va creciendo, se van uniendo todos los elementos restantes, incrementándose la posibilidad de que un elemento individual se una a un grupo preexistente en vez de que éste sirva de núcleo a un grupo nuevo.

7.3.2 Análisis de Resultados del Método de Unión-Exhaustiva

Para analizar la solución que se obtiene por el método jerárquico - de Unión-Exhaustiva se utiliza la Tabla (19) y el dendograma de la Figura (15).

Los grupos que se distinguen hasta la etapa 20, ya que a partir de ahí la fuerza de agrupamiento se comienza a crecer en mayor proporción, son:

1) $\P_7 = \{I_7, I_{20}\}$ este grupo se forma en la etapa 3, se considera aislado ya que se vuelve a unir hasta el final del proceso con el resto, con el mayor grado de disimilitud (.54).

2) $\P_4 = \{I_4, I_{31}, I_{10}, I_{24}, I_{25}\}$ este grupo se comienza a formar en la etapa 2 con $\{I_{10}, I_{24}\}$ al que se le une $\{I_4, I_{31}\}$ en la etapa 8 y finalmente se le une el elemento I_{25} en la etapa 13.

3) $\P_1 = \{I_1, I_{14}, I_5, I_8, I_{26}, I_{28}\}$ este grupo se forma en la etapa 4 con - $\{I_1, I_{14}\}$ a éste se le une I_5 en la etapa 10, el grupo $\{I_8, I_{26}\}$ en - la etapa 15 y en la etapa 17 I_{28}.

4) $\P_{13} = \{I_{13}, I_{21}, I_{16}, I_{27}, I_{18}, I_{30}\}$ este grupo se obtiene de la unión - del grupo $\{I_{18}, I_{30}\}$ de la etapa 1 con $\{I_{13}, I_{21}\}$ de la etapa 6 y con $\{I_{16}, I_{27}\}$ de la etapa 5.

5) $\P_6 = \{I_6, I_{17}, I_{23}\}$ este grupo se forma de la unión del grupo $\{I_6, I_{17}$

METODO DE UNION EXHAUSTIVA

ETAPA	CUMULO INICIAL	CUMULO FINAL	FZA AGRUPAMIENTO
1	18	30	2.148267E-02
2	10	24	2.240599E-02
3	7	20	3.287578E-02
4	1	14	3.373813E-02
5	16	27	3.593847E-02
6	13	21	3.729735E-02
7	4	31	3.994072E-02
8	4	10	4.042856E-02
9	8	26	4.158026E-02
10	1	5	4.344803E-02
11	6	17	4.552141E-02
12	6	23	4.598829E-02
13	4	25	4.673235E-02
14	13	16	4.727549E-02
15	1	8	5.331921E-02
16	13	18	5.864972E-02
17	1	28	5.882001E-02
18	4	6	7.780171E-02
19	12	13	7.972695E-02
20	12	32	8.285375E-02
21	1	15	9.507588E-02
22	1	3	.1237679
23	1	22	.1254316
24	4	29	.13301
25	2	19	.1421224
26	1	2	.1470325
27	4	12	.1586859
28	4	11	.1909428
29	1	4	.2163228
30	1	9	.2785187
31	1	7	.5394737

TABLA (19)

ETAPA	PRE-ETAPA II	PRE-ETAPA JJ	ETAPA II SIGUE
1	0	0	16
2	0	0	8
3	0	0	31
4	0	0	10
5	0	0	14
6	0	0	14
7	0	0	8
8	7	2	13
9	0	0	15
10	4	0	15
11	0	0	12
12	11	0	18
13	8	0	18
14	6	5	16
15	10	9	17
16	14	1	19
17	15	0	21
18	13	12	24
19	0	16	20
20	19	0	27
21	17	0	22
22	21	0	23
23	22	0	26
24	18	0	27
25	0	0	26
26	23	25	29
27	24	20	28
28	27	0	29
29	26	28	30
30	29	0	31
31	30	3	0

Cont. TABLA (19)

DENDOGRAMA DEL METODO DE UNION EXHAUSTIVA
(Etapas K=1,2,...17, para K>17 grupos fuera de rango)

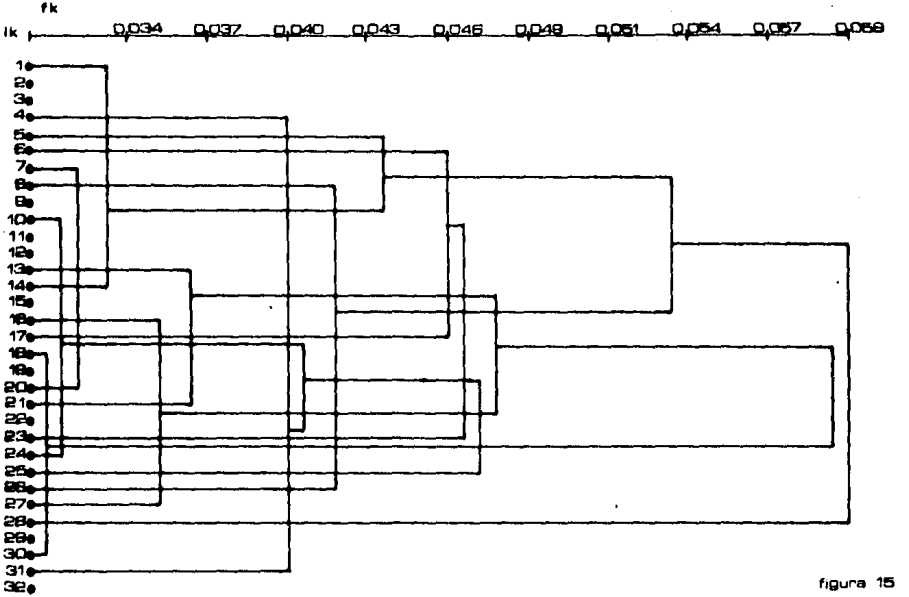


figura 15

con el elemento I_{23} en las etapas 11 y 12 respectivamente. Este grupo posteriormente (etapa 18) se une al grupo $\P_4$.

6) $\P_{12} = \{I_{12}, I_{32}\}$ este grupo se forma en la etapa 20, puede considerarse como parte del grupo, $\P_{13}$ ya definido, pero como su unión con este involucra un grado de disimilitud "alto" (.08) a comparación con el grado de disimilitud con el que se hicieron las demás uniones en este grupo.

7) $\P_2 = \{I_2, I_{19}\}$ este grupo se forma en la etapa 25 con el grado de disimilitud mayor (.14) al resto de los grupos, puede considerarse separado, o bien como parte del grupo $\P_1$ definido, ya que se une a éste en la etapa siguiente a su formación.

Por último, se tienen grupos de un sólo elemento, los cuales se pueden unir a los grupos ya existentes pero con un grado bastante "alto" de disimilitud, esto sería como sigue:

Grupo:	Se une a:	Con grado de disimilitud:
$\P_{15} = \{I_{15}\}$	$\P_1$	0.095
$\P_3 = \{I_3\}$	$\P_1$	0.124
$\P_{22} = \{I_{22}\}$	$\P_1$	0.125
$\P_{29} = \{I_{29}\}$	$\P_4$	0.133
$\P_{11} = \{I_{11}\}$	$\P_4$	0.191
$\P_9 = \{I_9\}$	$\P_1$	0.279

No se considera "buena" otra alternativa de agrupamiento ya que crece en gran proporción el grado de disimilitud (fuerza de agrupamiento) - al unir algunos de los grupos ya mencionados.

Se puede observar que en la solución obtenida por el método jerárquico de Unión-Exhaustiva, los grupos que se forman conforme van creciendo parecen alejarse de los elementos individuales que no están aún agrupados, por lo tanto están más propensos a formar el núcleo de un grupo nuevo - (ver 5.5.2).

7.3.3 Análisis de Resultados del Método de la Mediana

La solución obtenida por el método jerárquico de la Mediana es casi la misma que en los dos métodos anteriores (sobre todo en el de Unión-Exhaustiva) con diferencia en el comportamiento de los cúmulos o grupos de un sólo elemento que se obtienen al final. Esto se ve en la Tabla (20) y en el dendograma de la Figura (16).

Una observación importante acerca de la fuerza de agrupamiento (grado disimilitud) en este método (en general métodos de centroide) es que ésta puede incrementarse o decrementarse de etapa en etapa, lo que origina ramas que van en reversa en su dendograma.

Los grupos que forman la solución por el método de la Mediana son - los siguientes:

1) $\pi_{13} = \{I_{13}, I_{16}, I_{27}, I_{18}, I_{30}, I_{21}\}$ este grupo se forma de la unión del grupo $\{I_{18}, I_{30}\}$ en la etapa 1 con $\{I_{16}, I_{27}\}$ en la etapa 5 y con -

METODO DE LA MEDIANA

ETAPA	CUMULO INICIAL	CUMULO FINAL	FZA AGRUPAMIENTO
1	18	30	2.148267E-02
2	10	24	2.240599E-02
3	7	20	3.287578E-02
4	1	14	3.373813E-02
5	16	27	3.593847E-02
6	16	18	3.045296E-02
7	13	16	3.154816E-02
8	13	21	2.885306E-02
9	4	31	3.994072E-02
10	4	25	3.411613E-02
11	8	26	4.158026E-02
12	1	5	4.378875E-02
13	4	10	4.429911E-02
14	6	17	4.552141E-02
15	1	28	4.627485E-02
16	1	8	4.455839E-02
17	6	23	.0516081
18	4	6	5.849897E-02
19	13	29	6.575731E-02
20	1	3	6.600064E-02
21	12	32	8.285375E-02
22	1	4	9.554992E-02
23	12	13	9.175914E-02
24	1	22	9.661655E-02
25	15	19	.1107824
26	2	15	.1145349
27	1	11	.123807
28	1	12	.1502028
29	1	7	.2233815
30	2	9	.2296233
31	1	2	.3491708

TABLA (20)

ETAPA	PRE-ETAPA II	PRE-ETAPA JJ	ETAPA II SIGUE
1	0	0	6
2	0	0	13
3	0	0	29
4	0	0	12
5	0	0	6
6	5	1	7
7	0	6	8
8	7	0	19
9	0	0	10
10	9	0	13
11	0	0	16
12	4	0	15
13	10	2	18
14	0	0	17
15	12	0	16
16	15	11	20
17	14	0	18
18	13	17	22
19	8	0	23
20	16	0	22
21	0	0	23
22	20	18	24
23	21	19	28
24	22	0	27
25	0	0	26
26	0	25	30
27	24	0	28
28	27	23	29
29	28	3	31
30	26	0	31
31	29	30	0

Cont. TABLA (20)

DENDOGRAMA DEL METODO DE LA MEDIANA

(Etapas k=1,2,...20), para k>20 grupos fuera de rango)

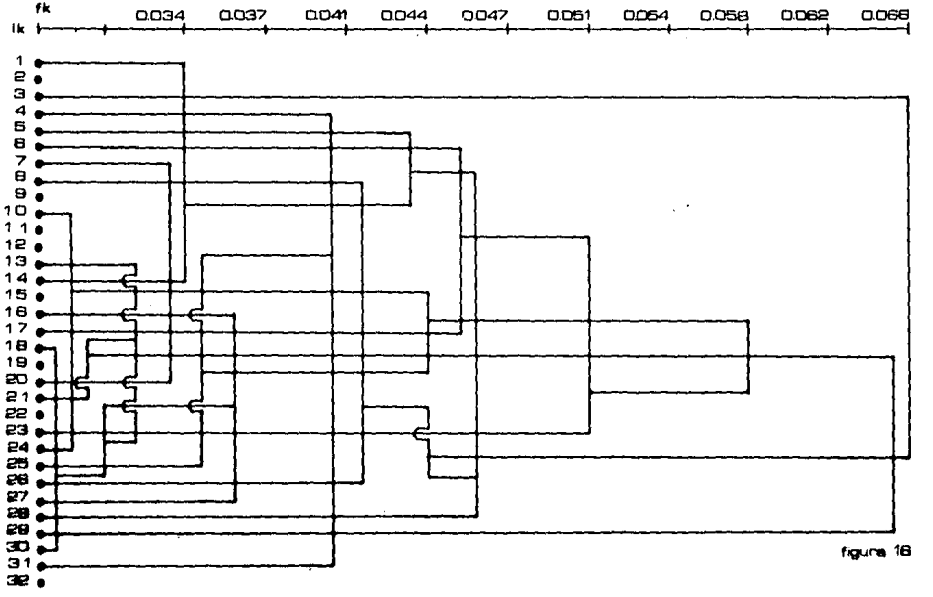


figura 16

$\{I_{13}, I_{21}\}$ en la etapa 8.

2) $\P_4 = \{I_4, I_{31}, I_{25}, I_{10}, I_{24}\}$ este grupo se forma de la unión del grupo $\{I_{10}, I_{24}\}$ en la etapa 2 con el grupo $\{I_4, I_{31}\}$ de la etapa 10 y finalmente el elemento I_{25}.

3) $\P_7 = \{I_7, I_{20}\}$ este grupo como en los métodos anteriores se considera aislado, ya que se forma en la etapa 3 y no se vuelve a unir hasta la etapa 29 con el grupo $\P_1$ con un grado de disimilitud de 0.22.

4) $\P_1 = \{I_1, I_{14}, I_5, I_{28}, I_8, I_{26}\}$ este grupo se forma de la unión de -- $\{I_1, I_{14}\}$ de la etapa 4 con los elementos $\{I_5\}$, e $\{I_{28}\}$ de la etapa 12 y 15 respectivamente y finalmente con el grupo $\{I_8, I_{26}\}$ de la - etapa 11.

5) $\P_6 = \{I_6, I_{17}, I_{23}\}$ este grupo se forma de la unión del grupo $\{I_6, I_{17}\}$ de la etapa 14 con el elemento $\{I_{23}\}$ en la etapa 17. Este grupo - puede formar parte del grupo $\P_4$ definido ya que se une a éste en - la siguiente etapa.

6) $\P_{12} = \{I_{12}, I_{32}\}$ este grupo se forma con un grado de disimilitud de 0.083 "alto" en la etapa 21, puede formar parte del grupo $\P_{13}$ definido ya que se une a este en la etapa 23 pero con un grado de disimilitud mayor al resto de los elementos de este grupo (0.092), por lo que se considera independiente.

7) $\P_2 = \{I_2, I_{15}, I_{19}\}$ es un grupo que se forma de la unión del grupo - $\{I_{15}, I_{19}\}$ de la etapa 25 con el elemento $\{I_2\}$ en la etapa 26, con - un grado de disimilitud mayor a los grupos ya definidos (0.114).

La solución tiene entonces 12 grupos, de las cuales 5 son grupos - de un sólo elemento que si se unen a los grupos ya existentes lo harán - con un grado de disimilitud (fuerza de agrupamiento) muy "alto", como sigue:

Grupo:	Se une a:	Con un grado de disimilitud:
$\P_{29} = \{I_{29}\}$	$\P_{13}$	0.065
$\P_3 = \{I_3\}$	$\P_1$	0.066
$\P_{22} = \{I_{22}\}$	$\P_1$	0.097
$\P_{11} = \{I_{11}\}$	$\P_1$	0.124
$\P_9 = \{I_9\}$	$\P_2$	0.230

Otra alternativa de agrupación no se considera "buena" ya que el - grado de disimilitud (fuerza de agrupamiento) conforme se van uniendo los grupos se incrementa en mayor proporción.

7.3.4 Comparación de los Resultados de los Tres Métodos

Como se puede ver en las secciones anteriores, la solución al problema de agrupar las 32 entidades federativas de acuerdo a sus 9 ramas de actividad económica, es casi la misma en los tres métodos (Unión-Simple, Unión-Exhaustiva y Mediana) con ligeras variaciones. Los tres métodos coinciden en:

- El grupo $\P_7 = \{I_7, I_{20}\}$ que se forma en la etapa 3 (de ambos métodos) es un grupo aislado.

- El grupo $\P_{13} = \{I_{13}, I_{21}, I_{16}, I_{27}, I_{18}, I_{30}\}$ se forma de la unión de tres grupos es decir: $\{I_{18}, I_{30}\} \cup \{I_{16}, I_{27}\} \cup \{I_{13}, I_{21}\}$.

- El grupo $\P_4 = \{I_4, I_{31}, I_{10}, I_{24}, I_{25}\}$ se forma de la unión de los grupos $\{I_{10}, I_{24}\} \cup \{I_4, I_{31}\} \cup \{I_{25}\}$.

- El grupo $\P_1 = \{I_1, I_{14}, I_5, I_8, I_{26}, I_{28}\}$ se forma de la unión de los grupos $\{I_1, I_{14}\} \cup \{I_5\} \cup \{I_8, I_{26}\} \cup \{I_{28}\}$.

- El grupo $\P_6 = \{I_6, I_{17}, I_{23}\}$ se forma de la unión de los grupos --- $\{I_6, I_{17}\} \cup \{I_{23}\}$ o bien, puede considerarse como parte del grupo $\P_4$.

- Los grupos de un sólo elemento son: $\{I_{29}\}$, $\{I_3\}$, $\{I_{22}\}$, $\{I_{11}\}$ y $\{I_9\}$.

- El grupo que se une con mayor grado de disimilitud (fuerza de agrupamiento) al grupo final es I_9.

Los tres métodos difieren en:

- El grupo $\P_{12} = \{I_{12}, I_{32}\}$ se forma en el método de Unión-Exhaustiva y en el método de la Mediana pero no en el de Unión Simple. Este grupo en ambos casos puede formar parte del grupo $\P_{13}$ definido pero con un grado de disimilitud "alto".

- El grupo $\P_2$ se forma en el método de la Mediana como $\P_2 = \{I_2, I_{15}, I_{19}\}$, en el método de Unión-Exhaustiva este grupo cambia, es decir --- $\P_2 = \{I_2, I_{19}\}$, en el método de Unión-Simple no se unen estos ele-

mentos, sino se tienen como grupos de un sólo elemento.

- El rango en que se encuentra el grado de disimilitud (fuerza de agrupamiento) con el que se unen los grupos en los tres métodos es muy diferente. Es decir, para el método de Unión Simple su rango es (0.021, 0.254) ya que en este caso se toma la mínima distancia entre los elementos que se unen; el método de Unión Exhaustiva tiene un rango de (0.021, 0.539), mayor al anterior por tomar la máxima distancia entre los elementos que se unen; el método de la Mediana tiene un rango intermedio a los dos anteriores que es (0.021, 0.349) al tomar el punto medio del lado más pequeño del triángulo que se forma de los dos grupos ya unidos con el que se van a unir.

- El método de Unión Simple forma un sólo grupo en la etapa 17, que contiene los grupos $\P_{13}$, $\P_4$, $\P_1$, $\P_6$ (ya definidos), al que en etapas posteriores se le unen los grupos aislados y los de un sólo elemento. A esa tendencia a dar cúmulos en forma de serpiente se le llama encadenada (ver 5.5.1). Sin embargo, los otros métodos van formando varios grupos que se unen hasta las últimas etapas del proceso.

La solución obtenida por el método jerárquico de la Mediana es la que se tomó como solución final al problema de agrupar las 32 entidades federativas, por ser la que menos grupos de un sólo elemento tiene.

7.4 Comentarios Respecto a la Concentración y Distribución de Actividades en México

Los grupos o cúmulos obtenidos de alguna manera se pueden asociar con la concentración de actividades en diferentes regiones de México. Para ello, resulta de interés analizar:

- La localización geográfica de estos grupos.
- Algunos factores demográficos de estos grupos, como son migraciones, tipo de asentamientos humanos (población urbana y rural).
- Relacionar estos grupos con un esquema de desarrollo.
- Los propósitos del Plan Nacional de Desarrollo para las entidades que forman los grupos obtenidos.

En las siguientes subsecciones se comentarán estos puntos.

7.4.1 Localización Geográfica de los Cúmulos Obtenidos

Los 12 cúmulos obtenidos como solución al problema de agrupar las 32 entidades federativas de acuerdo a la distribución porcentual de 9 ramas de actividad económica, son los siguientes:

$\P_7$ = {Chiapas, Oaxaca}

$\P_{13}$ = {{Nayarit, Veracruz}, {Michoacán, Tabasco}, {Hidalgo, Puebla}}

$\P_4$ = {{Durango, San Luis Potosí}, {Campeche, Yucatán}, {Sinaloa}}

$\P_1$ = {{Aguascalientes, Jalisco}, {Coahuila}, {Tamaulipas}, {Chihuahua, Sonora}}

$\P_6$ = {{Colima, Morelos}, {Quintana Roo}}

$\P_{12}$ = {Guerrero, Zacatecas}

$\P_2$ = {{Baja California Norte}, {México, Nuevo León}}

$\P_{29}$ = {Tlaxcala}

$\P_3$ = {Baja California Sur}

$\P_{22}$ = {Querétaro}

$\P_{11}$ = {Guanajuato}

$\P_9$ = {Distrito Federal}

Los elementos de cada uno de estos grupos tienen una distribución de actividad económica "muy semejante" ya que pertenecen a un mismo - grupo.

Los cúmulos o grupos de entidades federativas se representa geográficamente en el mapa de la Figura (17), de donde se pueden hacer las siguientes observaciones:

- El grupo $\P_7$ que comprende a Chiapas y Oaxaca se localiza en el sur de la República Mexicana, son estados vecinos.

- El grupo $\P_{13}$ se localiza en el este de la República Mexicana a excepción de Nayarit y Michoacán que se encuentran hacia el occidente.

- El grupo $\P_4$, los estados de Yucatán y Campeche son vecinos en el - sur, así como en el noroeste Sinaloa y Durango.

- Las entidades del grupo $\P_1$ se localizan en el norte de la Repúbli-

SITUACION GEOGRAFICA DE LOS CUMULOS OBTENIDOS

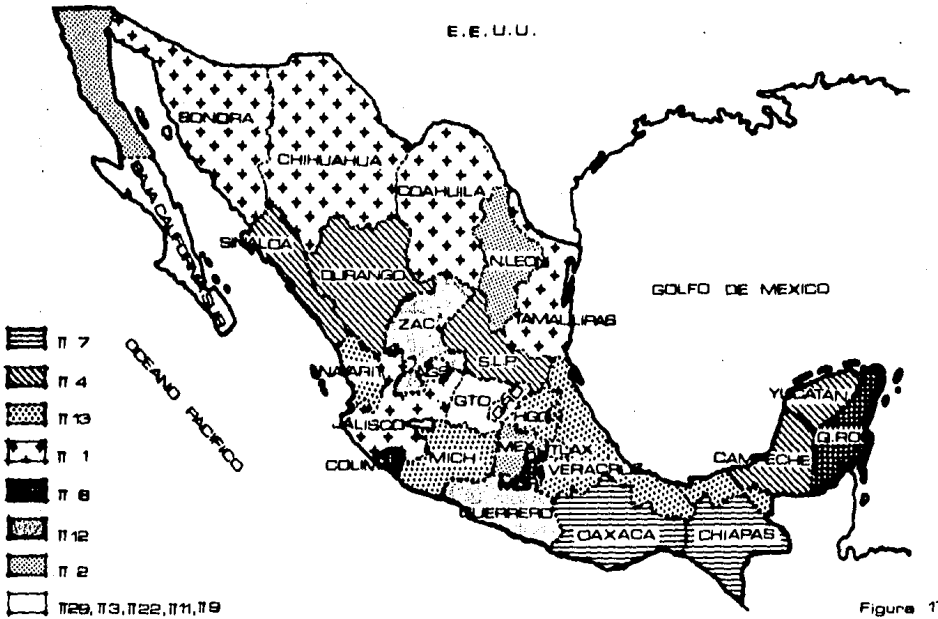


Figure 17

ca Mexicana, a excepción de Jalisco que se ubica en la región occidente. Sonora, Chihuahua y Coahuila son estados vecinos, así como Aguascalientes y Jalisco.

- Los estados que pertenecen a los grupos $\P_6$, $\P_{12}$ y $\P_2$ tienen una localización geográfica distante entre sí.

7.4.2 Esquema de Desarrollo

En la economía mundial los países desarrollados son aquellos que las actividades primitivas (agricultura, ganadería, silvicultura, minería, - caza y pesca) se encuentran mecanizadas y ocupan "menos" mano de obra - sin que la producción que se obtiene de estas actividades descienda por esto.

Por otro lado, en estos países la mano de obra se concentra en actividades industriales, de comercio y servicios.

En México una gran parte de la población (37% de la población económicamente activa (PEA)) se dedica a la agricultura, ganadería, silvicultura y pesca[1]. En algunas entidades federativas estas actividades son - preponderantes (más de un 50% de su PEA), pero para otros más desarrollados estas actividades pasan a un segundo o tercer término con respecto a las demás actividades.

---

(1) A estas 5 actividades: Agricultura, Ganadería, Silvicultura, Caza y Pesca se incluyen al referirse a la Agricultura.

En base a esto se propone un esquema sobre el nivel de desarrollo[1] en el que se da un orden de importancia a las ramas de actividad económica. Este esquema se aplica a los resultados de agrupar las 32 entidades federativas, para ordenar los grupos obtenidos.

ESQUEMA DE DESARROLLO

En este esquema, el grado de importancia de una actividad se mide por la proporción de la PEA ocupada en dicha actividad.

Nivel 1: Las actividades preponderantes son la industria y los servicios en segundo término el comercio, las actividades agrícola y de construcción son de baja importancia.

Nivel 2: La industria y los servicios por separado, son del orden de importancia de la actividad agrícola, el comercio y la construcción son importantes.

Nivel 3: Los servicios y el comercio en conjunto, son del orden de importancia de la actividad agrícola, la industria y la construcción son importantes.

Nivel 4: Un poco menos de la mitad de la población se dedica a la actividad agrícola, la industria, los servicios y el comercio son importantes.

---

(1) Este esquema sólo considera ramas de actividad económica cuya población (PEA) sea mayor o igual a un 10% de la PEA.

Nivel 5: Un poco más de la mitad de la población se dedica a la actividad agrícola, y el resto a la industria, servicios y comercio.

Nivel 6: Preponderantemente agrícola, entre las otras actividades destacan los servicios y el comercio.

Nivel 7: Preponderantemente agrícola, el resto de actividades carece de importancia.

7.4.3 Clasificación de Entidades Federativas de Acuerdo al Esquema de Desarrollo

Los 7 niveles del esquema de desarrollo planteado en la sección anterior, básicamente corresponden a los grupos de entidades federativas que se describen en la sección 7.4.1. Esto se comprueba ubicando más específicamente estos resultados dentro del esquema de la Tabla (21), a partir de lo cual se obtienen las siguientes observaciones:

- En el nivel 1 se puede ubicar al grupo $\P_2$, el cual corresponde a los estados de Baja California Norte, México y Nuevo León; en éstos, las actividades preponderantes son la Industria y los Servicios. Son estados de fuerte atracción migratoria [11], cuya población en su mayor parte es urbana [12].

- En el nivel 2 se puede ubicar al grupo $\P_1$, con los estados de Aguascalientes, Jalisco, Coahuila, Chihuahua, Sonora y Tamaulipas. Como ya se mencionó, con excepción de Jalisco, son estados del norte de la República Mexicana, cuyas actividades preponderantes son la agricultura, los servicios y la industria; se distinguen por ser

TABLA (21)

ESQUEMA DE DESARROLLO PARA LOS GRUPOS DE ENTIDADES FEDERATIVAS

(Porcentajes Promedio)

NIVEL	GRUPO	AGRICOLA	SERVICIOS	INDUSTRIA	COMERCIO	CONSTRUCCION	MINERA	OTRAS
	$\P_9$	10%	12%	22%	*	17%	18%	21%
1	$\P_2$	15%	22%	30%	18%	10%	*	5%
2	$\P_1$	25%	21%	22%	15%	10%	*	7%
	$\P_3$	27%	25%	10%	17%	10%	*	11%
3	$\P_6$	37%	20%	12%	15%	10%	*	6%
4	$\P_4$	45%	17%	12%	12%	*	*	14%
	$\P_{29}$	50%	14%	20%	*	*	*	16%
5	$\P_{13}$	53%	13%	10%	10%	*	*	14%
6	$\P_{12}$	60%	15%	*	10%	*	*	15%
7	$\P_7$	74%	*	*	*	*	*	26%
2 - 3 ó	$\P_{11}$	34%	*	15%	*	12%	13%	26%
3 - 4	$\P_{22}$	37%	15%	22%	10%	*	*	16%

* Actividades poco significativas ya que tienen menos del 10% de la población (PEA) y van sumadas en la columna de Otras.

estados de equilibrio migratorio [11], es decir, sus inmigraciones son equivalentes a sus emigraciones, por lo que se consideran estados estables. La mayor parte de su población es urbana [12].

- En el nivel 3 se ubica el grupo π_6 con Colima, Morelos y Quintana Roo. En este nivel comienza a ser importante la actividad agrícola; sin embargo, la suma de las actividades comerciales y las de servicios es del orden de ésta. Son estados con fuerte atracción migratoria [11].

- En el nivel 4 se puede ubicar al grupo π_4, que comprende los estados de Durango, San Luis Potosí, Campeche, Yucatán y Sinaloa. Casi la mitad de la población (PEA) de estos estados se dedica a la actividad agrícola; sin embargo, los servicios, el comercio y la industria en conjunto son del orden de importancia de ésta. Campeche y Sinaloa son estados de equilibrio migratorio; Durango, San Luis Potosí y Yucatán son de fuerte expulsión migratoria [11]. En estos estados la mitad de su población es urbana y la otra mitad rural [12].

- En el nivel 5 se puede clasificar al grupo π_{13}, con los estados de Nayarit, Veracruz, Michoacán, Tabasco, Hidalgo y Puebla. En estos estados más de la mitad de su población (PEA) se dedica a la actividad agrícola y la demás se reparte en el resto de actividades. Veracruz, Michoacán y Puebla se encuentran en los 5 primeros lugares de los estados más poblados [10] dentro de la República Mexicana. Los estados que pertenecen a este grupo tienen diferente categoría migratoria, es decir: Tabasco es de débil atracción migratoria; Veracruz es de equilibrio migratorio; Puebla y Nayarit

son de débil expulsión; y Michoacán e Hidalgo son estados de fuerte expulsión migratoria [11]. Son estados cuya población es su gran mayoría es rural [12].

- En el nivel 6 se puede ubicar al grupo η_{12} con Guerrero y Zacatecas. Son estados cuya población (PEA) es básicamente agrícola, aunque los servicios y el comercio son importantes. Son estados de expulsión migratoria [11] y su población es casi en su totalidad rural [12].

- En el nivel 7 se tiene el grupo η_7, formado por los estados de Chiapas y Oaxaca, los que son preponderantemente agrícolas sin ninguna otra actividad de importancia. Son estados de fuerte expulsión migratoria [11]. Cuya población casi en su totalidad es rural [12].

El resto de los grupos son los que comprenden una sola entidad federativa y aunque podrían clasificarse en alguno de los niveles propuestos, utilizando como base la proporción de su PEA dedicada a la agricultura, esta ubicación estaría forzada. Lo anterior se debe a que las proporciones relativas de su PEA muestran peculiaridades que salen fuera del esquema propuesto, ya que este esquema es solamente una conceptualización de un continuo de valores de la distribución de actividades que se han reducido a 7 niveles.

En estos niveles fue posible ubicar a la mayor parte de las entidades federativas, para clasificar a las restantes se requeriría un mayor número de niveles y una jerarquización de actividades más completa, por

ejemplo, que dé un orden de importancia a las industrias extractivas.

Los 5 grupos de una sola entidad federativa se han ubicado en la - Tabla (21) entre dos niveles de los propuestos, pensando que están en - una etapa de transición que podría localizarlos en alguno de ellos. La única excepción a ésto es el Distrito Federal que muestra un comporta-- miento parecido en parte, al grupo del nivel 1, pero con la peculiari-- dad de una alta proporción de su PEA dedicada a la minería. En resumen, para estos grupos se tiene que:

- El grupo π_9 que es el Distrito Federal cuya actividad principal es la industrial y destaca significativamente las actividades mineras y de construcción. Es la entidad con más población dentro de la - República Mexicana [10] y es una entidad de equilibrio migratorio [11] cuya población es totalmente urbana. Por su bajo nivel de actividad agrícola este grupo se ha ubicado antes del nivel 1.

- El grupo π_3 que es el estado de Baja California Sur cuyas actividades principales (de igual peso) son la agrícola y la de servicios; es un estado de atracción migratoria [11], cuya población en su mayoría es urbana [12]. Este grupo podría estar en el nivel 2; sin embargo, la pequeña proporción de PEA dedicada a la industria lo situarían en el nivel 3, por lo tanto, se ha puesto entre dos niveles.

- El grupo π_{11} es el estado de Guanajuato en el que se destaca así como en el Distrito Federal su actividad minera, que en conjunto a su actividad industrial son del orden de su actividad agrícola.

Este estado se encuentra en el 7º lugar de mayor población en la - República Mexicana [10], es una entidad de equilibrio migratorio - [11], cuya población se divide por igual en rural y urbana [12]. - Por sus actividades agrícola e industrial podría ubicarse en el nivel 3; sin embargo, las actividades de servicios y de comercio corresponden al nivel 7. En forma tentativa se podría ubicar entre los niveles 3 y 4, o bien, entre los niveles 2 y 3.

- El grupo π_{22} es el estado de Querétaro, cuya industria y servicios en conjunto son del orden de su actividad agrícola; es un estado - de débil expulsión migratoria [11], cuya población en su mayoría - es urbana. Este grupo tiene una actividad industrial importante; sin embargo, la construcción, los servicios y el comercio tienen - niveles bajos. Es de esperarse que este grupo pase al nivel 2, podría clasificarse entre los niveles 3 y 4 o bien, entre los niveles 2 y 3.

- El grupo π_{29} que es el estado de Tlaxcala, la mitad de su PEA se - dedica a la actividad agrícola y tiene una actividad industrial importante. Es un estado de fuerte expulsión migratoria [11], cuya población se divide por igual en rural y urbana [13]. Este grupo se clasificó entre los niveles 4 y 5, aunque su actividad industrial lo situaría en niveles más altos.

7.4.4 Comentarios Sobre el Plan Nacional de Desarrollo por Entidad Federativa

En esta sección se buscará una relación de los lineamientos del -

Plan Nacional de Desarrollo en su parte de Política Regional, con los resultados obtenidos en el esquema de desarrollo propuesto en la sección anterior y así analizar, las acciones de dicho plan en las actividades económicas de cada entidad federativa.

LINEAMIENTOS DEL PLAN NACIONAL DE DESARROLLO

En la parte de Política Regional del Plan Nacional de Desarrollo, se aspira a la obtención de un desarrollo estatal integral que incluya a la totalidad de las entidades federativas sobre todo a las menos desarrolladas, con la intención de que cuenten con la capacidad económica y administrativa que les permita alcanzar mayores niveles de bienestar y progreso. Algunas de las acciones que se emprenderán para el logro de este objetivo, son las siguientes:

- Desarrollo agrícola y pesquero en Sonora y Sinaloa que permitirán sustentar la integración de actividades industriales y de servicios. (Nivel 2 y 4).

- La ampliación de recursos se orientará al desarrollo de la agricultura de temporal, de la explotación forestal y minera, en Chihuahua, Durango y Zacatecas. También se impulsará la fruticultura como opción de creación de empleos. (Niveles 2, 4 y 6).

- En el noreste existen condiciones de integración favorables dado el tamaño económico de Monterrey y de las potencialidades agrícolas, ganaderas, mineras e industriales de Coahuila. Por lo que resulta necesario aprovechar estas perspectivas para la descentralización regional. (Nivel 1 y 2).

- Se intensificará la explotación nacional de recursos pesqueros, mineros y turísticos en las costas de Baja California (Norte y Sur). (Nivel 1 y 2).

- Se racionalizará el crecimiento urbano de las ciudades del sureste afectados por el auge petrolero y se dará prioridad a la previsión de infraestructura, vivienda y servicios. (Niveles 4 y 5).

- Se fomentará la integración entre las regiones de Nayarit y Oaxaca; la integración de circuitos turísticos de los centros a la costa, con el objeto de incrementar la afluencia de visitantes. (Nivel 2, 3, 5, 6, 7).

- Manzanillo consolidará sus funciones comerciales y de servicios convirtiéndose en el puerto principal del centro y occidente del país. (Nivel 3).

- Nuevas infraestructuras de transporte y abasto para fomentar la integración de Jalisco y Puebla. (Nivel 2 y 5).

- Se establecerá una estrategia común de apoyo a las comunidades campesinas de las sierras de Guerrero y de Oaxaca, tendientes a racionalizar la explotación de recursos naturales. (Nivel 6 y 7).

- La reordenación de la zona metropolitana de la Cd. de México para asegurar un desarrollo regional más equilibrado. (Nivel 1).

- Se fomentará el crecimiento de las industrias de bienes de consumo tradicional en áreas urbanas que ya presentan un desarrollo signifi

cativo de Guanajuato, Jalisco, Puebla, Tlaxcala y Veracruz. (Nivel 2 y 5).

- Se promoverá el desarrollo de los servicios profesionales y técnicos en las ciudades medias y ciudades mayores – Guadalajara, Puebla y Monterrey – reforzando el contrapeso de estas ciudades frente a la capital. (Nivel 1, 2 y 5).

- Con el objeto de frenar las migraciones hacia la zona metropolitana se buscará fortalecer las condiciones de desarrollo rural en zonas de expulsión migratoria. (Nivel 1).

- Se hará una integración de las economías regionales en el occidente incorporando el estado de Jalisco con los centros de Aguascalientes y San Luis Potosí para aprovechar más el puerto de Manzanillo para las relaciones con otras regiones y con el exterior. (Nivel 2, 3 y 4).

- En la región del Golfo se integrarán las economías de los puertos - con el área de Puebla y Tlaxcala y se reforzará esta última como articulación estratégica entre la costa y el altiplano. (Nivel 5).

- Se deberá restringir en forma severa pero selectiva el crecimiento de actividades en la ciudad de México. (Nivel 1).

- Se impulsará el aprovechamiento de recursos naturales y el desarrollo agroindustrial en los estados de Hidalgo, México y Morelos, dando prioridad a la consolidación y desarrollo de cuencas lecheras, -

la avicultura, producción de forrajes, producción de hortalizas y - su industralización. (Nivel 1, 3 y 5).

- Se reordenará la urbanización en la zona metropolitana de la ciudad de México y su periferia para albergar en forma adecuada el futuro crecimiento demográfico, para ello, las políticas de transporte, suelo y localización industrial se considerarán en forma conjunta. (Nivel 1).

- En el Distrito Federal se establecerán medidas que orienten la actividad industrial en forma selectiva. (Nivel 1).

- En el Distrito Federal se pretende disminuir los altos índices de - rezago y pobreza, el crecimiento demográfico y las migraciones tratando de revertirlos, para lograr un equilibrio en todo el país, por medio de la descentralización de la vida nacional. (Nivel 1).

COMENTARIOS

El esquema de desarrollo propuesto en la sección anterior se considera limitado, ya que los datos censales corresponden a una fecha determinada y en base a ésto se clasificó a las entidades federativas en los diferentes niveles. La obtención de esta información coincide con el - año (1982) en que fue propuesto el Plan Nacional de Desarrollo, por lo que se pretendió buscar una relación de este esquema con los lineamientos del Plan Nacional de Desarrollo.

En términos generales, no se considera haber encontrado una rela-

ción directa entre el nivel de desarrollo de cada entidad federativa con las acciones del Plan Nacional de Desarrollo, ya que una misma acción se propone para entidades de diferentes niveles o bien, acciones diferentes para entidades de un mismo nivel. Sin embargo, en algunos lineamientos del Plan Nacional de Desarrollo se proponen para entidades de un mismo - nivel, como en los siguientes casos:

- Racionalizar el crecimiento urbano de las ciudades del sureste afectadas por el auge petrolero, estas ciudades se localizan en los estados de los niveles 4 y 5.

- Se fomentará el crecimiento de las industrias de bienes de consumo tradicional en áreas urbanas que presentan desarrollo significativo de Guanajuato y Jalisco en el nivel 2; Puebla, Tlaxcala y Veracruz en el nivel 5.

- Frenar las migraciones y reordenar la urbanización en el Distrito Federal y Estado de México que son entidades que pertenecen al - nivel 1.

- Integrar las economías de los puertos con los estados de Puebla y Tlaxcala, que pertenecen al nivel 5.

7.4.5 El Esquema de Desarrollo y la Probabilidad de Muerte Infantil

La dinámica poblacional se encuentra determinada por la acción de - diversos fenómenos, entre los que destacan los llamados demográficos. Estos se constituyen en factores fundamentales del cambio en el tamaño y - composición de la población. Un factor de este tipo que ha sido estudiado ampliamente es la mortalidad, ya que esta representa uno de los elementos prioritarios e indispensables en el conocimiento de la evolución poblacional de cualquier país.

Varios estudios han mostrado una estrecha relación entre los indicadores de mortalidad y los del desarrollo económico y social. En esta línea, la mortalidad infantil se ha identificado como el parámetro de mayor sensibilidad a los cambios socioeconómicos. Por esta razón, se compara el nivel de desarrollo (ver sección 7.4.3) de los grupos de entidades federativas propuestos en el esquema de desarrollo por ocupación, con el - nivel de desarrollo de dichos grupos determinando por su probabilidad - de muerte infantil[1]

Cabe mencionar, que los datos de probabilidad de muerte infantil para cada entidad federativa [13], son para 1970 ya que no se dispone de - la información para 1980, en que los datos de población por rama de actividad económica [10] fueron obtenidos. Por consiguiente, se toma en consideración que en 10 años el desarrollo económico de un cierto estado o

(1) Defunciones de menores de un año entre nacidos vivos.

entidad federativa puede cambiar, aunque es de esperarse que este cambio no sea radical.

La probabilidad de muerte infantil para cada uno de los grupos de entidades federativas obtenidos en la sección 7.3 se calcula, considerando su población y su tasa de natalidad[1] para el año de 1970 [14], de la siguiente manera:

Sea: $P_M(\pi_k)$ la probabilidad de muerte infantil para el grupo π_k.

$P_M(I_i)$ la probabilidad de muerte infantil para el estado I_i.

$TN(I_i)$ tasa de natalidad para el estado I_i.

n_k número de estados que pertenecen al grupo π_k.

$Pob(I_i)$ población del estado I_i.

entonces se tiene:

$$P_M(\pi_k) = \frac{\sum_{i=1}^{nk} TN(I_i)\ Pob(I_i)\ P_M(I_i)}{\sum_{i=1}^{nk} TN(I_i)\ Pob(I_i)}$$

De lo anterior se obtiene:

$P_M(\pi_2) = 58.62$

$P_M(\pi_1) = 58.08$

$P_M(\pi_6) = 63.56$

---

(1) Nacimientos por cada mil habitantes.

$P_M(\P_4) = 62.09$

$P_M(\P_{13}) = 73.79$

$P_M(\P_{12}) = 65.27$

$P_M(\P_7) = 103.37$

y para los grupos de una sola entidad federativa se tiene:

$P_M(\P_9) = 54.85$

$P_M(\P_3) = 54.55$

$P_M(\P_{11}) = 65.00$

$P_M(\P_{22}) = 69.25$

$P_M(\P_{29}) = 64.70$

Con estos datos, se establecen niveles de desarrollo para los grupos en base a su probabilidad de muerte infantil, y se comparan con los niveles ya propuestos en el esquema de desarrollo de la sección 7.4.3. Esto se muestra en la tabla (22), en la que se observa que el nivel de desarrollo determinado por la probabilidad de muerte infantil, se intercambia en pares, con respecto al nivel obtenido de la ocupación de los grupos, sólo el grupo $\P_7$ permanece en el último nivel para ambos casos. Como ya se mencionó, se puede deber a que en 10 años el nivel de desarrollo para algunos estados ha mejorado, como son los grupos $\P_2$, $\P_6$ y $\P_{13}$, ya que de los niveles 2, 4 y 6 en 1970 han pasado a los niveles 1, 3 y 5 respectivamente, Sin embargo, ésta es sólo una posible razón.

Los grupos de una sola entidad federativa algunos mejoran su nivel como $\P_{11}$ y $\P_{22}$; otros permanecen en el mismo nivel $\P_9$ y $\P_{29}$; y el grupo

NIVEL DE DESARROLLO DE LOS GRUPOS POR OCUPACION Y POR PROBABILIDAD DE MUERTE INFANTIL

GRUPO	NIVEL POR OCUPACION (1980)	NIVEL POR PROBABILIDAD DE MUERTE INFANTIL (1970)
$\P_2$	1	2
$\P_1$	2	1
$\P_6$	3	4
$\P_4$	4	3
$\P_{13}$	5	6
$\P_{12}$	6	5
$\P_7$	7	7
$\P_9$	1	1
$\P_3$	2	1
$\P_{11}$	2-3 ó 3-4	4-5
$\P_{22}$	2-3 ó 3-4	5-6
$\P_{29}$	4	4

TABLA (22)

$\P_3$ pasa del nivel 1 al 2.

Estos resultados son relativos ya que por un lado la población por rama de actividad se obtiene en una fecha dada (del censo), y los datos sobre mortalidad infantil se obtienen en un cierto intervalo de tiempo.

Sin embargo, el nivel de desarrollo medido por la mortalidad infantil y el medido por la proporción que representan las diversas ocupaciones, guardan una "buena" congruencia, i.e., no hay cambios radicales. - Lo anterior da un cierto grado de confianza al uso de la distribución de la ocupación por ramas para determinar el nivel de desarrollo de los grupos de entidades.

CAPITULO 8

COMENTARIOS Y CONCLUSIONES FINALES

En este capítulo se dará un resumen general de los temas desarrollados en los capítulos precedentes, haciendo énfasis en lo logrado, y mencionando sus limitaciones. También, se proponen temas abiertos a investigación que se desprenden de este trabajo.

8.1 Comentarios Sobre el Trabajo Realizado

Como ya se expuso en el transcurso de esta tesis, el problema de cúmulos puede ser resuelto de varias maneras; ya sea usando métodos propios del análisis de cúmulos como son los métodos jerárquicos, o bien, usando técnicas de programación matemática. Para poder aplicar estos últimos, el problema de cúmulos se planteó como uno de optimización. El planteamiento no es único ya que se puede elegir desde la medida de asociación que determina el grado de semejanza entre dos individuos o elementos, hasta el criterio que se desea utilizar para encontrar la partición óptima.

Debido a lo anterior, se proporcionaron diferentes medidas de asociación y criterios de optimalidad. Cabe mencionar, que aún cuando se presentaron medidas y criterios de tipo estadístico, éstos no se analizan con detenimiento, ya que el enfoque de la tesis es más bien de Investigación de Operación que estadístico.

Uno de los objetivos centrales de la tesis fue proporcionar métodos de programación matemática para resolver el problema de cúmulos; la mecánica de éstos se presentó y se ilustró mediante la solución de ejemplos sencillos. Se comentó el hecho de que el tamaño del problema de cúmulos crece aceleradamente conforme aumenta el número de elementos que se desea agrupar, sobre todo en los métodos de Enumeración Exhaustiva, Programación Dinámica y Ramificación y Acotamiento. Debido a lo anterior, los problemas de interés práctico deben ser resueltos con la ayuda de un computador, ya que es prácticamente imposible resolverlos "a mano"; incluso, se encuentran problemas considerables para su solución con microcomputadoras personales.

Para ilustrar la utilidad del análisis de cúmulos, se decidió resolver un problema práctico; sin embargo, la utilización de un método de optimización requiere del uso de paquetería, de la que la autora no pudo disponer, o bien, de un esfuerzo de programación que iba más allá de los objetivos de esta tesis. Debido a lo anterior, se optó por utilizar un método heurístico, el presentado en este trabajo como método jerárquico.

Otra razón por la que se decidió programar un algoritmo para métodos jerárquicos fue, el que es posible programarlos para ser usado en microcomputadores personales, ya que éstos tienen suficiente capacidad para manejar el problema jerárquico.

Usando el algoritmo programado se resolvió una aplicación, en la que se busca agrupar las 32 entidades federativas de la República Mexica

na de acuerdo a la ocupación de sus habitantes, por rama de actividad - económica. Se presenta un esquema de desarrollo basado en la importan-- cia relativa de las distintas ramas de actividad, se compara con un esquema comúnmente aceptado, basado en la mortalidad infantil, y por último, se clasifican los grupos obtenidos del análisis en base a éste. Lo anterior, parece a la autora ser de utilidad, ya que permite asignar a - cada grupo un nivel de desarrollo.

8.2 Posibles Extensiones del Trabajo Realizado

De lo expuesto en este trabajo, es evidente que el problema de agrupamiento es sumamente complejo; aún técnicas aparentemente "bien establecidas", como el análisis de cúmulos, presentan un sinnúmero de alternativas y problemas aún no resueltos. Debido a ésto, no es difícil mencionar temas de investigación sobre este problema. A continuación se enumeran sólo unos cuantos problemas, que están íntimamente relacionados con este trabajo y que no fueron resueltos en el mismo, ya que aquí sólo se pretendió sentar las bases para iniciar investigaciones futuras en una - faceta del problema; su solución utilizando programación matemática.

Así pues, se pueden mencionar como posibles temas de investigación futura, los siguientes:

- Desarrollar un análisis más exhaustivo del comportamiento de las variables que definen a los elementos en el problema de cúmulos, ya - que en este trabajo sólo se consideraron variables cuantitativas y continuas.

- Dar un enfoque estadístico al análisis de cúmulos. Como por ejemplo: obtener funciones de distribución para los grupos y hacer pruebas de hipótesis que comprueben que un elemento dado, pertenece a un cierto grupo.

- Hacer una comparación práctica de diferentes algoritmos para resolver el problema de cúmulos, y decidir por ejemplo cúal es el más eficiente, el más rápido (en computadora), el que utiliza menos memoria, etc.

- Utilizar diferentes medidas de asociación que determinen el grado de semejanza entre los elementos a agrupar, y así, analizar y comparar los diferentes agrupamientos que se obtuvieran como resultado.

ANEXO A

La generalización del comportamiento de las funciones de distancia para p dimensiones es el siguiente:

Todo vector X_i en R^p puede expresarse en término de una base. Sea $\beta=\{e_1,e_2,\ldots,e_p\}$ una base de R^p a la que se le llama Canónica, cuando sus vectores son definidos:

$$e_1 = (1,0,0,\ldots,0)$$
$$e_2 = (0,1,0,\ldots,0)$$
$$\vdots$$
$$e_k = (0,0,\ldots,0,1,0,\ldots,0) \quad (k\text{-ésimo})$$
$$\vdots$$
$$e_p = (0,0,\ldots,0,1)$$

tales vectores son linealmente independientes y son unitarios --- ($|e_i| = 1 \;\forall\; i$), por lo tanto $\beta=\{e_i\}_{i=1}^{p}$ es un conjunto de vectores ortonormales.

Por lo tanto, todo vector $X_i \in R^p$ se puede expresar en términos de la Base Canónica β, es decir

$$X_i = \sum_{k=1}^{p} e_k X_{ki}$$

Al rotarse X_i, su sistema de referencia cambia, por lo que cambia de Base X_i' denota el vector rotado que se determina por una base --- $\beta'=\{e_k'\}_{k=1}^{p}$ formada de vectores ortonormales $\{e'_k\}_k$ representará un sis

tema de referencias nuevo, que se originó de la rotación del sistema original un ángulo ϕ.

En la Figura (18) se observa que X_i equivale a X'_i pero en diferentes sistemas de refencias.

Por lo tanto $X'_i = \sum_{k=1}^{p} X'_{ki}\, e'_k$

Como se dijo, $X_i = X'_i$ pero con diferentes sistemas de referencias.

Por lo tanto $\sum_k X_{ki}\, e_k = \sum_k X'_{ki}\, e'_k$

$$\sum_k X_{ki}\, e_k\, e'_\ell = \sum_k X'_{ki}\, e'_k\, e'_\ell$$

como $e'_k \cdot e'_\ell = \delta_{k\ell}$ donde $\delta_{k\ell} = \begin{cases} 1 & k = \ell \\ 0 & k \neq \ell \end{cases}$

$$\Longrightarrow \sum_k X_{ki}\, e_k \cdot e'_\ell = X'_{\ell i}$$

$$\therefore\ X'_{\ell i} = \sum_{k=1}^{p} [e'_\ell \cdot e_k]\, X_{ki} \quad \forall\ \ell = 1, \ldots, p$$

$$\Longrightarrow X'_i = \underline{E} \cdot X_i$$

$$X'_i = \underline{E} \cdot X_i = \begin{bmatrix} e'_1 \cdot e_1 & e'_1 \cdot e_2 & \cdots & e'_1 \cdot e_p \\ e'_2 \cdot e_1 & e'_2 \cdot e_2 & \cdots & e'_2 \cdot e_p \\ \vdots & \vdots & & \vdots \\ e'_p \cdot e_1 & e'_p \cdot e_2 & \cdots & e'_p \cdot e_p \end{bmatrix} \begin{bmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{pi} \end{bmatrix}$$

$\underline{E}$ será entonces la matriz de Rotación en un espacio p-dimensional.

En términos generales la Rotación es una transformación rígida ya que preserva la Distancia Euclidiana (norma ℓ_2).

DIFERENTES SISTEMAS DE REFERENCIAS PARA UN MISMO VECTOR

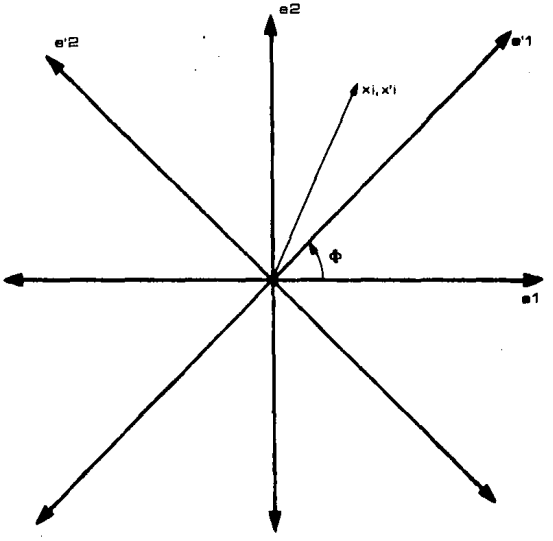
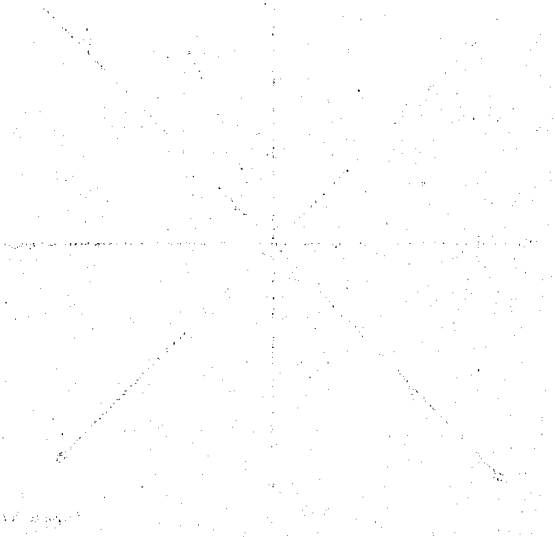


Figura 18

Para las otras normas no se puede decir nada, sólo que dependen del ángulo ϕ de Rotación.



ANEXO B

Para un problema más general, una métrica apropiada corresponde a - la que transforma las variables a no correlacionadas y de igual varianza, es decir, mapear la matriz de varianza covarianza a la matriz identidad.

Sea la matriz de varianza covarianza:

$$S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})\ (X_i - \bar{X})^T$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Se desea una matriz de transformación W (de pesos) tal que:

$$S \xrightarrow{W} I$$

Si $Y_i = WX_i$ su matriz de varianza covarianza para Y_i es $WSW^T = I$.

Las transformaciones que obtienen variables no correlacionadas son las que diagonalizan la matriz de varianza covarianza.

Si $X_i' = \phi^T X_i$ cuya matriz de varianza covarianza (simétrica) será - $\phi^T S \phi = \Lambda = diag(\lambda_1, \ldots, \lambda_p)$ donde ϕ^T es la matriz de transformación - ortonormal (ortogonales y unitarios).

Después de una transformación ortonormal, las varianzas de las variables que describen a los elementos están dadas por los valores característicos $(\lambda_1 \ldots \lambda_p)$.

Para que dichas variables tengan igual varianza sólo se necesita - una transformación que reescale $\Lambda^{-1/2}$ matriz diagonal con $\lambda_i^{-1/2}$ como - el i-ésimo elemento diagonal.

En conclusión, $\Lambda^{-1/2}$ es una transformación ortonormal 'blanca' ya que en primer lugar rota las coordenadas de la matriz de varianza covarianza de las variables para obtener la matriz diagonal de var-cov y - luego escalar dichas coordenadas y obtener la matriz identidad:

1) $X_i' = \Phi^T X \qquad S \Longrightarrow \Phi^T S \Phi = \Lambda$

2) $Y_i' = \Lambda^{-1/2} X_i' \;,\quad \Lambda \Longrightarrow \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I$

La transformación combinada es:

$$Y_i = \Lambda^{-1/2} \Phi^T X_i = WX_i$$

donde W es la matriz de 'peso', la métrica generalizada en el espacio de X_i corresponde a la métrica Euclidiana en el espacio de Y_i y es una medida apropiada de distancia para agrupar

$$d_2^2 (Y_i, Y_j) = d_2^2 (W X_i, W X_j)$$

$$= (W(X_i - X_j))^T W(X_i - X_j)$$

$$= (X_i - X_j)^T W^T W(X_i - X_j)$$

La matriz $W^T W = Q$ de lo que se deduce lo siguiente:

$$Q = (\Lambda^{-1/2}\ \Phi^T)^T\ \Lambda^{-1/2}\ \Phi^T$$

$$= \Phi\ \Lambda^{-1/2\,T}\ \Lambda^{-1/2}\ \Phi^T$$

$$= \Phi\ \Lambda^{-1}\ \Phi^T$$

$$\Longrightarrow Q = S^{-1}$$

Q es la inversa de la matriz de varianza covarianza, por lo que - la métrica de Mahalanobis para X_i y X_j es:

$$D_M(X_i, X_j) = [(X_i - X_j)^T\ S^{-1}\ (X_i - X_j)]^{1/2}$$

ANEXO C

Demostración de la igualdad entre las siguientes medidas de asociación:

$$\frac{1}{2n_k} \sum_{i=1}^{nk} \sum_{j=1}^{nk} d^2(X_i,X_j) = \frac{1}{2n_k} \sum_{i} \sum_{j} (X_i-X_j)^T (X_i-X_j)$$

Si $(X_i-X_j)^T (X_i-X_j) = X_i^T X_i - X_i^T X_j - X_j^T X_i + X_j^T X_j$

entonces se tiene:

$$= \frac{1}{2n_k} \sum_{i} [n_k X_i^T X_i - \sum_{j} X_i^T X_j - \sum_{j} X_j^T X_i + \sum_{j} X_j^T X_j]$$

$$= \frac{1}{2} [\sum_{i} X_i^T X_i - \sum_{i} X_i^T \bar{X}^k - \sum_{j} X_j^T \bar{X}^k + \sum_{j} X_j^T X_j]$$

$$= \frac{1}{2} \sum_{i} [X_i^T X_i - X_i^T \bar{X}^k - X_i^T \bar{X}^k + X_i^T X_i]$$

$$= \sum_{i} [X_i^T X_i - X_i^T \bar{X}^k] = \sum_{i} X_i^T X_i - (\sum_{i} X_i^T) \bar{X}^k$$

$$= \sum_{i} X_i^T X_i - n_k \bar{X}^{k\,T} \bar{X}^k = \sum_{i} X_i^T X_i - n_k (\bar{X}^k)^2$$

por lo tanto:

$$\frac{1}{2n_k} \sum_{i} \sum_{j} d^2(X_i,X_j) = \sum_{i} X_i^T X_i - n_k (\bar{X}^k)^2$$

Por otro lado:

$$\sum_{i=1}^{nk} d^2(X_i,\bar{X}_i^k) = \sum_{i=1}^{nk} (X_i-\bar{X}^k)^T (X_i-\bar{X}^k)$$

$$\sum_{i=1}^{nk} (X_i-\bar{X}^k)^T (X_i-\bar{X}^k) = \sum_i [X_i^T X_i - X_i^T \bar{X}^k - \bar{X}^{kT} X_i + \bar{X}^{kT} \bar{X}^k]$$

$$= \sum_i X_i^T X_i - (\sum_i X_i^T) (\bar{X}^k) - \bar{X}^{kT} \sum_i X_i + \bar{X}^{kT} \bar{X}^k \, n_k$$

$$= \sum_i X_i^T X_i - n_k \bar{X}_i^{kT} \bar{X}^k - \bar{X}^{kT} \bar{X}_i^k \, n_k + n_k \bar{X}^{kT} \bar{X}^k \, n_k$$

Por lo tanto:

$$\sum_{i=1}^{nk} (X_i - \bar{X}^k)^T (X_i - \bar{X}^k) = \sum_i X_i^T X_i - n_k (\bar{X}_i^k)^2$$

lo que significa que:

$$\frac{1}{2n_k} \sum_{i=1}^{nk} \sum_{j=1}^{nk} d^2(X_i,X_j) = \sum_{i=1}^{nk} d^2(X_i,\bar{X}_i^k)$$

ANEXO D

PROGRAMAS COMPUTACIONALES

D.1 Programa que calcula la matriz de Disimilitud.

D.2 Programa para Métodos Jerárquicos.

```
10 REM
20 REM
30 REM ******** PROGRAMA QUE CALCULA LA MATRIZ DE DISIMILITUD ********
40 REM
50 REM
60 REM
70 REM
80 REM     LECTURA DE VARIABLES Y ELEMENTOS
90 REM
100 REM
110 DIM Y(50,20),X(50,20),S(50),D(50,50),S(50,50),T(50)
120 DIM R(50),TT(50),RR(50),INV(20,20)
130 OPEN "I",#2,"B:DAT"
140 PRINT " # VARS, # ELEMS"
150 INPUT M,N
160 FOR I=1 TO N
170 S(I)=0
180 FOR J=1 TO M
190 INPUT #2 Y(I,J)
200 S(I)=S(I)+Y(I,J)
210 NEXT J
220 FOR J=1 TO M
230 X(I,J)=Y(I,J)/S(I)
240 NEXT J,I
250 REM
260 REM    ELECCION DE LA MEDIDA DE ASOCIACION
270 REM
280 REM
290 PRINT
300 PRINT
310 PRINT "DISTANCIA EUCLIDIANA (-1)"
320 PRINT
330 PRINT "COEFICIENTE DE CORRELACION (0)"
340 PRINT
350 PRINT "METRICA DE MAHALANOBIS (1)"
360 PRINT
370 INPUT MA
380 IF MA>=0 GOTO 530
390 REM
400 REM    ****** CALCULO DE LA DISTANCIA EUCLIDIANA *****
410 REM
420 OPEN "O",#1,"B:MAT"
430 FOR K=2 TO N
440 FOR I=1 TO K-1
450 T=0
460 FOR J=1 TO M
470 T=T+(X(K,J)-X(I,J))^2
480 NEXT J
490 D(K,I)=T^.5
500 PRINT #1,D(K,I)
510 NEXT I,K
520 GOTO 1130
530 IF MA=1 GOTO 930
540 REM
550 REM
560 REM
570 REM
580 REM
590 REM
600 REM
610 REM
```

```
620 REM
630 REM
640 REM
650 REM
660 REM
670 REM ***** CALCULO DEL COEFICIENTE DE CORRELACION LINEAL *****
680 REM
690 REM
700 REM
710 OPEN "O",#1"B:MAT"
720 FOR K=2 TO N
730 FOR I=1 TO K-1
740 S(K,I)=0
750 T(K)=0
760 R(I)=0
770 FOR J=1 TO M
780 S(K,I)=S(K,I)+Y(K,J)*Y(I,J)
790 T(K)= T(K)+Y(K,J)^2
800 R(I)=R(I)+Y(I,J)^2
810 NEXT J
820 TT(K)=T(K)^.5
830 RR(I)+R(I)^.5
840 D(K,I)=S(K,I)/(TT(K)*RR(I))
850 PRINT #1,D(K,I)
860 NEXT I,K
870 GOTO 1130
880 REM
890 REM
900 REM ***** CALCULO DE LA METRICA DE MAHALANOBIS *****
910 REM
920 REM
930 OPEN "I",#2,"B:INV"
940 REM
950 REM  LECTURA DE LA MATRIZ INVERSA
960 REM
970 PRINT "LECTURA DE LA MATRIZ SINGULAR INVERSA"
980 PRINT "DE VARIANZA COVARIANZA POR RENGLON"
990 FOR J=1 TO M
1000 FOR L=1 TO M
1010 INPUT #2,INV(J,L)
1020 NEXT L,J
1030 OPEN "O",#1,"B:MAT"
1040 FOR K=2 TO N
1050 FOR I=1 TO K-1
1060 D(K,I)=0
1070 FOR J=1 TO M
1080 FOR L=1 TO M
1090 D(K,I)=D(K,I)+(X(K,L)-X(I,L))*INV(J,L)*(X(K,J)-X(I,J))
1100 NEXT L,J
1110 PRINT #1,D(K,I)
1120 NEXT I,K
1130 STOP
1140 END
```

```
10 REM
20 REM
30 REM    *****  PROGRAMA PARA RESOLVER CUMULOS JERARQUICOS  *****
40 REM
50 REM
60 DIM X(2775),TITLE(20),II(50),JJ(50),IL(50),SS(50),JL(50)
70 DIM MEXT(50),NEAR(50),SREF(50),PIST(50),LLAST(50)
80 LIMIT=2775
90 PRINT "DE EL TITULO DEL PROBLEMA"
100 INPUT TITLE$
110 PRINT "DE LAS VARIABLES NE,ISIGN,NTSV,NTIN,INOPT"
120 INPUT NE,ISIGN,NTSV,NTIN,INOPT
130 PRINT TITLE$
140 PRINT "NE="NE,"ISIGN="ISIGN,"NTSV="NTSV,"NTIN="NTIN,"INOPT="INOPT
150 REM LECTURA DE LA MATRIZ DE SIMILITUD
160 GOSUB 230
170 REM LISTO PARA AGRUPAR
180 GOSUB 930
190 END
200 RETURN
210 REM
220 REM
230 REM      ***   SUBRUTINA DE LECTURA   ***
240 REM
250 REM
260 OPEN "I",#NTIN,"B:MAT"
270 IOPT=INOPT
280 IF IOPT<=0 GOTO 410
290 REM LECTURA DE LA MATRIZ DE SIMILITUD EN BLOCKS
300 FIRST=1
310 LAST=IOPT
320 FOR I=FIRST TO LAST
330 INPUT #NTIN,X(I)
340 NEXT I
350 REM USE EL FIN DEL DISCO COMO FIN DE LA MATRIZ DE SIMILITUD
360 IF EOF(NTIN)THEN 530
370 FIRST=FIRST+IOPT
380 LAST=LAST+IOPT
390 GOTO 320
400 REM LECTURA DE LA MATRIZ DE SIMILITUD COMO MATRIZ TRIANG INF
410 FIRST=1
420 LAST=1
430 FOR K=2 TO NE
440 FOR I=FIRST TO LAST
450 IF EOF(NTIN) THEN 650
460 INPUT #NTIN,X(I)
470 NEXT I
480 FIRST=LAST+1
490 LAST=LAST+K
500 NEXT K
510 REM PASO DEL EOF
520 IF NOT EOF(NTIN) THEN 670
530 RETURN
540 REM
550 REM
560 REM
570 REM
580 REM
590 REM
600 REM
610 REM
```

```
620 REM
630 REM MENSAJES DE ERROR
640 REM
650 PRINT "SE ENCONTRO EOF CUANDO NINGUNO SE ESPERABA"
660 GOTO 680
670 PRINT "NO HAY EOF CUANDO SE ESPERABA UNO"
680 PRINT K,FIRST,LAST,Z
690 FOR I=FIRST TO LAST
700 PRINT X(I)
710 NEXT I
720 END
730 REM
740 REM
750 REM
760 REM
770 REM
780 REM
790 REM
800 REM
810 REM
820 REM      ***   SUBRUTINA DE AGRUPAMIENTO   ***
830 REM
840 REM
850 REM
860 REM
870 REM
880 REM
890 REM
900 REM
910 REM
920 REM
930 PRINT "DE EL METODO QUE DESEE UTILIZAR: "
940 PRINT "    1  SIMPLE-LINK"
950 PRINT "    2  COMPLETE-LINK"
960 PRINT "    3  MEDIAN"
970 INPUT MET
980 REM INICIALIZACION DE VARIABLES Y CONSTANTES
990 NCL=NE
1000 K=1
1010 SIGN=ISIGN
1020 BIG=SIGN*1E+30
1030 REM INICIALIZACION DE ARREGLOS
1040 FOR J=1 TO NE
1050 LLAST(J)=0
1060 NEXT(J)=0
1070 PIST(J)=J
1080 SREF(J)=BIG
1090 NEXT J
1100 REM
1110 REM
1120 REM
1130 REM
1140 REM
1150 REM
1160 REM ENCONTRAR LA ENTRADA EXTREMA  EN CADA RENGLON"
1170 REM
1180 REM
1190 L=0
1200 FOR I=2 TO NE
1210 I1=I-1
1220 FOR J=1 TO I1
1230 L=L+1
1240 REM EN EFECTO X(L)=X(I,J)
1250 IF((X(L)-SREF(I))*SIGN)>0 GOTO 1280
1260 NEAR(I)=J
1270 SREF(I)=X(L)
```

```
1280 NEXT J,I
1290 REM
1300 REM
1310 REM
1320 REM
1330 REM
1340 REM CICLO PRINCIPAL .ENCUENTRA EL VALOR EXTREMO ENEL ARREGLO SREF
1350 REM
1360 REM
1370 REM
1380 SREFX=BIG
1390 FOR I=2 TO NCL
1400 LISTI=PIST(I)
1410 IF ((SREF(LISTI)-SREFX)*SIGN)>0 GOTO 1450
1420 IREF=I
1430 LREF=LISTI
1440 SREFX=SREF(LISTI)
1450 NEXT I
1460 REM
1470 REM
1480 REM
1490 REM LREF ES EL NUMERO DE RENGLON QUE CONTIENE  LA ENTRADA EXTREMA EN S
1500 REM
1510 REM
1520 NREF=NEAR(LREF)
1530 REM SE GENERAN LOS DATOS PARA EL ARBOL
1540 II(K)=NREF
1550 JJ(K)=LREF
1560 SS(K)=SREFX
1570 IL(K)=LLAST(NREF)
1580 JL(K)=LLAST(LREF)
1590 LLAST(NREF)=K
1600 IF IL(K)=0 GOTO 1630
1610 ILK=IL(K)
1620 NEXT(ILK)=K
1630 IF JL(K)=0 GOTO 1660
1640 JLK=JL(K)
1650 NEXT(JLK)=K
1660 K=K+1
1670 REM
1680 REM
1690 REM TERMINA SI SE HAN HECHO N-1 MEZCLAS
1700 REM
1710 IF K=NE GOTO 1860
1720 REM DATOS PARA EL CICLO SIGUIENTE
1730 NCL=NCL-1
1740 IF IREF>NCL GOTO 1790
1750 FOR I=IREF TO NCL
1760 PIST(I)=PIST(I+1)
1770 NEXT I
1780 REM DATOS PARA EL SIGUIENTE CICLO
1790 GOSUB 2230
1800 GOTO 1380
1810 REM TERMINA EL AGRUPAMIENTO
1820 REM SE GUARDAN LOS RESULTADOS COMO SE DESEE
1830 IF MET=1 THEN LPRINT "          METODO DE UNION SIMPLE"
1840 IF MET=2 THEN LPRINT "          METODO DE UNION EXAHUSTIVA"
1850 IF MET=3 THEN LPRINT "          METODO DE LA MEDIANA"
1860 K=K-1
1870 IF NTSV<=0 THEN RETURN
1880 LPRINT "          ",TITLE$
1890 LPRINT
1900 LPRINT
1910 LPRINT
1920 LPRINT "        ETAPA      CUMULO INICIAL     CUMULO FINAL     FZA AGRUP"
1930 LPRINT
```

```
1940 LPRINT
1950 FOR I=1 TO K
1960 LPRINT "          ",I,II(I),JJ(I),SS(I)
1970 NEXT I
1980 LPRINT CHR$(12)
1990 LPRINT
2000 LPRINT "          ETAPA     PRE-ETAPA II  PRE-ETAPA JJ   ETAPA II SIGUE"
2010 LPRINT
2020 LPRINT
2030 FOR I=1 TO K
2040 LPRINT "          ",I,IL(I),JL(I),MEXT(I)
2050 NEXT I
2060 RETURN
2070 REM
2080 REM
2090 REM
2100 REM
2110 REM
2120 REM
2130 REM
2140 REM
2150 REM     ***   SUBRUTINA CON METODOS DE AGRUPAMIENTO   ***
2160 REM
2170 REM
2180 REM
2190 REM
2200 REM
2210 REM
2220 REM
2230 IF MET=1 GOTO 2260
2240 IF MET=2 GOTO 2680
2250 IF MET=3 GOTO 3380
2260 REM INICIALIZACION
2270 BIG=SIGN*1E+30
2280 REM UPDATE FOR NEXT ROUND
2290 FOR J=1 TO NCL
2300 I=PIST(J)
2310 IF I=NREF GOTO 2850
2320 IF I>LREF GOTO 2350
2330 LL=((LREF-1)*(LREF-2))/2+I
2340 GOTO 2360
2350 LL=((I-1)*(I-2))/2+LREF
2360 IF I>NREF GOTO 2390
2370 LN=((NREF-1)*(NREF-2))/2+I
2380 GOTO 2400
2390 LN=((I-1)*(I-2))/2+NREF
2400 IF ((X(LL)-X(LN))*SIGN)>=0 GOTO 2680
2410 X(LN)=X(LL)
2420 IF I>NREF GOTO 2490
2430 REM I<=NREF
2440 REM CHECAR SI S(LN) TINE MEJOR VALOR QUE SREF(NREF)
2450 IF((X(LN)-SREF(NREF))*SIGN)>0 GOTO 2850
2460 NEAR(NREF)=I
2470 SREF(NREF)=X(LN)
2480 GOTO 2850
2490 IF I>LREF GOTO 2820
2500 REM
2510 REM
2520 REM
2530 REM
2540 REM
2550 REM I>NREF AND I<LREF
2560 REM
2570 REM
2580 REM CHECAR SI X(LN) TIENE MEJOR VALOR QUE SREF(I)
2590 REM
```

```
2600 REM
2610 REM
2620 REM
2630 REM
2640 IF ((X(LN)-SREF(I))*SIGN)>=0 GOTO 2850
2650 SREF(I)=X(LN)
2660 NEAR(I)=NREF
2670 GOTO 2850
2680 IF I<LREF GOTO 2850
2690 REM
2700 REM
2710 REM I>LREF
2720 REM
2730 REM
2740 REM ACTUALIZAR EL ARREGLO CON ELEMENTO EXTREMO LREF
2750 REM
2760 REM
2770 REM
2780 REM
2790 REM
2800 REM
2810 REM
2820 IF NEAR(I)<>LREF GOTO 2850
2830 NEAR(I)=NREF
2840 SREF(I)=X(LN)
2850 NEXT J
2860 RETURN
2870 REM
2880 REM INICIALIZACION
2890 REM
2900 BIG=SIGN*1E+30
2910 FOR J=I TO NCL
2920 I=PIST(J)
2930 IF I=NREF GOTO 3040
2940 IF I>NREF GOTO 2970
2950 LL=((LREF-1)*(LREF-2))/2+I
2960 GOTO 2980
2970 LL=((I-1)*(I-2))/2+LREF
2980 IF I>NREF GOTO 3010
2990 LN=((NREF-1)*(NREF-2))/2+I
3000 GOTO 3020
3010 LN=((I-1)*(I-2))/2+NREF
3020 IF ((X(LL)-X(LN))*SIGN)<=0 GOTO 3040
3030 X(LN)=X(LL)
3040 NEXT J
3050 REM
3060 REM SI EL ELEM EXTREMO EN EL RENGLON I FUE TAMBIEN LREF O NREF ES
3070 REM NECESARIO ENCONTRAR UN ELEM EXTREMO NUEVO SIN CONSIDERAR LOS
3080 REM RENGLONES ANTERIORES A NREF
3090 REM
3100 REM
3110 FOR J=1 TO NCL
3120 I=PIST(J)
3130 IF I=NREF GOTO 3150
3140 NEXT J
3150 IF J=1 GOTO 3280
3160 SREF(I)=BIG
3170 J1=J-1
3180 FOR L=1 TO J1
3190 LISTL=PIST(L)
3200 IF I>LISTL GOTO 3230
3210 LL=((LISTL-1)*(LISTL-2))/2+I
3220 GOTO 3240
3230 LL=((I-1)*(I-2))/2+LISTL
3240 IF ((X(LL)-SREF(I))*SIGN)>=0 GOTO 3270
3250 NEAR(I)=LISTL
```

```
3260 SREF(I)=X(LL)
3270 NEXT L
3280 J=J+1
3290 IF J>NCL THEN RETURN
3300 I=PIST(J)
3310 IF NEAR(I)=LREF OR NEAR(I)=NREF GOTO 3160
3320 GOTO 3280
3330 RETURN
3380 REM INICIALIZACION
3430 BIG=SIGN*1E+30
3440 REM
3450 REM
3460 REM DATOS PARA LA VUELTA SIGUIENTE
3470 REM
3480 REM
3490 IF LREF>NREF GOTO 3520
3500 LBET=((NREF-1)*(NREF-2))/2+LREF
3510 GOTO 3530
3520 LBET=((LREF-1)*(LREF-2))/2+NREF
3530 FOR J=1 TO NCL
3540 I=PIST(J)
3550 IF I=NREF GOTO 3650
3560 IF I>LREF GOTO 3590
3570 LL=((LREF-1)*(LREF-2))/2+I
3580 GOTO 3600
3590 LL=((I-1)*(I-2))/2+LREF
3600 IF I>NREF GOTO 3630
3610 LN=((NREF-1)*(NREF-2))/2+I
3620 GOTO 3640
3630 LN=((I-1)*(I-2))/2+NREF
3640 X(LN)=(X(LN)+X(LL))/2-X(LBET)/4
3650 NEXT J
3660 FOR J=1 TO NCL
3670 I=PIST(J)
3680 IF I=NREF GOTO 3700
3690 NEXT J
3700 IF J=1 GOTO 3830
3710 SREF(I)=BIG
3720 J1=J-1
3730 FOR L=1 TO J1
3740 LISTL=PIST(L)
3750 IF I>LISTL GOTO 3780
3760 LL=((LISTL-1)*(LISTL-2))/2+I
3770 GOTO 3790
3780 LL=((I-1)*(I-2))/2+LISTL
3790 IF ((X(LL)-SREF(I))*SIGN)>=0 GOTO 3820
3800 NEAR(I)=LISTL
3810 SREF(I)=X(LL)
3820 NEXT L
3830 J=J+1
3840 IF J>NCL THEN RETURN
3850 I=PIST(J)
3860 IF NEAR(I)=LREF OR NEAR(I)=NREF GOTO 3710
3870 GOTO 3830
3880 RETURN
```

B I B L I O G R A F I A

La Bibliografía se dividirá en 2 partes, como sigue:

1) La Bibliografía de consulta que se utiliza en este trabajo.

2) La Bibliografía de referencia, para el lector que desee profundizar acerca de los temas tratados en este trabajo. Se encuentran precedidos por un paréntesis cuadrado [ ].

BIBLIOGRAFIA DE CONSULTA

Duran, B. y Odell, P.	"Cluster Analysis". Lecture Notes in Economics and Mathematical Systems, Vol. 100. 1974.
Anderberg, M.R.	"Cluster Analysis for Applications". Academic Press, INC. 1973.
Garfinkel, R.S. y Nemhauser, G.L.	"Integer Programming". John Wiley and Sons. 1972.
Hillier, F.S. y Lieberman, G.J.	"Operations Research". Holden-Day, INC. 1974, pp. 697-720.
Hansen, P. y Delattre, M.	"Bicriterion of Cluster Analysis with Utility Function". North-Holland Publishing Company. 1979. pp. 975-986.

Jernigan, M.E. "Pattern Recognition". Course Notes. March 1979.

Bijmen, J.E. "Cluster Analysis". Survey and Evaluations of Techniques. Tiburg University Press the Netherlands. 1973, pp. 4-10.

BIBLIOGRAFIA DE REFERENCIA

[1] Jensen, R.E. "A Dynamic Programming Algorithm for Cluster Analysis". Operations Research, Vol. 12, December 1969, pp. 1034-1057.

[2] Gower, J.C. y Ross, G.J.S. "Minimum Spanning Trees and Single Linkage Cluster Analysis". Applied Statistics, Vol. 18, 1969, pp. 54-64.

[3] Hansen, P. "Subdegrees and Chromatic Numbers of Hypergraphs". Annals of Discrete Mathematics, Vol. 1, 1977, pp. 287-292.

[4] Hansen, P y Delattre, M. "Complete-Link Cluster Analysis by Graph Coloring". Journal of the American Statistical Association, Vol. 73, Num. 362. Jun. 1978, - pp. 397-403.

[5] Lance, G.N. y Williams, W.T. "A General Theory of Classificatory Sorting Strategies". 1. Hierarchical Systems, Computational Journal, Vol. 9, Num. 4. 1967, pp. 373-380.

[6] Anderberg, M.R. "Cluster Analysis for Applications". Academic Press, INC. 1973, pp. 278-290.

[7] Johnson, S.C. "Hierarchical Clustering Schemes". Psychometrika, Vol. 32. Num 3. - September 1967, pp. 241-247.

[8] Johnson, S.C. "Hierarchical Clustering Schemes". Psychometrika, Vol. 32. Num. 3. - September 1967, pp. 248-254.

[9] Gower, J.C. "A Comparasion of Some Methods of Cluster Analysis". Biometrics, - Vol. 23. Num. 4. December 1967. pp. 623-637.

[10] Instituto Nacional de Estadística, Geografía e Informática "X Censo General de Población y Vivienda, 1980". Resumen General - Abreviado. México 1984, pp. 75-98.

[11] Consejo Nacional de Población (CONAPO) "Política Demográfica Nacional y - Regional". Objetivos y Metas, - 1978-1982.

[12] Consejo Nacional de Población (CONAPO) "México Demográfico". Breviario - 1978, pp. 38-42.

[13] Corona, R.
Jiménez, R.
Minujin, A.

"La Mortalidad en México". Tablas Abreviadas de Mortalidad para las Entidades Federativas de la República 1940, 1950, 1960, 1970. Instituto de Investigaciones Sociales, UNAM. 1981.

[14] Consejo Nacional de Población (CONAPO)

"México Demográfico". Breviario - 1980-1981, pp. 49.