UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES
"Z A R A G O Z A"

EL ANALISIS ESTADISTICO CLUSTER, METODOS Y APLICACIONES EN BIOLOGIA

T E S I S

QUE PARA OBTENER EL TITULO DE

B I O L O G O

P R E S E N T A :

SUSANA OCEGUEDA CRUZ

MEXICO, D. F. 1991

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

RESUMEN

RESUMEN

El incremento en el uso de técnicas de estadística multivariada no sólo a la biología sino a muchos otros campos ha sido considerable. Dentro de éstas técnicas el análisis estadístico cluster es una herramienta útil para una de las tareas fundamentales en una ciencia como la biología, la clasificación.

En este trabajo se hace una revisión de las principales técnicas jerárquicas de análisis cluster, el propósito es dar una mayor difusión de las características de cada una de las técnicas y mostrar un estudio de caso donde se aplica a un conjunto de datos reales.

Se revisa cuales son las aplicaciones que tiene esta técnica en diferentes campos de la biología como en taxonomía y ecología; que son quizá las dos áreas donde ha sido más aplicado este tipo de análisis estadístico.

Además se dan algunas recomendaciones generales que el usuario de este tipo de técnicas debe tomar en consideración antes, durante y después del análisis.

CAPITULO 1

INTRODUCCION

El avance en la aplicación de los métodos estadísticos a la biología de campo, ha sido dominado por los requerimientos de científicos que trabajan con experimentos controlados. En estos, los factores no controlados se minimizan al grado de no interferir en el buen desarrollo del experimento. Un ejemplo de estudio de este tipo, sería la respuesta en el crecimiento del maíz a la aplicación de diferentes dosis de fertilizante. En este caso la dosis de fertilizante sería la variable independiente "x" o variable controlada. Los posibles factores no controlados que pudieran alterar la respuesta (crecimiento), como el ataque al cultivo por plagas, se minimizan con el uso de plaguicidas para que no interfieran con el experimento (Digby, 1987). En este caso se tiene uno de los experimentos más sencillos, pues se tiene bajo control, pero muchas veces no es así.

Los experimentos controlados han conducido a la creación y desarrollo de modelos estadísticos que sirven para obtener información de la población de la que se han extraído un conjunto de datos. En este caso se requiere que los datos cumplan con una serie de condiciones para que se aproximen a una distribución conocida como por ejemplo la distribución normal (Digby, 1987).

En otros casos se presentan situaciones donde el cambio de una variable (x) está relacionado con el cambio de la otra variable (y). En este caso existe una correlación simple entre variables. Si se trata de una variable (y) que está relacionada con varias variables independientes (x) se trata de una correlación múltiple.

Al contrario, los datos obtenidos en estudios ecológicos, casi nunca cumplen los supuestos de los modelos estadísticos tradicionales, pues generalmente se tienen grandes registros de datos donde se midieron una serie de características para un sitio de muestreo o individuo; en este caso se desconoce cual sería la variable independiente, porque

todas las variables medidas son de respuesta y pueden ser o no, interdependientes. Para el análisis de este tipo de datos, los métodos clásicos univariados resultan inefectivos y es necesario el uso de métodos multivariados (Gauch, 1982).

Los métodos de estadística multivariada responden a diferentes necesidades, se encuentran aquellos cuyo objetivo es reducir el número de variables originales a un número considerablemente menor frecuentemente de dos o tres variables, y donde 1) se pueden representar gráficamente todas las variables en dos dimensiones (análisis de correspondencias); 2) se buscan nuevas variables que aporten la mayor cantidad de información que sirva para analizar las interrelaciones entre variables en términos de factores (análisis de factores) ó 3) cada nueva variable es una combinación (lineal o no) de las variables originales (análisis de componentes principales) (Gauch, 1982).

También están las técnicas en las que se buscan las combinaciones óptimas de cada conjunto de variables dependientes como función de un conjunto de variables independientes, que maximicen la correlación entre estos conjuntos (correlación canónica), o donde se tiene una variable categórica dependiente que pertenece a dos o más clases como función de dos o más variables numéricas independientes y se trata de asignar la máxima probabilidad de que una entidad pertenezca a una clase en función de algunas variables numéricas independientes (análisis discriminante) (Gauch, 1982).

Finalmente está el análisis cluster que sirve para clasificar en grupos un conjunto de características medidas para un individuo o entidad y cuyo objetivo es únicamente describir la estructura del conjunto de datos. No permite hacer inferencias, pero tampoco requiere de supuestos de independencia, aleatoriedad, homocedasticidad, linearidad o normalidad de los datos y es una técnica que se puede usar de forma combinada con algunas de las anteriores (Anderberg, 1973).

La descripción del comportamiento de los datos implica el reconocimiento de clases o grupos similares. En ecología es importante reconocer clases ecológicas que corresponden a comunidades. Del mismo modo la taxonomía numérica busca la identificación de grupos de organismos relacionados por su máxima semejanza que puedan corresponder con alguna categoría taxonómica: especie, género, familia, etc.

Durante las últimas tres décadas los métodos de clasificación numérica han aumentado considerablemente. Este avance se muestra por ejemplo, en la cantidad de literatura en la que se hace mención del uso del análisis cluster para clasificación de datos de tipo ecológico. Una de las razones que ha propiciado este desarrollo es la disponibilidad de programas de cómputo que indudablemente facilitan el manejo de grandes cantidades de datos.

En nuestro país es escaso el número de trabajos donde se aplican este tipo de técnicas y generalmente se muestra el conjunto de datos, la medida de asociación entre variables y el método de agrupamiento que se usó. Sin embargo, pocos son los que hacen un análisis de los métodos usados y la interpretación que podría darse en cada uno de los casos al conjunto de datos.

Este trabajo pretende apoyar el uso del análisis estadístico cluster, así como dar a conocer sus ventajas y limitaciones en diferentes áreas de la biología. Básicamente se analizan técnicas de tipo jerárquico pues son las más usadas en biología, las cuales se aplican a un conjunto de datos de campo, analizando cada una de ellas con base en estos datos.

Por lo expresado anteriormente se plantearon los siguientes objetivos:

Revisar los fundamentos teóricos de las técnicas jerárquicas de análisis cluster más usadas en diferentes campos como son: el enlace simple, enlace completo, centroide, grupo promedio y de Ward.

Analizar un estudio de caso (de tipo ecológico) donde se aplican las diferentes técnicas y se discute sobre los alcances de ellas.

Describir brevemente de que manera se aplica el análisis cluster en diferentes áreas de la biología como en ecología y taxonomía.

Para lograr dichos objetivos se ha dado una división por capítulos para separar diferentes aspectos importantes a conocer antes de aplicar al análisis cluster.

En el capítulo 2 se dá la definición de cluster, los usos que tiene y la forma en que opera el análisis de cluster. Su finalidad es principalmente responder a la pregunta: ¿qué es el análisis cluster?.

En el capítulo 3 se presentan los diferentes tipos de arreglos matriciales que se pueden observar al realizar un análisis cluster.

El capítulo 4 está dedicado a la revisión de diferentes medidas de asociación entre variables.

El capítulo 5 está destinado a la descripción de las diferentes técnicas de análisis cluster. Con ejemplos numéricos se analizan brevemente los fundamentos de cada una de ellas.

En el capítulo 6 se presenta un estudio de caso, donde se analiza la aplicación del análisis cluster en la clasificación de tierras, y se discuten los resultados obtenidos.

En el capítulo 7 se discute sobre las ventajas del análisis cluster en la ecología y en la taxonomía.

Finalmente en el capítulo 8 se dan las conclusiones.

CAPITULO 2

GENERALIDADES SOBRE EL ANALISIS DE CLUSTER

Existen muchas definiciones del término cluster, pues en mayor o menor medida, el reconocimiento de clusters depende del tipo de datos que se analicen y del juicio del investigador. Se considera que para definir un cluster, éste debe tener una asociación natural; ambos términos, cluster y asociación natural, son tratados en este capítulo.

2.1 DEFINICION DE UN CLUSTER.

El término cluster tiene un significado vago. La razón de esto es que existen muchas características específicas, que son todas facetas, de lo que generalmente se reconoce como un cluster. La definición más comúnmente encontrada en diccionarios de términos estadísticos para cluster es: "un grupo de elementos contiguos de una población estadística"; por ejemplo, un grupo de gente viviendo en una casa (Everitt,1981).

Muchos autores proponen definiciones como: "un cluster es un grupo de entidades que son semejantes, y entidades de diferentes clusters no son semejantes". Otros sugieren que las entidades dentro de un cluster son más similares, entre sí, que con las entidades de otros clusters. Gengerelli (1963) define un cluster como un agregado de puntos en un espacio de prueba, tal que la distancia entre cualquiera de dos puntos en un cluster sea menor que la distancia entre cualquier punto en el cluster y cualquier punto que no pertenezca a él (Gengerelli, 1963, en Everitt, 1981).

Todas estas definiciones son válidas. Con estas ideas se puede dar una definición más general. "Un cluster es un conjunto de entidades con características similares que pertenecen a una población estadística. Si cada entidad se representa como un punto en un espacio geométrico de prueba, la distancia entre dos puntos de un cluster es menor que la distancia entre un punto en el cluster y otro fuera de este".

Un cluster puede ser tan amplio o tan restringido, según sea el marco de referencia del que se parta. Es decir, si

se considera que un cluster es un conjunto de entidades que tienen propiedades comunes; un cluster serían los seres vivos, otro más específico, los animales acuáticos, y otro aún menor los peces. A medida que se consideran mayor cantidad de atributos o características los grupos van siendo más restringidos y al final se puede obtener una representación jerárquica.

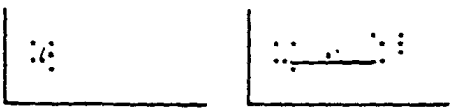


Figura 2.1. Interpretación gráfica de la definición de un cluster. Explicación en el texto.

Imagine un espacio geométrico de prueba de dos dimensiones como el de la figura 2.1, en el primer caso los puntos están muy cercanos entre sí, y constituyen un cluster, en el segundo caso existen dos cluster pues la distancia entre cualquiera de los puntos del cluster de la izquierda (a) es menor, que la distancia entre un punto en ese cluster y otro fuera de él (a').

2.2 PROPOSITO DE LAS TECNICAS DE CLUSTER.

El análisis cluster comprende diversas técnicas que sirven para describir la estructura de un conjunto de datos. Estas técnicas se han desarrollado para solucionar un problema: dada una muestra de unidades de datos (individuos, elementos, observaciones) descritos por resultados de variables seleccionadas (atributos, características, medidas). El objetivo es agrupar estas unidades de datos en clusters tales que los elementos dentro de un cluster tengan una "asociación natural" entre sí y todas aquellas excepciones queden en clusters distintos de este (Anderberg, 1973; Everitt, 1981).

El siguiente problema sería definir lo que es una "asociación natural"; este término no está claramente definido y la mayoría de los autores coinciden en que es un término subjetivo que queda a juicio del investigador.

Una asociación natural puede entenderse como la unión de dos observaciones o grupos de observaciones, las cuales no han sufrido ningún tipo de transformación. También se le llama asociación natural a la estabilidad que tienen los grupos, aún cuando se prueben diferentes métodos de agrupamiento o bien diferentes medidas de asociación. Dado que la mayor o menor estabilidad de un grupo puede tener diferentes apreciaciones; esta es la razón por la que se considera a la asociación natural un término subjetivo.

Sin embargo, para fines operativos diré que aquellos clusters que tienen una "asociación natural" son los que proporcionan la información que el investigador puede percibir aún antes de aplicar el análisis.

2.3 USOS DEL CLUSTER.

A la aplicación del análisis cluster se le conoce también como análisis Q, análisis tipológico, análisis de agrupamiento, análisis de cúmulos, clasificación, taxonomía numérica y reconocimiento de modelos no dirigidos. Los términos análisis de cúmulos y análisis de agrupamiento, actualmente son usados en textos en idioma español. Sin embargo, se conoce más como análisis cluster por lo que en este trabajo se seguirá manejando éste término. Esta variedad de nombres se debe a la diversidad de los campos en los que tiene aplicación, como: Psicología, Biología, Sociología, Inteligencia Artificial y Recuperación de Información (Everitt, 1981).

El análisis cluster como herramienta estadística tiene diversos usos. El principal uso del análisis cluster, que más bien es el objetivo de la técnica, es la clasificación de entidades; esto indirectamente dá idea de la importancia que tiene cada una de las variables en el análisis, puesto que la distancia a la que éstas se unen es un reflejo del peso que tienen para el estudio.

En ocasiones se tiene una gran cantidad de observaciones que no dan grupos manejables como unidades. Las técnicas cluster pueden ser usadas para reducción de datos porque en ocasiones se tiene que al excluir una o más variables del análisis se obtienen dendrogramas similares estando o no presentes estas variables. Es por lo tanto, más práctico agrupar todas las variables que se supone tienen el mismo efecto en el dendrograma resultante y así reducir la información de un grupo con "N" individuos a un grupo "k" (donde "k" es mucho más pequeño que "N") (Anderberg, 1973).

El análisis cluster también puede usarse antes de las técnicas tradicionales inferenciales para observar la estructura de los datos, esto puede conducir al reconocimiento de patrones de comportamiento importantes en los datos que no se habían reconocido, y de este modo puede servir para generar hipótesis. Cuando tiene este propósito la hipótesis debe probarse y se debe considerar que cualquier prueba nueva dependerá de nuevas observaciones y no del uso de los datos a partir de los que la hipótesis fué generada.

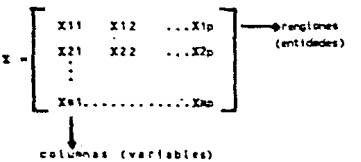
CAPITULO 3

MATRICES DE DATOS Y MATRICES DE ASOCIACION

El análisis cluster opera sobre una matriz de similitud o distancias a partir de la que se obtienen grupos de entidades las cuales son más parecidas. A continuación se presentan los diferentes tipos de matrices y las características de cada una de ellas.

3.1 MATRICES DE DATOS.

El primer paso en un análisis cluster es convertir una matriz de datos originales X, en una matriz de similitud o disimilitud entre individuos, o bien de medidas de distancia (Pielou, 1984). Una matriz de datos X, es un arreglo ordenado de los datos que se compone de renglones (i) y columnas (j) de la siguiente forma:



Esta matriz constituye un arreglo de los datos para un muestreo determinado. En ella X_{ij} es la variable, o atributo j-ésimo del i-ésimo individuo o entidad bajo estudio. Las medidas pueden ser cuantitativas, por ejemplo, altura, peso, edad. En otras pueden ser cualitativas o categóricas, por ejemplo: la presencia o ausencia de un síntoma, color de cabello, color de ojos; en muchos casos los datos pueden implicar una variedad de diferentes tipos de variables.

Es importante saber que según el tipo de agrupamiento deseado ya sea de variables o de localidades, es la estructura que debe tener la matriz de datos originales X.

Es común que los datos registrados por ecólogos animales sea el conteo del número de veces que ocurre un evento. Los eventos pueden ser:

1. El número de individuos de una especie particular que se encuentra en una localidad.
2. El número de veces que una especie se encuentra en la vecindad de otra especie.

En el primer caso el conteo involucra una medida de abundancia y en el segundo caso el conteo es una medida de asociación.

Para cada uno de estos casos los datos que conforman la matriz X tienen significados diferentes y por lo tanto las matrices de datos también.

En un análisis cluster se comparan renglones de datos, se calculan las distancias entre ellos y por último se unen o separan renglones de datos. Si se desea comparar entidades, éstas se deben colocar en los renglones y sus atributos medidos en las columnas. Ahora, si se desean agrupar variables se tendría que trabajar como la transpuesta de la matriz original; es decir, las variables quedarían en los renglones y sus mediciones en las columnas.

3.2 MATRICES DE ASOCIACION.

Existen diferentes maneras de construir una matriz de asociación entre entidades. Pueden originarse a partir de coeficientes de similitud o de medidas de distancia.

3.2.1 MATRIZ DE SIMILITUD.

Las matrices de similitud son el resultado del cálculo del grado de asociación entre variables a partir de coeficientes de similitud.

La similitud es el grado de semejanza existente entre dos individuos y por lo tanto sus valores van de 0 a 1. El cero indica una nula similitud y el 1 una similitud total. En una matriz de similitud, los valores de la diagonal siempre van a ser 1.

La disimilitud es la contraparte de la similitud y se puede decir que mientras una busca parecidos para unir individuos (similitud), la otra (disimilitud) considera diferencias para separarlos de un conjunto. En una matriz de disimilitud, los valores de la diagonal siempre van a ser 0.

3.2.2. MATRIZ DE DISTANCIA.

Las matrices de distancia son generadas a partir de cualquiera de las muchas medidas de distancia (métricas). Estas medidas mapean los puntos en un espacio geométrico y definen un grado de separación entre dos puntos. (En el capítulo 5 se presentan ejemplos de matrices de distancias).

3.2.3 MATRIZ DE CORRELACION.

La correlación entre dos vectores de variables es también una medida de similitud entre individuos. Al igual que la similitud puede tomar valores entre -1 y 1.

La matriz de correlación se construye a partir de los valores obtenidos por el coeficiente de correlación producto momento r ; donde r es el coseno del ángulo formado entre dos vectores de variables estandarizadas (Anderberg, 1973).

3.3 METODO GENERAL EN EL ANALISIS DE CLUSTER.

El análisis cluster es una herramienta estadística que implica que el investigador esté constantemente tomando una serie de decisiones.

Estas se pueden resumir en tres grandes pasos:

1. Selección de variables.
2. Medidas de asociación.
3. Método de cluster.

3.3.1 LA SELECCION DE VARIABLES.

Para cualquier grupo de entidades es claro que puede existir una variedad de clasificaciones posibles. Es necesario considerar que algunas clasificaciones son en algún sentido más valiosas o usuales que otras y que esto depende del contexto del problema y del conocimiento que tenga el investigador de este contexto.

Es importante realizar una selección inicial de variables que van a formar parte del estudio. La siguiente cuestión es conocer el número de variables deseables en la aplicación de las técnicas de cluster. Esto es importante pues la mayor o menor similitud entre entidades depende no sólo de los atributos medidos sino también de su número.

Los datos básicos para un análisis de cluster es un grupo de N entidades en las que se ha realizado la medición de p variables. Esta selección inicial, presumiblemente, refleja el juicio del investigador y la relevancia del propósito de clasificación.

Una consideración adicional es si los datos deben ser estandarizados de alguna manera. Se recomienda estandarizar las variables a una media cero y varianza uno, usando la desviación estándar que resulta del grupo completo de entidades. Esto puede tener el serio efecto de dilución de las diferencias entre grupos sobre las variables que son los mejores discriminadores (Everitt, 1981).

Al estandarizar las variables, se están ajustando a un modelo lo que hace que las diferencias entre los valores extremos se minimicen. Por ejemplo, si se quisieran agrupar 5 variables y los valores fueran 1, 100, 1000, 100,000 y 10,000,000. Las dos variables más parecidas son 1 y 100. Si se observan estos valores aisladamente no parecerían cercanos, es necesario entonces, considerar el marco de referencia que proporcionan los datos. Si se tomaran ahora los valores estandarizados, restando a cada variable el valor de la media y dividiendo esto entre la desviación estándar, el conjunto de datos se vuelva más homogéneo; por esta razón se dice que en ocasiones no es recomendable estandarizar pues se pierden las

diferencias que son los mejores discriminadores. El efecto que tendría esto en el análisis sería la disminución de la distancia en la que se dá la unión de grupos. Sin embargo es muy útil observar que efecto tiene la estandarización en el conjunto de datos, antes de decidir si es conveniente hacerla o no. El análisis cluster por si mismo, permite el reconocimiento de casos extremos o aberrantes, estos se pueden identificar en un dendrograma pues, en primer lugar, son las entidades que se unen hasta el final en la jerarquía y en segundo lugar, aún variando las técnicas de agrupamiento se mantiene su comportamiento extremo. Después de reconocer casos aberrantes, la mejor opción sería sacarlos del análisis y observar el efecto que tiene su ausencia en la jerarquía resultante y quizás no convenga, en estos casos, realizar una estandarización.

3.3.2 MEDIDAS DE ASOCIACION.

Una vez que se ha generado el marco de referencia sobre el que se quiere aplicar en análisis cluster; el siguiente paso es la elección de la medida de asociación.

Esta selección es difícil ya que existe una gran variedad de medidas que pueden emplearse.

Este tema se tratará en el siguiente capítulo, pues merece consideración particular.

3.3.3 METODO DE CLUSTER.

La elección del método de agrupamiento es otro problema que merece atención especial y dado que es el tema central de este trabajo, se tratará en el capítulo 5.

CAPITULO 4

MEDIDAS DE ASOCIACION ENTRE VARIABLES

4.1 MEDIDAS DE SIMILITUD Y DISTANCIA.

La mayoría de las técnicas de cluster comienzan con el cálculo de una matriz de similitudes o distancias entre entidades.

Existen tres conceptos de similitud y distancia que deben considerarse: entre entidades individuales, entre una entidad individual y un grupo de entidades y entre dos grupos de entidades.

En este trabajo básicamente se manejan las distancias entre grupos de entidades, se asume que al inicio existen grupos de una entidad. Esta es la distancia que se busca en cada uno de los métodos (distancia mínima, máxima o promedio).

La distancia para unir entidades, es siempre para los métodos aglomerativos la mínima y para los métodos divisivos, la distancia máxima se usa para separar individuos del conjunto.

Los coeficientes de similitud, es decir aquellos que sirven para evaluar la semejanza entre individuos, han sido conocidos como coeficientes de asociación y toman valores entre 0 y 1. En muchos casos las variaciones son del tipo "presencia" o "ausencia", esto se puede representar en una tabla de 2 x 2 en la cual la presencia de una variable se denota con + y su ausencia con -.

		individuo i		
		+	-	
individuo j	+	a	b	a+b
	-	c	d	c+d
		a+c	b+d	p

donde p= a+b+c+d.

Existen muchos tipos de coeficientes pero los más usados son el coeficiente de igualación simple y el coeficiente de Jaccard's (Pielou, 1984).

Los diferentes tipos de coeficientes de similitud dan origen a diferentes resultados a partir de los mismos datos.

Los datos cualitativos, por ejemplo, el color de ojos, que puede ser desde azul, café, verde, pueden ser tratados de manera similar como datos binarios, es decir, como presencia o ausencia de color verde, o bien la presencia o ausencia de color café, de forma que se sigan trabajando ceros y unos. Para variables cuantitativas, la medida de similitud más comúnmente usada entre individuos es el coeficiente de correlación producto momento.

En muchas aplicaciones de las técnicas cluster cada individuo es descrito por un grupo de variables que incluyen medidas binarias, cualitativas y cuantitativas. En tales casos el coeficiente de similitud propuesto por Gower (1971) puede ser ideal para estas mezclas de variables. El coeficiente se define como:

$$S_{ij} = \sum_{k=1}^{p} S_{ijk} \; / \; \sum_{k=1}^{p} W_{ijk}$$

El peso W_{ijk} es un grupo igual a 1 ó 0 dependiendo de si la comparación es considerada valida para la variable k y, excepto para el caso de variables dicotómicas, este peso sólo puede ser cero cuando la variable k es desconocida para uno o ambos individuos. Con las variables dicotómicas W_{ijk} es también colocada en 0 cuando la variable k es conocida o ausente para ambos individuos (Everitt, 1981).

Siempre que $W_{ijk}=0$, entonces S_{ijk} es llevada a cero, y si $W_{ijk}=0$ para todas las variables, S_{ij} es no definida.

4.2 MEDIDAS DE DISTANCIA.

Una función numérica d(x,y) de pares de puntos de un grupo E está dada como la métrica para E si satisface las siguientes condiciones:

(i) $d(x,y) \geq 0$; $d(x,y)=0$ si $x=y$;

(ii) $d(x,y) = d(y,x)$;

(iii) $d(x,y) \leq d(x,z) + d(y,z)$

La tercera condición es la única que diferencia las medidas de distancia y las medidas de similitud. Esta condición es conocida como la desigualdad del triángulo.

Probablemente la medida de distancia más comúnmente usada y la más familiar, es la métrica Euclidiana, donde la distancia entre puntos i y j denotada por d_{ij} es definida como

$$d_{ij} = \left[\sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \right]^{1/2}$$

donde X_{ik} es el valor de la k-ésima variable para la i-ésima entidad.

Existen también otras posibles medidas como la métrica absoluta, que se define como:

$$d_{ij} = \sum_{k=1}^{p} | X_{ik} - X_{jk} |$$

la de Mahalanobis (1936), conocida como Mahalanobis D^2, se define como:

$$d_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

donde Σ^{-1} es la transformación dentro de los grupos en una matriz de varianza-covarianza, y X_i y X_j son los vectores de los resultados para las entidades i y j. Mahalanobis D^2 tiene la ventaja sobre las medidas Euclidiana y de Manhattan que sirven para correlaciones entre variables. Cuando estas correlaciones son cero, es equivalente a la medida de distancia Euclidiana usando las variables estandarizadas.

Medidas de distancia que incluyen como casos especiales las distancias Euclidiana y de Manhattan son las metricas de Minkowski definidas por

$$d_{ij} = \left[\sum_{k=1}^{p} |X_{ik} - X_{jk}|^r \right]^{1/r}$$

Cuando r=1 esto se convierte en la métrica de Manhattan, y cuando r=2 en la distancia Euclidiana.

4.3 DIFERENCIAS ENTRE MEDIDAS DE SIMILITUD Y DISTANCIA.

La diferencia más obvia entre los dos tipos de medidas es que mientras la similitud toma valores entre 0 y 1, las medidas de distancia pueden tomar cualquier valor positivo. Sin embargo, se puede transformar a un grupo de valores de distancia un grupo correspondiente de valores para una función de similitud, simplemente restando el valor de similitud a uno.

4.4 SIMILITUD ENTRE GRUPOS Y MEDIDAS DE DISTANCIA.

Un método para construir medidas de similitud y de distancia entre grupos es substituir las medias del grupo para las p variables en la fórmula de coeficientes inter-individuales. Por ejemplo, si el grupo X tiene el vector media $\bar{X}=(\bar{x}1, \bar{x}2,\ldots,\bar{x}p)$ y el grupo Y el vector media $\bar{Y}=(\bar{y}1,\bar{y}2,\ldots,\bar{y}p)$, entonces una medida de distancia entre X y Y puede ser

$$d_{xy}=\left[\sum_{i=1}^{p}\left\{\bar{X}i-\bar{Y}i\right\}^2\right]^{1/2}$$

que es la distancia Euclidiana.

CAPITULO 5

TECNICAS DE CLASIFICACION POR ANALISIS CLUSTER

Las técnicas de clasificación de mayor uso en biología pueden considerarse dentro de tres grupos: clasificaciones jerárquicas, clasificaciones no jerárquicas y arreglos tabulares (Gauch, 1982). Básicamente se presentan técnicas jerárquicas, pues son las más usadas en la investigación biológica.

Las clasificaciones de tipo jerárquico operan sobre una matriz de distancias o similitudes para construir un diagrama de árbol o dendrograma que muestra las relaciones entre entidades.

Las clasificaciones no jerárquicas proporcionan soluciones que no necesariamente representan relaciones jerárquicas entre las entidades, es decir sus soluciones no dan dendrogramas.

Los arreglos tabulares son usados cuando las variables son binarias del tipo "presencia" o "ausencia". Para este tipo de clasificación se busca encontrar la asociación o independencia entre las variables comparadas.

5.1 CLASIFICACIONES JERARQUICAS.

Las clasificaciones de tipo jerárquico pueden ser de muy diversos tipos. Por ejemplo, de acuerdo a su forma de operación, pueden ser divisivas o aglomerativas. Las técnicas divisivas comienzan con la población completa y por subdivisiones sucesivas se forman grupos cada vez más pequeños. En cada etapa de la subdivisión se buscan las diferencias dentro de los grupos para separar subgrupos que difieren entre sí. En cambio las técnicas aglomerativas comienzan con todos individuos, los que se combinan por su semejanza hasta agotar las posibilidades de unión o hasta que no queden individuos aislados (Matteucci y Colma, 1982; Manly, 1986).

La derivación de un sistema de clasificación jerárquico a partir de una medida de asociación es un proceso en dos etapas. La primera etapa es la derivación de un dendrograma o

diagrama de relaciones (Fig. 5.1a). Un dendrograma puede ser descrito como una jerarquía con niveles numéricos. Los niveles a los cuales cada par de objetos se reúnen en un dendrograma son los niveles de división o nodos, que son determinados por el coeficiente de similitud (o el valor de distancia) a partir del que se derivó el dendrograma. En el dendrograma la abscisa no tiene un significado particular, excepto para espaciar las entidades originales empleadas en el estudio. La ordenada, por otra parte, representa los valores de similitud, que pueden hacerse sobre una escala.

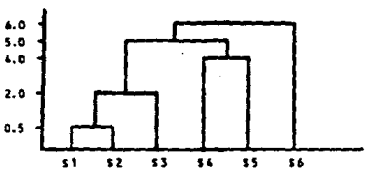


Fig. 5.1a. Ejemplo de dendrograma hipótetico. En el eje de las abscisas se representan diferentes sitios de muestreo (S1, S2,...S6) y en la ordenada los valores de las distancias a las cuales se dió la unión de entidades (sitios de muestreo).

La segunda etapa, la clasificación, consiste en la identificación de los niveles ordinarios (rangos) de la jerarquía con niveles numéricos en el dendrograma. Los grupos de entidades que están agrupados o debajo de algún nivel numérico en un dendrograma se llaman clusters (Fig. 5.1b). Los grupos de entidades que están agrupados en algún rango (nivel ordinario) establecido "a priori" en un sistema clasificatorio deben ser llamadas clases. Este rango o nivel ordinario se establece por el investigador y su identificación por lo tanto está sujeta a ciertos criterios de cada tipo de estudio. Las clases de un rango dado en una jerarquía consisten de aquellos objetos del cluster o del correspondiente nivel inferior en el dendrograma (Jardine y Sibson, 1968).

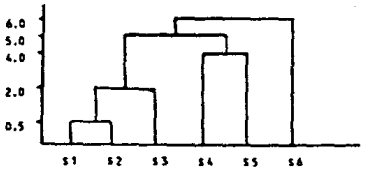


Figura 5.1b. Considerando una distancia = 4.0 se tienen dos clusters, uno con miembros (S1,S2 y S3) unidos a una distancia d=2.0; y otro con miembros (S4 y S5) unidos a d=4.0. También puede verse, como un solo cluster con 5 miembros, para a una d=5.0

Por otra parte, si se establece un rango d=2.0, se tendría una clase constituida por 6 entidades, (S1,S2,S3,S4,S5 y S6) que son las que se reunen en este nivel.

5.1.1 METODOS AGLOMERATIVOS

Las técnicas jerárquicas aglomerativas parten de n grupos donde existen tantos clusters como individuos. Cada grupo contiene a un solo miembro y mediante una serie de fusiones sucesivas se reducen los datos a un cluster que contiene todas las entidades. En cualquier estado particular los métodos unen grupos de individuos más similares, por lo que la diferencia básica en los diferentes métodos aglomerativos es la medida de distancia que se considera.

5.1.1.1 Metodo de enlazamiento simple

A este método también se le conoce como método del mínimo o método del vecino más cercano. Este método fué creado por Florek (1951; en SAS, 1985) y posteriormente modificado por McQuitty (1957) y Sneath (1957; en SAS, 1985).

Los grupos al inicio del análisis están formados por entidades solitarias es decir cada entidad constituye un grupo y existen tantos grupos como entidades. Estos grupos son fusionados en base a la distancia entre sus elementos más cercanos (Anderberg, 1973; Everitt, 1981).

El método de enlazamiento simple se considera el más fácil de las técnicas jerárquicas. En su forma unidimensional, es decir cuando sólo se considera una sóla variable, es aún más

simple. El algoritmo del método se puede resumir en tres etapas:

1. Ordenar las observaciones en secuencia ascendente y tratar cada observación como un grupo.

2. Examinar todos los pares de grupos adyacentes y encontrar los dos individuos más cercanos.

Es importante aclarar que si la medida de asociación es una distancia. Entonces la distancia elegida entre dos miembros es la mínima. Si por el contrario es una correlación, la distancia será la máxima (los individuos con mayor valor de correlación son los más cercanos).

3. Repetir el paso 2 hasta que exista un sólo grupo.

Ejemplo: Dado un conjunto de 10 observaciones para agruparse con la métrica absoluta D = wxj - xtw, que es la más sencilla para analizar el caso; el dendrograma resultante se muestra en la figura 5.2.

En la figura, en el eje de las abscisas se representan las distancias obtenidas por la métrica absoluta, y en la ordenada se encuentran los 10 individuos que se van a agrupar. Se puede observar que los individuos con valores 26 y 27 son los más cercanos de todo el conjunto y por lo tanto los primeros que se unen a una distancia de 1. Después a d=6.0, se unen los elementos con valores 9 y 15, además de 73 y 79. A d=9.0 se fusionan los miembros cuyos valores son 42 y 51. A estas 4 parejas de entidades se unen las demás, hasta que todos los elementos forman un sólo grupo.

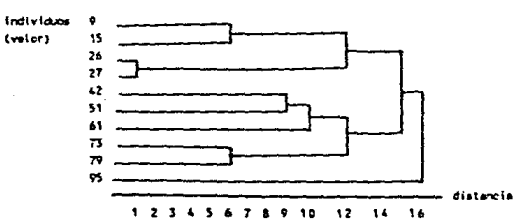


Fig. 5.2. Dendrograma obtenido por el método de enlazamiento simple para un grupo de 10 individuos y una variable.

Al considerar más de una variable y más de 1 entidad el método es un poco más complejo.

Por ejemplo, dados cinco individuos a ser clasificados, y la matriz de distancias entre los individuos:

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & 3.0 & 7.0 & 10.0 & 8.0 \\ 3.0 & 0.0 & 6.0 & 8.0 & 6.0 \\ 7.0 & 6.0 & 0.0 & 4.0 & 5.0 \\ 10.0 & 8.0 & 4.0 & 0.0 & 2.0 \\ 8.0 & 6.0 & 5.0 & \boxed{2.0} & 0.0 \end{bmatrix} \end{array}$$

En la primera etapa del método, los individuos 4 y 5 son fusionados para formar un primer grupo, $d_{45} = 2.0$ es la entrada con el valor más pequeño en la matriz D_1. Las distancias entre este grupo y los individuos restantes 1, 2 y 3 son obtenidas de D_1, como sigue:

$$d_{(45)1} = \{\min d_{14}, d_{15}\} = d_{15} = 8.0$$
$$d_{(45)2} = \{\min d_{24}, d_{25}\} = d_{25} = 6.0$$
$$d_{(45)3} = \{\min d_{34}, d_{35}\} = d_{34} = 4.0$$

ahora se forma una nueva matriz de distancias D_2 dando las distancias entre-individuos, y las distancias grupo-individuos.

$$D_2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & (45) \end{array} \\ \begin{bmatrix} 0.0 & 3.0 & 7.0 & 8.0 \\ \boxed{3.0} & 0.0 & 6.0 & 6.0 \\ 7.0 & 6.0 & 0.0 & 4.0 \\ 8.0 & 6.0 & 4.0 & 0.0 \end{bmatrix} \end{array}$$

Dado que la matriz es simétrica en lo sucesivo sólo se darán los valores de la triangular inferior.

La entrada más pequeña en D_2 es d_{12} que es 3.0, y así los individuos 1 y 2 son fusionados en un segundo grupo, ahora las distancias son:

$$d_{(12)3} = (\min d_{13}, d_{23}) = d_{23} = 6.0$$
$$d_{(12)(45)} = (\min d_{14}, d_{15}, d_{24}, d_{25}) = d_{25} = 6.0$$
$$d_{(45)3} = 4.0 \text{ (como antes)}$$

Esto puede ser representado en una matriz D_3,

$$D_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{bmatrix} 0.0 & & \\ 6.0 & 0.0 & \\ 6.0 & \boxed{4.0} & 0.0 \end{bmatrix} \end{array}$$

La distancia más pequeña es $d_{3(45)}$ que es 4.0 y entonces el individuo 3 es sumado a el grupo que contiene los individuos 4 y 5. Finalmente se unen estos dos grupos para formar un grupo sencillo conteniendo los 5 individuos.

El dendrograma que representa las fusiones entre los individuos es el siguiente (figura 5.3):

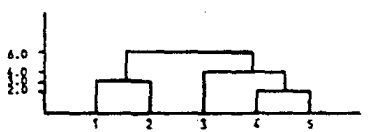


Figura 5.3. Dendrograma por el método de enlace simple para 5 individuos y más de una variable.

Este método es útil para observar la estructura de un conjunto de datos los cuales originan grupos bien separados, cuando los grupos son muy semejantes unos con otros; como se considera la distancia mínima; entonces el método enlaza preferentemente una entidad a un cluster ya existente, en lugar de formar un nuevo cluster. A esta propiedad se le conoce con el nombre de "encadenamiento" y es muchas veces criticada

porque finalmente se enlazan entidades muy disimilares y no se pueden visualizar diferencias marcadas en el conjunto de datos. (Anderberg, 1973).

Para estudios ecológicos donde interesa identificar regiones, este método resulta inadecuado ya que une todos los sitios y no se pueden apreciar regiones dentro del conjunto.

5.1.1.2. Método de enlazamiento completo

Este método es también conocido con el nombre de método del máximo o vecino más lejano y fué creado por Sorensen en 1948 (citado por SAS, 1985). En el la distancia entre grupos se ahora definida como la distancia entre sus pares de individuos más distantes (Anderberg, 1973; Everitt, 1981).

Para el caso unidimensional, tomando el ejemplo anterior se tiene un dendrograma que se muestra en la figura 5.4.

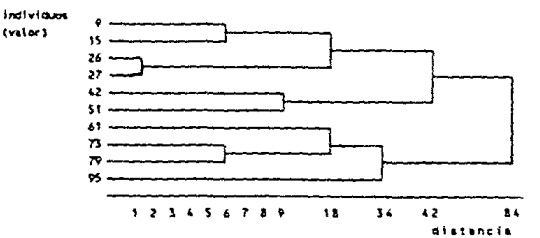


Figura 5.4. Dendrograma por el método de enlace completo para 10 individuos y una variable.

Comparando este diagrama con el de la fig. 5.2, se observa que la unión de los individuos es diferente. Ahora la distancia entre individuos considerada es la máxima y posteriormente se elige la menor de ellas para la unión de individuos. De esta manera, las uniones engloban a un mayor número de individuos en contraste con el método de enlace simple.

Para el caso donde se consideran más de una variable se tiene lo siguiente:

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & 3.0 & 7.0 & 10.0 & 8.0 \\ 3.0 & 0.0 & 6.0 & 8.0 & 6.0 \\ 7.0 & 6.0 & 0.0 & 4.0 & 5.0 \\ 10.0 & 8.0 & 4.0 & 0.0 & 2.0 \\ 8.0 & 6.0 & 5.0 & \boxed{2.0} & 0.0 \end{bmatrix} \end{array}$$

Usando esta técnica para la matriz de distancias D_1 anterior se comienza como con el método de enlazamiento simple, por la fusión de los individuos 4 y 5. las distancias entre este grupo y los tres individuos sobrantes 1, 2 y 3 son obtenidas a partir de D_1 como sigue:

$$d_{(45)1} = (\max d_{14}, d_{15}) = d_{14} = 10.0$$
$$d_{(45)2} = (\max d_{24}, d_{25}) = d_{24} = 8.0$$
$$d_{(45)3} = (\max d_{34}, d_{35}) = d_{35} = 5.0$$

Enseguida se calcula la nueva matriz de distancias D_2 :

$$D_2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & (45) \end{array} \\ \begin{bmatrix} 0.0 & & & \\ \boxed{3.0} & 0.0 & & \\ 7.0 & 6.0 & 0.0 & \\ 10.0 & 8.0 & 5.0 & 0.0 \end{bmatrix} \end{array}$$

La menor distancia entre individuos es ahora d_{12} que es 3.0, por lo que estos individuos se unen para formar un grupo. Las nuevas distancias ahora son:

$$d_{(12)3} = (\max d_{13}, d_{23}) = d_{13} = 7.0$$
$$d_{(12)(45)} = (\max d_{14}, d_{15}, d_{24}, d_{25}) = d_{14} = 10.0$$
$$d_{3(45)} = (\max d_{34}, d_{35}) = d_{35} = 5.0$$

Esto puede ser representado en una matriz D_3,

$$D_3 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{bmatrix} 0.0 & & \\ 7.0 & 0.0 & \\ 10.0 & \boxed{5.0} & 0.0 \end{bmatrix} \end{array}$$

La más pequeña entrada es $d_{3(45)} = 5.0$ y el individuo 3 se une al grupo con los individuos 4 y 5. Por último todos los individuos forman parte de un sólo grupo.

El dendrograma resultante es el de la figura 5.5.

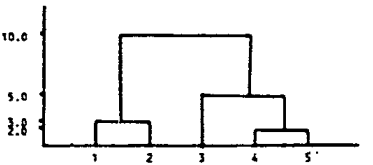


Fig. 5.5. Dendrograma por el método de enlace completo para 5 individuos y más de una variable.

El método de enlazamiento completo genera grupos grandes y proporciona diagramas que muestran las diferencias básicamente dentro de los grupos. Es decir, es un método que sirve para definir grupos grandes, ya que se toma la distancia máxima y con ella se engloban un mayor número de individuos en comparación al método de enlace simple. En contraste con este último método, la interpretación de los grupos puede hacerse sólo en términos de relaciones entre individuos dentro de cada cluster. Esto implica que se pueden llegar a conocer las variaciones internas de un grupo y describirlo de forma separada de todo el conjunto. Permite un mayor grado de análisis que el método de enlace simple.

En los métodos de enlace simple y completo, los dos grupos a ser unidos en cualquier paso son determinados por la distancia entre dos puntos individuales de datos, uno en cada grupo. Así un cluster es siempre representado por uno sólo de sus puntos; además, el punto representativo es siempre extremo

más que típico del grupo al que representa.

5.1.1.3. Método del centroide

La distancia entre grupos es definida como la distancia entre el grupo de centroides o vectores medios más similares. Si pensamos en un agregado de puntos en un espacio geométrico de prueba, el centroide es el punto que se encuentra en el centro de la nube de puntos. El método del centroide fué creado por Sokal y Michener (1958; en SAS, 1985).

El procedimiento entonces une los grupos de acuerdo a la distancia entre sus centroides, los grupos con la distancia más pequeña comienzan a unirse primero.

Por ejemplo, suponga que la técnica se aplica al siguiente grupo de datos, representados por 5 individuos cada uno teniendo valores para 2 variables.

		VARIABLE	
		1	2
	1	2.0	2.0
	2	1.0	3.0
INDIVIDUO	3	5.0	4.0
	4	6.0	6.0
	5	8.0	1.0

Para estos datos se calcula la matriz D_1 con las distancias entre individuos. Los valores numéricos son obtenidos usando la métrica Euclidiana:

$$d_{ij} = \sum_{k=1}^{p} \left[(X_{ik} - X_{jk}) \right]^{1/2}$$

Si se toma el cuadrado de la distancia Euclidiana, se tiene:

$$d_{12} = (2-1)^2 + (2-3)^2 = 1 + 1 = 2$$
$$d_{13} = (2-5)^2 + (2-4)^2 = 9 + 4 = 13$$
$$d_{14} = (2-6)^2 + (2-5)^2 = 16 + 9 = 25$$
$$d_{15} = (2-8)^2 + (2-1)^2 = 36 + 1 = 37$$
$$d_{23} = (1-5)^2 + (3-4)^2 = 16 + 1 = 17$$
$$d_{24} = (1-6)^2 + (3-5)^2 = 25 + 4 = 29$$
$$d_{25} = (1-8)^2 + (3-1)^2 = 49 + 4 = 53$$
$$d_{34} = (5-6)^2 + (4-6)^2 = 1 + 4 = 5$$
$$d_{35} = (5-8)^2 + (4-1)^2 = 9 + 9 = 18$$
$$d_{45} = (6-8)^2 + (5-1)^2 = 4 + 16 = 20$$

La matriz D_1 es:

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & & & & \\ \boxed{2.0} & 0.0 & & & \\ 13.0 & 17.0 & 0.0 & & \\ 25.0 & 29.0 & 5.0 & 0.0 & \\ 37.0 & 53.0 & 18.0 & 20.0 & 0.0 \end{bmatrix} \end{array}$$

La primera etapa, como en todos los métodos aglomerativos, es la elección de la distancia menor. La más pequeña entrada es $d_{12} = 2.0$. Esto forma un cluster con los individuos 1 y 2. Ahora son reducidos los datos y calculado el centroide para la pareja (12).

		VARIABLE	
		1	2
INDIVIDUO	(12)	1.5	2.5
	3	5.0	4.0
	4	6.0	6.0
	5	8.0	1.0

Las nuevas distancias son calculadas.

$$d_{(12)3} = (1.5-5.0)^2 + (2.5-4.0)^2 = 12.25 + 2.25 = 14.5$$
$$d_{(12)4} = (1.5-6.0)^2 + (2.5-6.0)^2 = 20.25 + 12.25 = 32.5$$
$$d_{(12)5} = (1.5-8.0)^2 + (2.5-1.0)^2 = 42.25 + 2.25 = 44.5$$

La nueva matriz D_2 es ahora:

$$D_2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & & & \\ 14.5 & 0.0 & & \\ 32.5 & \boxed{5.0} & 0.0 & \\ 44.5 & 18.0 & 20.0 & 0.0 \end{bmatrix} \end{array}$$

La distancia $d_{34} = 5.0$ es la mínima, así que se unen los individuos 3 y 4. Se tiene ahora:

		VARIABLE	
		1	2
INDIVIDUO	(12)	1.5	2.5
	(34)	5.5	5.0
	5	8.0	1.0

Las distancias son:

$$d_{(12)(34)} = (1.5-5.5)^2 + (2.5-5.0)^2 = 16.0 + 6.25 = 22.25$$
$$d_{(34)5} = (5.5-8.0)^2 + (5.0-1.0)^2 = 6.25 + 16.0 = 22.25$$

La matriz reducida D_3 es:

$$D_3 = \begin{array}{c} \\ (12) \\ (34) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} (12) & (34) & 5 \end{array} \\ \begin{bmatrix} 0.0 & & \\ \boxed{22.25} & 0.0 & \\ 44.5 & \boxed{22.25} & 0.0 \end{bmatrix} \end{array}$$

En este caso existe un empate por lo que el método puede tomar cualquiera de las dos opciones.

Tomemos la distancia entre el individuo 5 y el grupo (34). Los nuevos centroides son calculados:

		VARIABLE	
		1	2
INDIVIDUO	(12)	1.5	2.5
	(345)	6.75	3.0

$$d_{(12)(345)} = (1.5-6.75)^2 + (2.5-3.0)^2 = 27.56 + 0.25 = 27.81$$

Y por último se unen los dos grupos nuevos formando uno solo. El dendrograma que muestra las fusiones para los individuos del ejemplo quedaría (figura 5.6):

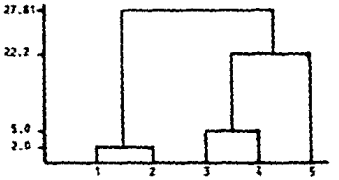


Fig. 5.6. Dendrograma por el método del centroide para 5 individuos y dos variables.

El agrupamiento por centroide es uno de los métodos diseñados para encontar un término medio entre los extremos del agrupamiento por enlace simple en un lado y el enlace completo en otro lado.

En el método del centroide la distancia entre dos clusters es tomada como la distancia entre sus centroides. El centroide de un cluster es el punto que representa el punto promedio del grupo.

5.1.1.4. Método del grupo promedio

Este método define la distancia entre un individuo y un grupo que tiene varios miembros como el promedio de las distancias entre el individuo y los miembros del cluster. Este método fué creado por Sokal y Michener (1958; en SAS, 1985), quienes usan este promedio como una medida de distancia entre un individuo y un grupo de individuos, mientras que Lance y Williams (1966) lo usan como medida de distancia entre grupos (Lance y Williams, 1966; en Colm, 1977). La manera más común de calcular el promedio, es con el método de pares de grupos no

ponderados usando promedios aritméticos (UPGMA), el cual usa el promedio aritmético simple no ponderado (Sokal y Sneath, 1963; Sneath y Sokal, 1973).

Para analizar este método supóngase que se tiene una matriz de datos X, donde se han registrado cuatro observaciones para cinco individuos.

$$X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 2.0 & 1.0 & 3.0 & 4.0 \\ 5.0 & 3.0 & 4.0 & 6.0 \\ 8.0 & 6.0 & 0.0 & 5.0 \\ 4.0 & 3.0 & 2.0 & 1.0 \\ 3.0 & 1.0 & 7.0 & 0.0 \end{bmatrix} \end{array}$$

La primera etapa consiste en convertir esta matriz de datos en una matriz de distancias. Estas se calculan usando la métrica Euclidiana.

$$d_{12} = [(2-5)^2+(1-3)^2+(3-4)^2+(4-6)^2]^{1/2} = (9+4+1+4)^{1/2} = 4.24$$
$$d_{13} = [(2-8)^2+(1-6)^2+(3-0)^2+(4-5)^2]^{1/2} = (36+25+9+1)^{1/2} = 8.43$$
$$d_{14} = [(2-4)^2+(1-3)^2+(3-2)^2+(4-1)^2]^{1/2} = (4+4+1+9)^{1/2} = 4.24$$
$$d_{15} = [(2-3)^2+(1-1)^2+(3-7)^2+(4-0)^2]^{1/2} = (1+0+16+16)^{1/2} = 5.74$$
$$d_{23} = [(5-8)^2+(3-6)^2+(4-0)^2+(6-5)^2]^{1/2} = (9+9+16+1)^{1/2} = 5.92$$
$$d_{24} = [(5-4)^2+(3-3)^2+(4-2)^2+(6-1)^2]^{1/2} = (1+0+4+25)^{1/2} = 5.48$$
$$d_{25} = [(5-3)^2+(3-1)^2+(4-7)^2+(6-0)^2]^{1/2} = (4+4+9+36)^{1/2} = 7.28$$
$$d_{34} = [(8-4)^2+(6-3)^2+(0-2)^2+(5-1)^2]^{1/2} = (16+9+4+16)^{1/2} = 6.71$$
$$d_{35} = [(8-3)^2+(6-1)^2+(0-7)^2+(5-0)^2]^{1/2} = (25+25+49+25)^{1/2} = 11.13$$
$$d_{45} = [(4-3)^2+(3-1)^2+(2-7)^2+(1-0)^2]^{1/2} = (1+4+25+1)^{1/2} = 5.57$$

La matriz de datos D_1 queda:

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & & & & \\ 4.24 & 0.0 & & & \\ 8.43 & 5.92 & 0.0 & & \\ 4.24 & 5.48 & 6.71 & 0.0 & \\ 5.74 & 7.28 & 11.13 & 5.57 & 0.0 \end{bmatrix} \end{array}$$

La más pequeña distancia es 4.24, sólo que en este caso existen dos parejas de individuos que poseen este valor, por lo que se puede elegir cualquiera de ellos, tomemos a los individuos 1 y 2.

El siguiente paso es calcular una nueva matriz de distancia que muestre las distancias interindividuales o individuo-grupo.

$$d_{(12)3} = \frac{1}{2}(d_{13} + d_{23}) = \frac{1}{2}(8.43 + 5.92) = 7.175$$
$$d_{(12)4} = \frac{1}{2}(d_{14} + d_{24}) = \frac{1}{2}(4.24 + 5.48) = 4.86$$
$$d_{(12)5} = \frac{1}{2}(d_{15} + d_{25}) = \frac{1}{2}(5.74 + 7.28) = 6.51$$

$$D_2 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.0 & & & \\ 7.17 & 0.0 & & \\ \boxed{4.86} & 6.71 & 0.0 & \\ 6.51 & 11.13 & 5.57 & 0.0 \end{bmatrix} \end{array}$$

La menor distancia es 4.86 y se unen el individuo 4 con el grupo (12). Las nuevas distancias y la matriz se calculan:

$$d_{(124)3} = \frac{1}{3}(d_{13} + d_{23} + d_{43}) = \frac{1}{3}(8.43 + 5.92 + 6.71) = 7.02$$
$$d_{(124)5} = \frac{1}{3}(d_{15} + d_{25} + d_{45}) = \frac{1}{3}(5.74 + 7.28 + 5.57) = 6.20$$

$$D_3 = \begin{array}{c} \\ (124) \\ 3 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} (124) & 3 & 5 \end{array} \\ \begin{bmatrix} 0.0 & & \\ 7.02 & 0.0 & \\ \boxed{6.20} & 11.13 & 0.0 \end{bmatrix} \end{array}$$

La distancia nueva es 6.20 y el grupo (124) se une el individuo 5. La nueva distancia es:

$$d_{(1245)3} = \frac{1}{4}(d_{13}+d_{23}+d_{43}+d_{53}) = \frac{1}{4}(8.43+5.92+6.71+11.13) = 8.05$$

Finalmente el diagrama de relaciones es el de la figura 5.7.

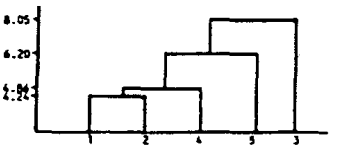


Fig. 5.7. Dendrograma por el método del grupo promedio para 5 individuos y 4 variables.

En el agrupamiento del grupo promedio, a diferencia del método del centroide, los clusters no pueden ser identificados con puntos representativos precisos y, por lo tanto, el concepto de distancia entre grupos es inevitablemente difuso.

El agrupamiento por grupo promedio dá grupos con centros indefinidos y el agrupamiento por centroide da grupos con centros exactamente definidos.

5.1.1.5. Método de Ward's

Ward (1963) propone que en cualquier estado de un análisis la pérdida de información que resulta del agrupamiento de individuos en clusters puede ser medida por la suma total de las desviaciones cuadradas de cada punto con respecto a la media del cluster al que pertenecen. En las etapas del análisis, se considera la unión de un posible par de clusters y los dos clusters cuya fusión genera en el incremento mínimo en la suma del cuadrado del error, son combinados. La mínima varianza es un caso especial del método de Ward, y es un ejemplo de técnica de optimización ya que se elige aquella fusión que dá un incremento mínimo en la suma del cuadrado del error como criterio (Everitt, 1981; SAS, 1985).

Por ejemplo, suponga 5 individuos a ser agrupados a partir de los valores sobre una variable simple usando este método. Los valores de la variable para cada uno de los 5 individuos son:

	VARIABLE
	VALOR
1	2
2	4
INDIVIDUO 3	5
4	8
5	10

La suma de los cuadrados del error (E.S.S.) está dada por:

$$E.S.S. = \sum_{i=1}^{p} Xi^2 - \frac{1}{n} \left(\sum_{i=1}^{p} Xi \right)^2$$

donde X_i es el resultado del i-ésimo individuo. En la etapa uno cada individuo es considerado como un simple miembro del grupo y así E.S.S. es cero. Los dos individuos cuya fusión resulta en el incremento mínimo en E.S.S. forman el primer grupo.

Por lo que se deben calcular los valores de E.S.S. para todas las combinaciones posibles.

1-2 $E.S.S. = (2^2 + 4^2) - \frac{1}{2}(2+4)^2 = 20 - \frac{1}{2}(36) = 2.0$

1-3 $E.S.S. = (2^2 + 5^2) - \frac{1}{2}(2+5)^2 = 29 - \frac{1}{2}(49) = 4.5$

1-4 $E.S.S. = (2^2 + 8^2) - \frac{1}{2}(2+8)^2 = 68 - \frac{1}{2}(100) = 18.0$

1-5 $E.S.S. = (2^2 + 10^2) - \frac{1}{2}(2+10)^2 = 104 - \frac{1}{2}(144) = 32.0$

2-3 $E.S.S. = (4^2 + 5^2) - \frac{1}{2}(4+5)^2 = 41 - \frac{1}{2}(81) = 0.5$

2-4 $E.S.S. = (4^2 + 8^2) - \frac{1}{2}(4+8)^2 = 80 - \frac{1}{2}(144) = 8.0$

2-5 $E.S.S. = (4^2 + 10^2) - \frac{1}{2}(4+10)^2 = 116 - \frac{1}{2}(196) = 18.0$

3-4 $E.S.S. = (5^2 + 8^2) - \frac{1}{2}(5+8)^2 = 89 - \frac{1}{2}(169) = 4.5$

3-5 $E.S.S. = (5^2 + 10^2) - \frac{1}{2}(5+10)^2 = 125 - \frac{1}{2}(225) = 12.5$

4-5 $E.S.S. = (8^2 + 10^2) - \frac{1}{2}(8+10)^2 = 164 - \frac{1}{2}(324) = 2.0$

El más pequeño valor para E.S.S. es 0.5, por lo que los individuos 2 y 3 se unen primero.

Calculando los nuevos valores para E.S.S. tenemos:

$$E.S.S.(23)1 = (4^2+5^2+2^2) - \frac{1}{3}(4+5+2)^2 = 45 - \frac{1}{3}(121) = 4.67$$

$$E.S.S.(23)4 = (4^2+5^2+8^2) - \frac{1}{3}(4+5+8)^2 = 105 - \frac{1}{3}(289) = 8.67$$

$$E.S.S.(23)5 = (4^2+5^2+10^2) - \frac{1}{3}(4+5+10)^2 = 141 - \frac{1}{3}(361) = 20.67$$

Como el valor más pequeño para E.S.S. es 4.67, el individuo 1 se une al grupo anterior que contiene a los individuos 2 y 3, incrementando el E.S.S. a 5.17. Por último se calculan los E.S.S. para las dos últimas combinaciones:

$$E.S.S.(123)4 = (2^2+4^2+5^2+8^2) - \frac{1}{4}(2+4+5+8)^2 = 109 - \frac{1}{4}(361) = 18.75$$

$$E.S.S.(123)5 = (2^2+4^2+5^2+10^2) - \frac{1}{4}(2+4+5+10)^2 = 145 - \frac{1}{4}(441) = 34.75$$

El individuo 4 ahora se une al grupo de 3 miembros y el error aumenta a 23.92. Por último se une también el individuo 5 a una distancia de 40.8.

El dendrograma que muestra la secuencia de uniones queda así (figura 5.8):

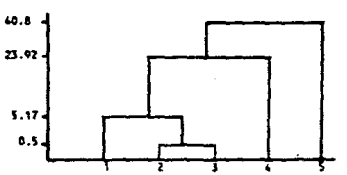


fig. 5.8. Dendrograma por el método de Ward para 5 individuos y una variable.

El método de la mínima varianza es adecuado cuando se sospecha que alguno (o todos) los elementos pertenecen a una o más clases homogéneas.

El agrupamiento por mínima varianza, así como el agrupamiento del vecino más lejano, tienden a dar grupos de justamente igual tamaño.

5.1.2 METODOS DIVISIVOS

Los métodos divisivos parten de un sólo grupo que contiene un total de N entidades y por sucesivas divisiones se originan grupos cada vez más pequeños.

Uno de los métodos divisivos es el propuesto por MacNaughton-Smith (1964). Este es un método politético pues intervienen varias características en la división. En este método un grupo dividido es acumulado por adición secuencial de la entidad cuya disimilitud total con el grupo dividido es la máxima. (MacNaughton-Smith, 1964; en Everitt, 1981).

5.1.2.1 Metodo de MacNaughton-Smith

La medida de disimilitud usada es el promedio de la distancia Euclidiana entre cada entidad y las otras entidades en el grupo. Por ejemplo, considere la matriz de distancia D, que tiene las distancias entre 7 individuos:

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{bmatrix} 0 & 10 & 7 & 30 & 29 & 38 & 42 \\ 10 & 0 & 7 & 23 & 25 & 34 & 36 \\ 7 & 7 & 0 & 21 & 22 & 31 & 36 \\ 30 & 23 & 21 & 0 & 7 & 10 & 13 \\ 29 & 25 & 22 & 7 & 0 & 11 & 17 \\ 38 & 34 & 31 & 10 & 11 & 0 & 9 \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{bmatrix} \end{array}$$

El promedio de las distancias para cada individuo es:
1) 22.28 2)19.28 3)17.71 4)14.86 5)15.86 6)19 7)21.86

Estos individuos son divididos en dos grupos usando este método. El individuo usado para iniciar la partición es aquel cuyo promedio de distancia a partir de sus individuos remanentes es la máxima. Se encuentra que es el individuo 1, dando los grupos:

(1) y (2,3,4,5,6,7).

Enseguida la distancia promedio de cada individuo en el grupo principal a los individuos en el grupo dividido es encontrada, siguiendo por la distancia promedio de cada

individuo en el grupo principal a los otros individuos en ese grupo. La diferencia entre estos dos promedios es encontrada. Esto dá el siguiente resultado:

Individuo	Distancia promedio a el gpo. partido (1)	Distancia promedio a el gpo. principal (2)	(2-1)
2	10.0	25.0	15.0
3	7.0	23.4	16.4
4	30.0	14.8	-15.2
5	29.0	16.4	-12.6
6	38.0	19.0	-19.0
7	42.0	22.2	-19.8

La diferencia máxima es 16.4 para el individuo 3, que es por lo tanto acumulada en el grupo partido dando los dos grupos:

(1,3) y (2,4,5,6,7).

Repitiendo el análisis tenemos los siguientes resultados:

Individuo	Distancia promedio a el gpo. partido (1)	Distancia promedio a el gpo. principal (2)	(2-1)
2	10+ 7/2 = 8.5	23+25+34+36/4 =29.5	21.0
4	30+21/2 = 25.5	23+7+10+13/4 =13.2	-12.3
5	29+22/2 = 25.5	25+7+11+17/4 =15.0	-10.5
6	38+31/2 = 34.5	34+10+11+9/4 =16.0	-18.5
7	42+36/2 = 39.0	36+13+17+9/4 =18.7	-20.3

Esto lleva a la acumulación del individuo 2 en el grupo dividido, y así en esta etapa los dos sub-grupos son

(1,3,2) y (4,5,6,7).

Individuo	Distancia promedio a el gpo. partido (1)		Distancia promedio a el gpo. principal (2)	(2-1)
4	30+21+23/3=24.3	7+10+13/3 =	10.0	-14.3
5	29+22+25/3=25.3	7+11+17/3 =	11.7	-13.6
6	38+31+34/3=34.3	10+11+9/3 =	10.0	-24.3
7	42+36+36/3=38.0	13+17+9/3 =	13.0	-25.0

Como todas las diferencias son ahora negativas, el proceso puede ahora continuar (si se desea) separadamente en cada uno de los sub-grupos.

Este método tiene la ventaja que el cálculo requerido es considerablemente menor que para "un método con todas las subdivisiones posibles".

Como en otras técnicas divisivas, una ineficiente partición temprana puede no ser corregida en una etapa posterior. Este es también el caso para las técnicas aglomerativas.

Las técnicas *monotéticas* son usadas generalmente en casos donde se tienen datos binarios.

5.1.2.2 Análisis de asociación

Suponga que m atributos binarios son medidos para una población de N objetos. Entonces dado cualquier cluster de digamos n objetos desearíamos encontrar la mejor división del cluster en dos sub-grupos en términos de la presencia o ausencia de uno de los caracteres binarios, digamos el caracter T. Esto es, un sub-grupo debe contener aquellos objetos que poseen T y el otro debe contener aquellos objetos que carecen de T, y el atributo T es seleccionado así tal que la disimilitud entre los dos clusters es maximizada en términos del criterio de selección. Este tipo de técnica es usualmente conocido como análisis de asociación, y varias variantes de la técnica han sido dados por (Lambert y Williams, 1962, 1966; y MacNaughton-Smith, 1965; en Everitt, 1981), que difieren en el criterio de división adoptado.

Las técnicas de análisis de asociación tienen aplicación en biogeografía, específicamente en la clasificación de sitios de censos de especies para definir provincias

bióticas. En estos casos se trata de dividir grandes extensiones de territorio y por lo tanto, sólo es importante el registro de la presencia o la ausencia de las especies (Pielou, 1979). Sin embargo, sus aplicaciones originales fueron todas ecológicas, donde las variables fueron las especies de plantas presentes en n cuadrantes. Estos cuadrantes se dividen en dos subgrupos sobre la base de las especies que mejor los separan (Gower, 1976).

Ejemplo: Dados 6 individuos para los que se midieron 4 variables, el objetivo es encontrar subgrupos que posean una de las variables y otros que carezcan de esta. Se procede entonces a realizar tablas de contingencia comparando todas las posibles combinaciones.

En la primera etapa se registran los valores de cada variable para los individuos.

		VARIABLES			
		1	2	3	4
INDIVIDUOS	1	0	1	0	1
	2	1	1	1	0
	3	0	0	1	0
	4	1	0	0	1
	5	0	1	1	1
	6	0	0	1	0

Enseguida se construyen las tablas de contingencia para todas las combinaciones entre variables.

var. 1 \ var. 2	1	0	
1	1	1	2
0	2	2	4
	3	3	6

var. 2 \ var. 3	1	0	
1	1	1	2
0	3	1	4
	4	2	6

var. 1 \ var. 4	1	0	
1	1	1	2
0	2	2	4
	3	3	6

var. 3 \ var. 4	1	0	
1	2	1	3
0	2	1	3
	4	2	6

var. 2 \ var. 4	1	0	
1	2	1	3
0	1	2	3
	3	3	6

var. 3 \ var. 4	1	0	
1	1	3	4
0	2	0	2
	3	3	6

Posteriormente se calculan los valores de ji-cuadrada para cada tabla de contingencia, de acuerdo a la siguiente expresión:

$$X_{jk}^2 = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

$$X_{12}^2 = \frac{((1*2)-(1*2))^2 6}{(1+1)(1+2)(1+2)(2+2)} = \frac{0}{72} = 0$$

$$X_{13}^2 = \frac{((1*1)-(1*3))^2 6}{(1+1)(1+3)(1+1)(3+1)} = \frac{24}{64} = 0.375$$

$$X_{14}^2 = \frac{((1*2)-(1*2))^2 6}{(1+1)(1+2)(1+2)(2+2)} = \frac{0}{72} = 0$$

$$X_{23}^2 = \frac{((2*1)-(1*2))^2 6}{(2+1)(2+2)(1+1)(2+1)} = \frac{0}{72} = 0$$

$$X_{24}^2 = \frac{((2*2)-(1*1))^2 6}{(2+1)(2+1)(1+2)(1+2)} = \frac{54}{27} = 2.0$$

$$X_{34}^2 = \frac{((1*0)-(3*2))^2 6}{(1+3)(1+2)(3+0)(2+0)} = \frac{216}{72} = 3.0$$

$X_{12}^2 + X_{13}^2 + X_{14}^2 = 0 + 0.375 + 0 = 0.375$

$X_{21}^2 + X_{23}^2 + X_{24}^2 = 0 + 0 + 2.0 = 2.0$

$X_{31}^2 + X_{32}^2 + X_{34}^2 = 0.375 + 0 + 3.0 = 3.375$

$X_{41}^2 + X_{42}^2 + X_{43}^2 = 0 + 2.0 + 3.0 = 5.0$ Máximo

Como ésta técnica es monotética, se usará el criterio de la presencia o la ausencia de un solo atributo. Para este caso, después de la obtención de los valores de ji-cuadrada para todas las tablas de contingencia, se hace la suma de los valores de ji-cuadrada para cada asociación de cada variable. Se identifica el valor máximo, este valor indica que la variable que tuvo la mayor asociación con las demás es la 4; entonces, se tomará la presencia o la ausencia de la variable 4 como criterio para que se inicie la separación entre todo el grupo. Se formaría un primer grupo (individuos 1,4 y 5), que son los que presentan la variable 4 y el segundo grupo (individuos 2,3 y 6), que son los que no la tienen.

5.2 CLASIFICACIONES NO JERARQUICAS.

Estas técnicas proporcionan soluciones que no necesariamente representan relaciones jerárquicas entre las entidades. Dentro de estas se encuentran las técnicas de densidad, donde los clusters son formados por el exámen de reçiones que contienen una concentración relativamente densa de entidades. También existen las técnicas de amontonamiento donde las clases o montones pueden sobreponerse; y las técnicas de optimización que son así llamadas porque sé optimiza alguna medida numérica predefinida, la cual es indicadora de una solución del agrupamiento (o división) deseable. Las técnicas de optimización permiten la relocación de las entidades, así que existe la posibilidad de que una partición inicial pueda ser corregida en una etapa posterior del análisis si se considera necesario (Everitt, 1981).

La idea central en muchos de estos métodos es seleccionar alguna partición inicial de los datos y entonces alterar los miembros del cluster para así obtener la mejor partición (Anderberg, 1973).

CAPITULO 6

ESTUDIO DE CASO: ANALISIS CLUSTER COMO HERRAMIENTA ANALITICA PARA LA CLASIFICACION DE TIERRAS.

En el capítulo anterior se revisaron algunas de las técnicas jerárquicas del análisis cluster y se ilustraron con ejemplos sencillos. Ahora se muestra un estudio de caso donde se obtuvieron dendrogramas para un conjunto de datos reales y donde el objetivo fué la generación de una clasificación de tierras.

Se realizó un análisis cluster para respaldar los criterios biológicos y fisiográficos de la clasificación de tierras de una zona con fines de conservación. Se seleccionaron los datos de mayor importancia para la clasificación y se realizaron análisis comparativos entre los métodos del enlace simple, enlace completo, centroide, método del grupo promedio y el método de Ward. Mediante el análisis cluster y por fotointerpretación se definieron 17 unidades homogéneas y se asignó una simbología a cada una de ellas; esto se representó en un mapa topográfico para la zona.

6.1 METODO CLASICO EN LA CLASIFICACION DE TIERRAS.

Es necesario distinguir el concepto de tierra del de suelo. El concepto de tierra es más amplio pues se define como una área específica de la superficie terrestre; cuyas características se refieren a todos los atributos razonablemente estables o cíclicamente predecibles de la biósfera como los de la atmósfera, el suelo, la geología, la hidrología, la vegetación, la fauna y los resultados de la actividad humana pasada y presente, así como las interacciones de todos ellos. (FAO, 1976; Ponce y Cuanalo, 1976).

Cuando se realizan clasificaciones de tierras generalmente se considera la aptitud de ellas a la agricultura de manera implícita, pues se dá mayor peso a los criterios agrícolas la mayoría de las veces. Así se tiene por ejemplo, la clasificación de tierras de la USDA usada por el INEGI en la

elaboración de mapas de uso potencial, donde se establecen ocho clases y se dan las características de cada una de ellas. Como resultado se tiene que aquellas zonas de más difícil acceso se proponen como no aptas ni para la agricultura ni para la explotación forestal, y se designan para la conservación, con este criterio sólo coinciden en esta clase los pedregales (INEGI, 1981). Si bien los pedregales tienen su importancia, existen infinidad de zonas ideales para conservarse por su potencial biológico, como es el caso de los bosques.

En una clasificación, desde el inicio se dá un enfoque particular de acuerdo a los objetivos que se persiguen. Estos objetivos determinan los criterios de clasificación y los criterios determinan los atributos que deben ser medidos (Usher, 1986).

Una clasificación objetiva de tierras debe procurar el uso de atributos medibles en campo. Estos pueden ser de dos tipos: los bióticos, es decir aquellos que consideran las abundancias de organismos tanto vegetales como animales; y los fisiográficos, donde se miden las formas del terreno y las características físicas y químicas de sus suelos.

En este estudio se planteó clasificar las tierras de una zona con fines de conservación y crianza de organismos en semicautiverio, por lo que se propuso la realización de una clasificación de tierras basada en criterios biológicos y fisiográficos.

Por ser una región con características propias, y los objetivos que se plantearon fueron muy específicos; no se usó una clasificación ya existente, puesto que no representaría de forma adecuada a cada clase de tierra. Debido a que no existe un sistema de clasificación que se ajuste a los objetivos de las áreas protegidas, se decidió realizar un análisis de datos que generara sus propias clases de tierras.

6.2 DESCRIPCION DEL AREA DE ESTUDIO.

La Estación de Aprovechamiento de la Vida Silvestre "Ing. Luis Macías Arellano", se localiza en el municipio Villa de Allende, distrito Valle de Bravo, en el Estado de México

(Contreras y Melo, 1974).

Colinda al norte con Villa Victoria, al sur con Valle de Bravo, al este con Amanalco y al oeste con Donato Guerra (figura 6.1).

La estación está situada en la Cordillera Neovolcánica, sobre la Sierra de Zitácuaro. Se considera al Eje Neovolcánico como una "zona de transición", ya que en su perímetro sur, oriente y poniente, limita con áreas tropicales.

Le corresponde un clima templado subhúmedo con lluvias de verano (Cw).

Su vegetación está compuesta por bosque secundario de pinos que alterna con bosque primario de encinos donde se encuentra una comunidad vegetal muy variada (Sánchez, 1970).

De acuerdo a la carta edafológica (INEGI, 1978), los suelos de la Estación son andosoles; que son suelos con alofanos formados sobre escorias o cenizas volcánicas e incluso sobre roca volcánica consolidada (Duchaufour, 1975).

6.3 METODO.

La primera parte del trabajo fué la fotointerpretación de la zona y la elaboración de un mapa provisional de unidades homogéneas.

Se realizaron los muestreos de vegetación con la técnica de Relevé. Esta técnica sirve para hacer estimaciones semicuantitativas de coberturas de acuerdo a la escuela fitosociológica de Zurich-Montpellier (Braun-Blanquet, 1979). (Relevé significa abstracción). Se usó la escala de coberturas de Domin-Krajina (Müller-Dombois, 1974). Además se tomaron muestras de suelos en diferentes unidades delimitadas.

En la fase de laboratorio se determinaron algunas de las propiedades de los suelos que fueron: color, densidad aparente, densidad real, textura, pH activo, pH potencial, porcentaje de materia orgánica y capacidad de intercambio catiónico total.

Posteriormente se realizó una selección de los datos que se tomarían en cuenta para la realización del análisis cluster. De los datos de vegetación se tomaron las especies de

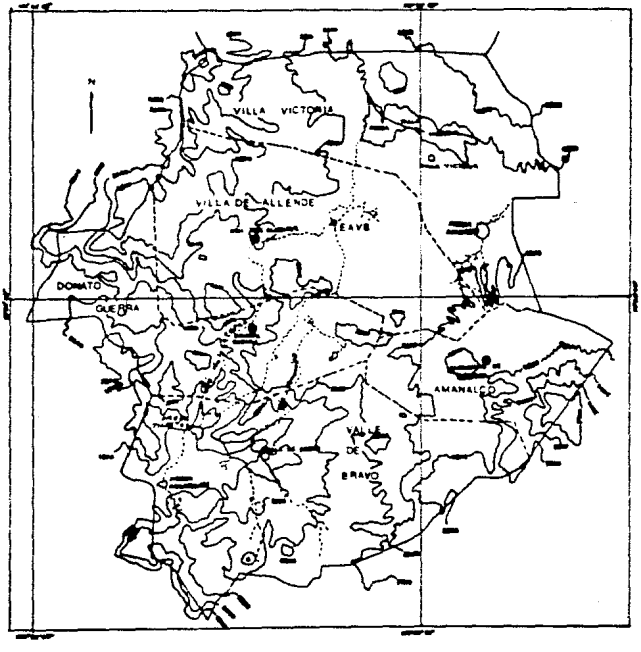


FIG. 6.1 Localización de la EAVS "Ing. Luis Madrid Anaya"

Equidistancia 200 m.

Tomado de Cartografía y Mapa (1981)

mayor distribución en la estación y aquellas especies muy importantes en la fisonomía de ésta como las especies arbóreas.

De suelos se incluyeron los datos de pH, porcentaje de materia orgánica y capacidad de intercambio catiónico sólo de la capa superficial del suelo pues se considera que ésta es la que interviene más intimamente en la cubierta vegetal y en las geoformas. Dentro de éstas se consideraron la pendiente y la pedregosidad.

En la cuarta etapa se hicieron gráficas bidimensionales para algunas especies con el fin de observar su comportamiento. Se realizó el análisis cluster para vegetación, suelos y fisiografía. Para esto se usaron dos paquetes de computadora (STATGRAPHICS Y SAS) tanto de análisis numérico como gráfico.

Con los resultados obtenidos se definieron nuevamente las unidades en las fotos y se vaciaron los límites de cada una de ellas en un mapa topográfico.

Con base en todo lo anterior se elaboró un mapa base de clases de tierras para la Estación de Aprovechamiento de la Vida Silvestre.

6.4 RESULTADOS.

En la figura 6.2 se muestra el par estereoscópico donde se definieron las unidades de paisaje por fotointerpretación; para trazarlas no sólo se tomaron criterios de textura y tono de gris, sino también por características geomorfológicas (laderas, valles, cimas o planicies).

Con respecto al muestreo de vegetación, se hicieron en cada una de las unidades definidas un total de 3 cuadrantes de 512 m^2 y se sumaron los valores de sus coberturas. Se establecieron diferentes clases de estratos, así que se asignó un valor de cobertura especie-estrato por el uso de la escala de Domin Krajina. Los resultados se muestran en el cuadro 6.1. (ver anexo).

Se determinaron algunas propiedades del suelo que se presentan en el cuadro 6.2. (ver anexo).Obsérvese la semejanza que presentan las propiedades del suelo entre los sitios.

Con el fin de que las diferencias entre los suelos de

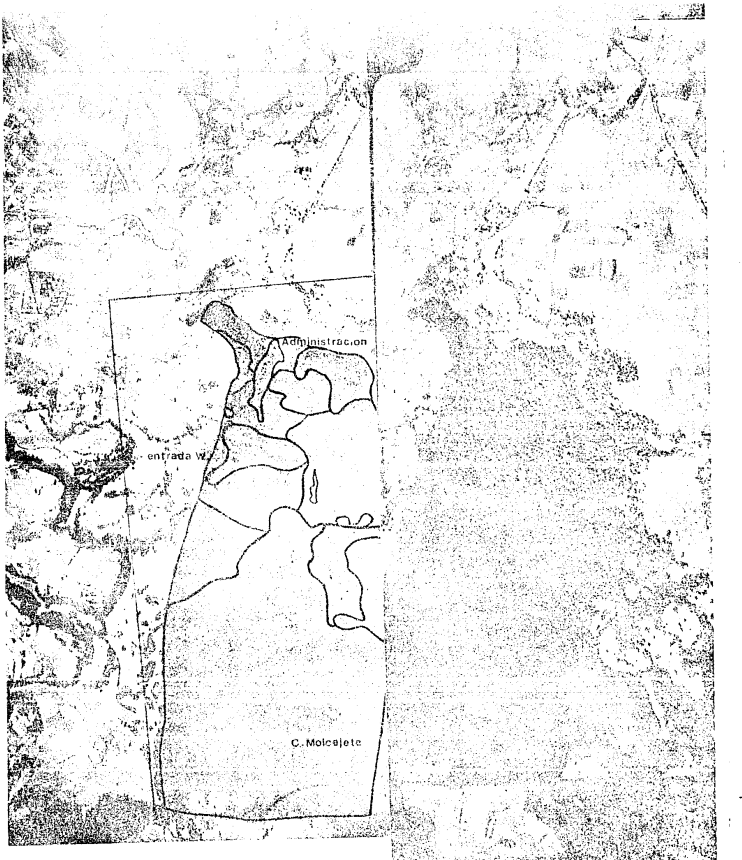


FIG. 6.2 Par estereoscópico del área de estudio

cada sitio se hicieran más notorias, se establecieron 4 clases tanto para suelos como para fisiografía, como se muestra en el cuadro 6.3. (ver anexo).

Posteriormente se hicieron gráficas bidimensionales de especies para observar los grupos que se formaban. Se observó que al considerar parejas de especies arbóreas el agrupamiento era distinto que cuando se graficaban las dominancias de especies arbustivas y herbáceas; aún usando el mismo método y el mismo número de grupos.

Durante el análisis cluster se probaron varios métodos debido a que cada uno de ellos agrupa de forma diferente. Los métodos ensayados fueron el del vecino más cercano, promedio, centroide, vecino más lejano y el de la mínima varianza.

Los datos de suelos y fisiografía que se consideraron en el análisis fueron los que se muestran en el cuadro 6.4. Los datos de vegetación que se eligieron se presentan en los cuadros 6.5 y 6.6. (ver anexo).

Con estos resultados, mediante la fotointerpretación y observaciones de campo, se definieron finalmente las unidades en el mapa.

Se realizó un perfil de vegetación para la estación (figura 6.7), donde se representan algunas unidades definidas y sus principales características.

Se elaboró y asignó simbología al mapa final de clases de tierras (figura 6.8).

6.5 EL ANALISIS CLUSTER COMO ALTERNATIVA EN LA CLASIFICACION DE TIERRAS.

Del conjunto de datos obtenidos tanto en campo como en laboratorio, se realizó una selección con el fin de que no todos se incluyeran en el análisis de grupos, ya que muchos de ellos no sirven como entidades que permitan la clasificación de tierras, es decir son datos comunes a todas las zonas de la estación y en una clasificación se deben emplear atributos singulares, no generales (Pielou, 1984).

Se realizó el análisis cluster para todas las especies vegetales, después para el estrato arbóreo solamente y por

último para el estrato arbustivo y herbáceo. Esto se hizo con el propósito de establecer si algún estrato en particular determinaba las clases de tierras, o bien si la clasificación debiera incluir a todas las especies vegetales. Para estos tres tipos de casos se probaron los métodos de agrupamiento por enlace sencillo, enlace completo, centroide, promedio y por el método de Ward. En primer término se tomaron los valores originales de las variables, se repitió el análisis para variables normalizadas y para variables estandarizadas.

El agrupamiento para todas las especies vegetales fué el mismo para variables originales (el programa SAS normaliza las variables, a menos que se especifique que no lo haga) y para variables sin normalizar, la única diferencia es en la escala. (figura 6.3 a y b) Por lo tanto se pueden analizar los resultados con cualquiera de estas dos pruebas y las conclusiones a las que se llegaría serían las mismas.

Se recomienda estandarizar las variables cuando estas cambian de manera drástica o bien cuando se tienen variables medidas en distintas escalas, pues lo que se busca es disminuir las variaciones entre variables. En este estudio las variables fueron medidas en la misma escala y no varían demasiado. Para variables estandarizadas se observó un agrupamiento distinto de los anteriores que no refleja las observaciones de campo y de fotointerpretación, por estos motivos se considera que no es necesario trabajar con variables estandarizadas (figura 6.3 c).

En estas tres pruebas, (variables normalizadas, sin normalizar y estandarizadas), el único método en el que se observa un comportamiento anómalo en el agrupamiento, es el del centroide, pues existen ciertas etapas del agrupamiento donde la unión se dá a una distancia dada x, y en una etapa posterior la distancia nueva resulta ser menor que la distancia anterior x. A esta característica se le conoce con el nombre de reversos que produce el método y se considera un error de la técnica, pues no se puede establecer una jerarquía en el dendrograma.

Los reversos que produce el método del centroide se repiten tanto para árboles como para arbustos y también con variables originales, normalizadas y estandarizadas

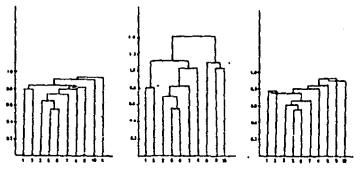


FIGURA 4.3.a.

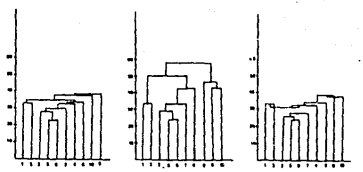


FIGURA 4.3.b.

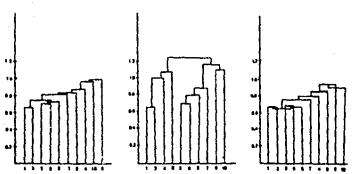


FIGURA 4.3.c.

FIGURA 4.3. Dendrogramas para todos los especies vegetales obtenidos por cinco metodos de agrupamiento diferentes. Observe las semejanzas entre cada método tanto para variables estandarizadas (b), como sin estandarizar (a). (c) corresponde a los cinco mismos metodos usando variables estandarizadas.

(figuras 6.3, 6.4 y 6.5).

Se considera que el método del centroide es adecuado cuando se tienen variables que estan medidas en diferentes escalas, pues lo que se busca es la generación de un punto central imaginario que sea el promedio de las observaciones y aunque los valores de las variables cambien drásticamente o esten medidas en diferentes escalas se puede establecer ese punto imaginario.

Al descartar el método del centroide por su característica de producir reversos, restan 4 métodos en el análisis; al observar los agrupamientos por el método de enlace simple, completo, promedio y de Ward (figura 6.3 a y b), resulta que el método de enlace simple dá una clasificación más limitada en comparación con los otros tres.

El agrupamiento para árboles (figura 6.4 a y b), origina dos grupos bien definidos, esto se puede observar en los métodos de enlace completo, promedio y de Ward. Con enlace completo un grupo está formado por los sitios 1, 3, 4, y 8 y el otro por 2, 5, 6, 7, 10 y 9. Con los métodos de enlace promedio y de Ward un grupo se forma por los sitios 1 y 3 y el otro por todos los demás sitios.

Al observar esto resulta que el método que permite tener una clasificación más detallada, es el método de enlace completo, pues dentro de esos dos grupos generados se aprecian subgrupos.

Se advierte una vez más un agrupamiento diferente para variables estandarizadas (figura 6.4 c).

Cuando se consideran sólo árboles, el método de enlace completo agrupa a los sitios 5 y 6 (figura 6.4 a y b) que comparten la mayoría de las especies arbóreas como se observa en el cuadro 6.5. Tienen pendientes de 16% y 21% (clase 3) respectivamente y tienen la misma exposición, aunque en el primero la pedregosidad es casi nula y en el segundo se tiene una pedregosidad de 25%. A esta pareja de sitios se unen después el 2, 7, 10 y 9. Todo esto al final forma un conjunto.

Por otra parte une los sitios 4 y 8, que sólo comparten dos especies, *Pinus leiophylla y Crataegus pubescens*, lo que no

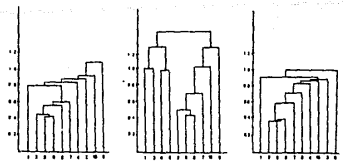


FIGURA 6.6.a.

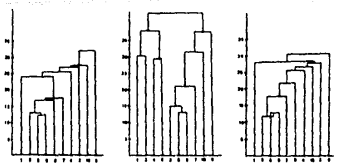


FIGURA 6.6.b.

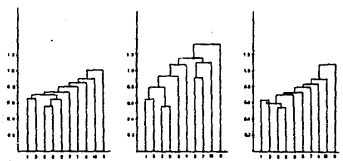


FIGURA 6.6.c.

FIGURA 6.6. Dendrogramas para grupos obtenidos por cinco métodos de agrupamiento diferentes. Observe que el método de unión completa es el que muestra una clasificación más detallada. (a) usando variables normalizadas; (b) usando variables sin normalizar y (c) usando variables estandarizadas.

los hace dignos de agrupación, ya que una de esas dos especies *Crataegus pubescens* es una especie que se distribuye de manera uniforme en toda la estación. Por lo tanto estos sitios son entidades distintas. Lo mismo ocurre cuando se comparan los sitios 1 y 3, que si bien comparten el mismo rango en cuanto a pendientes y pedregosidad; el primero corresponde a un bosque de *Pinus montezumae* y el sitio 3 a un bosque de *Pinus patula*.

En el dendrograma para árboles (figura 6.4 a y b, enlace completo), los sitios 5 y 6 son quienes tienen una menor distancia, es decir son tan parecidos entre sí, que se pueden considerar la misma unidad. Lo mismo pasaría con los sitios 2 y 7 que se unen a esta pareja (5 y 6). Todos estos se agrupan en las primeras etapas del método de análisis cluster, pues guardan cierta semejanza. Sin embargo, en la medida que el agrupamiento continúa los sitios restantes van siendo cada vez menos parecidos entre sí. Los sitios 4 y 8 por ejemplo, se agrupan en un cuarto ciclo o iteración del programa de cómputo, y el 1 y 3 hasta la quinta secuencia. Por agruparse después de varias etapas y además por comprobar que existen características que los hacen diferentes, cada uno de estos sitios corresponden a zonas individuales. Desde el momento que se reconocieron fusiones entre grupos muy disimiles, se consideró terminado el enlazamiento. Es decir, se trataron de reconocer clusters que tuvieran una asociación natural, este término ya fué discutido en el capítulo 2.

Lo mismo ocurre cuando se comparan los sitios 9 y 10 que si bien comparten algunas de sus especies arbóreas, sus coberturas son diferentes; en un sitio se tiene una cobertura muy alta para el encino (sitio 9) y menor para el pino, y en el sitio 10, pinos y encinos están más o menos en la misma proporción.

Cuando se consideraron los estratos arbustivo y herbáceo, la agrupación fué muy diferente con respecto a las anteriores.

En los dendrogramas para arbustos y hierbas (figura 6.5 a y b), los sitios 1 y 3 se agrupan en un primer lugar y como se observa en el cuadro 6.6 comparten todas sus especies

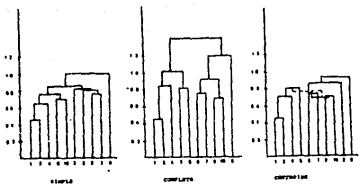


FIGURA 4.9.a.

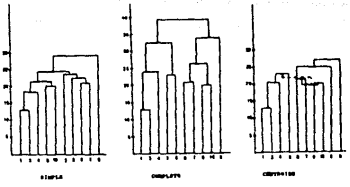


FIGURA 4.9.b.

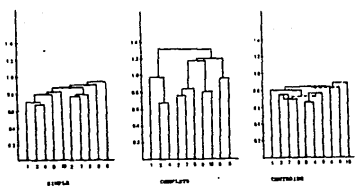


FIGURA 4.9.c.

FIGURA 4.9. Dendrogramas para bitmaps y arboles obtenidos por cinco metodos de agrupamiento diferentes. (a) usando variables normalizadas; (b) usando variables sin normalizar y (c) usando variables estandarizadas.

arbustivas y herbáceas excepto 3 de ellas, y además sus coberturas son muy parecidas.

Después se fusionan los sitios 9 y 10 que ya no son tan comunes en estos estratos; posteriormente los sitios 6 y 7 que tampoco son muy parecidos. Además estos 4 sitios mediante fotointerpretación y observaciones de campo tienen algunas características que los hacen diferentes uno del otro, y entonces se puede afirmar que se trata de zonas distintas.

Al observar este comportamiento en los datos, resulta que el análisis cluster que más concuerda con la fotointerpretación de vegetación es el que se hace con el estrato arbóreo pues los sitios se agrupan de acuerdo a las especies compartidas y sus coberturas. Además para este estudio, parece ser más adecuado clasificar con base en este estrato pues al ser de distribución espacial localizada, existen condiciones ecológicas que determinan esta distribución, lo que no ocurre con arbustos y herbáceas que crecen en cualquier sitio.

Es el estrato arbóreo entonces, quien determina las distintas clases de tierras. Sin embargo, si se realiza un análisis de relaciones entre los sitios de muestreo, los estratos bajos aportan mayor información acerca de las condiciones microambientales. Estas puedan ser, los nutrimentos que se encuentran en el suelo, la humedad del sitio, la presencia de pequeñas hondonadas y otras condiciones que serían de utilidad al estudiar por ejmplo, requerimientos de hábitat para algún organismo.

Para suelos y fisiografía se establecieron 4 clases tomando como límite sus valores extremos. De esta manera las variables se hicieron mayoras y fué más fácil establecer las clases de tierras.

El análisis de grupos para suelos originó un dendrograma (figura 6.6) donde se observa que los sitios con pendientes bajas de 0 a 3% (clase 1) y poca pedregosidad (clase 1), quedan unidos en un grupo y son los sitios 2,5,4,7 y 9; de todos sólo el sitio 2 tiene pendientes de 4 a 10% y pertenece a la clase 2.

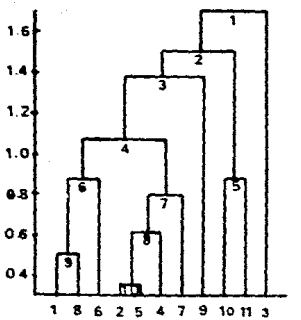


FIGURA 6.6. Dendrograma mediante la técnica de enlace completo para suelos y fisiografía.

Con respecto a las propiedades del suelo seleccionadas (pH, materia orgánica y C.I.C.T.), la semejanza entre estos sitios no es tan aparente como sucede con la fisiografía, aunque ciertos sitios comparten algunas de las propiedades químicas como puede observarse en el cuadro 6.4.

Por otra parte se unen los sitios 1,8 y 6 que tienen pH de 5.1 a 5.5 (clase 3). Los sitios 1 y 6 presentan pendientes de 11 a 40% (clase 3), y el sitio 1 con el 8 se encuentran en el mismo intervalo de C.I.C.T. (clase 3) y tienen pedregosidad baja (clase 1). Se tiene entonces que los sitios 1 y 8 comparten pH, C.I.C.T. y pedregosidad. El sitio 1 y 6 pendientes y los tres pH entre sí.

Finalmente se unen los sitios 10 y 11 que tienen la misma clase de pendiente (clase 3; 11-40%) y pedregosidad (clase 1; <10%). El sitio 3 se une hasta el final a todos los anteriores y es el único que tiene pedregosidad de 21 a 30% (clase 3). Este es el sitio menos parecido a los demás y por esta razón se une al final, ya que el método va agrupando a los sitios de acuerdo con su semejanza.

Se observa con esto que el análisis cluster dá mayor peso a los atributos fisiográficos para la unión de sitios en comparación con las propiedades químicas incluídas en el análisis, pues los sitios que se unen a una distancia menor (sitios muy similares) en lo que más se parecen entre sí es en sus atributos fisiográficos.

Mediante el resultado de este análisis se puede dividir a la estación en tres porciones. Una correspondería a la zona donde se ubican los sitios 2,5,4,7 y 9 que es la parte más plana de la estación (figura 6.8). Otra serían zonas separadas una de otra pero que son laderas de pendiente moderada, donde se localizan los sitios 1,8 y 6; y una tercera zona donde se tiene una mayor heterogeneidad de geoformas (sitios marcados como de verificación en el mapa de la figura 6.8).

Se estableció entonces como primer criterio de clasificación de tierras, el resultado del análisis para suelos-fisiografía. Un segundo criterio de clasificación fué el

análisis de grupos para vegetación que dió origen a un mayor número de regiones.

En la figura 6.7 se muestra un diagrama de perfil de vegetación del extremo sur de la estación a el extremo norte. Se observa que en el extremo sur predominan los pastizales; en la parte aledaña al cerro "El Molcajete" se tiene una mayor dominancia de encinos; en la siguiente unidad se presenta la mayor heterogeneidad de geoformas en la estación tanto en relieve como en vegetación. Y por último están las zonas donde predominan los pinos.

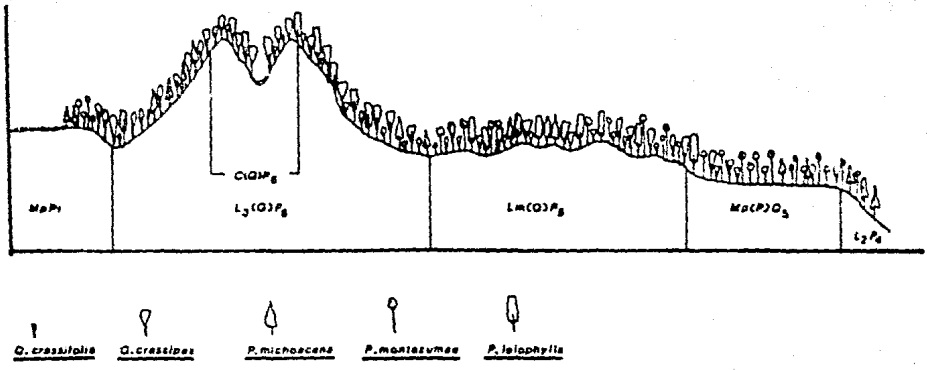


FIGURA 6.7 PERFIL DE VEGETACION PARA LA ESTACION.

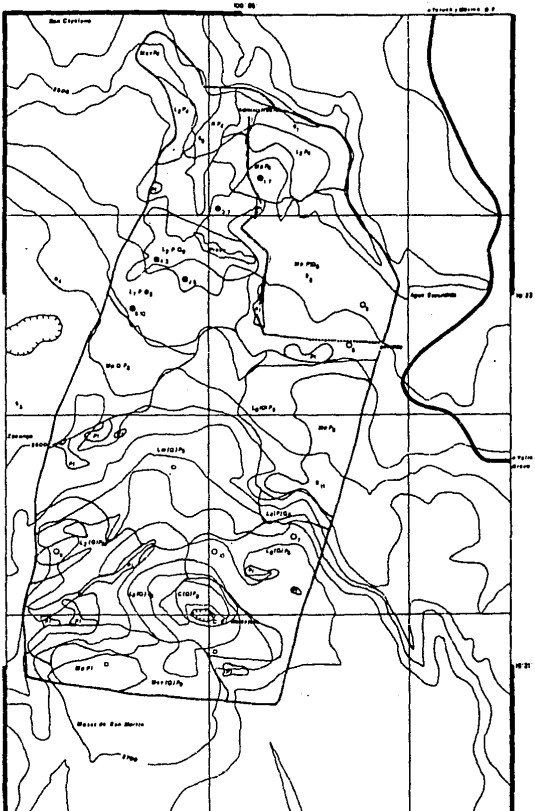


FIGURA 6.8 ESTACION DE APROVECHAMIENTO DE LA VIDA SILVESTRE "LUIS MACIAS ROLLADO", SEDUE

Mapa de Clases de Tierras Escala 1:10000 Equidistancia entre Curvas de Nivel 20m

Elaboró: Antonio Delgadillo Cruz*; Dibujo: Alejandro Estévez G

*SAHOP ZACATECAS Departamento de Biología, UAZ

CAPITULO 7

APLICACIONES DEL ANALISIS CLUSTER

En este apartado se hace un análisis de los alcances y limitaciones del análisis cluster en sus aplicaciones a la biología. Esta discusión se enfoca a las aplicaciones del cluster a la taxonomía y a la ecología de comunidades vegetales.

7.1 APLICACIONES DEL CLUSTER A LA TAXONOMIA.

La aplicación del análisis cluster a la taxonomía, recibe también los nombres de taxonomía numérica o fenética. Esta ha sido definida como la evaluación numérica de la afinidad o similitud entre unidades taxonómicas y su agrupamiento en taxa, con base en el estado de sus caracteres (Crisci, 1983).

El enfoque planteado por la taxonomía numérica comprende dos aspectos: uno filosófico, basado en la teoría clasificatoria denominada "feneticismo", y el otro, el de las técnicas numéricas, que son el camino operativo para aplicar dicha teoría.

El feneticismo sostiene los siguientes principios metodológicos:

a) Las clasificaciones deben hacerse con un gran número de caracteres, que deben ser tomados de todas las partes del cuerpo de los organismos y de todo su ciclo vital. (Sneath y Sokal, 1973).

b) Todos los caracteres utilizados tienen la misma significación e importancia (mismo peso) en la formación de los grupos.

c) La similitud total (o global) entre dos entidades es la suma de la similitud en cada uno de los caracteres utilizados en la clasificación.

d) Los grupos de taxones a formar se reconocen por una correlación de caracteres diferentes.

e) La clasificación es una ciencia empírica, en la cual la experiencia sensible desempeña el papel preponderante y, por

lo tanto, está libre de inferencias genealógicas.

f) Las clasificaciones deben basarse sólo en la similitud fenética. Se entiende por "fenético" cualquier tipo de carácter utilizable en la clasificación, incluyendo los morfológicos, fisiológicos, ecológicos, etológicos, moleculares, anatómicos, citológicos y otros.

g) El número de taxones establecido en cualquier rango es arbitrario, aunque siempre debe ser coherente con los resultados obtenidos.

Para el feneticismo es imposible llevar a cabo clasificaciones que expresen la filogenia o sean consecuentes con ella. La filogenia o historia evolutiva de los organismos, tiene varios componentes, entre ellos, la genealogía, la dirección de cambio o polaridad evolutiva, las relaciones temporales o cronísticas, las relaciones espaciales o biogeográficas y la cantidad de cambio o divergencia evolutiva (Wiley, 1981; Llorente, 1989). Para los feneticistas, que conciben a la sistemática como una ciencia empírica, el desconocimiento de detalles suficientes acerca de la historia evolutiva de la mayoría de los organismos, es una razón importante para que no se considere a la filogenia en sus clasificaciones. En muchos grupos de organismos, debido a la falta de fósiles o de otro tipo de información no se conoce la genealogía respectiva o ésta es en alto grado especulativa. En contraposición a la Sistemática Evolucionista y a la Sistemática Cladista, de acuerdo con los feneticistas; las clasificaciones deben basarse sobre hechos observables, por eso no es válido inferir filogenia a partir de caracteres. La polémica se centra en que para la fenética el objetivo que debe seguir la sistemática es la recuperación o descubrimiento de patrones naturales de relación entre especies (OTU's) a partir de las afinidades estimadas con el mayor número de caracteres, mientras que para evolucionistas y cladistas la búsqueda de patrones naturales implica la búsqueda también de los procesos que los generan. Si el patrón de distribución de estados de carácter entre grupos representa el resultado de una historia evolutiva común, la pregunta fundamental aquí es: ¿Es posible a

partir del patrón suponer una historia evolutiva? Para feneticistas y cladistas de patrón, la respuesta es no. Para evolucionistas y cladistas tradicionales, la búsqueda de patrones naturales, la estimación de su consistencia, su contrastación con otros patrones hipotéticos, es solo un recurso, no un fin. Ambas posiciones teóricas, parten de bases epistemológicas también distintas: el inductivismo de la fenética *versus* el deductivismo del cladismo y el evolucionismo.

Se establece también que todos los estados de un mismo carácter deben ser homólogos. El feneticismo propone la homología operativa: dos estados son homólogos cuando se corresponden en su composición y en su estructura. Por correspondencia en la composición se entiende la similitud cualitativa, desde el punto de vista biológico y/o químico de sus constituyentes. Por correspondencia estructural se entienda la similitud, en cuanto a orden de sus partes u orden espacio-temporal, en la estructura de sus fenómenos bioquímicos, o en el orden secuencial de las sustancias o estructuras organizadas.

En un estudio sobre taxonomía numérica y a diferencia de los estudios clásicos, deben utilizarse numerosos caracteres. ¿Pero cuál es la cifra de caracteres que satisface este requisito?. No existe una respuesta absoluta a este interrogante. En una época se estimó que este número no debía ser inferior a 60 (Sokal y Sneath, 1963), pero esta recomendación carece de bases teóricas o empíricas. El número de caracteres posibles es casi ilimitado y, por lo tanto, desde el punto de vista estadístico no puede sostenerse que el conjunto de todos los caracteres posibles forman una población al azar en la cual se puede indicar el tamaño de una muestra representativa. Sneath (1978) señaló que en la práctica los caracteres se comportan como variables "cuasialeatorias" y que, según su experiencia, 50 era el número mínimo a utilizar. Esto es todavía una afirmación intuitiva (Sneath, 1978; en Crisci, 1983).

Los feneticistas consideran que a mayor número de caracteres es mayor la estabilidad de sus clasificaciones. Sin embargo, se ha comprobado ampliamente que al aumentar el número de caracteres la clasificación se vuelve más inestable (Wiley, 1981). Además una clasificación de este tipo es muy difícil de interpretar.

¿Existe algún carácter que sea más importante que los otros?. En caso afirmativo, ¿cómo reconocerlo? La respuesta del feneticismo a estas preguntas es motivo de controversia. Para el feneticismo no existen *a priori* caracteres más importantes que otros y, si así fuera, es imposible reconocerlos.

Esta escuela taxonómica ha sido ampliamente criticada por otra corriente, la cladista. Una de las críticas está en relación al número de caracteres usados en una clasificación y a la ponderación de los mismos, mientras que los feneticistas opinan que debe ser la mayor cantidad y todos tienen la misma importancia, los cladistas sostienen que es más importante seleccionar caracteres homólogos, es decir aquellos que tienen significado para el reconocimiento de relaciones genealógicas o de ancestría-descendencia.

A pesar de que la controversia entre estas dos corrientes persiste, es indudable que la taxonomía numérica ha sido una gran motivación en general para los taxónomos; pues en la búsqueda de la "mejor" clasificación, se han desarrollado los principios de ambas escuelas, así como los algoritmos de cómputo que generan clasificaciones jerárquicas.

Robert Sokal, que es uno de los seguidores del feneticismo junto con otros autores como Rohlf y Mickevich, han desarrollado algoritmos computacionales de análisis cluster para aplicarlo a caracteres taxonómicos. Rohlf (1989), elaboró un programa llamado NT-SYS (Numerical Taxonomy System), que cuenta con una serie de métodos de análisis cluster además de otros métodos multivariados como los métodos de análisis de componentes principales y análisis de correspondencias entre otros.

7.2 LA IMPORTANCIA DE LA JERARQUIA EN LAS CLASIFICACIONES BIOLOGICAS.

Si se considera un grupo de entidades o muestras para ser clasificadas como un grupo de puntos en un espacio multidimensional, es claro que cualquier subdivisión de este espacio genera una clasificación.

Cuando se realiza una clasificación en muchos casos es importante que en ella se reflejen relaciones no solo entre una entidad y otra o entre un grupo de entidades, sino dentro de un sistema jerárquico, es decir, un sistema que represente relaciones de subordinación entre entidades o entre grupos; de tal manera que el grupo más grande contenga a los demás grupos.

En las clasificaciones biológicas muchas veces se requiere de una jerarquía. En ecología por ejemplo, al clasificar comunidades vegetales, interesa el reconocimiento de grupos y subgrupos donde la información contenida sea diferente. En taxonomía es fundamental clasificar a los organismos dentro de una jerarquía, donde un grupo o taxón representa a una familia, uno de menor rango es un género y dentro de éste existen varias especies.

Aunque las clasificaciones jerárquicas son las más usadas, también son útiles las clasificaciones no jerárquicas, la elección entre uno y otro tipo de clasificación dependerá de los objetivos particulares del estudio que se lleve a cabo.

Gauch (1982) opina que si el grupo de datos es grande, es recomendable el uso de una clasificación no jerárquica que también representa la variación entre las muestras sin que exista una subordinación entre grupos, lo cual además, no es estrictamente necesario en un estudio ecológico. Si el número de datos es pequeño, se prefiere usar la clasificación jerárquica, pues los resultados son más fáciles de interpretar que cuando se tiene un número grande de muestras.

Para Gauch (1982) una clasificación jerárquica es complementaria a una clasificación no jerárquica debido a que una clasificación no jerárquica es ideal con grandes grupos de datos y una jerárquica con pocos datos; la clasificación jerárquica tiene el propósito de revelar relaciones en los

datos y la no jerárquica no. Así una clasificación no jerárquica puede usarse inicialmente para resumir los datos y después con pocas muestras compuestas puede ser útil una clasificación jerárquica para reflejar las relaciones entre entidades o muestras.

7.3 CLASIFICACION Y ORDENACION.

La comunidad se concibe como un grupo de poblaciones biológicas que concurren en una área. Esta definición no es polémica en realidad, la discusión radica más bien en torno a la naturaleza de la comunidad. Esto es, para algunos como Clements, (citado por Whittaker, 1975), ese grupo de poblaciones mantienen interacciones entre sí; una población con todas las demás. Clements llevó su concepción de la Naturaleza de la Comunidad al extremo de considerarla análoga a un organismo individual. El propuso que, de la misma forma como sucede entre los organos y aparatos de un organismo, entre las poblaciones operan interacciones que en conjunto mantienen una condición de homeostásis. En el extremo opuesto Gleason (1926) y sus seguidores sostuvieron que las diferentes poblaciones de una comunidad, concurren en una área porque tienen requerimientos de hábitat semejantes; cada población tiene su propio espectro ecológico y, por lo tanto, el reconocimiento de comunidades es un proceso de abstracción (Gleason, 1926; en Müeller-Dombois y Ellemberg, 1974). De entonces a la fecha, la gran cantidad de investigación en ecología de comunidades se ha apoyado en alguna de las dos hipótesis extremas, en ocasiones de forma matizada: 1) El holismo de Clements, en el que todas las partes están conectadas y 2) El reduccionismo de Gleason, en el que todo es solamente la suma de las partes. Ahora bien, ¿Cómo se ha usado el análisis cluster en torno a esta polémica?

Para quienes el reconocimiento de las comunidades es un proceso de abstracción, es decir, un resultado y no una suposición *a priori*, la tipología o estructura de los dendrogramas puede significar una herramienta heurística para distinguir entre dos condiciones opuestas: 1) donde se reconocen grupos de sitios que pueden considerarse como

comunidades y 2) donde se observa un patrón de escalamiento que puede sugerir un comportamiento de cambios ecológicos estructurales graduales.

Figura 7.1. (1). Holismo de Clements vs. (2) Reduccionismo de Gleason, como criterios para definir comunidades biologicas, idealizados a traves de dos tipologias de dendrogramas diferentes.

Según la teoría que se considere de interés para un estudio ecológico, será el tipo de análisis que requiera. Si se estudia la variación de la vegetación a lo largo de un gradiente, se recomienda aplicar un método de ordenación, en donde se trata de ordenar especies y muestras en un espacio de pocas dimensiones, generalmente dos o tres; con el fin de que se pueda obtener una representación gráfica. El método de ordenación más común es el análisis de componentes principales (Gauch, 1982; Criaci, 1983).

Las estructuras de datos que se agrupan naturalmente agrupados son raros en ecología de comunidades; generalmente la variación de comunidades es continua y la clasificación la mayoría de las veces es impuesta. (Goodall, 1953; en Gauch, 1982). Sin embargo, si el objetivo es reconocer límites para la identificación de unidades homogéneas, la clasificación sería el método de análisis más apropiado.

7.4 IMPORTANCIA DEL ANALISIS CLUSTER EN LAS CLASIFICACIONES BIOLOGICAS.

El análisis cluster es una técnica cuyo objetivo es la agrupación de entidades, para originar una clasificación que muestre las relaciones entre éstas entidades. Se pueden obtener por medio de éste análisis tanto clasificaciones jerárquicas como no jerárquicas. Los algoritmos que dan relaciones jerárquicas se han desarrollado quizás más que los que

proporcionan clasificaciones no jerárquicas, pues las primeras tuvieron una mayor aceptación y uso en diversos campos del conocimiento como en Psicología, Sociología, Biología, y en todas aquellas áreas en donde se generaran grandes cantidades de datos que requirieran ser tratados por medio de computadoras, y donde el conocimiento de la jerarquía entre grupos fuera la mayor necesidad.

Este avance en los algoritmos que producían clasificaciones jerárquicas fué ampliamente estudiado por Sokal, quien junto con Sneath, principalmente; incorporó estos algoritmos a la taxonomía numérica o fenética, generándose así una corriente de pensamiento ampliamente apoyada en éstos. El rápido desarrollo de la taxonomía numérica ha contribuído paralelamente al desarrollo de métodos en ecología vegetal (Whittaker, 1980).

El aumento en el uso del análisis cluster se demuestra también al observar el gran número de trabajos donde se hace uso de esta técnica. Este auge se dió principalmente entre los años 60's y 70's, que fué también cuando el uso de las computadoras aceleró el desarrollo de algoritmos que generaran clasificaciones. Las primeras opciones de software requerían de la capacidad de memoria de una main-frame. En la actualidad existen paquetes estadísticos como SAS y NT-SYS diseñados para Microcomputadoras PC, que requieren solo 640 K y tienen capacidad de procesar matrices de 400 datos.

7.5 LA "LIBERTAD" DEL ANALISIS CLUSTER.

Los métodos de estadística tradicional aunque son muy importantes, no son aplicables a todo tipo de datos. Además requieren del cumplimiento de varios supuestos estadísticos. El análisis cluster tiene la gran ventaja de no requerir de ninguna condición para su uso, como por ejemplo los supuestos de aleatoriedad e independencia de la muestra, es decir, el caso de que cualquier individuo tenga la misma probabilidad de ser elegido para formar parte de la muestra y que su elección no afecte la de otro (Anderberg, 1973). Sin embargo, al igual que otras técnicas de Análisis Multivariado, el análisis

cluster requiere del supuesto de ortogonalidad de las variables. Cuando se utiliza una matriz de distancias, se debe considerar que estas se determinan bajo el supuesto de que las variables no están correlacionadas entre sí ($r = 0$) (figura 7.2). Si dos o más variables están altamente correlacionadas, ($r \rightarrow -1$ o $r \rightarrow 1$), entonces las distancias son falsas. En este caso, hay dos opciones de resolver este problema. Uno consiste en reemplazar el uso de distancias por medidas de asociación. El otro, consiste en reducir la dimensionalidad del problema, esto es, detectar los casos de colinearidad entre variables y reducirlas a una sola o, a través de otras técnicas como análisis de componentes principales convertir la totalidad de las variables a un grupo más pequeño de variables (Componentes Principales), que resultan de combinaciones de las variables originales, que sean ortogonales entre sí.

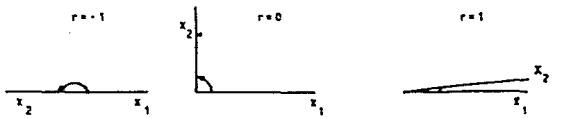


Figura 7.2. Dos variables X_1 y X_2 son ortogonales cuando r=o (b). Cuando hay alta correlación $r \rightarrow -1$ (a) o $r \rightarrow 1$ (c) el supuesto de ortogonalidad se viola.

Para estudios ecológicos, como es el caso de la clasificación de tierras, es hasta cierto punto ilógico tratar de que la muestra sea totalmente elegida al azar debido a que existe a *priori* un juicio para regionalizar previamente la zona y considerar unidades diferentes.

Esta "libertad" que tiene el análisis cluster con respecto a otras técnicas estadísticas, implica que el investigador tome una serie de decisiones antes, durante y al final del análisis, de manera que realmente revele características importantes de los datos (Everitt, 1981).

7.6 SIGNIFICADO GEOMETRICO DE CADA TECNICA DEL ANALISIS CLUSTER.

Es importante dejar en claro que cada técnica de agrupamiento tiene significado geométrico distinto y que no todas son aplicables al mismo tipo de datos y/o problema. A continuación se mencionan las características principales de cada una de ellas con el fin de que se consideren antes de seleccionar aquella que será aplicada a un conjunto de datos determinado.

El método de enlace simple tiene una característica que se conoce con el nombre de encadenamiento, es decir, une dos entidades y a medida que avanza el agrupamiento une una a una cada entidad al grupo formado al principio de dos entidades (figura 7.3a). Esta característica se considera una desventaja del método ya que al buscar aquella pareja de individuos cuya distancia es la mínima, fuerza mucho el agrupamiento y contrae las distancias a las que se unen las entidades, por esta razón el método de enlace simple tiende a producir dendrogramas cortos. El enlazamiento simple se recomienda para datos que originen grupos claramente definidos, pues se puede observar bien esta separación.

El método de enlace completo, por el contrario, alarga mucho las distancias a las que se dá el agrupamiento y no enlaza uno a uno a cada individuo, sino que considera la distancia máxima entre ellos y entonces forma grupos grandes que involucran a la mayor cantidad posible de entidades cuya distancia resulta ser menor que la distancia máxima localizada por el algoritmo (figura 7.3b). El método de enlace completo genera dendrogramas alargados. Se recomienda cuando se quieran observar diferencias dentro de los grupos, pues forma grupos grandes.

Algunos autores, por tratar de evitar llegar a situaciones extremas como las producidas por los métodos de enlace simple y completo; proponen trabajar con las distancias promedio, es decir aquellas distancias que de alguna manera caractericen a cada grupo de entidades (figura 7.4). El dendrograma producido por el método de enlace promedio

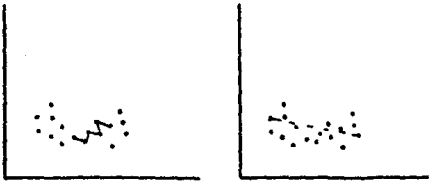


FIGURA 7.3. Representación gráfica de los métodos de (a) enlazamiento simple y (b) enlazamiento completo . Observe que el enlace simple produce "encadenamiento" entre las entidades comparadas y el método de enlace completo forma grupos grandes de entidades.

FIGURA 7.4. (a). Diagrama que muestra la fusión de un individuo con un cluster usando el método del grupo promedio. La distancia del individuo al cluster es calculada como $D_{iD-(ABC)}=(D_{iDA}+D_{iDB}+D_{iDC})/3$. (b). Fusión de dos clusters usando el método del grupo promedio. La distancia calculada incluye a todos los posibles pares de miembros de los dos grupos: $D_{i(FD)(ABC)}=(D_{iAD} + D_{iBD} + D_{iCD} + D_{iAF} + D_{iBF} + D_{iCF})/3$.

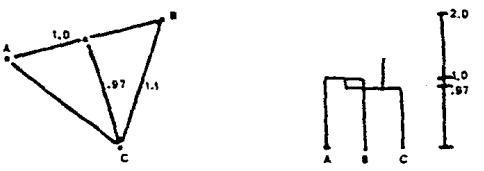


FIGURA 7.5. Diagramas que muestran como se originan los reversos en el método del centroide . (a). Distancias calculadas por el método del centroide. (b). Dendrograma. Observe que A y B se unen a una distancia d= 1.0 y (AB) se une a C a una distancia menor (d= 0.97), lo que produce un reverso en la jerarquía.

representa el comportamiento promedio para las entidades estudiadas.

En el método de enlace promedio se puede también tener el problema de que la media es una medida que se ve afectada por la presencia de datos extremos y los puntos promedio considerados puedan no ser los mejores; el método del centroide trata de evitar este problema.

El método del centroide busca los centros de gravedad de los grupos que se van formando y aunque el centroide es el promedio de las entidades comparadas, gráficamente no representa lo mismo que el método del promedio, pues el método del centroide genera un punto hipotético dentro de una esfera.

El método del centroide generalmente no produce clasificaciones jerárquicas, pues las distancias a las que se dá la formación de grupos no siempre va en aumento. Tiene como característica, la producción de reversos; muchos autores consideran que los reversos que produce el método del centroide son una desventaja, pues producen dendrogramas no jerárquicos y esto muchas veces dificulta la interpretación de los datos (figura 7.5).

Tanto el método del promedio como el del centroide, son valiosos cuando se quiere observar el comportamiento promedio de los datos, aunque cada uno de ellos, tiene un significado geométrico diferente.

El método de la mínima varianza es importante cuando se quiere conocer la relación que existe entre las varianzas de dos o más entidades. Tiene la característica de producir dendrogramas cortos, en los que es muy pequeña la escala que representa las distancias a las que se dá la fusión entre grupos.

7.7 RECOMENDACIONES GENERALES PARA EL USO DEL ANALISIS CLUSTER.

Es muy importante tener una idea clara del modo de operación de las técnicas del análisis cluster, pues cada una puede producir dendrogramas diferentes aún con los mismos datos. Del mismo modo la medida de distancia elegida aún con el mismo grupo de datos y el mismo método de agrupamiento, puede

proporcionar resultados diferentes.

El usuario de este tipo de técnicas en ocasiones se interesa en alimentar a la computadora con los datos de sus estudios y espera obtener resultados satisfactorios. Aunque esto puede hacerse, no es recomendable, pues el análisis cluster es una herramienta estadística que necesita, para que sea más poderosa, de una gran interrelación entre el usuario y sus datos y en la medida que se conozca mejor el fenómeno estudiado, el investigador estará en la posibilidad de darle un uso digno al análisis cluster.

Aunque existen actualmente muchos paquetes de cómputo que realizan este análisis rápidamente y que manejan por lo general una sola medida de distancia; si nosotros como usuarios requerimos del uso de otro tipo de medida, es importante conocerlo para transformar los datos de manera que aunque calcule la euclidiana, una transformación puede resultar en otro tipo de medida, que sea en realidad la que nos interesa.

Si es usted un usuario potencial del análisis cluster debe considerar lo siguiente:

Si los datos seleccionados para el análisis no son importantes para el estudio, cualquier tipo de herramienta estadística, ya sea univariada o multivariada revelará resultados irrelevantes para el estudio.

Cada una de las medidas de distancia tiene significado estadístico diferente, así que es recomendable investigar sobre ellas para elegir la más adecuada para sus datos.

Se debe considerar que cada método de cluster (enlace simple, enlace completo, etc.) tiene significado geométrico diferente y por lo tanto, los resultados pueden variar mucho de un método a otro. Si no se conocen a fondo los datos, o aún conociéndolos bien no se sabe que método elegir; la mejor solución es indudablemente probar varios y una vez que se tengan los resultados, realizar la selección del método que aparentemente refleje mejor las relaciones entre los datos.

Si las variables están medidas en diferentes escalas, es recomendable realizar una estandarización de las variables y

probar a su vez con los datos originales; si se observa que el agrupamiento fué exactamente el mismo para uno y otro caso; entonces no es necesaria la estandarización.

Generalmente el análisis cluster se usa junto con alguna otra técnica multivariada. Es común encontrar trabajos donde se usa primero el método de componentes principales, cuyo objetivo es reducir el número de dimensiones a dos o tres, las cuales se pueden representar gráficamente y se obtienen aquellos factores que explican el mayor porcentaje de variabilidad de los datos; una vez que se obtiene una representación gráfica de los puntos dentro de los componentes se agrupan mediante análisis cluster y se presentan los resultados.

El análisis cluster tiene como desventaja que no es una técnica estadística inferencial, sino totalmente descriptiva, sin embargo, su aplicación puede servir para generar hipótesis acerca de los datos y posteriormente se puede aplicar una técnica inferencial.

CAPITULO 8

CONCLUSIONES

En las ciencias biológicas es muy común la tarea de clasificación, se clasifican organismos, comunidades, o variables. El análisis cluster es una técnica multivariada que genera clasificaciones, sus ventajas son:

1) Es una técnica que no está limitada por supuestos estadísticos, como los de aleatoriedad, independencia, homocedasticidad o linearidad de los datos; y por lo tanto la única condición es que el estudio sea multivariado, es decir, que el conjunto de variables sean interdependientes.

2) Es una técnica descriptiva, que no llega a conclusiones absolutas, esto permite poder usarla en combinación con otras técnicas inferenciales o con otras técnicas multivariadas. Puede considerarse como una técnica de tipo exploratorio, más que inferencial.

3) El análisis cluster, por ser descriptivo puede revelar características importantes de los datos y se pueden generar hipótesis que requieran nuevos estudios, su valor es eminentemente heurístico.

4) El hecho de que requiere que el investigador esté constantemente tomando decisiones acerca de sus datos (medida de distancia elegida, y método de agrupamiento), obliga a la búsqueda de la mayor información posible acerca del tipo de estudio y del tratamiento estadístico de los datos. Quizás ésta sea una de las mayores ventajas, pues el buen o mal uso de las técnicas del análisis cluster dependen del grado de conocimientos que tenga el usuario de su problema. En el caso de las aplicaciones a los diferentes campos de la biología (taxonomía, ecología, biogeografía, entre otros), es necesario que se conozcan con profundidad las hipótesis que están en disputa, así como reconocer el grado con que estas hipótesis pueden ser rechazadas o aceptadas por los resultados del análisis cluster.

Es importante destacar que el análisis cluster no debe usarse como una técnica aislada, quizás para algunos casos sea válido usarlo así. Sin embargo, dentro del contexto de la estadística multivariada, el análisis cluster se usa la mayoría de las veces de manera combinada con el análisis de componentes principales.

A N E X O

CUADRO 6.1. VALORES DE COBERTURAS PARA LAS ESPECIES EN CADA SITIO DEMUESTREO. LAS LITERALES INDICAN EL TIPO DE ESTRATO. A: EMBUEGUES (> 30 mts.); B: ARBOREO (15-30 mts.); C: ARBOREO (< 15 mts.); D: ARBUSTIVO (lenosos de 1 a 3 mts.); E: HERBACEO (no lenosos de 25 cm. a 1.5 mts.) y F: RASANTE (hasta 25 cm.).

ESTRATO / FAMILIA / Especie	1	2	3	4	5	6	7	8	9	10
A										
PINACEAE										
Pinus leiophylla	-	-	-	12	-	-	-	-	-	-
Pinus montezumae	-	-	-	-	-	-	-	-	-	18
B										
BETULACEAE										
Alnus jorullensis	-	-	-	-	-	-	-	-	-	6
ERICACEAE										
Arbutus xalapensis	-	-	-	-	-	-	-	2	3	-
FAGACEAE										
Quercus crassifolia	-	-	-	-	-	-	-	6	24	-
Quercus crassipes	-	3	-	-	18	9	-	16	6	18
Quercus laurina	-	-	-	-	-	-	-	-	15	12
PINACEAE										
Pinus leiophylla	-	10	-	12	7	15	6	12	8	6
Pinus michoacana	-	6	-	-	-	-	-	-	-	-
Pinus montezumae	-	9	-	-	17	10	22	-	5	15
Pinus patula	-	-	22	-	-	-	-	-	-	-
ROSACEAE										
Prunus serotina	-	-	-	-	2	-	-	-	-	-
C										
BETULACEAE										
Alnus jorullensis	-	-	-	-	-	-	9	-	-	6
CLETHRACEAE										
Clethra mexicana	-	-	-	-	-	-	2	-	-	-
ERICACEAE										
Arbutus xalapensis	-	-	-	-	-	-	-	-	3	-
FAGACEAE										
Quercus crassifolia	-	2	2	6	2	3	6	-	12	6
Quercus crassipes	-	6	4	18	3	2	10	-	5	6
Quercus laurina	-	-	-	-	-	-	-	-	9	-
PINACEAE										
Pinus leiophylla	-	2	-	-	8	-	3	-	-	-
Pinus montezumae	24	8	5	-	3	-	3	3	5	-
ROSACEAE										
Crataegus pubescens	-	-	-	12	-	-	-	6	-	6
Prunus serotina	-	-	-	-	-	-	-	6	-	-
D										
BERBERIDACEAE										
Berberis moranensis	5	2	9	9	-	2	-	4	3	-
BETULACEAE										
Alnus jorullensis	-	-	-	-	-	3	9	-	-	-
CLETHRACEAE										
Clethra mexicana	-	-	-	-	-	-	2	-	-	-

ESTRATO	B		I	T		O	B		
FAMILIA									
Especie	1	2	3	4	5	6	7	8	9	10
COMPOSITAE										
Baccharis sp.	19	22	12	9	19	9	8	19	–	6
Brickellia pendula	5	2	–	6	–	–	–	–	–	–
Eupatorium areolare	–	–	6	6	–	–	3	–	3	–
Eupatorium glabratum	–	–	–	–	–	3	10	–	–	–
Eupatorium resinose	15	22	15	6	6	6	12	17	5	–
Senecio angulifolius	6	–	2	–	–	–	2	–	6	12
Senecio argutus	6	–	3	6	6	–	2	2	–	–
Senecio roldana	6	–	5	6	–	2	–	–	–	–
Senecio salignus	2	–	–	–	–	–	–	–	–	–
Stevia rhombifolia	5	2	7	6	–	–	9	–	–	9
Stevia sp.	–	–	–	–	–	–	–	5	3	–
Stevia tomentosa	–	–	–	–	–	–	2	–	–	–
CORNACEAE										
Cornus excelsa	2	2	6	6	2	–	–	8	–	–
ERICACEAE										
Arbutus xalapensis	1	6	2	–	2	6	6	6	–	–
FAGACEAE										
Quercus crassifolia	–	–	–	–	–	2	8	–	–	6
Quercus crassipes	–	12	–	–	3	3	6	–	–	6
Quercus rugosa	–	–	–	–	–	–	–	–	–	6
GARRYACEAE										
Garrya laurifolia	–	–	7	6	–	–	–	6	–	–
ONAGRACEAE										
Fuchsia minimifolia	5	2	9	6	6	2	–	12	8	12
Fuchsia thymifolia	–	9	–	9	6	6	6	7	–	6
PINACEAE										
Pinus sp. (plantulas)	–	–	5	–	–	–	7	–	–	21
POLYGONACEAE										
Nonnina xalapensis	–	6	1	6	–	2	16	–	–	9
ROSACEAE										
Acaena elongata	–	–	–	–	–	–	–	2	–	–
Crataegus pubescens	6	–	6	–	2	6	–	6	–	–
Prunus serotina	6	6	6	6	2	8	2	8	3	6
SOLANACEAE										
Cestrum terminale	–	–	–	–	2	–	2	–	–	–
VERBENACEAE										
Lantana sp.	6	6	–	–	–	–	–	–	–	–
B										
ACANTHACEAE										
Dicliptera peduncularis	–	–	–	–	–	–	–	6	–	–
COMPOSITAE										
Bidens triplinervia	–	–	2	–	–	–	–	–	3	–
Cirsium sp.	1	2	–	3	2	–	2	–	–	–
Erigeron maximus	–	–	–	–	–	–	–	–	3	–
Gnaphalium sp.1	2	6	2	6	–	–	2	–	–	–
Gnaphalium sp.2	2	2	7	–	–	–	–	–	–	–
Senecio angulifolius	1	–	–	–	–	–	–	–	–	–
Senecio tolucanus	–	–	–	9	–	–	–	–	–	–

ESTRATO										
FAMILIA										
Especie	1	2	3	4	5	6	7	8	9	10
ERICACEAE										
Arbutus xalapensis	-	-	-	-	-	-	-	-	-	6
GERANIACEAE										
Geranium potentillaefolium	-	2	-	-	-	-	2	-	-	6
GRAMINEAE										
Avena fatua	24	10	17	15	11	14	-	2	9	-
Muhlenbergia macroura	9	17	9	-	21	8	6	3	-	-
Panicum sp.	-	-	4	8	-	-	-	-	-	-
Setaria macrostachya	2	-	10	12	-	-	-	-	-	-
IRIDACEAE										
Orthorosanthus chimborocensis	-	2	-	-	-	4	2	21	6	6
LABIATAE										
Salvia elegans	-	-	-	-	-	-	-	2	-	9
LEGUMINOSAE										
Lupinus elegans	3	2	-	-	-	-	6	-	-	6
LILIACEAE										
Smilax moranensis	-	-	-	-	-	-	-	8	-	-
ONAGRACEAE										
Cuphea aequipetala	2	-	-	-	-	-	-	-	-	-
ORCHIDACEAE										
Spiranthes sp.	-	-	-	-	-	-	2	-	-	-
ROSACEAE										
Prunus serotina	-	-	-	-	-	-	-	3	-	-
SCROPHULARIACEAE										
Castilleja tenuiflora	-	-	-	-	2	-	2	-	-	-
Penstemon campanulatus	-	3	-	-	-	-	4	2	3	-
UMBELLIFERAE										
Arracacia atropurpurea	-	-	-	-	-	-	-	-	-	9
Eringium cymosum	-	-	-	-	-	-	2	-	-	-
VALERIANACEAE										
Valeriana subincisa	2	-	-	-	-	-	-	-	-	-
?										
LABIATAE										
Stachys sp.	8	5	6	6	-	-	2	-	-	-
LEGUMINOSAE										
Desmodium venustrum	2	2	2	6	-	-	-	-	-	-
RANUNCULACEAE										
Ranunculus sp.	-	-	2	6	-	-	-	-	-	-
ROSACEAE										
Acaena elongata	-	4	2	-	-	2	2	8	-	-
Alchemilla procumbens	-	-	-	-	-	2	3	-	-	-
Fragaria sp.	-	-	4	6	-	-	2	-	-	-
RUBIACEAE										
Galium praetermissum · helechos	4	2	-	9	-	-	-	-	6	-
Adiantum sp.	-	2	-	-	-	-	2	-	5	-
Cheilantes sp.	-	-	-	-	-	-	-	-	3	-
Polipodium sp.	-	-	-	-	-	-	-	-	3	-
Pteridium sp.	6	2	-	6	-	-	-	-	6	9

CUADRO 6.2. PROPIEDADES DE LOS SUELOS PARA CADA MUESTRA Y CADA HORIZONTE.D.AP.= DENSIDAD APARENTE; D.R.= DENSIDAD REAL; pH ACT.= pH ACTIVO; pH POT.= pH POTENCIAL; %M.O.= % DE MATERIA ORGANICA Y CICT= CAPACIDAD DE INTERCAMBIO CATIONICO.

MUESTRA	D.AP.	D.R.	TEXTURA AR AC L	pH ACT.	pH POT.	%M.O.	CICT
1 0-10	0.64	1.14	40.7 20.0 35.3 FRANCO	5.4	4.7	20.22	25.05
1 10-27	0.65	1.09	40.7 24.0 35.3 FRANCO	5.5	5.0	13.59	26.98
1 > 27	0.65	1.09	50.7 16.0 33.3	5.2	5.2	6.28	26.80
2 0-30	0.65	1.08	49.4 10.4 40.2 FRANCO	5.4	5.1	9.76	19.62
2 30-60	0.73	1.06	floculo	5.6	5.7	3.58	31.28
2 > 60	0.81	1.11	58.0 16.9 25.1 MIGAJON ARENOSO	5.7	5.1	2.76	28.28
3 0-20	0.74	1.14	56.4 14.4 29.3 MIGAJON ARENOSO	4.9	4.5	28.10	30.83
3 20-40	0.77	1.11	52.0 14.2 33.8 FRANCO	5.1	4.8	9.62	28.38
3 > 40	0.74	1.08	54.4 14.2 31.5 MIGAJON ARENOSO	5.9	5.4	2.16	30.83
4 0-30	0.67	1.73	49.4 11.8 38.7 FRANCO	4.7	4.4	11.61	18.57
4 30-60	0.56	1.97	46.7 15.0 38.2 FRANCO	5.6	5.3	5.25	23.39
4 > 60	0.57	1.83	floculo	6.0	5.4	5.25	32.19
5 0-25	0.70	2.00	42.2 16.0 41.8 FRANCO	5.2	4.7	5.88	23.59
5 > 25	0.59	1.98	floculo	6.1	5.5	3.34	26.04
6 0-19	0.75	2.38	37.2 23.6 39.2 FRANCO	5.3	4.3	7.14	21.22
6 19-36	0.98	2.11	35.4 28.0 36.6 MIGAJON ARCILL.	5.5	4.3	2.52	19.00
7 0-30	0.61	1.80	51.3 12.9 35.8 FRANCO	5.7	5.0	11.50	27.82
7 30-60	0.68	1.63	floculo	5.2	5.3	2.16	30.65
7 > 60	0.64	1.48	floculo	5.7	5.2	5.00	33.54
8 0-13	0.67	2.17	44.7 16.7 38.6 FRANCO	5.5	4.5	13.94	25.86
8 13-37	0.62	1.83	48.7 10.7 40.6 FRANCO	6.0	5.1	9.62	21.62
9 0-15	0.56	1.34	48.6 14.8 36.6 FRANCO	5.6	4.8	22.66	21.78
9 15-25	0.66	1.77	48.6 18.9 36.6 FRANCO	5.3	4.5	16.02	18.79
9 25-40	0.69	1.93	46.6 22.9 30.6 FRANCO	6.0	4.7	7.49	17.78
9 > 40	0.80	2.17	30.0 42.2 27.8 ARCILLOSO	6.0	4.9	6.78	19.27
10 0-30	0.52	2.00	36.0 17.4 46.6 FRANCO	5.6	4.8	12.47	39.65
10 30-60	0.55	1.92	53.3 8.5 38.2 MIGAJON ARENOSO	5.9	5.0	7.32	29.32
10 > 60	0.57	1.92	floculo	6.3	5.5	1.09	46.87
11 0-16	0.51	1.61	46.9 12.9 40.2 FRANCO	6.8	6.0	9.62	27.88
11 16-40	0.54	2.00	48.0 13.4 38.6 FRANCO	6.7	5.8	5.36	50.79
11 > 40	0.57	2.38	floculo	6.3	5.6	3.96	44.41

pH ACTIVO (H O, 1:2.5). pH POTENCIAL (KCl, 1:2.5).

CUADRO 6.2. PROPIEDADES DE LOS SUELOS PARA CADA MUESTRA Y CADA HORIZONTE. (continuación).

MUESTRA	COLOR (10 YR)	MUESTRA	COLOR (10 YR)
1 0-10	3/3 S 2/1 H	7 0-30	5/3 S 3/1 H
1 10-27	3/3 S 3/1 H	7 30-60	5/3 S 3/2 H
1 > 27	5/4 S 3/4 H	7 > 60	5/3 S 3/2 H
2 0-30	5/3 S 3/3 H	8 0-13	5/2 S 3/1 H
2 30-60	5/4 S 3/3 H	8 13-37	5/4 S 3/3 H
2 > 60	5/4 S 3/4 H	8 > 37	5/4 S 4/4 H
3 0-20	5/3 S 3/2 H	9 0-15	5/3 S 3/2 H
3 20-40	5/3 S 3/2 H	9 15-25	5/2 S 2/1 H
3 > 40	5/4 S 3/4 H	9 25-40	5/2 S 2/1 H
4 0-30	5/3 S 3/3 H	9 > 60	5/1 S 2/1 H
4 30-60	5/4 S 3/6 H	10 0-30	4/3 S 3/1 H
4 > 60	5/4 S 4/4 H	10 30-60	4/3 S 3/2 H
5 0-25	5/3 S 3/2 H	10 > 60	5/4 S 4/4 H
5 > 25	5/4 S 4/4 H	11 0-16	4/2 S 3/1 H
6 0-19	5/3 S 3/2 H	11 16-40	4/3 S 3/3 H
6 > 19	6/4 S 3/4 H	11 > 60	5/4 S 4/3 H

CLASES	pH	% M.O.	CICT	% PEND.	% PEDREG.
1	> 6.0	22.55-28.10	34.39-39-65	0-3	< 10
2	5.6-6.0	17.00-22.54	29.12-34.38	4-10	11-20
3	5.1-5.5	11.44-16.99	23.85-29.11	11-40	21-30
4	< 5.0	5.88-11.43	18.57-23.84	> 40	31-40

CUADRO 6.3 Clases establecidas para algunas propiedades del suelo y variables fisiográficas.

MUESTRAS	pH	M.O.	CICT	PERD.	PEDREG.
1	3	2	3	3	1
2	3	4	4	2	1
3	4	1	2	3	3
4	4	3	4	1	1
5	3	4	4	1	1
6	3	4	4	3	2
7	2	3	3	1	1
8	3	3	3	2	1
9	2	1	4	1	1
10	2	3	1	3	1
11	1	4	3	3	1

CUADRO 6.4 Datos de suelo-fisiografía que se consideraron en la realización del análisis de grupos. Estos datos no corresponden con los valores reales, ya que los valores numéricos fueron asignados de acuerdo a las clases establecidas en el cuadro 6.3.

CUADRO 6.5. COBERTURAS DE LAS ESPECIES ARBOREAS CONSIDERADAS EN EL ANALISIS DE GRUPOS.

ESTRATO			S	I	T	I	O	S		
FAMILIA *Especie*	1	3	4	8	2	5	6	7	10	9
A										
PINACEAE										
Pinus leiophylla	-	-	12	-	-	-	-	-	-	-
Pinus montezumae	-	-	-	-	-	-	-	-	18	-
B										
BETULACEAE										
Alnus jorullensis	-	-	-	-	-	-	-	-	6	-
ERICACEAE										
Arbutus xalapensis	-	-	-	2	-	-	-	-	-	5
FAGACEAE										
Quercus crassifolia	-	-	-	6	-	-	-	-	-	24
Quercus crassipes	-	-	-	16	3	10	9	-	18	6
Quercus laurina	-	-	-	-	-	-	-	-	12	15
PINACEAE										
Pinus leiophylla	-	-	12	12	10	7	13	6	6	8
Pinus michoacana	-	-	-	-	6	-	-	-	-	-
Pinus montezumae	-	-	-	-	9	17	10	22	15	5
Pinus patula	-	22	-	-	-	-	-	-	-	-
ROSACEAE										
Prunus serotina	-	-	-	-	-	2	-	-	-	-
C										
BETULACEAE										
Alnus jorullensis	-	-	-	-	-	-	-	9	6	-
CLETHRACEAE										
Clethra mexicana	-	-	-	-	-	-	-	2	-	-
ERICACEAE										
Arbutus xalapensis	-	-	-	-	-	-	-	-	-	3
FAGACEAE										
Quercus crassifolia	-	2	6	-	2	2	3	6	6	12
Quercus crassipes	-	4	18	-	6	3	2	10	6	5
Quercus laurina	-	-	-	-	-	-	-	-	-	9
PINACEAE										
Pinus leiophylla	-	-	-	-	2	8	-	3	-	-
Pinus montezumae	26	5	-	3	8	3	-	3	-	5
ROSACEAE										
Crataegus pubescens	-	-	12	6	-	-	-	-	6	-
Prunus serotina	-	-	-	6	-	-	-	-	-	-

CUADRO 6.6. COBERTURAS DE LAS ESPECIES ARBUSTIVAS Y HERBACEAS CONSIDERADAS EN EL ANALISIS DE GRUPOS.

ESTRATO			S	I	T	I	O	S	
FAMILIA *Especie*	1	3	5	4	2	8	6	7	9	10
D										
BERBERIDACEAE										
Berberis moranensis	5	9	-	9	2	4	2	-	3	-
COMPOSITAE										
Baccharis sp.	19	12	19	9	22	19	9	8	-	6
Eupatorium areolare	-	4	-	6	-	-	-	3	3	-
Eupatorium resinosa	15	15	4	6	22	17	6	12	5	-
Senecio angulifolius	4	2	-	-	-	-	-	2	6	12
Senecio argutus	6	3	6	6	-	2	-	2	-	-
Senecio salignus	2	-	-	-	-	-	-	-	-	-
CORNACEAE										
Cornus excelsa	2	4	2	6	2	8	-	-	-	-
ERICACEAE										
Arbutus xalapensis	1	2	2	-	4	4	6	6	-	-
FAGACEAE										
Quercus crassipes	-	-	3	-	12	-	3	6	-	6
ONAGRACEAE										
Fuchsia minimifolia	5	9	4	6	2	12	2	-	8	12
Fuchsia thymifolia	-	-	4	9	9	7	6	4	-	6
POLYGONACEAE										
Monnina xalapensis	-	1	-	6	4	-	2	14	-	9
ROSACEAE										
Crataegus pubescens	6	6	2	-	-	6	6	-	-	-
Prunus serotina	6	6	2	6	4	8	8	2	3	6
H										
ERICACEAE										
Arbutus xalapensis	-	-	-	-	-	-	-	-	-	6
GRAMINEAE										
Avena fatua	24	17	11	15	10	2	14	-	9	-
Muhlenbergia macroura	9	9	21	-	17	3	8	-	-	-
IRIDACEAE										
Orthorosanthus chimboracensis	-	-	-	-	2	21	4	-	6	6
LABIATAE										
Stachys sp.	8	6	-	6	5	-	-	2	-	-

LITERATURA CITADA

Abbott, L. A., P. A. Bisley y D. J. Rogers (1985) Taxonomic Analysis in Biology. Columbia University Press, New York.

Andenberg, M. R. (1973) Cluster Analysis for applications. Academic Press, New York. 359 p.

Braun-Blanquet, J. (1979) Fitosociología. Bases para el estudio de las comunidades vegetales. Blume ediciones. Madrid, España. 820 p.

Cola A. O' Muircheartaigh y P. Clive (1977) The analysis of survey data. Vol.I. EXPLORATING DATA STRUCTURES. John Wiley and Sons.

Contreras y Melo (1974) La importancia biológica y social de las reservas naturales. Ed. INERNAR, México, D. F. 90 p.

Crisci, J.V. y M. F. L. Armengol (1983) Introducción a la teoría y práctica de la Taxonomía Numérica. Depto. de Asuntos Cient. Y Tec. de la Sec. Gral. de la O.E.A., Washington, D.C. Serie de biología. Monografía No. 26. 132 p.

Digby, P. G. y R. A. Kempton (1987) Multivariate analysis of ecological communities. London. Chapman and Hall.

Duchaufour, P. (1975) Manual de Edafología. Toray-Masson, Barcelona, España. 476 p.

Everitt, B.S. (1981) Cluster Analysis. 2nd. ed. Halsted, New York. 137 p.

FAO (1976) Esquema para la evaluación de tierras. Organización de las Naciones Unidas para la Agricultura y la Alimentación, Roma.

Gauch, H. G. (1982) Multivariate analysis in community ecology. Cambridge University Press. 298 p.

Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857-872.

Gower, J. C. (1976) A comparison of some methods of cluster analysis. *Biometrics*, 23, 623-628.

INEGI (1981) Manual para la interpretación de la carta de Uso Potencial. Escala 1:50,000. Instituto Nacional de Estadística, Geografía e Informática, S.P.P., México, D.F.

Jardine, N. y R. Sibson (1968) The construction of hierarchic and non-hierarchic classifications.

Kendall, M. (1980) Multivariate Analysis. 2da. Ed. Macmillan, New York.

Llorente, J. (Compilador) (1989) Patrones de la Sistemática y Evolución en México. *Ciencias* (Número especial 3). 112 p.

Mahalanobis, P.C. (1936) On the generalized distance in statistics. *Proc. Natn. Inst. Sci. Calcutta*, 12, 49-55.

Manly, B. (1986) Multivariate Statistical Methods. A PRIMER. Chapman and Hall, New York.

Matteucci, S. y A. Colma (1982) Metodología para el estudio de la vegetación. Depto. de Asuntos Cient. y Tec. de la Sec. Gral. de la O.E.A., Washington, D.C. Serie de biología. Monografía No. 23, 163 p.

Mueller-Dombois, D. y H. Ellemberg (1974) Aims and methods of vegetation ecology. Ed. John Wiley and Sons., New York.

Ortiz, C. y M. Cuanalo (1984) Metodología del levantamiento fisiográfico. Colegio de Postgraduados. Universidad Autónoma de Chapingo.

Pielou, E. C. (1979) Biogeography. John Wiley and Sons., New York. 351 p.

Pielou, E. C. (1984) The interpretation of ecological data. Ed. John Wiley and Sons., New York. 263 p.

Rohlf, F. J. (1989) NTSYS-pc. Numerical Taxonomy and Multivariate Analysis System. Versión 1.50. Exeter Pub. LTD. New York.

Sanchez, S. E. (1970) Flora de San Cayetano. *Tesis*. Secretaría de Educación Pública. Escuela Normal Superior, México, D. F.

SAS Institute Inc, (1985) SAS/STAT Guide for Personal Computers, Versión 6 Edition; U. S. A.

Sneath, P. H. A. y R. R. Sokal (1973) Numerical Taxonomy. W. H. Freeman, San Francisco.

Sokal, R. R. y P. H. Sneath (1963) The Principles of Numerical Taxonomy. W. H. Freeman, San Francisco.

STATGRAPHICS (1989) STATGRAPHICS User's Guide. Statistical Graphics Corporation. U. S. A.

Tamhane, R. (1978) Suelos: su química y fertilidad en zonas tropicales. Ed. Diana, México, D. F. 483 p.

Usher, M. B. (1986) Wildlife conservation evaluation: attributes, criteria and values. *In*: Usher. M. B. (Ed.), WILDLIFE CONSERVATION EVALUATION. Chapman and Hall, London.

Wiley, E. (1981) Phylogenetics. The theory and practice of Phylogenetic Systematics. John Wiley and Sons inc. New York. 439 p.

Whittaker, R.H. (1975) Communities and Ecosystems. MacMillan Publishing CO., INC. 385 p.

Whittaker, R.H. (Ed.) (1980) Classification of Plant Communities. Dr.W. Junk bv Publishers, The Hague.