UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES
A C A T L A N

ANALISIS DE CONGLOMERADOS COMO UNA ALTERNATIVA EN LA FORMACION DE ESTRATOS.

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

BEATRIZ ALCANTARA GONZALEZ

NAUCALPAN, EDO. DE MEXICO 1996

OBJETIVOS

Objetivo General

Establecer la metodología para resolver en general el problema de la división de una población en subpoblaciones homogéneas y heterogéneas.

Objetivos Específicos

Aplicar la metodología en una población formada por instituciones encargadas de la coordinación de la política de población en los estados.

Aplicar la metodología en una población formada por todas las entidades federativas en relación con 10 principales causas de mortalidad.

Aplicar la metodología en la división del estado de Chiapas en subpoblaciones de municipios con base en variables del tipo socioeconómicas.

INDICE

INTRODUCCIÓN

En diversas investigaciones de tipo social se han detectado variaciones importantes en las variables de análisis, de tal manera que diseñar investigaciones que no tomen en cuenta estas diferencias, produciría sesgos en las estimaciones. En estos casos la teoría del muestreo propone dividir a la población en estratos, los cuales son subpoblaciones que son homogéneas al interior de ellas y, heterogéneas entre ellas con respecto a las variables que interesa medir. Sin embargo, la mayoría de la gente que ha escrito sobre temas de muestreo no trata el problema de como construir esas subpoblaciones.

Es por esto que, se hace necesario tener una herramienta alternativa para dividir a la población en estratos. Sin embargo la división implica resolver dos grandes problemas; en primer lugar, es necesario determinar en cuantas subpoblaciones es necesario dividir a la población total de estudio y en segundo, determinar qué elementos de la población de estudio forman cada uno de los estratos.

En este trabajo se propone una solución alternativa para ambos problemas. En el Capítulo I, se abordan los temas del muestreo y la estratificación de tal manera que se plantean los problemas que se presentan al diseñar una muestra y al estratificar una población, también se plantean las ventajas y desventajas que presenta una muestra estratificada.

El Capítulo II trata de manera detallada la técnica estadística del Análisis de Conglomerados, en la cual se presentan dos alternativas para determinar el número de subpoblaciones en que hay que dividir a una población dada y, además se establecen diferentes maneras para asignar los elementos de una población a las diversas subpoblaciones.

Por último, en el Capítulo III se hace la aplicación en tres casos de algunas de las técnicas mencionadas en el Capítulo II, con el fin de determinar el número de estratos y la formación de los mismos, pero con la característica de ser homogéneos al interior y heterogéneos al exterior de ellos. Uno de estos casos es un estudio realizado con las secretarías técnicas de los consejos estatales de población de todos los estados, donde se incluyen tres variables: nivel de estructura, de presupuesto y de capacitación. El segundo caso se realizó para las causas de mortalidad a nivel estatal, la tercera y última de las aplicaciones se realizó con variables de tipos socioeconómico de todos los municipios del Estado de Chiapas.

Aunque las aplicaciones se hicieron con variables de tipo numérico, es importante remarcar que la metodología propuesta se puede utilizar también con variables de tipo nominal, ordinal o combinaciones de ambas.

El Muestro y la Estratificación

I.1 Los Dos Problemas Fundamentales del Muestreo

En cualquier investigación ya sea del tipo social, económica, demográfica, de salud, etc., existe siempre una población de estudio que está determinada por sus elementos y, su ubicación en el espacio y en el tiempo.

El proceso de investigación empieza con el establecimiento de los objetivos del estudio y, con base en éstos se determinan los elementos a estudiar. La especificación del espacio determina el lugar en el cual se encuentran los elementos de estudio y el tiempo determina el momento en que van a ser válidos los resultados de la investigación y obviamente el momento en que van a ser medidos los elementos de la población de estudio.

De los elementos de la población se obtienen datos que se utilizan para la toma de decisiones. Una de las maneras de medir las características de los elementos de la población es mediante una encuesta. Este proceso de medición se puede delimitar en tres grandes etapas: 1) La preparación del instrumento de medición, 2) La obtención de datos y 3) El procesamiento y análisis de la información.

Existen dos opciones para llevar a cabo la obtención de datos, una es mediante la medición de todos los elementos de la población de estudio lo cual se conoce como *censo* y la otra es mediante la medición de sólo un subconjunto de los elementos de la población y que se conoce como una *muestra*.

Tanto en el censo como en la muestra, los parámetros de interés son en general totales, medias y proporciones. Con el censo se obtienen los valores reales, mientras que con la muestra se obtienen estimadores. Por ejemplo, supóngase que los elementos son los comercios y que las características de interés son: el número de empleados, la renta del local, el número de empleados del sexo femenino, el capital y las ventas. Algunos de los parámetros de interés en este caso pueden ser: el total de empleados, el promedio de empleados por comercio y la proporción de comercios que se dedican a la venta de zapatos.

Es importante dejar claro que en ambos casos existen problemas inherentes. Por ejemplo si se decide realizar el estudio de la población completa (censo) surgirían diversos problemas como la falta de tiempo para realizar la investigación, la lejanía existente entre los elementos a estudiar; provocando de esta manera un elevado presupuesto, entre otros problemas; y es por estas razones que los investigadores generalmente optan por la obtención de datos mediante la elaboración de una encuesta por muestreo.

Algunas de las ventajas de una muestra con respecto al censo son: menos costo, mayor rapidez en la obtención de resultados, más posibilidades de utilizar personal altamente calificado y mayor exactitud. Sin embargo, en caso de que se decida estudiar a la población mediante una muestra, el investigador tiene que resolver entre otros problemas, lo siguiente:

1) ¿Determinar cuántos elementos de la población de estudio se deben de incluir en la muestra para obtener conclusiones válidas?

2) ¿Determinar cuáles elementos son los que se deben elegir para que formen parte de la muestra?

El primer problema se conoce como el tamaño de la muestra, mientras que el segundo se conoce como la selección de los elementos de medición y ambos son conocidos como los dos problemas fundamentales del muestreo.

I.2 El Tamaño de la Muestra

El cálculo del número de elementos que conformarán la muestra es muy importante, ya que si N denota el total de elementos de la población, el tamaño de una muestra puede ser tan pequeño como un solo elemento o tan grande como N-1 elementos. Es obvio por sentido común que la estimación de alguna característica de la población total con un solo elemento tendrá asociado, un error más grande que si ésta se hace con casi todos los elementos de la población. Por supuesto que ambos casos son los extremos, pero dan idea de que entre más grande es la muestra, el error de las estimaciones es más pequeño.

Es así que resulta interesante cuestionarse sobre la existencia de un tamaño de muestra que no sea ninguno de estos extremos y que permita obtener estimaciones que tengan asociado un error entre el caso de un solo elemento y el de N-1.

Para abordar la solución de este problema asociado con el estudio del tamaño de la muestra, supóngase que P denota alguna proporción de la población que se desea estimar. Una pregunta que surge de manera natural en este momento es la siguiente:

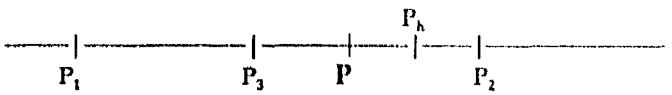
> ¿Existirá alguna relación entre el número de elementos de la población de estudio y el parámetro P?

Para contestar a esta pregunta, supóngase primeramente que con 1, 2, 3, ..., h elementos de la población se obtienen respectivamente estimadores del parámetro P, los cuales se denotarán como P_1, P_2, ..., P_h. En segundo lugar supóngase también que los respectivos estimadores están más cerca[1] del parámetro poblacional P entre más grande sea el número de elementos considerados en la estimación (Figura I.1).

---

[1] A la distancia entre el estimador P_n para n=1,2,...,n=h y el parámetro P se le conoce como el error muestral.

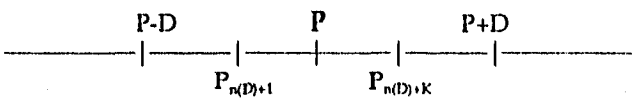
Figura I.1



Es decir que, entre más grande es el número de elementos de la población que se toman en cuenta para la estimación, más cerca se encuentra ésta del parámetro poblacional P y por lo tanto el error muestral es más pequeño. Pero lo más importante que hay que resaltar, es que esto es cierto en general. Es decir, existe un resultado de la teoría de la probabilidad conocido como **La Ley de los Grandes Números** (Gredenko, 1968) que afirma lo siguiente:

Para cualquier error muestral de tamaño D que se esté permitido aceptar, se puede encontrar un número n que depende de D y denotaremos entonces como n(D), tal que cualquier estimación del parámetro P que se haga con números de elementos de la población mayores o iguales que esta n(D), tendrá un error muestral a lo más de tamaño D (Figura I.2).

Figura I.2



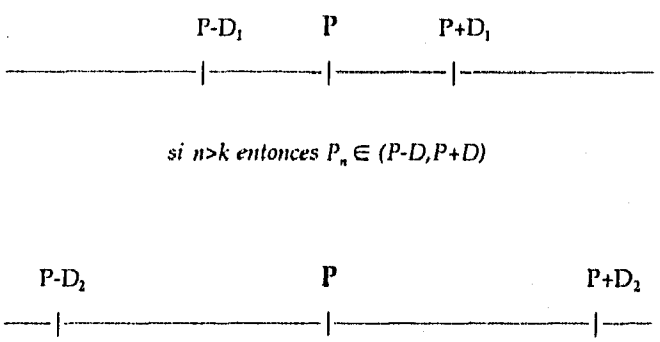
Si al error D se le interpreta como una vecindad de radio D del parámetro poblacional P, entonces el resultado anterior se puede interpretar en términos de convergencia diciendo que, a partir del número n(D) todas las estimaciones que se hagan con un número mayor estarán contenidas en la vecindad de radio D (Figura I.2).

Ahora bien, qué sucederá si consideramos una D_1 mayor que la D anterior y una D_2 menor que D. Como D_1 es más grande, entonces existen resultados de la teoría de convergencia que

aseguran que se puede encontrar un $n(D_1)$ menor que la n(D) tal que también cualquier estimación que se haga con un número de elementos de la población mayor o igual que $n(D_1)$ tendrán asociado un error muestral a lo más de tamaño D_1. Con respecto a D_2 los resultados de la teoría de convergencia aseguran también que, se puede encontrar una $n(D_2)$ mayor que n(D) que tiene características análogas.

El resultado anterior, se resume diciendo que si el radio de la vecindad se hace más pequeño, entonces la n a partir de la cual las estimaciones están contenidas en la vecindad es más grande que la n asociada con la vecindad más grande (Figura I.3).

Figura I.3



En palabras el error muestral de un estimador se puede interpretar como la distancia máxima que se permite que el estimador se desvíe del parámetro poblacional. También es conocida como la precisión del estimador, esto tiene implicaciones en el sentido de que entre más pequeña sea la precisión del estimador más grande es el número mínimo de elementos de la

población de estudio que se necesitan para que a partir de este número los estimadores estén contenidos en la vecindad. También se puede interpretar diciendo que entre más grande sea la n, el estimador del parámetro P es más preciso y por lo tanto el error muestral es más pequeño.

Al número más pequeño de elementos de la población de estudio, a partir del cual todos los estimadores están contenidos en la vecindad definida por la precisión deseada, se conoce como el **tamaño de la muestra.**

Lo anterior se ha ejemplificado solamente en términos determinísticos. Esto quiere decir que a partir de la n que determina el tamaño de la muestra, todas las estimaciones estarán contenidas en la vecindad. Sin embargo, cuando se trata de fenómenos aleatorios puede suceder que algunas estimaciones estén fuera de la vecindad de radio D a pesar de realizarse con una n mayor o igual que el tamaño de la muestra.

Este hecho representa un gran problema para el investigador si no se sabe cuantas veces puede suceder esto. Es decir, si sucede muchas veces implica que las estimaciones no son muy precisas. Lo ideal es que sucediera pocas veces.

La misma Ley Débil de los Grandes Números, asegura también que lo anterior sucede muy pocas veces entre más grande sea el tamaño de la muestra, debido a que los diferentes estimadores calculados con n cada vez más grandes ***convergen en probabilidad*** al parámetro P. En términos probabilísticos, la convergencia en probabilidad afirma que, la probabilidad de que dadas las estimaciones de P para n mayor o igual que el tamaño de la muestra caigan fuera de la vecindad se hace pequeña entre más grande sea la n considerada para las estimaciones. Implícitamente, lo anterior afirma que, la probabilidad de que las estimaciones estén contenidas en la vecindad dada, se hace grande entre más grande sea la n considerada para las estimaciones. En la teoría del muestreo a esta probabilidad se le conoce como la ***confiabilidad*** del estimador. En la práctica se calcula la n que se necesita para una

confiabilidad dada. Por lo tanto se puede concluir hasta este momento que:

> Entre más grande es el tamaño de la muestra, el estimador es más preciso y la confiabilidad es también más grande.

De esta manera se concluye también que para calcular el tamaño de la muestra se necesita una precisión y una confiabilidad que el investigador debe establecer. Así, una vez que el investigador establece la precisión y la confiabilidad del estimador de interés, se puede obtener la fórmula para calcular el tamaño de la muestra. Si el parámetro poblacional de interés es una proporción entonces la fórmula está dada por:

$$n = PQ\left(\frac{Z_{\alpha/2}}{D}\right)^2 \qquad (I.1)$$

donde:

n es el tamaño de la muestra

P es la proporción poblacional que se desea estimar

Q = 1-P es el complemento de P

1-α es la confiabilidad del estimador

D es la precisión del estimador; y

$Z_{\alpha/2}$ es el valor de la Distribución Normal estandarizada que deja a su derecha una probabilidad de $\alpha/2$.

En ocasiones cuando la población es pequeña puede resultar que el tamaño de la muestra sea más grande que la población, en este caso Cochran (1977) propone utilizar una fórmula corregida que está dada por:

$$n = \frac{PQ\left(\frac{Z_{\alpha/2}}{D}\right)^2}{1 + \frac{1}{N}\left(\frac{PQ\,(Z_{\alpha/2})^2}{D^2} - 1\right)} \qquad (I.2)$$

donde:

n es el tamaño de la muestra

N es el número de elementos de la población

P es la proporción poblacional que se desea estimar

Q = 1-P es el complemento de P

1-α es la confiabilidad del estimador

D es la precisión del estimador; y

$Z_{\alpha/2}$ es el valor de la Distribución Normal estandarizada que deja a su derecha una probabilidad de $\alpha/2$.

Si el parámetro que se desea estimar es una media entonces la fórmula para calcular el tamaño de muestra está dado por:

$$n = \frac{\frac{\delta^2}{\mu^2}\left(\frac{Z_{\alpha/2}}{D}\right)^2}{1 + \frac{1}{N}\left(\frac{\delta^2}{\mu^2}\,\frac{(Z_{\alpha/2})^2}{D^2}\right)} \qquad (I.3)$$

donde:

n es el tamaño de la muestra

N es el número de elementos de la población

μ es la media poblacional a estimar

δ^2 es la varianza de la característica de interés

D es la precisión del estimador

$1-\alpha$ es la confiabilidad del estimador; y

$Z_{\alpha/2}$ es el valor de la Distribución Normal estandarizada que deja a su derecha una probabilidad de $\alpha/2$.

Ahora bien, tanto en la fórmula I.1 como en la fórmula I.2 se observa que el cálculo del tamaño de muestra depende de los parámetros P o Q que se desean estimar. Este es un problema complicado de resolver porque precisamente el desconocimiento de estos parámetros es lo que da origen a la investigación, de tal manera que si se conociera el valor de estos parámetros no tendría caso hacer la investigación. En la práctica este problema se ha resuelto de la siguiente manera:

1) Obtener información en estudios previos del parámetro de interés.
2) Por conocimiento del fenómeno se pueden establecer en ocasiones valores aproximados de los parámetros.
3) Recabar información de los parámetros de interés mediante pruebas piloto.
4) Establecimiento de supuestos en los parámetros P o Q que sobrecalculen el valor del

tamaño de muestra.

Por ejemplo, si el parámetro a estimar es una proporción, algo importante que hay que resaltar es el hecho de que en la Fórmula I.1 y Fórmula I.2 una vez que se establece la confiabilidad y la precisión, el cálculo depende de los valores de PQ el cual tiene un valor máximo cuando P=0.5. Si se trabaja bajo este supuesto se obtendrá un tamaño de muestra que cubre al investigador en el caso de que P sea más grande o más pequeño que 0.5. Para el caso en que el parámetro de interés es una media, los puntos 1 a 3 pueden ser una solución.

I.3 La Selección de los Elementos

Una vez calculado el tamaño de la muestra, el problema siguiente es determinar quienes van a ser parte de ella. En poblaciones humanas el problema es todavía más complejo, porque los elementos de la población no los tenemos en una lista para poder seleccionarlos y porque generalmente la selección de ellos, implica una selección de viviendas o de lugares en que éstos habitan. A su vez, la selección de viviendas implica una selección de localidades o zonas geográficas en las que están inmersas las viviendas, etc.

Ahora bien independientemente de la selección, en la teoría de muestreo existen dos maneras de seleccionar a los elementos:

1) El muestreo probabilístico y

2) El muestreo no probabilístico

En el muestreo probabilístico, todos los elementos de la población tienen una probabilidad conocida y no nula de ser seleccionados y se caracteriza porque los resultados pueden inferirse a la población total.

En el muestreo no probabilístico, la probabilidad de selección de algunos elementos es cero. Esto se traduce en que, en cualquier intento de selección, estos elementos de la población nunca van a ser parte de la muestra. Lo anterior implica que los resultados no pueden inferirse a toda la población porque esos elementos no están representados en la muestra. Sin embargo, su utilización se justifica en ocasiones por la comodidad y la economía para recopilar la muestra.

I.3.1 Tipos de Muestreo Probabilístico

Los métodos de selección probabilísticos son muy variados y generalmente se dividen en aquellos en donde la selección de los elementos se hace en una sola etapa y aquellos en donde la selección se hace en varias etapas.

Si se tiene evidencia de que los elementos de la población son muy parecidos en la característica de interés entonces la selección se puede hacer en una sola etapa. El muestreo aleatorio simple o el muestreo sistemático se pueden utilizar en este caso.

Por el contrario, cuando se sabe que la población está compuesta por elementos diferentes, es necesario dividir a ésta en subpoblaciones, de tal manera que los elementos que forman a una subpoblación sean lo más parecidos entre ellos y que entre subpoblaciones sean lo más diferentes posible.

En la teoría del muestreo existen dos métodos para dividir una población. Si la división se hace con base en criterios geográficos entonces se dice que la población está dividida en conglomerados, mientras que si la división se hace en términos conceptuales, entonces se dice que la población está dividida en estratos.

El muestreo en varias etapas o multietápico se caracteriza porque cada unidad de la población se puede dividir en cierto número de unidades más pequeñas o subunidades. Si se toma una muestra de n unidades y si las subunidades contenidas en una unidad seleccionada dan resultados semejantes, no parece económico medirlas todas. Una práctica acostumbrada consiste en seleccionar y medir una muestra de subunidades de alguna unidad elegida. Esta técnica se llama *submuestreo o muestreo en dos etapas* porque la muestra se obtiene en dos pasos, dado que la unidad no se mide completamente, sino que a su vez es objeto de un muestreo, donde lo primero que se hace es seleccionar una muestra de unidades, que a menudo se denominan unidades primarias, y después se selecciona una muestra de unidades de la segunda etapa. El procedimiento se puede generalizar para cualquier número de etapas en que se requiera subdividir a la unidad de análisis. Este tipo de muestreo se usa frecuentemente en poblaciones humanas.

I.3.2 Tipos de Muestreo No Probabilístico

Muestreo de juicio

En este tipo de muestreo un conjunto de expertos utiliza su experiencia en el conocimiento del fenómeno para identificar elementos potenciales de la población de estudio. Este muestreo es útil y aconsejable, si el muestreo probabilístico no es factible o resulta altamente caro o si el tamaño de la muestra es muy pequeño.

Muestreo de cuota

Es un muestreo de juicio, en donde la muestra se forma incluyendo un número mínimo de elementos para subgrupos específicos de la población de estudio. Es un muestreo que se basa frecuentemente en datos demográficos y/o en localizaciones de regiones. Por ejemplo, si el tamaño de muestra es n=1,000 y, si se sabe que la distribución de los elementos de la población de estudio es de 30%, 10%, 20% y 40% en el Norte, Este, Centro y Suroeste respectivamente, entonces se muestrean 300, 100, 200 y 400 elementos de las zonas geográficas respectivas.

Muestreo de bola de nieve

Este muestreo consiste en construir una lista especial mediante el empleo de un conjunto

inicial de miembros como informantes. La lista se construye preguntando a cada entrevistado inicial la identificación de otras personas para que posteriormente sean entrevistadas. Es muy útil para estudios en poblaciones especializadas, como por ejemplo, el estudio de una población de homosexuales, drogadictos o enfermedades muy especiales, etc.

I.4 LA ESTRATIFICACIÓN

Como se mencionó anteriormente, cuando los elementos de la población no se parecen es recomendable dividir a la población total en subpoblaciones, en el muestreo estratificado a cada subpoblación se le conoce como *estrato*. Los elementos de la población se dividirán en subpoblaciones y cada elemento deberá de estar en uno y solamente uno de los estratos. La estratificación difiere de la conglomeración no sólo por el hecho que se define en términos conceptuales, sino que además en el hecho de que todos los estratos son sujetos a medición, mientras que con los conglomerados, se obtiene una muestra del total de ellos.

La finalidad de la estratificación es que los elementos que pertenecen a un estrato sean lo más parecidos posible entre ellos, mientras que los elementos de un estrato a otro deben ser lo más diferentes posible. Si este es el caso se dice que se tienen estratos homogéneos al interior de ellos y heterogéneos entre ellos.

En el muestreo estratificado, la población total de N elementos de estudio se divide en subpoblaciones de $N_1, N_2,...,N_n$ elementos respectivamente. Como no se deben traslapar y en

su conjunto comprenden a toda la población, entonces:

$$N_1+N_2+...+N_n=N$$

Para obtener todo el beneficio de la estratificación, los valores de los N_n deben ser conocidos. La estratificación es una técnica muy común. Para su utilización hay muchas razones; las principales son las siguientes:

- Si los datos deseados deben tener una precisión conocida en algunas subdivisiones de la población, es aconsejable tratar cada subdivisión como una población por derecho propio.

- Por conveniencia administrativa, puede ser necesario el uso de la estratificación; así por ejemplo, la agencia que realiza una encuesta, podría tener sucursales en el campo, cada una de las cuales supervisaría la encuesta de una parte de la población.

- Los problemas de muestreo pueden tener marcadas diferencias en diversas partes de la población. Con poblaciones humanas, las personas que viven en instituciones (como hoteles, hospitales, cárceles) se colocan en un estrato diferente de las que viven en casas ordinarias, ya que otro método de muestreo es el apropiado para cada una de estas situaciones. Al muestrear negocios, podríamos tener una lista de las grandes firmas que se deben colocar en un estrato diferente. Quizá algún tipo de muestreo por áreas debe

usarse para las firmas pequeñas.

- La estratificación puede dar lugar a una ganancia en la precisión de las estimaciones de características de la población total. Quizá sea posible dividir una población heterogénea en subpoblaciones, en las que cada una sea internamente homogénea. Esto es lo que sugiere el nombre de *estratos*, con su implicación de una división de capas. Si cada estrato es homogéneo, en cuanto a que las medidas varíen ligeramente de una unidad a otra, una estimación precisa de cualquier medida de estrato se puede obtener a partir de una pequeña muestra en dicho estrato. Y posteriormente podrán combinarse estas estimaciones en una estimación precisa para toda la población.

I.5 Construcción de Estratos

Cockran (op. cit.), plantea el problema de la siguiente manera: ¿Cuál es la mejor característica para la construcción de los estratos? ¿Cómo se determinan los límites entre estratos? ¿Cuántos estratos debería haber?.

Para una sola característica o variable y, lo ideal es considerar por supuesto, la distribución de frecuencia de y. Otra opción es probablemente la distribución de frecuencia de alguna otra cantidad altamente correlacionada con y, porque generalmente el desconocimiento de la investigación origina la investigación.

Debe quedar claro que la estratificación basada en los valores de y es poco realista. En la práctica, se utiliza alguna otra variable x. Dalenius (1957) desarrolla ecuaciones para límites de x que minimizan la varianza, siempre y cuando se tenga cierto conocimiento de la regresión de y sobre x. Si esta regresión es no lineal, estos límites pueden diferir considerablemente de los óptimos, cuando x es la variable que se mide.

En la estratificación geográfica el problema es menos tratable matemáticamente porque existen muchas maneras diferentes de formar los límites de estratos. El procedimiento usual es estimar algunas variables de alta correlación con las características principales de la encuesta y usar una combinación de sentido común y prueba y error para construir buenos límites para estas variables seleccionadas. Dado que las ganancias en precisión con la estratificación es probable que sean modestas, no vale la pena hacer un gran esfuerzo para mejorar los límites.

Existen ejemplos concretos de este problema aplicados en el área económica y agraria, discutidos por Stephan (1941) y Hagood y Bernert (1945), y King y Mac Carty (1941), respectivamente.

ANÁLISIS DE CONGLOMERADOS

II.1 OBJETIVO DEL ANÁLISIS DE CONGLOMERADOS

Frecuentemente nos enfrentamos al problema de tratar de analizar una gran cantidad de datos que además tienen un comportamiento muy complejo. Es por esto que se sugiere utilizar métodos que resumen los datos en primer lugar para detectar las relaciones existentes y en segundo para identificar los patrones de comportamiento entre ellos. Si al aplicar lo anterior resultasen grupos claros de objetos, entonces éstos pueden ser estudiados y sus propiedades resumidas, obteniendo después resultados que le ayuden al investigador a hacer predicciones o descubrir hipótesis que expliquen la estructura de los datos observados, lo cual sería imposible si se estudia a la población completa (Gordon, 1981, p.6-9). Esto es posible si los "g" grupos de objetos que se forman son muy homogéneos al interior de cada grupo y además muy heterogéneos entre ellos. El número de grupos (g) es desconocido pero usualmente es mucho menor que el número de elementos de la población de estudio.

El análisis de conglomerados es un campo muy amplio que abarca diferentes disciplinas. Existen muchos ejemplos de campos tan diversos como son: la biología, la zoología, la botánica, la sociología, la psicología, percepción auditiva y visual, lingüística, etc. Su uso ha sido al igual que la mayor parte de otros procedimientos multivariados, facilitado y agilizado con el uso de las computadoras, sin ellas sus aplicaciones se hacen muy tediosas. El elemento común a estos problemas es que un gran número de objetos han sido medidos en cierta cantidad de variables y éstos necesitan agruparse.

Como antecedentes de esta técnica se tiene la clasificación de personas en tipos, esta clasificación era una técnica antigüa, por ejemplo, los nativos de la India usaban el sexo y las características físicas y de conducta para clasificar a la gente en 6 grupos o tipos, a los cuales se les asignaban nombres de animales. Los antigüos médicos de Grecia y Roma desarrollaron varias tipologías basadas en variaciones de características físicas que se generan desde una mixtura hasta 4 clases, una de las tipologías desarrolladas más importantes fue la efectuada por Galen (129-199 A.D.) donde definió 9 tipos de temperamentos que relacionaban la suceptibilidad de las personas a las enfermedades y a diferencias individuales de comportamiento. En el siglo XVIII Linnaeus desarrolló una clasificación del reino animal y vegetal (General Plantarum, publicado por primera vez en 1737).

Una de las aplicaciones sugerida por Morrison (1967) en el campo de la investigación de mercados, es que una vez formados los "g" grupos, entonces se pueden tomar unidades representativas de las demás y comparar los resultados de las otras unidades representativas de cada grupo. Por ejemplo, se puede tener un gran número de ciudades que se podrían usar como pruebas de mercado, pero debido a factores de presupuesto, la prueba se debe de restringir a un número más pequeño de estas ciudades; entonces si las ciudades se agruparan en un número pequeño de conglomerados, de manera que las ciudades dentro de un grupo fueran muy similares, entonces una ciudad de cada grupo se puede usar como una prueba del mercado de cierto producto.

Everitt (1974) menciona que el análisis de conglomerados se puede ver también como una reducción del manejo de la información, ya que la concerniente a N individuos se reduce a la información de únicamente "g" grupos, que en determinado momento se pueden tratar como unidades y, por lo tanto, puede ser posible dar una explicación más concisa y entendible de las observaciones que se estén considerando.

Otra aplicación de esta técnica es producir grupos que sean útiles para propósitos predictivos de algún tipo. Por ejemplo, un estudio realizado por Paykel (1972) en el que aplicó el análisis

de conglomerados a pacientes psiquiátricos, formando grupos que reaccionan de manera diferente al aplicarles un tipo de droga. Los resultados de la investigación se utilizaron para decidir el tipo de droga que deberá de aplicar a cada paciente según el grupo al que pertenezca.

II.2 MEDIDAS DE PROXIMIDAD

Los datos básicos para el análisis de conglomerados es un conjunto de N unidades de las cuales se tienen p medidas diferentes, que integran una matriz de Nxp, como se muestra a continuación:

$$X = \begin{vmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ x_{N1} & \cdots & x_{Np} \end{vmatrix}$$

donde:

X denota una matriz de orden Nxp

x_{ij} es la medición de la j-ésima variable para el objeto i-ésimo

En general los objetos que conforman los grupos, pueden ser ya sea humanos, ciudades, barcos, plantas, animales, o cualquier otro objeto; y las mediciones de las variables pueden ser la estatura, el peso en humanos, los tamaños de población, el porcentaje de los ingresos en una determinada ciudad, etc.

La mayoría de los métodos de clasificación suponen que todas las relaciones relevantes dentro del conjunto de objetos de la matriz X a clasificar, se pueden resumir en una matriz que contiene valores de proximidad entre cada par de objetos. Es decir, a cada par de objetos se le asocia un valor numérico p_{ij}, el cual representa una medida de que tanto el i-ésimo y el j-

ésimo objeto se parecen uno a otro. La palabra proximidad se usa para expresar *similaridad o disimilaridad*. Al investigar medidas de disimilaridad entre cada par de objetos es conveniente restringir nuestra atención a una clase particular de medidas de disimilaridad, llamadas coeficientes de disimilaridad. Si la $\mathfrak{L}$ denota el conjunto de objetos a clasificar, entonces un coeficiente de disimilaridad es una función definida de la siguiente manera:

d: $\mathfrak{L} \times \mathfrak{L} \Rightarrow R$ tal que:

$d_{ij} \geq 0 \qquad \forall\, i,j \in \mathfrak{L}$

$d_{ii} = 0 \qquad \forall\, i \in \mathfrak{L}$

$d_{ij} = d_{ji} \qquad \forall\, i,j \in \mathfrak{L}$

En palabras lo anterior se interpreta diciendo que la distancia que hay del objeto i-ésimo y el j-ésimo será mayor o igual a cero; la distancia entre el objeto i-ésimo y el mismo será nula, es decir, igual a cero; la distancia que existe del objeto i-ésimo y el j-ésimo es la misma que la del j-ésimo al i-ésimo objeto; siempre y cuando estos pertenezcan al conjunto de objetos a clasificar.

En ocasiones se hablará de *similaridad* y en otras de *disimilaridad*. Esto no debe ocasionar problema ya que las medidas de disimilaridad se pueden transformar en medidas de similaridad, mediante la siguiente fórmula:

$S_{ij} = 1/(1 + d_{ij})$

$S_{ij} = c - d_{ij}$ para alguna constante c de la característica de interés

Unicamente se presentan problemas al tratar de transformar S_{ij} en d_{ij} si la similaridad de un mismo objeto toma valores diferentes para diferentes objetos y si se desea obtener una medida de disimilaridad que posea propiedades de una métrica. Una manera de evitar este problema es decidir desde el inicio del estudio el uso de una de las dos medidas de proximidad.

Se define la *similaridad* de un elemento de tal manera que un valor grande indica que dos objetos están cercanos y uno pequeño que dos objetos están muy lejanos. Por lo tanto, la *similaridad* es el inverso lógico del concepto de distancia, en donde un valor grande indica que los objetos están muy alejados y uno pequeño que los objetos están cercanos.

Como el cálculo de la matriz de disimilaridades, o de similaridades es un paso preliminar en la clasificación de datos, se debe tener cuidado de obtener los conjuntos apropiados de similaridades (disimilaridades). Existe un gran número de medidas diferentes de proximidad. Las diferentes medidas dependen del tipo de variables que estén involucradas en el estudio.

En el caso de que las p variables de la matriz X sean todas binarias entonces cada variable toma los valores de 1 ó 0, dependiendo si posee o no el atributo de estudio, como se muestra a continuación:

$$X_{ij} = \begin{cases} 1 \text{ si está presente el atributo} \\ 0 \text{ si está ausente el atributo} \end{cases}$$

donde:

i = 1,2,...,N (número de objetos)
j = 1,2,...,p (número de variables)

Las relaciones entre el i-ésimo y el j-ésimo objeto en términos de variables binarias se pueden resumir en una tabla de 2x2, como se muestra a continuación:

Estado de la variable		Objeto j-ésimo	
		1	0
Objeto i-ésimo	1	a	b
	0	c	d

Los códigos utilizados en la tabla anterior en la i-ésima y j-ésima variables se utilizan para indicar si el atributo está presente o ausente; en este caso "1" nos indica que el atributo está presente y el "0" que el atributo está ausente. El número que aparece en la entrada (1,1) de la tabla anterior, significa que "a" de las p variables se encuentran en el estado 1 en ambos objetos; en la entrada (1,0) indica que "b" variables se encuentran en el estado 1 del objeto i y en el estado 0 en el objeto j; en la entrada (0,1) indica que "c" variables se encuentran en el estado 0 en el objeto i y en el estado 1 en el objeto j, por último la entrada (0,0) representa que "d" variables se encuentran en el estado 0 en ambos casos. Muchas medidas de similaridad entre el i-ésimo y el j-ésimo objeto se pueden definir con base en la tabla anterior.

Dos medidas comúnmente usadas en este caso son:

$$S_{ij} = \frac{a+b}{a+b+c+d}$$ (coeficiente de apareamiento)

$$S_{ij} = \frac{a}{a+b+c}$$ (coeficiente de Jacard)

En el caso de que las variables medidas sean nominales u ordinales se procede de la siguiente manera: como una variable nominal contiene S estados, puede ser recodificada en términos de S variables binarias. Si la variable nominal se encuentra en el estado m-ésimo ($1 \leq m \leq S$), entonces la m-ésima variable binaria se encontrará en el estado 1 y las otras (S-1) variables se encontrarán en el estado 0. Si dos objetos se comparan con base a estas S variables binarias, entonces se puede ver que el *coeficiente de Jacard* simplemente toma valores de 1 y 0 dependiendo de si el estado de la variable nominal es el mismo en ambos objetos y se obtendría una contribución de esta variable en el parecido de los objetos. Medidas completas de la disimilaridad entre un par de objetos se obtienen sumando la contribución de todas las variables.

Una de las formas más sencillas de determinar si dos observaciones son similares es

realizando el cálculo del número de características comunes, dicha técnica es denominada como *coeficiente de comparación*, el cual es la forma de determinar si dos observaciones son similares, tal medida es más apropiada cuando las variables son nominales. Por ejemplo, la podríamos utilizar para clasificar compradores de muebles, con base en su estilo de preferencia (americano antiguo, o americano moderno), estampado (normal, a cuadros o floral) y color (azul, amarillo, rojo o verde). En este caso los datos se muestran a continuación:

CLIENTE	PREFERENCIA		
	ESTILO	PATRON	COLOR
1	1	3	1
2	1	2	2
3	2	3	3
4	2	2	3

El ***coeficiente de comparación*** se puede aplicar tanto a variables ponderadas como a no ponderadas, a continuación se explicará cada uno de los casos.

No Ponderado: El coeficiente de comparación más sencillo cuenta el número de parejas iguales entre dos objetos:

$$S_{ij} = \sum_{c=1}^{p} Z_c$$

donde:

S_{ij} es la similaridad entre el objeto i y j

p es el número de características (variables) medidas

Z_c vale "1" si $x_{ic}=x_{jc}$ y vale "0" en caso contrario

x_{ic} es el valor del objeto i en la variable c

x_{jc} es el valor del objeto j en la variable c

Retomándo el ejemplo de los muebles, obtenemos las siguientes medidas de similaridad:

$$S_{12} = 1 + 0 + 0 = 1$$
$$S_{13} = 0 + 1 + 0 = 1$$
$$S_{14} = 0 + 0 + 0 = 0$$
$$S_{23} = 0 + 0 + 0 = 0$$
$$S_{24} = 0 + 1 + 0 = 1$$
$$S_{34} = 1 + 0 + 1 = 2$$

Ponderado: Los coeficientes simples de conciliación tienden a estar dominados por las variables con pocas categorías (es más sencillo encontrar parejas si existen sólo 2 categorías que si existen 20). Por esta razón, muchos investigadores prefieren ponderar parejas, ya que el realizar un estudio con variables no ponderadas implica mayores dificultades. Una forma común de hacer esto es ponderando una pareja en una característica por el número de categorías en dicha característica, mediante la siguiente fórmula:

$$S_{ij} = \sum_{c=1}^{p} V_c Z_c$$

donde:

S_{ij} es la similaridad entre el objeto i y j

p es el número de características (variables) medidas

Z_c vale "1" si $x_{ic}=x_{jc}$ y vale "0" en caso contrario

x_{ic} es el valor del objeto i en la variable c

x_{jc} es el valor del objeto j en la variable c

V_c es el número de valores de la característica c

El cálculo de la similaridad del ejemplo mencionado anteriormente realizando la ponderación de parejas, se ejemplifica enseguida, para lo cual es necesario conocer los valores de V_c que

se muestran en la siguiente tabla:

C	CARACTERISTICA FAVORITA				Nº DE CARACTERISTICA (V_c)
	1	2	3	4	
1: Estilo 2: Estampado 3: Color	Americano antiguo Normal Azul	Americano moderno A cuadros Amarillo	 Floral Rojo	 Verde	V_1=2 V_2=3 V_3=4

$$S_{12} = 2(1) + 3(0) + 4(0) = 2$$
$$S_{13} = 2(0) + 3(1) + 4(0) = 3$$
$$S_{14} = 2(0) + 3(0) + 4(0) = 0$$
$$S_{23} = 2(0) + 3(0) + 4(0) = 0$$
$$S_{24} = 2(0) + 3(1) + 4(0) = 3$$
$$S_{34} = 2(1) + 3(0) + 4(1) = 6$$

Esta técnica es útil cuando las variables son puramente nominales. Sin embargo, cuando los datos están dados en escalas a intervalos, tienden a producir resultados no separados. Para ejemplificar esta técnica con datos en escala se utiliza la siguiente información:

Objeto Observación Persona	Variable/Característica			
	1	2	3	4
1	2	4	6	32
2	5	2	5	38
3	3	3	7	30
4	1	2	3	16
5	4	3	2	30

De la variable 1 a la 3 son variables de actitud medidas en una escala de 7 puntos y la variable 4 es una escala de interés que tiene rangos entre 11 y 40. Los coeficientes de conciliación, utilizando las ecuaciones anteriores son los siguientes:

S_{ij}	No ponderado	Ponderado
S_{12}	0	0
S_{13}	0	0
S_{14}	0	0
S_{15}	0	0
S_{23}	0	0
S_{24}	1	7
S_{25}	0	0
S_{34}	0	0
S_{35}	2	37
S_{45}	0	0

Estos resultados indican que los objetos 3 y 5 son similares, debido a que son los que presentan los mayores valores de similaridad que son 37 para variables ponderadas y 2 para no ponderadas; mientras que los objetos 2 y 4 son algo similares, siguiendo el criterio anterior, y que todos los demás pares son completamente diferentes ya que todos presentaron similaridades de cero. Sin embargo, este resultado no es muy convincente ya que los objetos 2 y 4 parecen ser muy diferentes, salvo en cuanto a la variable 2 y los objetos 1 y 3 son muy similares en las cuatro variables. La razón para esto es que a pesar de estar cerca, no se le da crédito por el coeficiente de conciliación; dos objetos concuerdan exactamente o no concuerdan para nada. Por esta razón, muchos investigadores prefieren medidas de distancia que explícitamente incorporen cercanía. Algunas de las más populares son las siguientes:

- Suma de desviaciones absolutas
- Suma de diferencias al cuadrado

Suma de desviaciones absolutas:

$$D_{ij}=\sum_{c=1}^{p}|x_{ic}-x_{jc}|$$

donde:

D_{ij} representa la distancia existente entre el objeto i-ésimo y el j-ésimo

p es el número de características (variables) medidas

x_{ic} es el valor del objeto i-ésimo en la variable c

x_{jc} es el valor del objeto j-ésimo en la variable c

En este caso:

$$D_{12} = |2-5| + |4-2| + |6-5| + |32-38| = 12$$
$$D_{13} = |2-3| + |4-3| + |6-7| + |32-30| = 5$$
$$\vdots$$
$$D_{45} = |1-4| + |2-3| + |3-2| + |16-30| = 19$$

Suma de diferencias al cuadrado:

$$D_{ij}=\sum_{c=1}^{p}(x_{ic}-x_{jc})^2$$

donde:

D_{ij} representa la distancia existente entre el objeto i-ésimo y el j-ésimo

p es el número de características (variables) medidas

x_{ic} es el valor del objeto i-ésimo en la variable c

x_{jc} es el valor del objeto j-ésimo en la variable c

En este caso:

$$D_{12}=(2-5)^2+(4-2)^2+(6-5)^2+(32-38)^2=50$$
$$D_{13}=(2-3)^2+(4-3)^2+(6-7)^2+(32-30)^2=7$$
$$\vdots$$
$$D_{45}=(1-4)^2+(2-3)^2+(3-2)^2+(16-30)^2=207$$

Por otro lado para las variables numéricas, existen dos medidas de disimilaridad usadas comúnmente entre el objeto i y el j, estas son:

La Distancia Euclidiana:

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

La Métrica de Bloque:

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

donde:

d_{ij} representa la distancia existente entre el objeto i-ésimo y el j-ésimo

p es el número de características (variables) medidas

x_{ic} es el valor del objeto i-ésimo en la variable c

x_{jc} es el valor del objeto j-ésimo en la variable c

Para ambos coeficientes W_k es un conjunto de ponderaciones. Si no se tiene información acerca de la relevancia de las diferentes variables, se deben usar ponderaciones iguales. En muchas aplicaciones al inicio de la investigación se posee información de la importancia de las variables y entonces se debe incorporar esta información en la investigación. Los problemas asociados con la especificación precisa de ponderaciones se reducen si se encuentra que la clasificación se afecta muy poco al hacer cambios despreciables en los ponderadores. En muchos estudios de clasificación la ponderación precisa no es muy importante, pero tal robustez no siempre existe, por lo que en cada estudio de clasificación se debe considerar cuidadosamente el ejercicio de intentar especificar los factores de ponderación apropiados.

Es muy común que en estudios de clasificación los objetos se describan en términos de variables de tipo mixto. Gordon (1981) da tres enfoques al respecto.

Primera alternativa

Si la mayoría de las variables pertenecen a un tipo, lo más simple es convertir todas las variables al tipo de variable que predomina. Una variable se puede convertir en una variable binaria uniendo aconsejablemente los códigos. Una opción es mediante el análisis de la distribución frecuencial de la variable. Ejemplo:

	VARIABLE 1			
CODIGO	1	2	3	4
FRECUENCIA (%)	20	60	15	5

El código 2 presenta la mayor concentración de frecuencia y es por ese motivo que se puede definir una variable binaria de la siguiente manera:

	VARIABLE 1	
CODIGO	2	≠ DE 2
FRECUENCIA (%)	60	40

Una variable ordinal se puede considerar como nominal simplemente ignorando el orden de los estados. Una variable numérica se puede categorizar como una variable ordinal, ya sea tomando intervalos de igual longitud o que cada estado contenga un número similar de objetos. El cambio de una variable nominal a una ordinal únicamente se puede hacer si existe una variable asociada a la nominal, y que implique un orden en los códigos. Por ejemplo, supóngase que se tiene una muestra de mujeres que usan métodos anticonceptivos y que su distribución frecuencial es la siguiente:

	TIPO DE METODO USADO					
	RITMO	DIU	PAST.	INYEC.	LIG.	OTROS
CODIGO	0	1	2	3	4	5
FRECUENCIA (%)	25	28	32	15	12	8

Si ahora pensamos en términos de la efectividad del método usado, a la variable tipo de método se le pueden asignar números de tal manera que el orden entre ellos refleje cual es el más efectivo; la nueva codificación quedaría como se muestra a continuación:

	TIPO DE METODO USADO					
	OTROS	RITMO	PAST.	INYEC.	DIU	LIG.
CODIGO	0	1	2	3	4	5
FRECUENCIA (%)	8	25	32	15	28	12

Así, se deduce que el método con mayor efectividad es la LIGADURA y que el DIU es más efectivo que aquéllos clasificados como otros, pero menos efectivo que la ligadura.

Una variable ordinal se puede transformar en una numérica asignando el mismo punto de la recta real a todos los objetos que pertenecen al mismo código. Una manera de asignar el punto es usando estadísticas de orden de alguna distribución supuesta[1].

Segunda alternativa

Una segunda alternativa en la clasificación de objetos propuesta por Gordon, es definir un coeficiente de similaridad que incorpore información de los diferentes tipos de variables. Este coeficiente se llama *coeficiente de similaridad general* [2] y se define de la siguiente manera:

---

[1] Una discusión más completa con respecto a este tipo de transformación se presenta en Anderberg (1973, Capítulo 3).

[2] Definido por Gower (1971).

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

donde: $S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$

Donde S_{ijk} es la similaridad entre el i-ésimo y el j-ésimo objeto medido por la variable k-ésima y W_{ijk} vale uno o cero. Si S_{ijk} corresponde a una variable numérica entonces W_{ijk} vale 1, si el coeficiente S_{ijk} corresponde a una variable binaria entonces W_{ijk} vale cero cuando la variable k está ausente en ambos individuos, y en los otros casos vale 1. Si la variable es nominal u ordinal entonces se forman variables binarias y se procede como en el caso anterior; y R_k representa el rango de la variable k, R_k se calcula mediante la diferencia de los valores de x, mayor y menor.

En la siguiente tabla se presenta la forma en que se tomarán los valores de S_{ijk} y de W_{ijk} , de acuerdo a la presencia o ausencia del atributo de interés en los individuos.

Objeto I	+	+	-	-
Objeto J	+	-	+	-
S_{ijk}	1	0	0	0
W_{ijk}	1	1	1	0

- En la tabla anterior el signo negativo (-) indica la ausencia del atributo, y el positivo (+) indica la presencia de dicho atributo.

- En el caso de que los dos individuos presenten la característica de interés, tanto en el valor de S_{ijk} como en el de W_{ijk} serán iguales a 1, esto es lo que representa la primera columna de la tabla anterior.

- Cuando uno de los elementos presenta el atributo y el otro no, se considera su similaridad (S_{ijk}) como 0, ya que son diferentes; y el coeficiente de ponderación (W_{ijk}) sería igual a uno, ya que basta con que sólo uno de los dos presente el atributo.

- La tercera columna nos indica que el individuo i no presenta el atributo pero el j sí, por lo que su similaridad es nula (cero); mientras que su coeficiente de ponderación sería igual a 1, ocurre lo mismo que en el caso anterior.

- En caso de que ninguno de los dos individuos posean el atributo, tanto la similaridad como el coeficiente de ponderación valdrían 0.

Ejemplo: supongamos que se tienen 3 individuos, de los cuales se tienen medidas las siguientes variables: peso, estatura, color de ojos, color de cabello, fumador o no fumador. Los datos se presentan a continuación:

INDV	PESO (LIBR) K=1	ESTATURA (PULG) K=2	COLOR DE OJOS K=3	COLOR DE CABELLO K=4	FUMA O NO FUMA K=5
1	120	66	NEGRO	NEGRO	FUMADOR
2	130	72	VERDE	CASTAÑO	FUMADOR
3	150	70	NEGRO	NEGRO	NO FUMADOR

En este caso se utilizaron la medida de disimilaridad para variables numéricas, definida por Gower (1971), de la siguiente manera:

- En la tabla anterior el signo negativo (-) indica la ausencia del atributo, y el positivo (+) indica la presencia de dicho atributo.

- En el caso de que los dos individuos presenten la característica de interés, tanto en el valor de S_{ijk} como en el de W_{ijk} serán iguales a 1, esto es lo que representa la primera columna de la tabla anterior.

- Cuando uno de los elementos presenta el atributo y el otro no, se considera su similaridad (S_{ijk}) como 0, ya que son diferentes; y el coeficiente de ponderación (W_{ijk}) sería igual a uno, ya que basta con que sólo uno de los dos presente el atributo.

- La tercera columna nos indica que el individuo i no presenta el atributo pero el j sí, por lo que su similaridad es nula (cero); mientras que su coeficiente de ponderación sería igual a 1, ocurre lo mismo que en el caso anterior.

- En caso de que ninguno de los dos individuos posean el atributo, tanto la similaridad como el coeficiente de ponderación valdrían 0.

Ejemplo: supongamos que se tienen 3 individuos, de los cuales se tienen medidas las siguientes variables: peso, estatura, color de ojos, color de cabello, fumador o no fumador. Los datos se presentan a continuación:

INDV	PESO (LIBR) K=1	ESTATURA (PULG) K=2	COLOR DE OJOS K=3	COLOR DE CABELLO K=4	FUMA O NO FUMA K=5
1	120	66	NEGRO	NEGRO	FUMADOR
2	130	72	VERDE	CASTAÑO	FUMADOR
3	150	70	NEGRO	NEGRO	NO FUMADOR

En este caso se utilizaron la medida de disimilaridad para variables numéricas, definida por Gower (1971), de la siguiente manera:

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$$

donde: R_k toma el valor de 1 en las variables: color de ojos, color de cabello, fuma o no fuma.

$$S_{121} = 1 - \frac{|120 - 130|}{30} = 0.67, \; W_{121} = 1$$

$$S_{122} = 1 - \frac{|66 - 72|}{6} = 0, \; W_{122} = 1$$

$S_{123} = 0, \; W_{123} = 1$
$S_{124} = 0, \; W_{124} = 1$
$S_{125} = 1, \; W_{125} = 1$

$S_{12} = 0.334$

$S_{131} = 0 \quad W_{131} = 1$
$S_{132} = 0.333 \quad W_{132} = 1$
$S_{133} = 1 \quad W_{133} = 1$
$S_{134} = 1 \quad W_{134} = 1$
$S_{135} = 0 \quad W_{135} = 1$

$S_{13} = 0.466$

$S_{231} = 0.333 \quad W_{231} = 1$
$S_{232} = 0.667 \quad W_{232} = 1$
$S_{233} = 0 \quad W_{233} = 1$
$S_{234} = 0 \quad W_{234} = 1$
$S_{235} = 0 \quad W_{235} = 1$

$S_{23} = 0.200$

En este ejemplo la determinación del valor de W_{ijk} depende del criterio que haya asignado el investigador; y este valor será de 1 ó 0; el 1 representa la presencia del atributo y el 0 la ausencia de éste.

A continuación se presentan los criterios tomados para asignar los valores de W_{ijk}.

- *Peso*: es una variable numérica; por lo tanto siempre valdrá 1, ya que todos los individuos cuentan con el dato correspondiente a peso.

- *Estatura:* con esta variable ocurre exactamente lo mismo que con la variable anterior.

- *Color de ojos*: en lo que respecta a esta variable, se decidió que se considera como presencia del atributo si el individuo tiene los ojos negros y como ausencia si los ojos son de algún color diferente al negro.

- *Color de cabello*: en el caso de esta variable se consideró como presencia del atributo cuando el cabello del individuo es negro, y como ausencia de éste cuando el cabello es diferente al negro.

- *Fumador o no fumador*: en esta variable se consideró como la presencia del atributo si el individuo es fumador y como la ausencia si el individuo no fuma.

Tercera alternativa

El tercer método, es efectuar *análisis separados de objetos*, cada análisis se hace con respecto a un sólo tipo de variables y después tratar de resumir los resultados de los diferentes tipos. Este es un método alternativo que se recomienda utilizar si existen serios problemas para implementar alguna de las propuestas anteriores.

II.3 TÉCNICAS DE ANÁLISIS DE CONGLOMERADOS

Las técnicas de análisis de conglomerados pueden clasificarse en diversos tipos, los más usuales son los siguientes:

1) Técnicas jerárquicas, en estas técnicas los objetos son clasificados en grupos, el proceso será repetido a diferentes niveles para formar un tercer grupo.

2) Técnicas de optimización o partición; consisten en particionar a la población, en las cuales los grupos se forman por la optimización de un criterio de conglomerados. Las clases son exclusivas, así forman una partición del conjunto de entidades.

3) Técnicas de densidad; aquí los conglomerados se forman por la entrada a regiones contenidas en una relativa densidad de entidades.

Estos tipos no son necesariamente exclusivos, y varias técnicas de conglomerados podrían ser colocados en más de una categoría.

II.3.1 Técnicas jerárquicas

Métodos aglomerativos

El procedimiento básico en todos estos métodos es muy similar. Se calculan las medidas de similaridad o de distancias que existen entre los individuos de la matriz. El producto final de los métodos es un dendograma mostrando las fusiones de los individuos, los cuales terminan en un sólo grupo.

La diferencia entre los métodos se incrementa por las diferentes medidas de similaridad que se utilizan entre un individuo y un grupo de individuos, o entre dos grupos. Algunos de los

métodos son útiles cuando la matriz de distancias es usada como el punto inicial. Algunas técnicas aglomerativas se describirán a continuación.

a) Método del individuo más cercano o de la liga simple

Este método puede utilizarse tanto con medidas de similaridad como con medidas de distancia. El método consiste en fusionar en grupos a los individuos más cercanos, es decir los que presentan las distancias más pequeñas.

Por ejemplo, supongamos que existen 5 individuos que serán clasificados, y la matriz de distancias se denomina D_1, como se presenta a continuación:

$$D_1 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0.0 & 2.0 & 6.0 & 10.0 & 9.0 \\ 2 & & 0.0 & 5.0 & 9.0 & 8.0 \\ 3 & & & 0.0 & 4.0 & 5.0 \\ 4 & & & & 0.0 & 3.0 \\ 5 & & & & & 0.0 \end{array}$$

En esta matriz el elemento del i-ésimo renglón y la j-ésima columna dan la distancia, d_{ij}, entre los individuos i y j. Los individuos 1 y 2 son fusionados para formar un nuevo grupo, debido a que d_{12} es la entrada más pequeña en la matriz D_1. La distancia entre este grupo y los 3 individuos restantes 3, 4, y 5 se obtienen de la matriz D_1 como se muestra a continuación:

$$d_{(12)3} = \min\{d_{13}, d_{23}\} = \min\{6,5\} = d_{23} = 5.0$$

$$d_{(12)4} = \min\{d_{14} d_{24}\} = \min\{10,9\} = d_{24} = 9.0$$

$$d_{(12)5} = \min\{d_{15} d_{25}\} = \min\{9,8\} = d_{25} = 8.0$$

y podemos formar una nueva matriz D_2 dando las distancias de los inter-individuos, y las distancias grupos-individuos.

$$D_2 = \begin{array}{c|cccc} & (12) & 3 & 4 & 5 \\ \hline (12) & 0.0 & 5.0 & 9.0 & 8.0 \\ 3 & & 0.0 & 4.0 & 5.0 \\ 4 & & & 0.0 & 3.0 \\ 5 & & & & 0.0 \end{array}$$

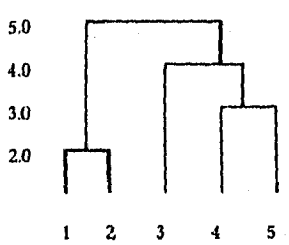
La entrada más pequeña en D_2 es d_{45} la cual es 3.0, y así los individuos 4 y 5 son fusionados en un segundo grupo, y sus distancias son las siguientes:

$$d_{(12)3} = 5.0$$
$$d_{(12)(45)} = \min(d_{14}, d_{15}, d_{24}, d_{25}) = d_{25} = 8.0$$
$$d_{3(45)} = \min(d_{34}, d_{35}) = d_{34} = 4.0$$

Con estos datos se forma la nueva matriz D_3, como se muestra a continuación:

$$D_3 = \begin{array}{c|ccc} & (12) & 3 & (45) \\ \hline (12) & 0.0 & 5.0 & 8.0 \\ 3 & & 0.0 & 4.0 \\ (45) & & & 0.0 \end{array}$$

Ahora la entrada más pequeña es $d_{(45)3}$ y así el individuo 3 es incluido en el grupo integrado por los individuos 4 y 5. Finalmente la fusión de los 2 grupos ha formado un sólo grupo conteniendo a los 5 individuos. El dendograma de la fusión se muestra a continuación:



b) El método del individuo más lejano o completa unión

Este método es exactamente lo opuesto al método de la liga simple, pues la distancia entre los grupos es definida como la distancia entre sus más remotas parejas de individuos. Usando esta técnica para la matriz de distancias D_1 de la sección anterior, iniciamos como el método de liga simple por fusionar a los individuos 1 y 2. La distancia entre este grupo y los restantes individuos 3, 4, y 5 son obtenidos de la matriz D_1 como se presenta a continuación:

$$d_{(12)3} = \max(d_{13}, d_{23}) = \max(6,5) = d_{13} = 6.0$$

$$d_{(12)4} = \max(d_{14}, d_{24}) = \max(10,9) = d_{14} = 10.0$$

$$d_{(12)5} = \max(d_{15}, d_{25}) = \max(9,8) = d_{15} = 9.0$$

generando una nueva matriz D'_2 como se muestra a continuación:

$$D'_1 = \begin{array}{c|cccc} & (12) & 3 & 4 & 5 \\ \hline (12) & 0.0 & 6.0 & 10.0 & 9.0 \\ 3 & & 0.0 & 4.0 & 5.0 \\ 4 & & & 0.0 & 3.0 \\ 5 & & & & 0.0 \end{array}$$

Observando que la entrada más pequeña en la matriz D'_2 es d_{45}, lo cual indica que estos individuos se deben fusionar en un nuevo grupo, realizando los cálculos de distancias obtenemos:

$$d_{(12)3} = \max\{ d_{13}, d_{23} \} = \max\{ 6, 5 \} = d_{13} = 6$$

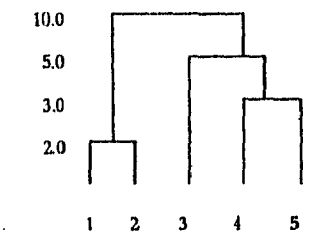
$$d_{(12)(45)} = \max\{ d_{14}, d_{15}, d_{24}, d_{25} \} = \max\{ 10, 9, 9, 8 \} = d_{14} = 10$$

$$d_{3(45)} = \max\{ d_{34}, d_{35} \} = \max\{ 4, 5 \} = d_{35} = 5$$

De los cálculos antes realizados se genera una nueva matriz, D'_3 como se muestra enseguida:

$$D'_3 = \begin{array}{c|ccc} & (12) & 3 & (45) \\ \hline (12) & 0.0 & 6.0 & 10.0 \\ 3 & & 0.0 & 5.0 \\ (45) & & & 0.0 \end{array}$$

La entrada más pequeña es de $d_{3(45)}$; por lo que se decide que el individuo 3 forme parte de un nuevo grupo con los individuos 4 y 5; quedando de esta manera 2 grupos, uno integrado por el individuo 1 y 2, y el otro por el 3, 4 y 5. Finalmente se fusionan todos los individuos en un sólo grupo, el resultado final se muestra en el siguiente dendograma:



c) Análisis del centroide

Este método fue propuesto originalmente por Sokal y Michener (1958), y por King (1966, 1967). Los grupos se encuentran en espacios euclidianos y son reemplazados por la información de las coordenadas de sus centros. La distancia entre los grupos es definida como la distancia entre los centros de los grupos. El procedimiento entonces es fusionar grupos de acuerdo a la distancia que existe entre sus centros, el grupo con la distancia más pequeña se fusiona primero. Para ejemplificar, presentamos los siguientes datos correspondientes a 5 individuos cada uno de ellos con dos variables, como se muestra enseguida:

		variable 1	2
individuo	1	1.0	1.0
	2	1.0	2.0
	3	6.0	3.0
	4	8.0	2.0
	5	8.0	0.0

Para estos datos se formó la matriz D_1 con sus distancias como se muestra a continuación:

$$D_1 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0.0 & 1.0 & 29.0 & 50.0 & 50.0 \\ 2 & & 0.0 & 26.0 & 49.0 & 53.0 \\ 3 & & & 0.0 & 5.0 & 13.0 \\ 4 & & & & 0.0 & 4.0 \\ 5 & & & & & 0.0 \end{array}$$

El procedimiento consiste en realizar primero una fusión de 2 individuos quienes son los más cercanos. La examinación de la matriz D_1 muestra que d_{12} es la entrada más pequeña y es así que los individuos 1 y 2 son fusionados para formar un grupo, y las coordenadas del centro del grupo son calculadas. Originando una nueva matriz de distancias D_2 la cual se calcula de la siguiente manera:

		variable 1	2
individuo	(12)	1.0	1.5
	3	6.0	3.0
	4	8.0	2.0
	5	8.0	0.0

	(12)	3	4	5
(12)	0.00	27.25	49.25	51.25
3		0.00	5.00	13.00
D_2= 4			0.00	4.00
5				0.00

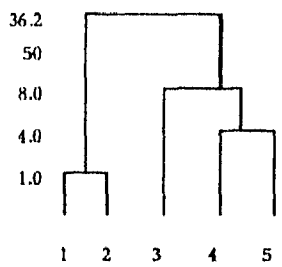
La entrada más pequeña en la matriz D_2 está entre los individuos 4 y 5, los que son fusionados en un 2º grupo, y los individuos reemplazados por las coordenadas del grupo de centros son los siguientes:

	variable 1	2
(12)	1.0	1.5
individuo 3	6.0	3.0
(45)	8.0	1.0

y la nueva matriz de distancias D_3 es la siguiente:

	(12)	3	(45)
(12)	0.00	27.25	49.25
D_3 = 3		0.00	8.00
(45)			0.00

En este caso la entrada más pequeña es la distancia entre el individuo 3 y el grupo de los individuos 4 y 5, y así este es fusionado para formar un grupo con tres miembros (3,4 y 5). El siguiente paso consiste en las fusiones de los 2 grupos restantes dentro de un sólo grupo, el dendograma de este ejemplo se muestra a continuación:



Este método como el anterior puede ser utilizado tanto con mediciones de similaridad como de distancia.

d) Método de la mediana

Una desventaja del método del centroide es que si el tamaño de los dos grupos que se fusionaran son muy diferentes, el centroide del nuevo grupo estará muy cercano al grupo más grande y permanecerá en ese grupo, las propiedades de las características del grupo pequeño virtualmente se pierden. Con la aplicación de este método la estrategia puede realizarse independientemente del tamaño del grupo, asumiendo que los grupos por fusionarse son de tamaños iguales, la posición aparente del nuevo grupo será siempre entre los dos grupos a ser fusionados. Más aun si representamos los centroides de los grupos por fusionar por (i) y (j), entonces la distancia del centroide a un tercer grupo (h) del grupo formado por la fusión (i), (j) se encuentra a lo largo de la mediana del triángulo definido por (i), (j) y (h), y es por esta razón que Gower (1969) que fue el primero en sugerir esta estrategia propuso el nombre de mediana y derivo sus propiedades del Teorema de Apollonius.

Aunque este método puede ser utilizado tanto para medidas de similaridad como de distancias, Lance y Williams (1967) sugirieron que debería considerarse como incompatibles para medidas tales como coeficientes de correlación ya que la interpretación en un sentido geométrico no es posible.

II.3.2 Técnicas de optimización o partición

Estas técnicas produce una partición de los objetos pero difiere de las técnicas jerárquicas por el hecho de que permite la reubicación de objetos, permitiendo de esta manera que particiones que inicialmente eran pobres sean corregidas en una etapa posterior.

La mayoría de estas técnicas hacen la partición optimizando un criterio predefinido. Por ejemplo, algunos intentan minimizar la traza de **W**, donde **W** es la matriz de sumas de cuadrados y productos cruzados dentro de los grupos. La mayoría de los métodos supone que el número de grupos ha sido decidido por el investigador[3]. Otros permiten que el número de grupos cambie durante el curso del análisis. La mayoría de estas técnicas emplean tres etapas en su ejecución, que son las siguientes:

a) Etapa de formación inicial de grupos
b) Etapa para asignar objetos a los grupos iniciales
c) Etapa para reasignar alguno, algunos o todos los objetos a otros grupos, una vez que el proceso de clasificación inicial se ha terminado

La diferencia entre los métodos, se encuentra entre a y c. En lo que sigue se describen estas etapas.

Para llevar a cabo la etapa de formación inicial de grupos, se supone que el número de grupos ya ha sido establecido por el investigador. La mayoría de las técnicas empiezan encontrando **K** puntos en el espacio de dimensión **P**, dichos puntos actúan como estimaciones del centro de los grupos. Por ejemplo Mac Queen (1967) selecciona los primeros **K** puntos muéstrales. Beal (1969) empieza con un valor de **K** más grande que el que se fijó de antemano y establece los centros de los grupos espaciados regularmente en intervalos de una desviación estándar

[3] Algunos criterios para esta técnica se plantean en Everitt (1974, Capítulo 3).

en cada variable. Después reduce el número de grupos hasta que un criterio basado en la suma residual de cuadrados se satisface. Thorndike (1954) escoge los K puntos mutuamente lejanos.

Una vez que se han seleccionado los K puntos iniciales, los objetos se asignan al grupo cuyo centro está más cercano (usualmente se utiliza la distancia euclidiana). En ocasiones la estimación del centro se actualiza después de que cada unidad ha entrado en el grupo o hasta que han entrado todas las entidades.

Una vez que la clasificación inicial se ha realizado, se buscan entidades que puedan reubicarse en otro grupo. La reubicación se lleva a cabo optimizando algún criterio de agrupación. En general la reubicación se efectúa si produce un incremento (o decremento en el caso de minimización) en el valor del criterio. El procedimiento continúa hasta que ya no se produce incremento con ninguno de los objetoss. Esta solución se puede aceptar o intentar mejorarla repitiendo el procedimiento usando una configuración inicial de grupos diferentes.

Ahora discutiremos los criterios para el agrupamiento, todos se derivan de la siguiente ecuación:

$$T = W + B$$

donde:

T es la matriz de dispersión total

W es la matriz de dispersión al interior de los grupos

$W = \sum_{i=1}^{g} W_i$ tal que W_i es la matriz de dispersión del grupo i

B es la matriz de dispersión al exterior de los grupos

g es el número de grupos

El primer criterio que discutiremos será la minimización de la traza de W. Esto es obvio ya que lo que se pretende es tener grupos homogéneos al interior de ellos, al minimizar la traza de W se maximiza la traza de B y lo que produce es que los grupos sean más heterogéneos entre si, lo cual es lo que se desea, es decir grupos lo más heterogéneos posibles entre si y al interior de ellos lo más homogéneos posibles.

Otro criterio que se ha considerado es la minimización del determinante de W, el cual es equivalente a la maximización del radio $|T| / |W|$. Otro criterio más es la maximización de la matriz (BW^{-1}). Ambos criterios, tanto la traza de (BW^{-1}) como $|T| / |W|$ se pueden expresar en términos de los valores característicos de la matriz BW^{-1}, es decir:

$$Tr\ (\ BW^{-1}\ ) = \sum_{i=1}^{p} \lambda_i$$

$$\frac{|T|}{|W|} = \prod_{i=1}^{p} (\ 1 + \lambda_i\ )$$

donde:

λ_i es el valor característico de la matriz (BW^{-1}).

p es el número de variables

II.3.3 Técnicas de densidad

Si los objetos se consideran como puntos de un espacio métrico, una manera para formar grupos sería considerar las partes del espacio en la cual los puntos son muy densos, separados por partes del espacio de baja densidad.

El método TAXMAP, compara las distancias relativas entre puntos, para buscar regiones del

espacio, densamente pobladas relativamente continuas, rodeadas por regiones vacías relativamente continuas. Los grupos se forman inicialmente de manera similar al método de la liga simple, pero se adaptan criterios para determinar cuando se debe de detener la asignación de unidades a los grupos. Un criterio tomado es terminar la asignación si el punto actualmente considerado es más favorable de alguna manera que el último punto admitido. El autor del método usa una medida que se obtiene con base en las similaridades promedio, la cual decrece ligeramente hasta que existe una discontinuidad. Un ejemplo puede aclarar más estos conceptos. Supóngase la siguiente matriz de similaridades entre 5 individuos.

$$S = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 1.0 & 0.7 & 0.9 & 0.4 & 0.3 \\ 2 & & 1.0 & 0.8 & 0.5 & 0.4 \\ 3 & & & 1.0 & 0.4 & 0.2 \\ 4 & & & & 1.0 & 0.7 \\ 5 & & & & & 1.0 \end{array}$$

Los dos individuos más parecidos se usan para iniciar el grupo. De la matriz S se obtiene que 0.9 es la similaridad más grande y corresponde a los individuos 1 y 3. El siguiente punto considerado para incluirse en el grupo es el más parecido a un punto que ya se encuentre en el grupo esto lleva a incluir al individuo 2, cuya similaridad con el individuo 3 es de 0.8. Ahora se calcula la similaridad promedio entre los tres individuos del grupo. Los resultados se presentan a continuación:

Miembros en el grupo:	1,3
Candidato a entrar al grupo:	2
Similaridad:	0.9
Similaridad promedio entre los individuos 1, 3 y 2:	1/3(0.9+0.7+0.8)=0.8

La diferencia en similaridad es (0.9-0.8)=0.1 y por lo tanto la medida de discontinuidad es (0.8-

0.1)=0.7. La similaridad con valores bajos indican que el candidato debe de incluirse al grupo, si se consideran los valores bajos aquéllos que son menores que 0.5, entonces el individuo 2 debe de agregarse al grupo, y considerar a otro individuo como candidato a incluirse. Es decir:

Miembros en el grupo:	1,3,2
Candidato a entrar al grupo:	4
Similaridad promedio entre los individuos 1, 3, 2 y 4:	1/6 [0.9 + 0.7 + 0.4 + 0.8 + 0.5 + 0.4] = 0.6

II.4 Número Optimo de Grupos

En el análisis de conglomerados el número de grupos que se deben formar es un problema que en algún momento se debe de resolver. Si la partición de los grupos se realiza tomando un criterio geográfico se puede pensar que la decisión es un poco ambigua, pero es importante dejar claro que el número de grupos que se pueden formar dado el conjunto de datos, no es arbitrario cuando se desea que sean homogéneos, depende necesariamente de la variabilidad interna de los posibles grupos o estratos que se formen. Si en los datos de interés no existe una estructura grupal, entonces no se deberá particionar a la población; es decir, no se trata de formar grupos simplemente por el hecho de formarlos, los datos deben de contener esa información.

Existen diversas técnicas para definir el número de grupos los más comunes son: el análisis de la variabilidad interna y el análisis del dendograma.

II.4.1 Variabilidad interna

Everitt (op. cit.) menciona que el análisis de la variabilidad interna es un posible criterio para determinar el número de grupos o regiones.

Este procedimiento consiste en formar con la distancia euclidiana diferentes grupos y medir la variabilidad interna de éstos. El número de grupos y la variabilidad interna se grafican para determinar el número óptimo de grupos.

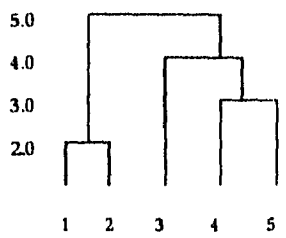
El número de grupos se determina de acuerdo a la gráfica resultante, en donde el número óptimo será aquél en el que el cambio en la variabilidad ya no sea significativa.

II.4.2 Análisis del dendograma

Otra metodología para determinar el número de grupos es mediante el dendograma generado por el análisis de conglomerados. En la técnica anterior se procede a incrementar el número de grupos, en esta se procede al revés. Esto, es primero se construyen el número de grupos igual al número de elementos que existan en la población después se calculan las distancias entre las variables de éstas y se agrupan los elementos más cercanos. Este procedimiento se continúa hasta quedar con un sólo grupo. La información del procedimiento se resume en un dendograma, el cual puede utilizarse también para determinar el número de regiones (Mardia op. cit. y Everitt op. cit.).

El dendograma es una gráfica en la cual el eje Y (vertical) representa a los objetos mientras el eje X (horizontal) representa distancias entre las distintas regiones o grupos. De tal manera que la longitud de las líneas horizontales representan la homogeneidad o similaridad interna de los grupos. Es decir, entre más pequeña sea la longitud de la línea horizontal significa que las entidades que forman los grupos están más cercanas entre sí y viceversa. En otras palabras, esto significa también que los grupos son más homogéneos. La longitud de la línea vertical indica los objetos que integran al grupo respectivo.

Para ejemplificar el uso del dendograma retomamos el ejemplo de la sección II.3.1 generado por el método del individuo más cercano o de la liga simple, el cual se presenta a continuación:



Este dendograma representa la union de los elementos con la utilización de la metodología elegida, primero se realizó la agrupación en un grupo con los individuos 1 y 2, en el siguiente paso se fusionaron los individuos 4 y 5 generando un nuevo grupo, quedando tres grupos el primero integrado con los individuos 1 y 2, el segundo con el 4 y 5 y el tercero con el individuo 3, en el siguiente paso se integro el individuo 3 al grupo de los individuos 4 y 5, por último se fusionaron los dos grupos y todos los individuos quedaron en un sólo grupo.

CONSTRUCCIÓN DE ESTRATOS

III.1 ESTRATIFICACIÓN Y ANÁLISIS DE CONGLOMERADOS

Antes de hablar de estratificación y análisis de conglomerados, es importante mencionar la gran diferencia que existe entre el muestreo por conglomerados y el análisis de conglomerados, ya que la semejanza que existe en los nombres, puede ocasionar confusiones, con lo que se entendería que los dos nombres se refieren a la misma técnica, siendo que son totalmente diferentes.

El muestreo por conglomerados, es una técnica de muestreo que se utiliza principalmente para reducir costos, ya que es uno de los métodos que proporciona mayor información minimizando su costo. Este método se efectúa realizando una división de la población total en grupos llamados conglomerados, los cuales son definidos de acuerdo a su localización geográfica y, que al reducir distancias se reducen costos de traslado del investigador. Al conglomerar, implícitamente se supone que al estar los elementos cercanos debería de existir una gran similaridad entre ellos, pero no siempre ocurre esto, ya que pueden estar muy cercanos pero ser completamente diferentes con respecto a una variable de interés para el investigador.

Algunas de las razones para emplear el método de muestreo por conglomerados son las siguientes:

- No se cuenta o es muy costoso obtener un buen marco que liste todos los elementos de la población, mientras que lograr un marco que liste todos los conglomerados es mucho más fácil.

- El costo para obtener observaciones, se incrementa con la distancia que separa los elementos de la población.

Por otro lado, tenemos que el análisis de conglomerados es una técnica estadística pero no de muestreo, que se emplea para definir y encontrar las similaridades que existen entre los elementos de una población, para de esa forma generar grupos cuyos elementos sean homogéneos al interior (muy parecidos) y heterogéneos al exterior del grupo (muy diferentes).

Es importante resaltar que la potencia del muestreo estratificado se basa en el hecho de que son homogéneos al interior de ellos y lo más heterogéneos posible entre ellos. Sin embargo, la mayoría de los autores de técnicas de muestreo nunca mencionan en sus obras cómo formar estratos homogéneos y heterogéneos. De tal manera que el uso de esta técnica se ve limitada mientras la formación de estratos no tenga la característica mencionada anteriormente.

Precisamente el objetivo de este trabajo, es mostrar al lector cómo la técnica de Análisis de Conglomerados puede ser una alternativa para la construcción de estratos con la característica de ser homogéneos al interior y heterogéneos al exterior.

Cabe resaltar también que, si la formación de estratos se realiza con base en la localización geográfica, entonces se está utilizando el procedimiento de la técnica de muestreo por conglomerados, ya que ese es el criterio utilizado en la formación de conglomerados, lo cual no satisface las propiedades requeridas en la estratificación, ya que se requiere que los elementos del estrato sean homogéneos al interior pero heterogéneos al exterior y, esta propiedad no está garantizada con este procedimiento dado que los elementos que se

encuentran cercanos geográficamente no necesariamente son parecidos.

III.2 Estratificación en Instituciones Responsables de la Política Demográfica en los Estados

La regionalización es un procedimiento que generalmente se ha utilizado para agrupar entidades federativas similares en algún sentido. La de uso más común en diferentes estudios ha sido la geográfica. Sin embargo, este tipo de regionalización puede no ser del todo adecuada, sobre todo cuando el fenómeno en estudio no está relacionado con las condiciones climatológicas y ambientales de la población. En este caso, la técnica de análisis de conglomerados puede ser una herramienta alternativa para la formación de dichos estratos o regiones.

El Consejo Nacional de Población (CONAPO) para coordinar, asesorar y capacitar a la Secretaría Técnica[1] de los consejos estatales de población[2] (COESPO's) realizó una regionalización de tipo geográfico. En el siguiente cuadro se presentan las entidades que conforman cada región.

[1] La Secretaría Técnica del Consejo Estatal de Población es la encargada de coordinar a las instituciones que forman parte de dicho consejo.

[2] El Consejo Estatal de Población está integrado por varias instituciones del gobierno estatal y es responsable de la política de población en el estado.

CUADRO III.1
ENTIDAD FEDERATIVA POR REGION GEOGRAFICA

REGION I	REGION II	REGION III
Baja California	Aguascalientes	Campeche
Baja California Sur	Colima	Chiapas
Coahuila	Distrito Federal	Guerrero
Chihuahua	Guanajuato	Michoacán
Durango	Hidalgo	Oaxaca
Sinaloa	Jalisco	Puebla
Sonora	Morelos	Quintana Roo
Nayarit	México	Tabasco
Nuevo León	Querétaro	Veracruz
Tamaulipas	San Luis Potosí	Yucatán
	Tlaxcala	
	Zacatecas	

Sin embargo, según los resultados de la Encuesta para la Detección de Oportunidades de Desarrollo de los COESPO's (EDODeC)[3], realizada en 1995 por el CONAPO, las necesidades de capacitación (Figura III.1), de presupuesto (Figura III.2) y de la estructura del personal de las secretarías técnicas (Figura III.3) son muy diferentes en cada uno de los COESPO's.

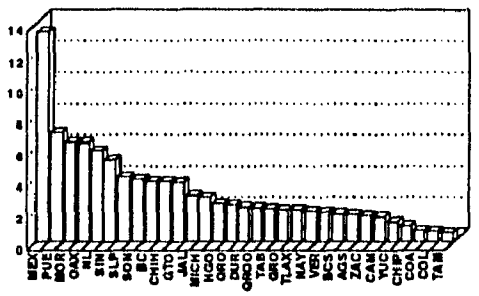
En la Figura III.1, se gráficaron los niveles de capacitación del personal de las secretarías técnicas, en donde se aprecian las diferencias entre las entidades federativas. El Estado de México es el que presenta el mayor nivel de capacitación de su personal, las entidades de Puebla, Morelos y Oaxaca también presentan altos niveles, los niveles medios pertenecen a los estados de Jalisco y Michoacán, mientras que Coahuila, Colima y Tamaulipas presentan niveles bajísimos.

En la Figura III.2 se observa que el presupuesto otorgado para el Estado de México sobresale entre las demás entidades, Nuevo León, Puebla y Oaxaca presentan presupuestos altos pero no son comparables con el de México (que es casi del doble), Tamaulipas, Chiapas, Colima,

[3] La EDODeC se llevó a cabo mediante un censo de las entidades federativas sin incluir al Distrito Federal. La información se captó vía un cuestionario de autoentrevista que se envió a cada una de las secretarías técnicas. Con los datos obtenidos en la encuesta se formaron los índices: nivel de estructura, nivel de capacitación y presupuesto.

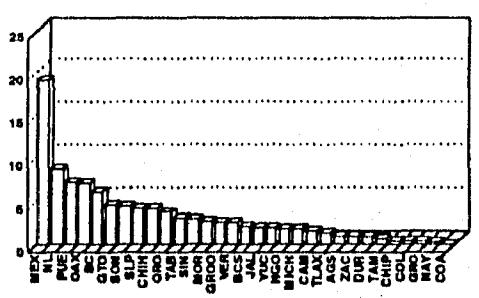
Guerrero, Nayarit y Coahuila presentan los niveles más bajos de apoyo presupuestal.

FIGURA III.1
NIVELES DE CAPACITACION DEL PERSONAL DE LAS SECRETARIAS TECNICAS POR ENTIDAD FEDERATIVA



FUENTE: Cuadro 1 del Anexo 1.

FIGURA III.2
NIVELES DE PRESUPUESTO DE LAS SECRETARIAS TECNICAS POR ENTIDAD FEDERATIVA

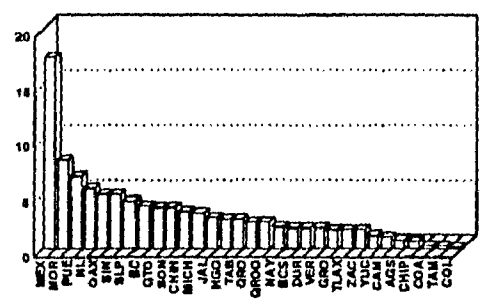


FUENTE: Cuadro 1 del Anexo 1.

En la Figura III.3 se observa que hay entidades que cuentan con altos niveles en cuanto a la estructura del personal de la Secretaría Técnica como es el caso otra vez del Estado de México

el cual es el que presenta el mayor nivel de todas las entidades federativas. Morelos y Puebla están por debajo del Estado de México y, en los niveles medios se encuentran Jalisco, Hidalgo y Tabasco, mientras que Coahuila, Tamaulipas y Colima presentan los más bajos niveles de estructura.

FIGURA II.3
NIVELES DE ESTRUCTURA DEL PERSONAL DE LAS SECRETARIAS TECNICAS POR ENTIDAD FEDERATIVA



FUENTE: Cuadro 1 del Anexo 1.

Cabe resaltar que la regionalización geográfica, agrupa entidades que no son similares con respecto a las variables de capacitación, presupuesto y estructura (Cuadro III.2). En el Cuadro, se puede observar que la Región I agrupa a Baja California Sur, Coahuila, Durango, Nayarit y Tamaulipas con otras entidades cuyos niveles son muy diferentes con respecto a los índices mencionados anteriormente. En la Región II resaltan por sus bajos niveles los estados de Aguascalientes, Colima, Tlaxcala y Zacatecas y por sus altísimos niveles en los tres índices el Estado de México. Finalmente, en la Región III contrastan de los otros estados por sus altos niveles Michoacán, Oaxaca y Puebla.

Con base en los datos anteriores se puede concluir entonces que, en este caso la regionalización geográfica no resulta muy adecuada y, dado que interiormente las regiones deben de estar formadas por entidades con niveles de estructura, capacitación y presupuesto

muy similares se aplicó el análisis de conglomerados. El primer criterio para determinar la similitud de los índices fue definir una distancia entre entidades federativas. De tal manera que cuando la distancia entre dos entidades federativas resulta "pequeña", se considera que son similares y que por lo tanto deben pertenecer a la misma región o estrato.

Ahora bien, dado que dos entidades pueden estar cercanas (ser similares) con respecto a un índice pero al mismo tiempo estar alejadas (no ser similares) con respecto a otro, fue necesario utilizar la distancia euclidiana, en primer lugar, porque considera a todas las variables al mismo tiempo y en segundo, porque las variables (índices) son de tipo numérico. Matemáticamente se define como:

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

donde:

d_{ij} representa la distancia existente entre la entidad i-ésima y la j-ésima

p es el número de características o índices

x_{jk} es el valor de la variable (índice) k-ésima en la entidad j-ésima

La distancia definida anteriormente, tiene la característica de que si dos entidades federativas están cercanas entre si, entonces estarán cercanas simultáneamente con respecto a todas las variables medidas en esta entidad. Es decir, serán similares tanto en el nivel de estructura, capacitación y presupuesto.

En este caso las entidades federativas representan las unidades de estudio. Las variables que se utilizaron para construir las regiones son las siguientes:

1) Nivel de Estructura
2) Nivel de Capacitación
3) Presupuesto

CUADRO III.2
INDICES DE LA SECRETARIA TECNICA POR ENTIDAD FEDERATIVA, SEGUN REGION GEOGRAFICA, 1995

ENTIDAD FEDERATIVA	INDICES		
	DE ESTRUCTURA	DE CAPACITACION	DE PRESUPUESTO
REGION I	**2.78**	**3.01**	**2.23**
Baja California	3.89	3.87	6.09
Baja California Sur	1.88	1.78	2.02
Coahuila	0.27	0.75	0.10
Chihuahua	3.35	3.86	4.14
Durango	1.88	2.21	0.87
Sinaloa	4.96	5.24	2.98
Sonora	3.75	4.00	4.44
Nayarit	2.01	1.97	0.28
Nuevo León	5.50	5.84	8.72
Tamaullpas	0.27	0.58	0.65
REGION II	**4.17**	**3.79**	**3.81**
Aguascalientes	0.80	1.76	0.98
Colima	0.13	0.67	0.46
Guanajuato	3.75	3.81	4.57
Hidalgo	2.68	2.45	1.80
Jalisco	2.82	2.95	1.99
México	17.43	13.43	19.26
Morelos	8.04	6.40	2.66
Querétaro	2.41	2.38	3.75
San Luis Potosí	4.29	4.12	4.22
Tlaxcala	1.74	2.01	1.25
Zacatecas	1.74	1.67	0.98
REGION III	**2.64**	**2.83**	**2.78**
Campeche	1.07	1.59	1.90
Chiapas	0.67	1.01	0.46
Guerrero	1.74	2.05	0.43
Michoacán	3.22	2.86	1.78
Oaxaca	4.96	6.35	7.12
Puebla	6.57	7.01	7.19
Quintana Roo	2.41	2.14	2.57
Tabasco	2.68	2.11	3.02
Veracruz	1.88	1.88	2.55
Yucatán	1.21	1.26	0.76

FUENTE: Cálculos con base en la EDODeC.

Número de Regiones o Estratos

Un problema común en cualquier tipo de regionalización es el de determinar el número óptimo de regiones en que se debe particionar, en este caso a la República Mexicana o, a la población total de estudio. En general este es un problema al que se enfrenta todo investigador cuando desea particionar una población en subpoblaciones.

Desde el punto de vista geográfico se puede pensar que la decisión del número de regiones es un poco ambigua, ya que es determinado por el investigador, mientras que desde el punto de vista del análisis de conglomerados el criterio se basa en medir la variabilidad interna propia de los datos de interés de cada grupo o estrato. El criterio básico es formar regiones homogéneas al interior de ellas y lo más heterogéneas posible entre ellas. Con este criterio el número de grupos que se pueden formar, ya no es arbitrario, depende necesariamente de la variabilidad interna de las posibles regiones que se formen.

Cabe resaltar que si en los datos de interés no existe una estructura grupal, entonces no se deberá particionar a la población. En otras palabras, dado un conjunto de datos de una población, no se trata de formar grupos simplemente por el hecho de formarlos, los datos deben de contener la información al respecto.

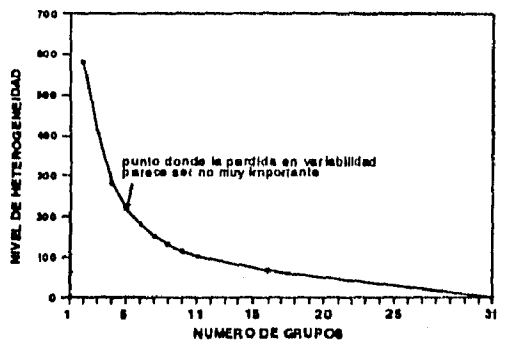
En la Figura III.4 se gráfico el número de regiones y la variabilidad interna o medida de homogeneidad. Se puede observar que:

- Realmente existe una estructura grupal en los datos puesto que la relación observada no es constante.
- A medida que se incrementa el número de regiones la variabilidad al interior de éstas disminuye, indicando que las regiones son más homogéneas al interior.

- Cuando el número de grupos es 31 la variabilidad interna es cero, ya que en este caso cada entidad es comparada con ella misma.

- Dado que el objetivo es formar regiones que tengan la menor variabilidad internamente (homogéneas), entonces se puede concluir que 5 regiones pueden ser las adecuadas, ya que en la figura se observa que la ganancia en la disminución de la variabilidad interna no es muy importante si se definen más de 5 regiones.

FIGURA III.4
NIVEL DE VARIABILIDAD INTERNA POR NUMERO DE GRUPOS



FUENTE: Cuadro 2 del Anexo 1.

Formación de Regiones

Una vez determinado el número de grupos o regiones, la siguiente etapa es la formación de los mismos. El método utilizado para realizar la formación de los grupos es el del centroide, en este método los grupos se encuentran en espacios euclidianos, la distancia entre los grupos es definida como la distancia entre los centros de los grupos. El procedimiento es fusionar grupos de acuerdo a la distancia que existe entre sus centros, las entidades u objetos con la distancia más pequeña se fusiona primero (Ver la sección II.3.1 del Capítulo II). Los cálculos se hicieron con el paquete estadístico Statiscal Package for Social Sciences (SPSS) versión 4.1

y los grupos que se formaron se presentan en el Cuadro III.3.

CUADRO III.3
INDICES DE LA SECRETARIA TECNICA POR ENTIDAD FEDERATIVA, SEGUN REGION, 1995

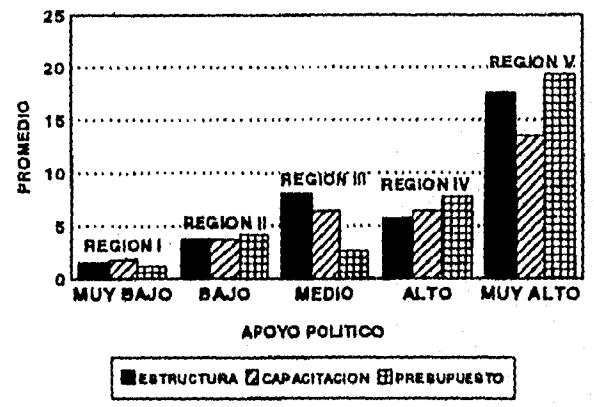
ENTIDAD FEDERATIVA	INDICES		
	DE ESTRUCTURA	DE CAPACITACION	DE PRESUPUESTO
REGION I	**1.58**	**1.76**	**1.21**
Aguascalientes	0.80	1.76	0.98
Baja California Sur	1.88	1.78	2.02
Campeche	1.07	1.59	1.90
Coahuila	0.27	0.75	0.10
Colima	0.13	0.67	0.46
Chiapas	0.67	1.01	0.46
Durango	1.88	2.21	0.87
Guerrero	1.74	2.05	0.43
Hidalgo	2.68	2.45	1.80
Jalisco	2.82	2.95	1.99
Michoacán	3.22	2.86	1.78
Nayarit	2.01	1.97	0.28
Querétaro	2.41	2.38	3.75
Quintana Roo	2.41	2.14	2.57
Tabasco	2.68	2.11	3.02
Tamaulipas	0.27	0.58	0.65
Tlaxcala	1.74	2.01	1.25
Veracruz	1.88	1.88	2.55
Yucatán	1.21	1.26	0.76
Zacatecas	1.74	1.67	0.98
REGION II	**3.64**	**3.67**	**4.15**
Baja California	3.89	3.87	6.09
Chihuahua	3.35	3.86	4.14
Guanajuato	3.75	3.81	4.57
San Luis Potosí	4.29	4.12	4.22
Sinaloa	4.96	5.24	2.98
Sonora	3.75	4.00	4.44
REGION III	**8.04**	**6.40**	**2.66**
Morelos	8.04	6.40	2.66
REGION IV	**5.68**	**6.40**	**7.68**
Nuevo León	5.50	5.84	8.72
Oaxaca	4.96	6.35	7.12
Puebla	6.57	7.01	7.19
REGION V	**17.43**	**13.43**	**19.26**
México	17.43	13.43	19.26

FUENTE: Cálculos con base en la EDODeC.

Identificación de Regiones

La última etapa en la estratificación o regionalización es identificar las características de las regiones, para esto se analizó el nivel de los tres índices y teniendo en mente que los niveles de estructura, capacitación y presupuesto, dependen del apoyo que otorgue el gobernador a la Secretaría Técnica se definió lo siguiente: como la Región I es la que tiene los niveles más bajos, se definió como la región con *Muy Bajo Apoyo Político*, mientras que la Región V por tener los niveles más altos, se definió como la región con *Muy Alto Apoyo Político*. La clasificación de las restantes regiones se puede observar en el Figura III.5.

FIGURA III.5
PROMEDIO DE INDICES DE LA SECRETARIA TECNICA POR REGION Y APOYO POLITICO



FUENTE: Cuadro III.3.

III.3 Estratificación y Mortalidad

La mortalidad es uno de los componentes fundamentales y determinantes del tamaño y de la composición por sexo y edad de la población.

La explicación del proceso de extinción de una generación a través de la edad concierne a la demografía, la medicina y la salud publica. El interés del demógrafo es conocer la forma en que las características físicas o biológicas, la organización social y el medio ambiente se relacionan con la mortalidad. Uno de los factores que influye en la mortalidad de las personas es la forma de vida, tales como la ocupación, los ingresos, los hábitos alimenticios y el tipo de comunidad en que vive.

En México, numerosos estudios han sido consagrados a la mortalidad, lo que ha permitido concluir que los niveles de mortalidad muestran diferencias en las regiones, estados, municipios, etc. En el cuadro III.4 se presentan los estados regionalizados de igual manera que el ejemplo anterior. Se pueden apreciar las grandes diferencias que existen en las entidades, por ejemplo Baja California Norte y Sur aunque se encuentran cercanas presentan tasas diferentes y pertenecen a la misma región; asimismo en la Región II se observa una gran diferencia entre la tasa de mortalidad del Distrito Federal y Aguascalientes, ya que existe una diferencia de aproximadamente 100 puntos; de igual forma ocurre en la Región III.

CUADRO III.4
TASAS DE MORTALIDAD GENERAL POR REGION GEOGRAFICA, 1993

ENTIDAD FEDERATIVA	TASA DE MORTALIDAD	ENTIDAD FEDERATIVA	TASA DE MORTALIDAD
REGION I			
Baja California	462.1	Nayarit	420.3
Baja California Sur	392.4	Nuevo León	416.2
Coahuila	450.4	Sinaloa	360.6
Chihuahua	515.2	Sonora	494.7
Durango	417.3	Tamaulipas	438.7
REGION II			
Aguascallentes	436.1	México	450.6
Colima	510.6	Morelos	463.7
Distrito Federal	534.7	Querétaro	465.0
Guanajuato	490.5	San Luis Potosí	464.6
Hidalgo	479.0	Tlaxcala	518.3
Jalisco	526.1	Zacatecas	462.2
REGION III			
Campeche	427.1	Puebla	571.5
Chiapas	421.5	Quintana Roo	317.5
Guerrero	348.8	Tabasco	396.6
Michoacán	472.1	Veracruz	430.8
Oaxaca	552.0	Yucatán	509.8

FUENTE: Secretaría de Coordinación y Desarrollo, Dirección General de Estadística, Informática y Evaluación. Mortalidad, 1993.
NOTA: Tasas por 10,000 habitantes.

De tal manera que cualquier investigación de la mortalidad general a nivel nacional, debe tener en cuenta estas diferencias y, por lo tanto resulta interesante la construcción de estratos con base en datos de mortalidad.

Es por tal razón que se consideró pertinente realizar el estudio de la mortalidad a nivel estatal

y generar grupos que se parezcan en los niveles de mortalidad, mediante la técnica de Análisis de Conglomerados.

El caso que se ejemplifica es el estudio de 10 principales causas de muerte en las 32 entidades federativas, en el cual no se considera pertinente realizar la regionalización de acuerdo a la localización geográfica de los estados, debido también a que diferentes entidades pueden presentar causas de muerte completamente distintas y estar geográficamente cercanas.

Nuevamente, la finalidad de esta sección es agrupar entidades que sean homogéneas en sus niveles de mortalidad. De tal manera que las tasas de mortalidad entre entidades de una misma región sean lo más similar posible, pero al mismo tiempo que entidades de distintas regiones difieran tanto como sea posible en sus tasas de mortalidad.

Se consideraron como unidades de estudio a las 32 entidades federativas. Los datos se refieren a las siguientes tasas de 10 principales causas de muerte:

1) Enfermedades del corazón
2) Tumores malignos
3) Accidentes
4) Diabetes mellitus
5) Enfermedades cerebrovasculares
6) Ciertas afecciones originadas en el período perinatal
7) Cirrosis y otras enfermedades crónicas del hígado
8) Neumonía e influenza
9) Homicidio y lesiones inflingidas intencionalmente por otra persona
10) Enfermedades infecciosas intestinales

Inicialmente se definió un criterio para determinar la similitud de las tasas entre entidades federativas. Para esto se utilizo nuevamente una distancia que considera a todas las variables

al mismo tiempo y que se utiliza solamente en el caso de variables numéricas. Dicha distancia es la euclidiana y se define de la siguiente manera:

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

donde:

d_{ij} representa la distancia existente entre la entidad i-ésima y la j-ésima

p es el número de causas de muerte, en este caso p valdría 10

x_{jk} es el valor de la variable (tasa de mortalidad) k-ésima en la entidad j-ésima

Esta distancia tiene la característica de que si dos entidades federativas están cercanas entre sí, entonces estarán cercanas simultáneamente con respecto a todas las variables medidas en esta entidad. Es decir, serán similares en las 10 principales tasas de mortalidad analizadas en este ejemplo (Ver Cuadro 3 del Anexo 1).

Número de Regiones o Estratos

El criterio para determinar el número de regiones fue nuevamente con base en la variabilidad interna de los datos de interés de cada entidad.

En la Figura III.6 se gráfico el número de regiones y la variabilidad interna o medida de homogeneidad. Donde se puede observar que:

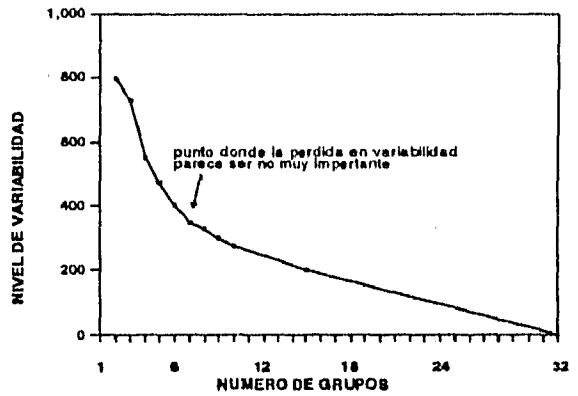
- Realmente existe una estructura grupal en los datos de mortalidad puesto que la relación observada no es constante.

- A medida que se incrementa el número de regiones la variabilidad al interior de éstas

disminuye, indicando que las regiones son más homogéneas al interior de ellas.

- Dado que el objetivo es formar regiones que tengan la menor variabilidad internamente (homogéneas), entonces se puede concluir que 7 regiones pueden ser las adecuadas, ya que en la figura se observa que la ganancia en la disminución de la variabilidad interna no es muy importante si se definen más de 7 regiones.

FIGURA III.6
NIVEL DE VARIABILIDAD INTERNA POR NUMERO DE GRUPOS



FUENTE: Cuadro 4 del Anexo 1.

Formación de Regiones

Una vez determinado el número de estratos o regiones, la siguiente etapa fue la formación de los mismos. El método utilizado para realizar la formación de los grupos es el del individuo más lejano o completa unión, que es el opuesto al de la liga simple donde la distancia entre los grupos es definida como la distancia entre sus remotas parejas de individuos (Ver la sección II.3.1 del Capítulo II). Los cálculos se hicieron con el paquete estadístico Statiscal Package for Social Sciences (SPSS) versión 4.1 (Ver Cuadro III.5).

CUADRO III.5
TASAS DE MORTALIDAD GENERAL POR REGION GEOGRAFICA, 1993

ENTIDAD FEDERATIVA	MEDIA REGIONAL	ENTIDAD FEDERATIVA	MEDIA REGIONAL
REGION I	**21.8**	**REGION IV**	**32.1**
Quintana Roo		Hidalgo México Puebla Querétaro Tlaxcala	
REGION II	**23.7**		
Guerrero			
REGION III	**29.4**		
Aguascalientes Baja California Sur Campeche Durango Guanajuato Michoacán Morelos Nayarit San Luis Potosí Sinaloa Tabasco Veracruz Zacatecas		**REGION V**	**32.2**
		Chiapas Oaxaca	
		REGION VI	**33.6**
		Coahuila Colima Distrito Federal Jalisco Nuevo León Yucatán	
		REGION VII	**34.8**
		Baja California Chihualtua Sonora Tamaulipas	

FUENTE: Cuadro 5 del Anexo 1.

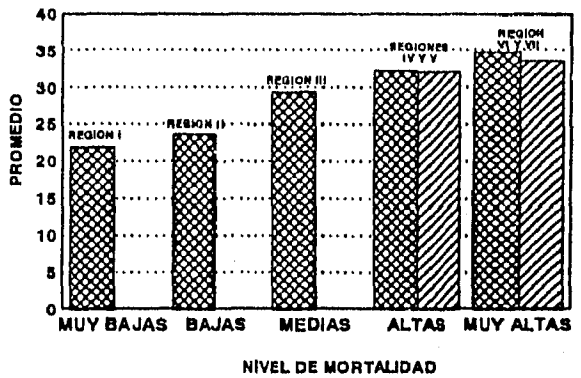
Identificación de Regiones

La última etapa fue la identificación de las características de las regiones, para lo cual se analizó la media total, determinando con ella el nivel de mortalidad para cada grupo.

Con base en el Cuadro III.5 se determinó que la Región I es la que presenta *Muy Baja Mortalidad;* la Región II se categóriza con *Baja Mortalidad;* la Región III concentra a las

entidades con *Media Mortalidad;* las regiones IV y V con *Alta Mortalidad* aunque estas regiones son distintas fueron clasificadas en la misma categoría; y por último las regiones VI y VII se consideran como las regiones que presentan *Muy Alta Mortalidad.* La clasificación de las regiones se puede observar en la Figura III.7.

FIGURA III.7
PROMEDIO DE TASAS DE MORTALIDAD GENERAL POR REGION SEGUN NIVEL DE MORTALIDAD, 1993



FUENTE: Cuadro III.8.

III.4 Estratificación y Marginación

"Los efectos sociales de la crisis económica de los años ochenta han convertido a la marginación social en uno de los grandes problemas nacionales. Con ello, la recuperación del crecimiento económico en la presente década tiene, entre sus propósitos más acuciantes, combatir las desigualdades sociales así como reducir en el corto plazo sus implicaciones socioeconómicas y demográfico-espaciales" (CONAPO-CONAGUA, 1990).

Las nuevas características de la economía mundial y la implementación nacional de un nuevo

modelo de crecimiento, fruto de la reforma económica en marcha, están obligando a encarar los problemas de la desigualdad y la marginación social en nuevos términos. En particular, ha ganado consenso la idea que el estado no puede ser el único actor en el combate a la marginación, y que ésta no disminuirá hasta que se consolide el proceso de crecimiento y se impulse el potencial productivo que la nación tiene en la propia población marginada.

En 1990, el CONAPO construyó un índice de marginación para cada municipio con el fin de asignar a cada municipio su grado de marginación. "El índice es una medida que valora dimensiones estructurales de la marginación social en México; identifica nueve de sus formas y mide su intensidad espacial como porcentaje de la población total no participante del disfrute de bienes y servicios accesibles a los ciudadanos marginados, cuyas cantidades y calidades se consideran mínimos de bienestar en atención al nivel de desarrollo alcanzado por el país. Por consiguiente, el índice permite un análisis integrado y comparativo del impacto global que las carencias tienen en cada uno de los municipios, los cuales son agrupados por grados de intensidad" (CONAPO-CONAGUA, 1990).

Para la construcción del Indice de Marginación, la fuente de información fue el XI Censo General de Población y Vivienda de 1990. Para el cálculo de dicho índice se incluyeron 4 dimensiones: vivienda, ingresos monetarios, educación y distribución de la población y, las variables analizadas fueron:

1) Porcentaje de población analfabeta
2) Porcentaje de población de 15 años y más sin primaria completa
3) Porcentaje de ocupantes en vivienda particular sin disponibilidad de drenaje ni excusado
4) Porcentaje de ocupantes en vivienda particular sin disponibilidad de energía eléctrica
5) Porcentaje de ocupantes en vivienda particular sin disponibilidad de agua entubada
6) Porcentaje de viviendas particulares con algún nivel de hacinamiento
7) Porcentaje de ocupantes en vivienda particular con piso de tierra
8) Porcentaje de población en localidades de menos de 5,000 habitantes

9) Porcentaje de población ocupada que gana hasta dos salarios mínimos

Cabe resaltar que, otra manera de abordar la problemática de la desigualdad social es mediante la construcción de grupos de municipios que se parezcan en los valores de las variables mencionadas anteriormente y que entre los grupos las diferencias sean lo más grande posible.

En esta sección se realizó la construcción de conglomerados para el Estado de Chiapas a nivel municipal con base en las nueve variables mencionadas anteriormente, mismas que fueron empleadas para la construcción de los índices de marginación. Los datos para los municipios de Chiapas se presentan en el Cuadro 6 del Anexo 1.

Inicialmente se definió un criterio para determinar la similitud de los porcentajes entre los municipios. Para esto también se empleo una distancia que considera a todas las variables al mismo tiempo, esta distancia es la euclidiana y se define de la siguiente manera:

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

donde:

d_{ij} representa la distancia existente entre el municipio i-ésimo y el j-ésimo
p es el número de características o variables, en este caso p valdría 9
x_{ik} es el valor de la variable (porcentaje) k-ésima en el municipio i-ésimo
x_{jk} es el valor de la variable (porcentaje) k-ésima en el municipio j-ésimo

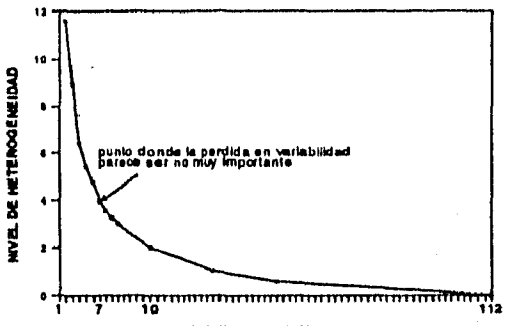
Esta distancia tiene la característica de que sí dos municipios están cercanos entre sí entonces estarán cercanas simultáneamente con respecto a todas las variables medidas en este municipio. Es decir, serán similares en los 9 porcentajes analizados en este ejemplo (Ver

Cuadro 6 del Anexo 1).

Número de Regiones o Estratos

Para determinar el número de regiones o estratos, se hizo un análisis de la homogeneidad al interior de los grupos. La medida utilizada representa el grado en que los grupos son diferentes en las variables socio-económicas mencionadas anteriormente. El número máximo de regiones que se pueden formar es 112, en cuyo caso cada municipio se compara con el mismo y por lo tanto el grado de diferencia en las variables es cero. Esto se interpreta diciendo que un municipio no puede ser diferente a el mismo y en cuyo caso se tiene el grado máximo de homogeneidad al interior de los grupos.

FIGURA III.8
NIVEL DE VARIABILIDAD INTERNA POR NUMERO DE GRUPOS



FUENTE: Cuadro 7 del Anexo 1.

En la Figura III.8 se gráfico el número de regiones y la variabilidad interna o medida de homogeneidad. Donde se puede observar:

- Una clara estructura grupal en los datos puesto que la relación observada no es constante.

- A medida que se incrementa el número de regiones la variabilidad al interior de éstas disminuye, indicando que las regiones son más homogéneas al interior.

- Existe un descenso muy importante en el nivel de homogeneidad entre dos y diez regiones.

- Con lo cual se puede concluir que 7 regiones pueden ser las adecuadas, ya que en la figura se observa que la ganancia en la disminución de la variabilidad interna no es muy importante si se definen más de 7 regiones.

Formación e Identificación de Regiones

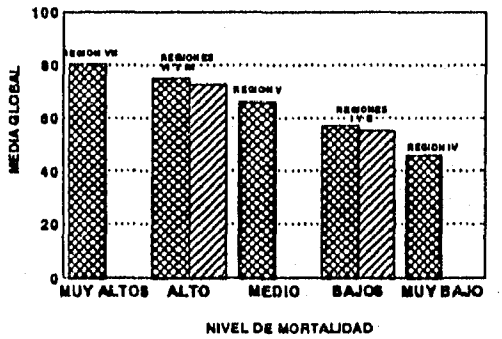
La etapa posterior a la determinación del número de regiones fue la formación de las mismas. Para realizar la formación de los grupos se utilizó el método de la mediana (Ver sección II.3.1). Los cálculos se hicieron con el paquete estadístico Statiscal Package for Social Sciences (SPSS) versión 4.1 y los grupos que se formaron se presentan en el Cuadro III.6

La última etapa en la estratificación o regionalización es identificar las características al interior de cada región, para esto se calculó la media para cada variable de los grupos y posteriormente una media global, con lo que se asignaron las siguientes categorías: la Región VII con *Muy Alto Grado de Marginalidad*, las regiones VI y III con *Alto Grado de Marginalidad*, la Región V con *Medio Grado de Marginalidad*, las regiones I y II con *Bajo Grado de Marginalidad* y las Región IV con *Muy Bajo Grado de Marginalidad*. La conformación de las regiones se presenta en la Figura III.9.

CUADRO III.6
MEDIA GLOBAL POR REGION

REGIONES	MEDIA GLOBAL
REGIO I	57.08
REGION II	55.05
REGION III	72.57
REGION IV	45.81
REGION V	65.74
REGION VI	74.79
REGION VII	80.32

FIGURA III.7
PROMEDIO DE TASAS DE MORTALIDAD GENERAL POR REGION



FUENTE: Cuadro III.6.

Conclusiones

Conclusiones Generales

A pesar de que en este trabajo no se hizo una aplicación de todas las técnicas de análisis de conglomerados utilizando los diferentes tipos de variables, se puede concluir que al menos metodológicamente el trabajo contiene los elementos necesarios para hacer la aplicación en un momento dado.

Cuando las variables son de tipo numérico el trabajo plantea un acercamiento a la solución del problema relacionado con el número de grupos en que se debe particionar a la población, así como la formación de grupos homogenéos y heterogéneos. Sin embargo debe trabajarse más sobre todo aplicando diferentes métodos de análisis de conglomerados en un mismo estudio para medir la similaridad entre los elementos de la población, ya que aquí soloamente se aplicó uno solo para generar los grupos.

Conclusiones Específicas

En los tres casos analizados se detectó una clara estructura grupal. Esto quiere decir que no sería del todo adecuado llevar a cabo investigaciones que no tomaran en cuenta las diferencias en aspectos relacionados con la mortalidad, marginalidad, así como para asesorar a las secretarías técnicas de los Consejos Estatales de Población.

La manera en cómo se determina el número de grupos se basa en el hecho de que la idea es trabajar con el menor número de grupos y, aunque el criterio se basa en el punto donde la medida de variabilidad no cambia de manera importante, la idea es no incrementar el número de grupos aunque todavía la medida de variabilidad sea más pequeña. Sin embargo, debe quedar claro que habrá que trabajar más este aspecto.

Los resultados del estudio de la estratificación en instituciones responsables de la política demográfica en los estados se pueden utilizar para establecer diferentes estrategias que permitan coordinar, asesorar y capacitar a los consejos de acuerdo a sus necesidades. Por ejemplo para las entidades que conforman la Región I, una opción es hacer un estudio detallado con una sola secretaría técnica ya que todas las demás serán muy similares.

En el caso de las regiones V y III que se integran por una sola entidad la estrategia aquí sería analizar las causas por las que cuentan con una buena estructura y capacitación con el fin de implementarlas de manera prioritaria en aquellas entidades de la Región I.

Cabe resaltar que los consejos de los estados de Nuevo León, Oaxaca y México que son los que respectivamente se clasificaron como alto y muy alto apoyo político, son además los que mejor desempeñan sus funciones, de acuerdo a lo observado en la coordinación que existe entre el Consejo Nacional de Población y estos consejos.

Los resultados de la estratificación de estados por causas de mortalidad se pueden utilizar seleccionando en cualquiera de las regiones a una entidad y entonces hacer estudios más específicos en ésta pues los resultados que se obtengan deben ser muy similares a las otras entidades de la misma región.

En el caso de la estratificación utilizando información de marginalidad, los resultados pueden usarse en cualquier investigación que este relacionada con empleo, educación y programas de salud. La idea sería seleccionar en cada estrato el municipio que tuviera la media más grande, el que tuviera la más pequeña y el que estuviera en un termino medio.

Finalmente se puede concluir que los estratos generados con la aplicación del análisis de conglomerados son diferentes a los que se forman mediante criterios geográficos esto se debe básicamente a que en los tres casos analizados las variables de análisis no están relacionadas con aspectos ambientales, es decir la mortalidad es baja o alta por otras causas y no por el lugar geográfico en que se localiza el estado. Algo análogo sucede con la marginalidad y las cuestiones que dependen de decisiones de tipo político.

ANEXO 1

CUADRO 1

INDICES DE LA SECRETARIA TECNICA POR ENTIDAD FEDERATIVA

ENTIDAD FEDERATIVA	INDICES		
	DE ESTRUCTURA	DE CAPACITACION	DE PRESUPUESTO
Aguascalientes	0.80	1.76	0.98
Baja California	3.89	3.87	6.09
Baja California Sur	1.88	1.78	2.02
Campeche	1.07	1.59	1.90
Coahuila	0.27	0.75	0.10
Colima	0.13	0.67	0.46
Chiapas	0.67	1.01	0.46
Chihuahua	3.35	3.86	4.14
Durango	1.88	2.21	0.87
Guanajuato	3.75	3.81	4.57
Guerrero	1.74	2.05	0.43
Hidalgo	2.68	2.45	1.80
Jalisco	2.82	2.95	1.99
México	17.43	13.43	19.26
Michoacán	3.22	2.86	1.78
Morelos	8.04	6.40	2.66
Nayarit	2.01	1.97	0.28
Nuevo León	5.50	5.84	8.72
Oaxaca	4.96	6.35	7.12
Puebla	6.57	7.01	7.19
Querétaro	2.41	2.38	3.75
Quintana Roo	2.41	2.14	2.57
San Luis Potosí	4.29	4.12	4.22
Sinaloa	4.96	5.24	2.98
Sonora	3.75	4.00	4.44
Tabasco	2.68	2.11	3.02
Tamaulipas	0.27	0.58	0.65
Tlaxcala	1.74	2.01	1.25
Veracruz	1.88	1.88	2.55
Yucatán	1.21	1.26	0.76
Zacatecas	1.74	1.67	0.98

FUENTE: Cálculos con base en la EDODeC, 1995.

CUADRO 2
VARIABILIDAD INTERNA,
POR NUMERO DE GRUPOS

NUMERO DE GRUPOS	VARIABILIDAD
2	581.74
3	408.45
4	279.69
5	218.67
6	180.04
7	150.44
8	130.23
9	112.96
10	101.61
15	66.13
31	0.00

CUADRO 3
TASAS DE MORTALIDAD GENERAL POR CAUSA Y ENTIDAD FEDERATIVA
1993

Entidad	Enfermedades del Corazón	Tumores Malignos	Accidentes	Diabetes Mellitus	Enfermedades Cerebrovas-culares	Ciertas Afecciones de Origen Perinatales	Cirrosis y Otras Enfermedades Crónicas del Hígado	Neumonía e Influenza	Homicidio y Lesiones Infingidas Intencionalmente por Otra Persona	Enfermedades Infecciosas Intestinales
Aguascalientes	56.70	48.50	45.30	37.00	25.10	33.10	13.50	12.30	4.30	10.30
Baja California	74.40	56.00	55.20	35.00	24.10	23.90	19.10	15.90	13.60	5.80
Baja California Sur	60.00	64.60	44.40	32.20	22.80	23.00	10.20	14.20	6.00	7.40
Campeche	43.30	42.80	41.70	26.00	25.40	23.50	23.70	11.10	13.50	11.10
Coahuila	74.70	58.20	44.20	44.90	28.00	14.70	12.20	12.40	10.00	6.50
Colima	72.40	61.00	47.90	34.80	30.00	18.80	25.70	11.80	14.60	10.60
Chiapas	37.10	39.50	39.50	15.50	16.20	23.60	15.20	24.40	14.40	46.50
Chihuahua	85.50	58.90	67.40	36.80	22.20	23.80	10.10	17.90	17.10	5.70
Distrito Federal	94.80	65.40	31.70	55.00	29.30	23.10	32.70	24.20	13.30	7.90
Durango	65.60	45.50	45.30	35.40	17.30	6.60	6.50	11.70	24.70	5.10
Guanajuato	61.60	44.30	41.80	37.10	25.70	37.20	14.80	28.40	8.90	17.20
Guerrero	35.20	30.20	43.10	17.50	16.10	6.40	10.70	9.40	46.40	21.60
Hidalgo	58.80	43.40	41.70	28.60	25.20	27.30	45.20	25.50	6.90	11.80
Jalisco	79.70	61.90	51.70	39.90	29.40	25.90	21.30	23.60	13.30	8.30
México	49.90	37.10	32.70	31.40	19.00	30.40	37.90	32.50	30.30	16.70
Michoacán	64.50	51.90	45.00	31.40	24.30	19.40	14.60	12.70	34.90	10.70
Morelos	65.70	51.20	39.50	34.40	27.20	25.50	34.10	13.80	37.70	13.30
Nayarit	67.80	57.00	56.50	26.30	25.20	12.30	13.20	13.00	28.30	7.00
Nuevo León	80.80	60.50	39.90	31.50	26.80	15.90	12.40	14.70	3.60	2.90
Oaxaca	40.10	43.80	39.60	19.80	25.50	17.80	25.10	25.30	41.20	57.30
Puebla	59.90	44.70	37.00	32.50	24.00	40.30	40.30	36.60	13.60	31.00
Querétaro	46.60	37.10	42.70	27.80	21.00	34.90	29.70	27.70	7.40	16.30
Quintana Roo	32.70	23.90	53.30	15.70	9.70	34.10	18.50	10.20	12.60	7.40
San Luis Potosí	64.90	51.10	38.90	27.00	26.60	25.70	12.50	22.80	10.90	34.20
Sinaloa	69.68	59.20	43.10	28.70	21.40	7.50	7.20	8.80	23.30	5.60
Sonora	103.60	68.60	49.30	36.60	25.20	20.90	13.70	15.30	10.90	9.30
Tabasco	57.40	45.30	47.20	28.80	20.80	28.30	13.10	12.10	9.40	7.50
Tamaulipas	78.60	58.80	54.00	37.80	25.00	16.00	11.50	10.90	14.20	4.10
Tlaxcala	30.50	43.70	36.60	35.80	28.50	34.30	35.60	37.40	6.60	14.60
Veracruz	55.50	51.10	29.80	28.80	26.90	16.90	28.30	11.10	10.50	12.90
Yucatán	76.10	57.60	34.50	51.50	35.60	28.70	31.10	20.50	4.60	23.10
Zacatecas	45.50	52.60	46.80	25.10	26.70	20.80	5.50	19.80	10.20	10.60

FUENTE: *Secretaría de Salud. Mortalidad 1993.*
NOTA: Tasas por 100,000.

CUADRO 4
VARIABILIDAD INTERNA,
POR NUMERO DE GRUPOS

NUMERO DE GRUPOS	VARIABILIDAD
2	797.71
3	731.26
4	553.56
5	469.42
6	399.42
7	346.41
8	326.71
9	297.54
10	273.67
15	201.10
32	0

CUADRO 5
TASAS DE MORTALIDAD GENERAL SEGUN CAUSA Y ENTIDAD FEDERATIVA, POR REGION
1993

Entidad	Media Global	Enfermedades del Corazón	Tumores Malignos	Accidentes	Diabetes Mellitus	Enfermedades Cerebrovas-culares	Ciertas Afecciones de Origen Perinatales	Cirrosis y Otras Enfermedades Crónicas del Hígado	Neumonía e Influenza	Homicidio y Lesiones Infligidas Intencio-nalmente por Otra Persona	Enfermedades Infecciosas Intestinales
REGION I	21.8	32.76	23.90	53.30	15.70	9.78	34.10	18.50	10.20	12.60	7.40
Quintana Roo		32.70	23.90	53.30	15.70	9.70	34.10	18.50	10.20	12.60	7.40
REGION II	23.7	35.28	30.28	43.10	17.50	16.18	6.48	10.78	9.40	46.40	21.68
Guerrero		35.20	30.20	43.10	17.50	16.10	6.40	10.70	9.40	46.40	21.60
REGION III	29.4	61.5	51.1	43.5	34.8	34.3	21.5	14.4	14.8	17.1	11.5
Aguascalientes		58.70	48.50	45.30	37.00	25.10	33.10	13.50	12.30	4.30	10.30
Baja California Sur		60.00	64.60	44.40	32.20	22.80	23.00	10.20	14.20	6.00	7.40
Campeche		43.30	42.80	41.70	26.00	25.40	23.50	23.70	11.10	13.50	11.10
Durango		65.60	45.50	45.30	35.40	17.30	6.60	6.50	11.70	34.70	5.10
Guanajuato		61.60	44.30	41.80	37.10	25.70	37.20	14.80	28.40	8.90	17.20
Michoacán		64.50	51.90	45.00	31.40	24.30	19.40	14.60	12.70	34.90	10.70
Morelos		65.70	51.20	39.50	34.40	27.20	25.50	24.10	13.80	37.70	13.30
Nayarit		67.80	57.00	56.50	36.30	25.30	12.30	13.20	13.00	28.30	7.00
San Luis Potosí		64.90	51.10	38.90	27.00	26.60	25.70	12.50	22.80	10.90	14.20
Sinaloa		69.40	59.20	43.10	28.70	21.40	7.50	7.20	8.80	23.30	5.40
Tabasco		57.40	45.30	47.20	28.80	20.80	28.30	13.10	12.10	9.40	7.50
Veracruz		55.50	51.10	29.80	28.80	26.50	16.50	28.30	11.10	10.50	12.90
Zacatecas		65.50	52.40	46.80	25.10	26.70	20.80	5.50	19.80	10.20	10.40
REGION IV	32.1	53.1	41.8	38.1	31.2	23.5	33.8	37.7	31.9	12.9	18.1
Hidalgo		58.80	43.40	41.70	28.60	25.20	27.30	45.20	25.50	6.90	11.80
México		49.90	37.10	32.70	31.40	19.00	30.40	37.90	32.50	30.30	16.70
Puebla		59.90	44.70	37.00	32.50	34.00	40.30	40.30	36.60	13.60	31.00
Querétaro		46.60	37.10	42.70	27.80	21.00	34.90	29.70	27.70	7.40	16.30
Tlaxcala		50.50	42.70	36.60	35.80	28.50	36.30	35.40	37.40	6.60	14.60

Entidad	Media Global	Enfermedades del Corazón	Tumores Malignos	Accidentes	Diabetes Mellitus	Enfermedades Cerebrovas-culares	Ciertas Afecciones de Origen Perinatales	Cirrosis y Otras Enfermedades Crónicas del Hígado	Neumonía e Influenza	Homicidio y Lesiones Infligidas Intencio-nalmente por Otra Persona	Enfermedades Infecciosas Intestinales
REGION V	33.3	44.6	41.7	47.5	17.7	2.85	28.7	28.3	24.9	27.8	51.9
Chiapas		37.10	39.50	39.50	15.50	16.20	23.60	15.20	24.40	14.60	66.50
Oaxaca		60.10	43.80	39.60	19.80	25.50	17.80	25.10	25.30	41.20	57.30
REGION VI	33.6	79.8	68.8	41.7	39.4	39.9	26.3	22.7	17.7	9.8	8.3
Coahuila		74.70	58.20	44.20	44.90	28.00	14.70	12.20	12.40	10.00	6.50
Colima		72.40	61.00	47.90	34.80	30.00	18.80	25.70	11.80	14.60	10.60
Distrito Federal		94.80	65.40	31.70	55.00	29.30	23.10	32.70	24.20	13.50	7.90
Jalisco		79.70	61.90	51.70	39.90	29.40	25.90	21.30	22.60	13.30	8.20
Nuevo León		80.80	60.50	39.90	31.50	26.80	15.90	12.60	14.70	3.40	2.90
Yucatán		76.10	57.60	34.50	32.90	35.60	28.70	31.10	20.50	4.40	13.10
REGION VII	34.8	85.5	68.6	71.6	36.6	24.1	21.2	13.6	15.8	14.8	6.2
Baja California		74.40	56.00	55.20	35.00	24.10	23.90	19.10	15.90	13.60	5.80
Chihuahua		85.50	58.90	67.40	36.80	22.20	23.80	10.30	17.90	17.10	5.70
Sonora		103.60	68.60	49.30	36.60	25.20	20.90	13.70	15.30	10.90	9.20
Tamaulipas		78.60	58.80	54.00	37.80	25.00	16.00	11.50	10.90	14.20	4.10

FUENTE: Secretaría de Salud. Mortalidad 1993.
NOTA: Tasas por 100,000.

CUADRO 6

CHIAPAS: POBLACION TOTAL, INDICADORES SOCIODEMOGRAFICOS, INDICE DE MARGINACION MUNICIPAL

	NOMBRE	% DE POB. ANALFABETA > 15 AÑOS	% DE POB. SIN PRIMARIA COMPLETA > 15 AÑOS	% DE OCU-PANTES EN VIVIENDAS SIN DRENAJE NI EXCUSADO	% DE OCU-PANTES EN VIVIENDAS SIN ENERGIA ELECTRICA	% DE OCU-PANTES EN VIVIENDAS SIN AGUA ENTUBADA	% DE VIVIENDAS CON HACINAMIENTO	% DE OCU-PANTES EN VIVIENDAS CON PISO DE TIERRA	% DE POB. EN LOCALIDADES CON < 5000 HABITANTES	% DE POB. OCUPADA CON INGRESO MENOR DE 2 SALARIOS MINIMOS	INDICE	GRADO
1	ACACOYAGUA	25,16	69,54	51,54	36,05	53,56	85,87	44,32	100,00	90,24	0,82125	ALTO
2	ACALA	30,05	64,88	24,97	10,01	32,21	76,23	31,61	47,44	82,79	-0,03644	MEDIO
3	ACAPETAHUA	25,78	65,94	56,24	36,04	72,61	77,96	50,30	100,00	81,78	0,76354	ALTO
4	ALTAMIRANO	51,79	83,31	43,67	75,01	48,75	79,95	79,56	100,00	93,53	1,55549	MUY ALTO
5	AMATAN	42,17	85,82	62,15	83,14	65,70	79,83	68,20	100,00	90,97	1,74354	MUY ALTO
6	AMATENANGO DE LA FRONTERA	32,49	74,24	51,55	37,07	20,50	84,77	56,27	100,00	86,80	0,80247	ALTO
7	AMATENANGO DEL VALLE	60,03	85,46	74,92	43,05	32,18	82,50	81,98	100,00	96,07	1,85389	MUY ALTO
8	ANGEL ALBINO CORZO	35,86	75,88	45,77	19,79	19,79	82,60	50,38	71,27	83,96	0,52599	ALTO
9	ARRIAGA	17,42	49,74	25,65	9,87	23,81	65,82	13,45	39,68	76,09	-0,68134	BAJO
10	BEJUCAL DE OCAMPO	20,66	60,80	45,89	83,51	72,93	91,29	78,74	100,00	85,89	1,46689	MUY ALTO
11	BELLA VISTA	17,36	70,48	64,52	63,45	56,99	84,73	66,24	100,00	90,57	1,11460	ALTO
12	BERRIOZABAL	29,04	68,12	38,26	23,29	50,27	75,96	43,45	36,36	83,50	0,26495	ALTO
13	BOCHIL	41,69	69,58	48,23	36,81	33,57	77,29	66,49	59,26	81,88	0,68795	ALTO
14	BOSQUE, EL	54,34	79,39	40,13	29,45	21,50	80,40	76,53	100,00	89,30	1,07006	ALTO
15	CACAHOATAN	23,66	58,20	19,26	20,42	19,35	77,46	36,11	69,78	82,81	-0,09056	MEDIO
16	CATAZAJA	22,46	61,39	38,80	19,70	70,39	72,16	41,92	100,00	80,67	0,35671	ALTO
17	CINTALAPA	21,69	59,49	29,07	21,28	39,60	70,09	29,39	50,46	80,24	-0,14957	MEDIO
18	COAPILLA	30,98	76,75	55,07	31,59	14,75	75,63	72,45	100,00	90,01	0,79094	ALTO
19	COMITAN DE DOMINGUEZ	22,96	53,50	35,46	16,97	33,68	63,45	38,00	38,78	79,06	-0,29493	MEDIO
20	CONCORDIA LA	33,70	73,87	40,70	24,22	35,43	80,18	41,40	79,96	84,07	0,51218	ALTO
21	COPAINALA	23,83	66,34	47,20	29,50	17,45	74,77	58,12	100,00	85,83	0,43290	ALTO
22	CHALCHIHUITAN	62,02	74,86	81,55	65,58	74,46	66,82	84,20	100,00	86,17	2,14446	MUY ALTO
23	CHAMULA	71,30	91,20	89,88	43,33	65,92	79,75	83,69	100,00	92,10	2,08829	MUY ALTO
24	CHANAL	54,30	78,12	84,34	53,92	94,17	86,33	87,85	100,00	92,49	1,99797	MUY ALTO
25	CHAPULTENANGO	42,70	81,96	60,52	71,00	45,04	74,27	86,23	100,00	92,67	1,46221	MUY ALTO
26	CHENALHO	51,38	69,24	68,35	78,12	56,19	87,23	90,49	100,00	93,03	1,87857	MUY ALTO
27	CHIAPA DE CORZO	24,84	57,74	32,24	10,73	28,13	70,12	32,76	58,56	78,87	-0,20045	MEDIO
28	CHIAPILLA	47,47	78,16	46,83	12,92	5,80	83,00	43,57	100,00	91,16	0,70976	ALTO
29	CHICOASEN	21,29	63,17	36,12	16,54	20,91	77,23	32,04	100,00	82,34	0,09544	ALTO
30	CHICOMUSELO	31,06	75,40	66,16	18,56	46,36	85,06	54,78	100,00	87,88	0,91738	ALTO
31	CHILON	56,17	82,26	79,74	78,04	41,96	84,93	83,21	100,00	93,31	1,88556	MUY ALTO
32	ESCUINTLA	24,07	65,09	45,21	36,80	29,47	78,23	36,61	70,20	83,48	0,31184	ALTO
33	FRANCISCO LEON	45,81	91,51	84,72	89,36	51,91	85,78	86,56	100,00	93,47	2,09631	MUY ALTO
34	FRONTERA COMALAPA	20,42	64,79	50,30	22,23	37,14	80,32	49,96	81,21	83,66	0,36869	ALTO
35	FRONTERA HIDALGO	27,84	58,90	40,70	37,16	67,66	76,24	58,62	100,00	84,29	0,66896	ALTO
36	GRANDEZA, LA	21,13	67,37	34,53	81,59	42,73	87,20	79,43	100,00	92,25	1,09752	ALTO
37	HUEHUETAN	26,25	60,50	56,47	39,49	64,52	79,8[illegible]	51,19	80,55	84,36	0,68952	ALTO
38	HUIXTAN	46,49	73,50	67,81	58,54	70,21	82,7[illegible]	83,05	100,00	93,76	1,60834	MUY ALTO
39	HUITIUPAN	45,49	81,44	56,49	88,83	43,09	85,94	91,51	100,00	90,34	1,57789	MUY ALTO
40	HUIXTLA	19,02	52,21	31,81	25,59	42,63	70,71	29,53	43,86	71,53	-0,30089	MEDIO
41	INDEPENDENCIA, LA	25,53	73,16	45,54	23,77	84,36	82,54	55,86	100,00	89,86	0,83480	ALTO
42	IXHUATAN	37,59	82,62	41,73	41,56	21,81	93,62	67,40	100,00	85,05	0,94413	ALTO
43	IXTACOMITAN	27,74	67,96	41,63	42,77	49,44	76,63	48,82	100,00	83,04	0,63534	ALTO
44	IXTAPA	29,88	69,43	60,84	12,35	14,09	80,90	46,79	100,00	87,72	0,51947	ALTO
45	IXTAPANGAJOYA	30,40	80,05	59,86	63,90	39,92	82,02	73,72	100,00	93,32	1,27409	MUY ALTO
46	JIQUIPILAS	19,37	60,45	27,89	14,25	29,46	69,74	27,43	77,79	83,14	-0,15436	MEDIO
47	JITOTOL	39,08	77,12	35,15	50,98	15,74	82,35	61,71	100,00	90,36	0,89856	ALTO
48	JUAREZ	24,68	64,27	38,46	41,83	78,19	72,59	43,43	73,35	78,86	0,46250	ALTO
49	LARRAINZAR	62,07	82,29	80,54	78,11	47,78	85,45	92,19	100,00	94,41	2,03301	MUY ALTO
50	LIBERTAD, LA	20,26	68,34	30,85	55,21	72,34	67,94	16,55	100,00	83,98	0,39865	ALTO
51	MAPASTEPEC	26,33	65,74	41,89	36,59	80,30	73,51	41,92	93,96	80,01	0,38437	ALTO
52	MARGARITAS, LAS	48,37	83,27	36,54	86,40	72,72	93,36	77,90	90,02	96,20	1,48228	MUY ALTO
53	MAZAPA DE MADERO	14,54	69,59	49,85	81,53	35,40	86,47	73,76	100,00	86,58	0,94580	ALTO
54	MAZATAN	22,77	57,89	45,89	32,41	85,30	77,86	61,14	100,00	85,10	0,73966	ALTO
55	METAPA	22,35	46,95	37,13	25,22	37,94	71,83	42,40	100,00	71,63	-0,02256	MEDIO
56	MITONTIC	69,30	86,19	91,22	83,75	76,26	85,51	96,37	100,00	98,11	2,34951	MUY ALTO
57	MOTOZINTLA	20,24	64,90	38,87	47,58	26,97	82,97	51,10	76,48	81,93	0,41411	ALTO
58	NICOLAS RUIZ	31,64	81,55	34,92	15,46	5,24	87,95	45,83	100,00	83,24	0,50564	ALTO
59	OCOSINGO	46,71	78,29	60,24	67,95	49,17	80,80	74,68	84,52	87,56	1,38012	MUY ALTO

CUADRO 6

CHIAPAS: POBLACION TOTAL, INDICADORES SOCIODEMOGRAFICOS, INDICE DE MARGINACION MUNICIPAL

	NOMBRE	% DE POB. ANALFABETA > 15 AÑOS	% DE POB. SIN PRIMARIA COMPLETA > 15 AÑOS	% DE OCU-PANTES EN VIVIENDAS SIN DRENAJE NI EXCUSADO	% DE OCU-PANTES EN VIVIENDAS SIN ENERGIA ELECTRICA	% DE OCU-PANTES EN VIVIENDAS SIN AGUA ENTUBADA	% DE VIVIENDAS CON HACINAMIENTO	% DE OCU-PANTES EN VIVIENDAS CON PISO DE TIERRA	% DE POB. EN LOCALIDADES CON < 5000 HABITANTES	% DE POB. OCUPADA CON INGRESO MENOR DE 2 SALARIOS MINIMOS	INDICE	GRADO
60	OCOTEPEC	59,91	91,89	85,42	53,35	48,94	81,72	83,85	100,00	89,82	1,73707	MUY ALTO
61	OCOZOCOAUTLA DE ESPINOSA	25,39	67,35	35,51	22,85	26,81	73,78	51,15	57,12	83,96	0,17542	ALTO
62	OSTUACAN	33,56	78,59	50,48	63,44	74,92	79,88	51,46	100,00	84,99	1,17226	MUY ALTO
63	OSUMACINTA	17,71	62,97	38,32	13,41	4,47	77,47	36,82	100,00	76,97	-0,05910	MEDIO
64	OXCHUC	34,81	64,79	78,31	67,81	58,33	84,77	89,06	100,00	92,06	1,82373	MUY ALTO
65	PALENQUE	31,60	65,22	50,41	43,16	58,92	77,07	49,79	73,01	79,52	0,63150	ALTO
66	PANTELHO	63,66	82,78	62,74	75,22	45,37	85,17	91,62	100,00	86,21	1,63423	MUY ALTO
67	PANTEPEC	55,94	86,53	64,81	66,16	32,56	82,00	85,77	100,00	84,89	1,59512	MUY ALTO
68	PICHUCALCO	28,58	64,08	42,65	45,41	53,19	72,44	40,29	61,18	76,06	0,33961	ALTO
69	PIJIJIAPAN	23,33	63,85	37,17	29,14	74,30	74,41	38,51	72,01	77,63	0,33203	ALTO
70	PORVENIR, EL	19,95	64,97	40,12	63,84	61,82	90,84	71,55	100,00	92,63	1,20622	MUY ALTO
71	VILLA COMALTITLAN	26,44	66,96	57,70	44,62	63,76	80,39	52,50	75,10	86,38	0,80275	ALTO
72	PUEBLO NUEVO SOLISTAHUACAN	50,80	82,75	55,10	56,43	27,29	79,98	79,98	100,00	88,18	1,34017	MUY ALTO
73	RAYON	45,37	76,77	54,27	47,26	36,09	77,26	82,61	100,00	83,91	1,03685	ALTO
74	REFORMA	18,36	51,84	21,84	20,85	36,75	80,99	25,33	36,24	57,24	-0,63979	BAJO
75	ROSAS, LAS	51,26	79,83	57,46	30,80	51,86	73,20	64,77	26,62	84,39	0,84418	ALTO
76	SABANILLA	42,96	82,73	63,59	80,06	44,32	86,85	90,89	100,00	86,56	1,85242	MUY ALTO
77	SALTO DE AGUA	46,33	75,88	78,99	63,43	61,82	84,81	78,88	100,00	85,97	1,56917	MUY ALTO
78	SAN CRISTOBAL DE LAS CASAS	24,99	44,79	21,72	16,95	27,47	80,06	33,99	17,85	71,25	-0,68946	BAJO
79	SAN FERNANDO	26,82	67,01	41,97	15,08	46,95	78,70	36,85	89,88	81,06	0,27598	ALTO
80	SILTEPEC	33,39	81,87	51,82	76,53	31,02	80,09	72,93	100,00	90,16	1,34694	MUY ALTO
81	SIMOJOVEL	56,21	78,81	41,69	85,74	27,84	81,11	78,89	77,39	83,97	1,21840	MUY ALTO
82	SITALA	71,36	93,32	75,02	86,13	69,16	85,49	92,24	100,00	90,55	2,32256	MUY ALTO
83	SOCOLTENANGO	40,19	78,22	45,87	26,92	41,40	78,06	49,32	100,00	89,98	0,85277	ALTO
84	SOLOSUCHIAPA	31,34	74,42	40,91	47,02	21,13	81,12	57,56	100,00	83,34	0,71749	ALTO
85	SOYALO	35,48	69,41	41,87	10,62	30,90	76,78	56,17	100,00	78,17	0,45054	ALTO
86	SUCHIAPA	29,52	58,18	32,12	9,44	20,27	77,51	35,23	24,59	79,32	-0,21030	MEDIO
87	SUCHIATE	28,18	59,55	35,04	23,47	66,05	74,69	52,32	81,56	76,81	0,29185	ALTO
88	SUNUAPA	39,16	85,63	64,34	80,16	66,06	84,25	83,35	100,00	83,77	1,65764	MUY ALTO
89	TAPACHULA	18,36	43,53	23,72	19,27	44,72	85,27	33,45	34,72	68,98	-0,58542	BAJO
90	TAPALAPA	29,43	62,81	60,21	62,10	35,79	77,35	81,40	100,00	90,26	1,21983	MUY ALTO
91	TAPILULA	33,67	66,79	39,76	30,21	21,70	78,12	53,65	39,59	79,47	0,25682	ALTO
92	TECPATAN	26,87	68,92	42,81	37,34	26,44	75,84	50,83	82,82	80,92	0,42330	ALTO
93	TENEJAPA	49,44	74,81	73,28	45,89	19,21	84,67	92,14	100,00	95,30	1,45256	MUY ALTO
94	TEOPISCA	49,75	79,20	48,93	28,34	24,06	75,30	65,77	53,36	89,50	0,78495	ALTO
95	TILA	50,50	78,09	60,14	65,94	49,17	82,41	86,12	89,67	86,51	1,56298	MUY ALTO
97	TONALA	19,22	55,11	25,32	16,40	46,47	71,46	18,85	52,23	73,89	-0,33915	MEDIO
98	TOTOLAPA	57,60	85,12	69,01	14,56	22,39	90,01	46,64	100,00	92,53	1,22679	MUY ALTO
99	TRINITARIA, LA	28,43	66,52	37,68	31,01	71,06	81,02	66,52	90,51	84,76	0,81828	ALTO
100	TUMBALA	53,81	78,25	71,30	67,99	40,16	82,53	86,01	100,00	93,04	1,67065	MUY ALTO
101	TUXTLA GUTIERREZ	10,71	29,56	7,26	3,05	15,81	54,50	14,29	2,02	60,17	-1,56401	BAJO
102	TUXTLA CHICO	27,35	56,30	53,83	28,59	75,82	78,14	50,40	81,16	80,36	0,57818	ALTO
103	TUZANTAN	24,48	69,23	59,89	31,78	45,24	81,19	42,01	100,00	90,78	0,73205	ALTO
104	TZIMOL	37,93	81,55	50,75	16,48	52,39	76,61	49,00	100,00	87,05	0,83888	ALTO
105	UNION JUAREZ	24,17	63,05	15,30	25,38	14,04	78,59	40,18	100,00	90,44	0,16144	ALTO
106	VENUSTIANO CARRANZA	37,75	70,36	36,04	14,09	22,08	72,87	41,52	73,34	81,78	0,23536	ALTO
107	VILLA CORZO	28,92	67,23	47,25	21,83	32,74	77,36	42,55	50,65	83,82	0,27538	ALTO
108	VILLAFLORES	22,66	59,07	33,29	11,19	25,18	73,99	31,39	59,46	81,44	-0,14807	MEDIO
109	YAJALON	45,74	69,97	48,21	55,21	30,19	77,42	70,08	54,86	86,04	0,85956	ALTO
110	SAN LUCAS	43,41	81,57	75,16	24,70	27,47	73,74	80,71	100,00	88,87	1,10948	ALTO
111	ZINACANTAN	63,72	81,87	87,22	21,07	46,75	82,43	88,35	100,00	90,70	1,85116	MUY ALTO
112	SAN JUAN CANCUC	88,54	82,82	85,70	80,42	86,96	90,86	96,81	75,82	96,28	2,48760	MUY ALTO

FUENTE: CONAPO-CONAGUA; Indicadores Sociodemográficos e Indice de Marginación Municipal, 1990.

CUADRO 7
VARIABILIDAD INTERNA,
POR NUMERO DE GRUPOS

NUMERO DE GRUPOS	VARIABILIDAD
2	11561.78
3	8947.15
4	6424.90
5	5430.77
6	4743.57
7	3934.10
8	3565.95
9	3265.71
10	3027.08
15	2021.64
112	0.00

CUADRO 6

CHIAPAS: INDICADORES SOCIODEMOGRAFICOS POR MUNICIPIO, SEGUN REGION

MUNICIPIO/MEDIA		% DE POB. ANALFABETA > 15 AÑOS	% DE POB. SIN PRIMARIA COMPLETA > 15 AÑOS	% DE OCU-PANTES EN VIVIENDAS SIN DRENAJE NI EXCUSADO	% DE OCU-PANTES EN VIVIENDAS SIN ENERGIA ELECTRICA	% DE OCU-PANTES EN VIVIENDAS SIN AGUA ENTUBADA	% DE VIVIENDAS CON HACINAMIENTO	% DE OCU-PANTES EN VIVIENDAS CON PISO DE TIERRA	% DE POB. EN LOCALIDADES CON < 5000 HABITANTES	% DE POB. OCUPADA CON INGRESO MENOR DE 2 SALARIOS MINIMOS
REGION I	57,09	27,07	60,57	44,71	32,89	58,70	76,14	46,13	83,91	81,87
1 ACACOYAGUA		25,16	69,54	51,54	38,05	63,98	85,67	44,32	100,00	90,34
3 ACAPETAHUA		25,78	65,84	56,24	38,04	72,61	77,88	80,30	100,00	81,78
16 CATAZAJA		22,46	61,39	38,80	19,70	70,39	72,18	41,82	100,00	80,67
35 FRONTERA HIDALGO		27,84	58,80	40,70	37,18	67,68	76,24	88,82	100,00	84,29
37 HUEHUETAN		38,25	80,30	66,47	38,48	84,82	78,84	51,19	80,55	84,38
41 INDEPENDENCIA, LA		25,53	73,16	46,54	23,77	64,38	82,54	86,86	100,00	80,86
43 IXTACOMITAN		27,74	67,86	41,63	42,77	48,44	76,10	48,82	100,00	83,04
48 JUAREZ		24,88	64,27	38,46	41,83	78,19	72,39	43,43	73,35	78,68
50 LIBERTAD, LA		20,28	68,34	30,85	85,21	72,34	87,84	16,65	100,00	83,38
51 MAPASTEPEC		26,33	65,74	41,89	38,58	80,50	73,51	41,82	83,86	80,01
54 MAZATAN		22,77	57,89	46,89	32,41	65,30	77,88	61,14	100,00	85,10
55 METAPA		22,35	46,85	37,13	25,22	37,84	71,83	42,40	100,00	71,63
85 PALENQUE		31,80	65,22	50,41	43,18	58,82	77,07	48,79	73,01	79,52
88 PICHUCALCO		26,68	64,08	42,85	46,41	53,19	72,44	40,38	61,16	76,08
89 PIJIJIAPAN		26,33	63,85	37,17	39,14	74,30	74,41	36,51	72,01	77,63
71 VILLA COMALTITLAN		36,44	66,88	57,70	44,82	83,76	80,38	82,50	75,10	88,36
83 SOCOLTENANGO		40,19	78,22	46,87	26,82	41,40	78,08	48,32	100,00	89,96
87 SUCHIATE		38,18	58,85	36,04	53,47	66,05	74,68	82,32	81,58	76,81
98 TOTOLAPA		37,60	65,12	69,01	14,58	32,30	90,01	46,84	100,00	82,53
101 TUXTLA GUTIERREZ		10,71	29,56	7,28	3,05	15,61	54,50	14,28	2,02	60,17
102 TUXTLA CHICO		27,35	56,30	53,83	28,59	75,82	78,14	85,40	81,18	80,36
103 TUZANTAN		24,48	69,23	59,89	31,78	45,24	81,18	42,01	100,00	90,78
REGION II	55,05	22,04	69,73	43,70	27,20	27,46	77,81	50,72	82,40	84,28
2 ACALA		30,05	64,89	24,87	10,01	32,21	76,23	31,81	47,44	82,79
6 AMATENANGO DE LA FRONTERA		32,49	74,24	61,55	37,07	29,50	84,77	58,27	100,00	86,80
7 AMATENANGO DEL VALLE		60,03	85,48	74,82	43,05	32,16	82,50	81,98	100,00	96,07
9 ANGEL ALBINO CORZO		35,86	75,86	45,77	19,79	18,79	82,80	52,36	71,27	83,98
14 BOSQUE, EL		54,34	78,38	40,13	28,45	21,50	80,40	76,53	100,00	89,30
15 CACAHOATAN		23,86	58,20	19,26	23,42	19,35	77,46	36,11	89,78	82,81
17 CINTALAPA		21,88	58,49	29,07	21,28	38,80	70,08	28,38	50,46	80,24
18 COAPILLA		30,96	78,75	55,07	31,58	14,75	75,63	72,46	100,00	80,01
19 COMITAN DE DOMINGUEZ		22,86	53,50	35,46	14,97	33,88	83,46	36,02	36,78	79,08
20 CONCORDIA LA		33,70	73,67	40,70	24,22	35,43	80,18	41,40	79,86	84,07
21 COPAINALA		23,83	68,31	47,20	28,50	17,45	74,77	58,12	100,00	85,85
27 CHIAPA DE CORZO		24,84	57,74	32,24	10,73	28,13	70,12	32,78	58,56	78,87
28 CHIAPILLA		47,47	78,16	46,63	12,82	5,80	83,00	43,57	100,00	91,18
29 CHICOASEN		21,29	63,17	36,12	16,54	20,91	77,23	32,04	100,00	82,34
30 CHICOMUSELO		31,08	75,40	66,16	16,58	46,36	85,06	54,78	100,00	87,88
32 ESCUINTLA		24,07	66,08	45,21	36,60	29,47	78,23	38,81	70,20	83,46
34 FRONTERA COMALAPA		20,42	64,78	50,30	22,23	37,14	80,32	48,86	81,21	83,66
40 HUIXTLA		18,02	52,21	31,81	25,59	42,83	70,71	29,53	43,86	71,53
42 IXHUATAN		37,59	82,82	41,73	41,58	21,81	83,82	67,40	100,00	85,05
44 IXTAPA		28,88	69,43	60,84	12,35	14,08	80,80	46,79	100,00	87,72
46 JIQUIPILAS		19,37	60,45	27,89	14,25	29,46	88,74	27,43	77,79	83,14
47 JITOTOL		30,08	77,12	35,15	50,98	15,74	82,35	81,71	100,00	90,56
57 MOTOZINTLA		30,24	64,80	38,87	47,59	38,97	82,97	51,10	76,48	81,83
58 NICOLAS RUIZ		31,84	61,55	34,82	15,48	5,24	87,85	45,83	100,00	83,24
61 OCOZOCOAUTLA DE ESPINOSA		25,39	67,35	35,61	22,85	28,81	73,78	51,15	57,12	83,86
63 OSUMACINTA		17,71	62,97	38,32	13,41	4,47	77,47	36,82	100,00	78,67
73 RAYON		46,37	78,77	54,27	47,28	36,08	77,26	62,61	100,00	83,91
79 SAN FERNANDO		26,62	67,01	41,87	15,08	46,85	78,70	36,85	69,68	81,06
84 SOLOSUCHIAPA		31,34	74,42	40,81	47,02	21,13	81,12	57,56	100,00	83,34
85 SOYALO		35,48	68,41	41,87	10,62	30,80	78,78	54,17	100,00	78,17
92 TECPATAN		28,87	66,82	42,61	37,34	26,44	75,84	50,83	82,62	80,82
93 TENEJAPA		48,44	74,81	73,28	46,89	19,21	84,87	92,14	100,00	85,30
96 TILA		50,50	78,09	68,14	66,84	48,17	82,41	86,12	88,57	86,51
97 TONALA		19,22	55,11	25,32	16,40	46,47	71,48	18,88	52,23	73,89
104 TZIMOL		37,83	61,56	50,75	16,48	52,38	78,61	48,00	100,00	87,05
105 UNION JUAREZ		24,17	63,05	15,30	25,38	14,04	78,59	40,18	100,00	80,44
106 VENUSTIANO CARRANZA		37,75	70,38	38,04	14,08	22,08	72,87	41,52	73,34	81,78
107 VILLA CORZO		28,82	67,23	47,26	21,83	32,74	77,36	42,55	50,85	83,82
108 YAJALON		45,74	69,97	48,21	55,21	30,19	77,42	70,06	54,88	86,04
110 SAN LUCAS		43,41	81,57	75,16	24,70	27,47	73,74	80,71	100,00	85,87

CUADRO 6

CHIAPAS: INDICADORES SOCIODEMOGRAFICOS POR MUNICIPIO, SEGUN REGION

MUNICIPIO/REGION		% DE POBL. ANALFABETA > 15 AÑOS	% DE POBL. SIN PRIMARIA COMPLETA > 15 AÑOS	% DE OCUPANTES EN VIVIENDAS SIN DRENAJE NI EXCUSADO	% DE OCUPANTES EN VIVIENDAS SIN ENERGIA ELECTRICA	% DE OCUPANTES EN VIVIENDAS SIN AGUA ENTUBADA	% DE VIVIENDAS CON HACINAMIENTO	% DE OCUPANTES EN VIVIENDAS CON PISO DE TIERRA	% DE POBL. EN LOCALIDADES CON < 5000 HABITANTES	% DE POBL. OCUPADA CON INGRESO MENOR DE 2 SALARIOS MINIMOS
REGION III	72,57	43,63	77,80	62,75	70,17	53,25	81,63	60,33	98,32	87,28
4 ALTAMIRANO		51,79	83,31	43,67	76,01	46,75	78,95	78,56	100,00	93,53
5 AMATAN		42,17	85,82	62,15	83,14	65,70	79,83	86,20	100,00	90,97
10 BEJUCAL DE OCAMPO		20,69	60,80	45,89	88,51	72,93	81,29	78,74	100,00	96,89
11 BELLA VISTA		17,38	70,45	64,52	83,45	86,89	84,73	86,24	100,00	90,57
22 CHALCHIHUITAN		62,02	74,86	81,85	85,56	74,46	86,82	84,20	100,00	96,17
23 CHAMULA		71,30	81,20	88,88	43,33	65,82	78,75	83,69	100,00	92,10
24 CHANAL		54,30	78,12	84,34	53,82	84,17	86,33	87,65	100,00	92,49
25 CHAPULTENANGO		42,70	81,38	60,52	71,00	45,04	74,27	86,23	100,00	92,67
26 CHENALHO		51,36	69,24	86,36	78,12	58,19	87,23	90,49	100,00	93,03
31 CHILON		58,17	82,26	78,74	78,04	41,86	84,83	83,21	100,00	93,31
33 FRANCISCO LEON		45,81	81,81	84,72	86,36	61,81	85,78	86,56	100,00	93,47
36 GRANDEZA, LA		21,13	67,37	34,53	81,86	42,73	87,20	78,43	100,00	92,25
38 HUIXTAN		46,48	73,50	67,81	86,54	70,21	82,79	83,05	100,00	93,76
39 HUITIUPAN		46,48	81,44	86,48	86,83	43,08	86,84	81,51	100,00	93,34
45 OCTAPANGAJOYA		30,40	80,05	86,86	63,80	38,82	82,02	73,72	100,00	93,32
48 LARRAINZAR		62,07	82,29	80,54	78,11	47,78	85,45	92,18	100,00	94,41
52 MARGARITAS, LAS		48,32	63,27	36,54	66,40	72,72	83,36	77,80	80,02	86,20
53 MAZAPA DE MADERO		14,84	68,88	46,86	61,83	35,40	88,47	73,78	100,00	96,51
56 MITONTIC		89,30	83,18	91,22	83,75	78,28	85,51	96,37	100,00	93,11
58 OCOSINGO		48,71	76,29	80,24	67,86	48,17	80,80	74,68	84,82	97,56
60 OCOTEPEC		88,91	81,89	85,42	53,36	46,84	81,72	83,86	100,00	96,62
62 OSTUACAN		33,56	78,98	50,48	63,44	74,92	78,98	51,46	100,00	94,89
64 OXCHUC		34,81	64,79	76,31	87,81	58,33	84,77	86,08	100,00	92,06
66 PANTELHO		63,88	82,76	62,74	75,22	46,37	85,17	81,82	100,00	96,21
67 PANTEPEC		55,94	86,53	84,81	86,16	32,56	82,00	85,77	100,00	94,89
70 PORVENIR, EL		19,85	84,97	40,12	83,84	61,82	80,84	71,55	100,00	92,63
72 PUEBLO NUEVO SOLISTAHUACAN		80,80	82,76	55,10	96,43	27,28	79,98	79,98	100,00	96,16
76 SABANILLA		42,98	62,73	63,58	80,06	44,32	86,85	90,89	100,00	96,56
77 SALTO DE AGUA		46,33	75,86	78,98	63,43	61,82	84,81	76,86	100,00	95,97
80 SILTEPEC		33,38	81,87	51,82	76,83	31,02	89,09	72,83	100,00	93,16
82 SITALA		71,36	80,32	75,02	86,13	89,16	85,49	92,24	100,00	93,55
86 SUNUAPA		38,16	85,63	84,34	80,16	86,06	84,25	83,36	100,00	93,77
90 TAPALAPA		29,43	82,61	80,21	82,10	35,78	77,35	81,40	100,00	93,28
89 TRINITARIA, LA		26,43	86,82	37,86	31,91	71,06	81,02	86,52	90,51	94,76
REGION IV	45,81	30,27	60,86	36,53	21,89	32,98	72,05	42,46	36,07	77,55
9 ARRIAGA		17,42	46,74	25,85	9,97	23,51	65,82	13,45	39,86	76,09
12 BERRIOZABAL		29,04	66,12	36,28	23,29	50,27	75,98	43,45	36,36	63,90
13 BOCHIL		41,88	69,36	46,23	39,81	33,57	77,29	66,49	69,26	91,86
74 REFORMA		18,36	51,84	21,84	20,85	36,75	66,99	35,33	36,24	57,24
75 ROSAS, LAS		51,26	79,63	57,46	30,80	81,86	73,20	84,77	36,62	84,39
78 SAN CRISTOBAL DE LAS CASAS		24,86	44,79	21,72	16,85	27,47	60,06	33,89	17,85	71,25
85 SUCHIAPA		29,82	68,18	32,12	9,44	20,27	77,61	35,23	24,59	78,32
88 TAPACHULA		16,36	43,53	23,72	18,27	44,72	85,27	33,45	34,72	66,86
91 TAPILULA		33,67	66,78	36,76	30,21	21,70	78,12	63,65	39,56	79,47
94 TEOPISCA		46,75	79,30	48,83	36,84	24,05	75,30	65,77	63,36	88,50
109 VILLAFLORES		22,68	58,07	33,39	11,19	25,18	73,89	31,36	58,46	81,44
REGION V	65,74	56,21	75,91	41,89	65,74	27,84	81,11	78,89	77,36	93,87
81 SIMOJOVEL		56,21	78,91	41,89	65,74	27,84	81,11	78,89	77,36	93,87
REGION VI	74,78	53,81	78,35	71,30	67,88	40,18	82,53	86,01	100,00	93,04
100 TUMBALA		53,81	78,35	71,30	67,88	40,18	82,53	86,01	100,00	93,04
REGION VII	80,32	66,13	82,96	81,46	55,74	67,85	86,86	82,46	87,81	93,49
111 ZINACANTAN		63,72	81,87	87,22	21,07	46,75	82,43	86,35	100,00	90,70
112 SAN JUAN CANCUC		66,54	82,82	86,70	80,42	86,96	90,86	86,61	75,62	94,29

FUENTE: CONAPO-CONAGUA; Indicadores Sociodemográficos e Indice de Integración Municipal, 1990.

BIBLIOGRAFÍA

Anderberg, M. R. (1973), *Cluster Analysis for Aplications*, Academic Press, New York.

Beale, E. M (1969 a), *Cluster Analysis*, London: Scientific Control Systems.

Centro Latinoamericano de Demografía (1972), *Métodos Demográficos para el Estudio de la Mortalidad*, Santiago de Chile.

CONAPO-CONAGUA (1993), *Indicadores Socioeconómicos e Indice de Marginación Municipal*, 1990, México.

Cochran William (1982), *Técnicas de Muestreo*, Compañía Editorial Continental, S.A. de C.V., México.

Everit, B. (1974), *Cluster Analysis*, Henemann Educational Books, London.

Gordon A. D. (1981), *Clasification*, Chapman and Hall, London, New York.

Lehmann R. Donald (1993), *Investigación y Análisis de Mercado*, Marketing Universitario, 3a. Edición en inglés y 1a. en español, Comañía Editorial Continental S.A. de C.V., México.

Morrison D. (1967), *Multivariate Statistical Methods*, Ed. McGraw-Hill Book Company. México.

Paykel, E.S. (1971), *Clasification of Depressed Patients: a Cluster Analysis Derived Group*. Br. J. Psychiat; 118, 275.

Scheaffer, R. Mendehall William y Ott L., *Elementos de Muestreo*, Grupo Editorial Iberoamérica, México.

Secretaría de Salud (1994), *Mortalidad 1992*, Subsecretaría de Coordinación y Desarrollo, México.

Secretaría de Salud (1995), *Mortalidad 1993*, Subsecretaría de Coordinación y Desarrollo, México.

Sokal, R.R, y Michenener, C.D. (1958). *A Statiscal Method for Evaluating Systematic Relationships*. Univ. Kansas Sci. Bull; 38, 1409-1438.