UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE CIENCIAS

TECNICAS DE GRAFICACION PARA DATOS MULTIVARIADOS

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I A

P R E S E N T A :

HORTENSIA MORENO MACIAS

MEXICO, D. F. 1992

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

INTRODUCCION

Aún hoy en día, es común asociar a la estadística con un análisis meramente numérico de los datos, motivo por el cual durante un largo período de tiempo se han desarrollado análisis estadísticos sin atribuir la debida importancia a un valioso elemento: las representaciones gráficas.

"Es más fácil encontrar la historia de métodos gráficos para la presentación de conclusiones que de gráficas *como herramienta de trabajo diario del científico*"[1].

Sin embargo, el trabajo de algunos investigadores ha demostrado la trascendencia de los métodos gráficos no sólo para *la presentación de datos o conclusiones* sino también para *resaltar* varias *características y descubrir patrones de comportamiento*. Muchos procedimientos estadísticos tienen fundamento en algunos supuestos y si los supuestos son falsos, los cálculos numéricos conducen a conclusiones erróneas. Sin la representación gráfica es latente el peligro de obtener, por ejemplo, la ecuación de un modelo lineal de regresión y hacer predicciones cuando quizá la relación lineal entre los datos no existe.[2]

En este sentido, las gráficas pueden pensarse como un *elemento intuitivo en la verificación de los supuestos y un punto de apoyo para el desarrollo numérico*. Las técnicas de graficación integran un laboratorio para el análisis de datos.

"Un despliegue bien seleccionado realzará el entendimiento de los datos y puede proveer un *antídoto parcial* ante el peligroso hábito de aplicar técnicas de análisis multivariado de manera descuidada y poco crítica".[3]

La utilidad de los despliegues gráficos no se limita al momento anterior al análisis. Algunas técnicas de Análisis Multivariado emiten, como parte de los resultados, una representación gráfica que facilita su interpretación. De esta manera es posible distinguir *puntos aislados, formación de cúmulos, relación entre variables y similaridad entre individuos.*

Dada la importancia de las representaciones gráficas en el análisis de datos, han surgido investigaciones con el objetivo de que el diseño y elaboración de esquemas dejen de ser materia de intuición y sentido común; por un lado, existen trabajos sobre el impacto que provocan los diferentes elementos básicos de percepción visual (longitudes, áreas, formas, colores, etc.) al observar una gráfica.[4] Por otro, el avance tecnológico de la computación ha significado un elemento indispensable en el progreso de la materia tanto en la precisión y velocidad de cálculo como en la elaboración de la gráfica y, hoy en día, representa un soporte trascendente en la búsqueda de técnicas aún más eficientes.

En este trabajo se exponen varias técnicas de graficación que si bien no son todas, si proporcionan una idea de su importancia y aplicación en el análisis de datos.

1.-COX,D.R. Journal Royal Statistical Society. Some remarks on the role in Statistics of graphical methods. 1978 Vol.27-1.

2.-ANSCOMBE, F.J. The American Statistician. Graphs in Statistical Analysis. 1973 Vol 27-1.

3.-CHAMBERS, KLEINER. Handbook of Statistics. "Graphical tech- niques for multivariate data and for clustering" 1982 . Vol. 2

4.-CLEVELAND,MC GILL. JASA. "Graphical perception:Theory, experimentation and application to the development of graphical methods. 1984 Vol. 79-387

Las técnicas de graficación más comunes consideran una o dos variables. Sin embargo, los problemas reales son por sí mismos multivariados por lo que se hace necesario emplear métodos cuyo objetivo es usar fundamentalmente gráficas en dos dimensiones para representar datos que tienen intrínsecamente más de dos variables. Para lograrlo, algunas técnicas usan los valores de los datos como parámetros para seleccionar o dibujar símbolos gráficos y otros generan diagramas de dispersión al reducir la dimensión de los datos por medio de la matriz original o a través de variables derivadas.

Debido a que en la literatura del análisis multivariado se dá una basta explicación teórica sobre cada técnica, en el presente se hace una pequeña alusión a sus características numéricas y se brinda mayor importancia al aspecto gráfico. En cada caso se proporcionan las referencias necesarias para ampliar la información.

En el desarrollo de este trabajo se muestran aplicaciones de las diferentes técnicas a problemas reales de diversa índole como son Salud, Educación, Contaminación, Demografía y Economía.

Esta tesis consta de cinco capítulos cuyos objetivos son :

- Describir y resaltar las características de diferentes técnicas de graficación para datos multivariados a fin de proporcionar un panorama sobre sus capacidades ante diversos problemas.
- Ilustrar el uso de éstas técnicas a partir de un conjunto de datos para valorar su utilidad en problemas reales.
-Realizar un análisis estadístico a un conjunto de datos para mostrar los atributos de cada técnica y enfatizar sus diferencias.

En el primer capítulo se desarrollan diagramas sencillos que ilustran las tendencias de distribución de los datos como son las *barras, tallos y hojas y las gráficas de caja con su respectívas extensiones.*

El segundo capítulo contiene las gráficas que permiten percibir la relación entre dos o más variables como es el caso de los *diagramas de dispersión y el despliegue de cuantiles.*

Las técnicas que ilustran datos multivariados por medio de símbolos como *soles, estrellas o caras de Chernoff* se describen en el capítulo 3.

El capítulo 4 es quizá el que contiene las técnicas más elaboradas y que requieren de mayor atención tanto en la selección de la técnica a aplicar como en la interpretación de los resultados. En él se incluyen el *Análisis de Cúmulos, Escalamiento Multidimensional, Análisis Discriminante, Análisis de Componentes Principales y Análisis de Correspondencias.*

El último capítulo es el análisis del conjunto de precios de algunos productos que forman parte de la denominada "Canasta Básica" en diferentes tiendas de autoservicio del Distrito Federal durante marzo de 1990 y abril de 1991. Aquí se hace una integración de las técnicas descritas en los capítulos previos a partir de las cuales se observan diferentes aspectos de los datos multivariados y, en estos términos, se hace la comparación entre las capacidades de cada técnica.

La estructura de los primeros cuatro capítulos es similar entre sí con cinco apartados por técnica en turno:

A. Descripción de la Técnica.-Es la presentación de la técnica.

B. Desarrollo de la Técnica.- Se explica brevemente la parte teórica del diseño de la gráfica.

C. Aplicaciones de la Técnica.-Incluye ejemplos de aplicación con sus interpretaciones.

D.Ventajas y desventajas .- Se especifican las ventajas y desventajas de aplicar la técnica.

E. Revisión Actualizada.-Contiene los esquemas generados a partir de algunas modificaciones a la idea original.

Las tablas de los datos que se analizan en este trabajo se encuentran en el apéndice A. En cada una de ellas se dá una pequeña explicación sobre los elementos y el tipo de variables que presenta cada problema, por lo que es recomendable consultarlos antes de la interpretación de resultados. Cada conjunto de datos es sometido a diferentes técnicas en distintas partes del texto empenzando con las más sencillas, de donde se obtienen resultados que pueden ser parciales en el sentido de que conforme se aplican otras técnicas a los mismos datos, lo observado puede confirmarse en todo caso, abrir nuevas posibilidades de intrepertación. Para facilitar la continuidad, se han asignado pequeños logotipos que identifican los datos en turno. De esta manera al integrar las partes correspondientes a un mismo símbolo, se obtiene un análisis para cada tabla de datos.

Para el desarrollo de este trabajo se emplearon los paquetes estadísticos SYSTAT y STATGRAPHICS. Por medio de HARVARD GRAPHICS se dió mayor resolución a algunos resultados.

En el caso de las caras de Chernoff se recurrió a tomar un ejemplo de la bibliografía por considerarse claro, completo e ilustrativo.

PLANTEAMIENTO DEL PROBLEMA

En el desarrollo de un experimento estadístico, las respuestas a las variables planteadas generan una secuencia de observaciones que forman el conjunto de datos. En esta dirección es importante distinguir entre las variables cuantitativas y las cualitativas.

El sexo, el lugar de nacimiento y el estado civil de un grupo de personas son ejemplos de variables cualitativas porque expresan una propiedad no numérica. Las variables cualitativas pueden clasificarse de acuerdo a su tipo en nominales u ordinales.

El grupo sanguíneo al que pertenece una persona es un ejemplo de variable nominal cuyas categorías están determinadas por los diferentes tipos de sangre: O, A, B, AB. El sistema psiquiátrico de agrupación por diagnóstico también constituye una escala nominal, desde el punto de vista en que un médico puede clasificar a un paciente como "esquizofrénico", "paranoico", "maniaco-depresivo" o "psiconeurótico".

En la escala ordinal, también se habla de grupos pero, los elementos de una categoría están relacionados 'ordinalmente' con los elementos de las otras. El sistema de grados en el ejército es un ejemplo claro de este tipo de escala: el sargento es superior al cabo y el cabo es a su vez superior al soldado raso. La clasificación de la sociedad en clases socio-económicas también forma parte de la escala ordinal.

Por otra parte, el número de personas que componen una familia y la edad de los integrantes de un equipo representan cantidades es decir, una propiedad numérica, por lo que se dice que son variables cuantitativas.

Las variables numéricas pueden ser continuas o discretas.Una variable continua es aquella cuyos valores posibles no tienen interrupción. Por ejemplo, el peso o la estatura de un conjunto de estudiantes. Al hablar del peso de dos personas, por ejemplo, 50 y 51 kilos, existe la posibilidad de encontrar entre el resto del conjunto a alguien que pese 50.5 Kg. o 50.25 Kg., es decir, entre 50 y 51 Kg., existen una infinidad de pesos posibles. En este sentido se dice que la variable es continua, o sea, "no se interrumpe".

Una variable discreta es aquella cuyos valores se interrumpen o separan. Por ejemplo, el total de integrantes de una familia,la cantidad de huevos recolectados en una granja al día o el número de materias reprobadas por un alumno durante un semestre.

Pensar en obtener 30.5 huevos en una granja al día no resulta lógico así como tampoco lo es expresar el número de habitantes en una población en términos de racionales. Entonces se dice que se trata de variables discretas.

Por otro lado, si en el experimento se observan dos variables x_1 y x_2 simultáneamente, se tiene como resultado un conjunto de parejas ordenadas (el número de parejas determina el tamaño de la muestra). Del mismo modo, si se trata de tres variables, x_1, x_2 y x_3, cada observación es una tercia ordenada. En general, para p variables, evaluadas a partir de n individuos, se tiene una matriz de datos de dimensión n x p.

$$x_{11}\, x_{12}\, x_{13} \cdots x_{1p}$$

$$x_{21}\, x_{22}\, x_{23} \cdots x_{2p}$$

$$x_{n1}\, x_{n2}\, x_{n3} \cdots x_{np}$$

donde el i-ésimo renglón dá cuenta de las p variables para el individuo i y la j-ésima columna indica el valor que toma la variable j en los n individuos.

El problema, entonces, es describir el conjunto de datos multivariados que forman el arreglo, empleando como herramienta descriptiva su representación gráfica en el plano cartesiano.

CAPITULO
I

I DIAGRAMAS PARA LA DISTRIBUCION DE DATOS.

Los diagramas que se presentan en este capítulo permiten distinguir *aspectos relevantes de la distribución y la densidad de los datos.*

I.1 DIAGRAMAS DE BARRAS.

A. Descripción de la Técnica.-

Suelen utilizarse con frecuencia por ser sencillos de elaborar e interpretar aún para las personas que no cuentan con alguna preparación sistemática al respecto.

La aplicación principal que se le dió a ésta técnica fue la de almacenar (con sus limitaciones) gráficamente la información. Más tarde se uso también *para representar la distribución y la densidad relativa de los datos* a lo largo de los intervalos.

B. Desarrollo de la Técnica.-

El diseño de las gráficas de barras está sujeto a una tabla de frecuencias.

Sobre el eje horizontal (vertical) se hacen particiones generalmente iguales que representan la amplitud del intervalo y que vienen a formar la base de las barras cuya altura dá cuenta de la frecuencia para cada clase. El diagrama de barras para las calificaciones de los estudiantes de Economía política (tabla 1 del apéndice) en siete intervalos es el siguiente:

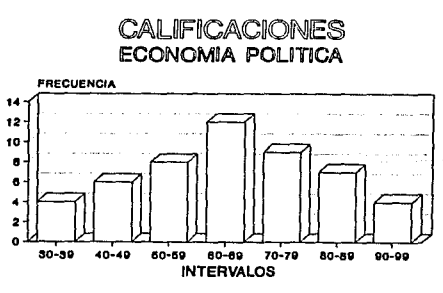


FIGURA I.1. Diagramas de barras para las calificaciones individuales de 50 estudiantes. (tabla 1)

A menudo es útil observar el diagrama sin el espacio que separa las barras. Para conseguirlo, lo más conveniente es dividir en partes iguales la distancia que hay entre los dos intervalos y ensanchar las barras. (fig.I.2)

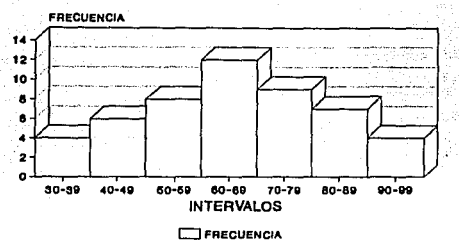


FIGURA I.2. Histograma que ilustra la tabla 1.

Para variables cualitativas, el trazo de la gráfica de barras es muy sencillo. Basta con hacer particiones iguales a lo largo del eje horizontal (vertical). Cada intervalo representa una categoría de la variable. La altura de las barras tiene el mismo significado, es decir, representa la frecuencia con que se presentó cada una de las diversas clases. (fig. I.3)

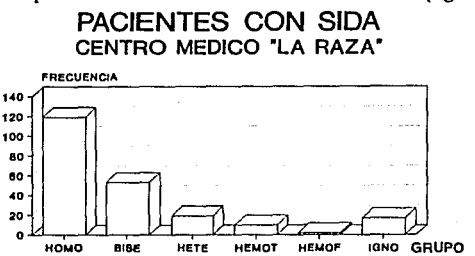


FIGURA I.3. Primeros pacientes con SIDA atendidos en el hospital La Raza. (tabla 2)

C. Aplicaciones de la Técnica.-

Las barras correspondientes a los grupos a los que pertenecen los primeros pacientes con SIDA atendidos en el Hospital "La Raza" (figura I.3), muestra de manera muy clara el predominio del grupo de homosexuales seguido por el de los bisexuales aunque, este último es escasamente la mitad de la frecuencia del primero (tabla 2 del apéndice A).

Debido a que la longitud de las barras esta apoyada en una escala común, no es difícil percatarse (imaginando líneas paralelas al eje horizontal que pasen por el borde superior de cada barra) de que los dolientes de SIDA que declararon ser heterosexuales son más que quienes se contagiaron al recibir una transfusión de sangre contaminada o que aquellos cuya fuente de contagio se ignora.

D. Ventajas y Desventajas.-

La arbitrariedad en el diseño de un histograma, es decir, en la selección del número y el tamaño de los intervalos e incluso su colocación en el eje de apoyo (si no se tiene definida

una jerarquía) lo hace una técnica con desventajas.

La imagen proporcionada por la figura I.4 es muy diferente a la observada en la figura I.2 a pesar de tratarse exactamente de los mismos datos. La agrupación en pocos intervalos impide captar características o comportamientos relevantes. De la misma manera, un número grande de intervalos produce una imagen en la que se pierden las características que tienen como grupo (fig. I.5).

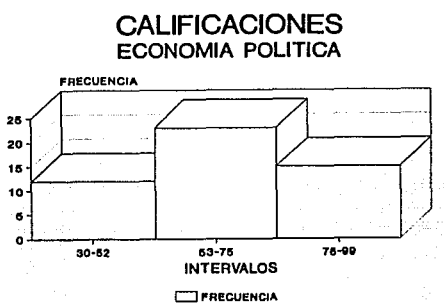


FIGURA I.4. Histograma para los datos de la tabla 1 agrupados en 3 intervalos.

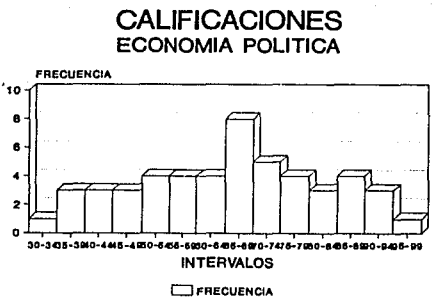


FIGURA I.5. Histograma para los datos de la tabla 1 agrupados en 14 intervalos.

Una ventaja radica en que por este medio se percibe la distribución de los datos, la tendencia de concentración y posibles simetrías. Es decir, se obtiene una idea general de esta distribución a partir de una presentesación gráfica sencilla.

E Revisión Actualizada.

Los elementos básicos del diseño de los diagramas de barras facilitan una variedad de presentaciones que incrementan las posibilidades de aplicación de ésta técnica como las que se presentan a continuación:

I.1.1 GRUPOS DE BARRAS.

A. Descripción de la Técnica.-

Con las mismas cualidades de los diagramas de barras, se pueden armar grupos de barras que permiten observar el comportamiento de una variable a lo largo de otra. Es decir, por este medio se grafícan 2 variables preferentemente discretas, aunque también es aplicable en problemas que incluyen variables continuas.

B Desarrollo de la Técnica.-

Se elaboran tantos grupos de barras como categorías tenga la segunda variable. Cada grupo dá cuenta del comportamiento de la primer variable en la categoría correspondiente a la segunda variable en cuestión.

C. Aplicaciones de la Técnica.-

En la figura I.6 se consigue representar las tendencias de consumo de drogas entre estudiantes por región durante 1986.(tabla 3). Sin duda, en las tres zonas, las drogas que más se consumieron durante 1986 fueron los inhalantes, la mariguana, las anfetaminas y los tranquilizantes, aunque sus proporciones varían de una región a otra.

Es decir, en general en el país, los inhalantes significaban el principal enervante entre los jóvenes estudiantes.

En el norte, la diferencia de proporciones entre mariguana, inhalantes y anfetaminas no es tan notable como en la zona Centro donde además, las anfetaminas tenían mayor incidencia que la mariguana.

Por otro lado, en el sur, los tranquilizantes eran la segunda fuente de drogadicción superando casi en el doble a la mariguana.

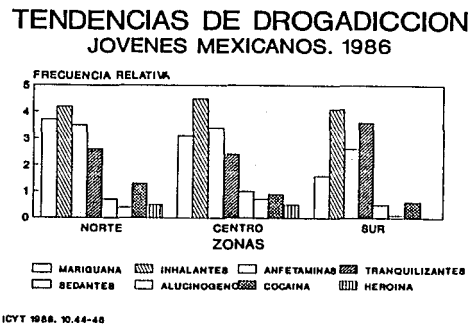


FIGURA I.6. Grupos de barras para las tendencias de drogadicción (tabla 3)

Comparando las longitudes de las barras entre zonas, *el sur se distingue por tener los índices de drogadicción más bajos* llegando incluso a no registrar casos de consumo de heroína.

C. Ventajas y Desventajas.-

Esta técnica constituye un buen recurso para la observación de dos variables en forma simultánea. Siguiendo el método de las líneas imaginarias paralelas al eje de apoyo, las frecuencias son percibidas sin dificultad.

Si se tiene un número grande de categorías en cada variable, el diagrama crece y puede llegar a ser poco ilustrativo.

I.1.2 BARRAS ENCIMADAS.

A. Descripción de la Técnica.-

Una técnica similar a la anterior, la constituyen las barras encimadas. El objetivo es el mismo: presentar el comportamiento de 2 variables. Los factores visuales relevantes también son las longitudes de las barras aunque, quizá en este caso es mejor recurrir a la comparación de las áreas de los rectángulos.

B. Desarrollo de la Técnica.-

Los intervalos de la primer variable, permanecen sobre el eje común, en tanto la segunda variable se grafica como la frecuencia para cada una de sus categorías distinguiéndose por el color o la densidad del dibujo. Cada barra formada representa la suma de los casos para cada intervalo. En la figura I.7, los tipos de droga empleados en 1986,se localizan en el eje horizontal, formando la base de las barras. En cada barra se distinguen tres densidades de dibujo diferente lo que representa las tres regiones en que se clasificó la República Mexicana.

C. Aplicaciones de la Técnica.-

Por medio de esta técnica es fácil obtener una *idea general del grado de consumo de las diferentes drogas en la República Mexicana* y, a partir de las densidades, detectar la zona con mayor o menor incidencia.

Durante 1986, las proporciones que corresponden a los inhalantes acumulan el porcentaje mayor lo que los ubica como la principal fuente de adicción en el territorio nacional. (fig. I.7).

Por otro lado, los alucinógenos y la heroína demostraron la menor demanda. Además, las *barras que representan a la zona sur, en general son las de menor dimensión en cada intervalo*, confirmándose así la observación que se hizo a partir del grupo de barras. El único caso que integra la excepción es el porcentaje de adicción a los tranquilizantes.

En cuanto a las zonas Norte y Centro, la primera tiene superioridad en lo que se refiere al consumo de mariguana, anfetaminas, tranquilizantes y cocaína. Los porcentajes en el cosumo de heroína son similares en ambas regiones.

La problemática durante 1976 era un tanto diferente (fig. I.8); por un lado, los índices de drogadicción eran inferiores. Por otra parte, los tranquilizantes eran el recurso con mayor demanda entre los jovenes, además,los alucinógenos eran mas comunes en este período que diez años después.Ya desde 1976, la zona sur se declaraba como la región con los índices mas bajos de drogadicción en la República Mexicana y su principal fuente eran los tranquilizantes.

En general, al comparar las figuras I.7 y I.8 se distingue claramente un cambio en las tendencias de drogadicción en los jóvenes mexicanos quienes en el período más reciente consumían principalmente inhalantes, mariguana, anfetaminas y tranquilizantes.

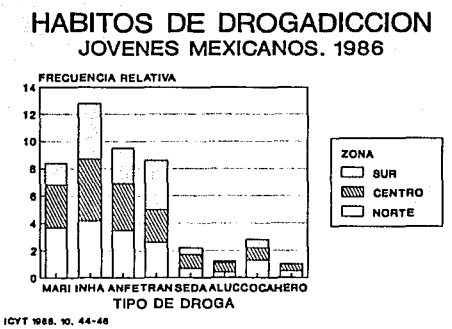


FIGURA I.7. Barras encimadas, tendencias de drogadicción. Las diferentes densidades distinguen las zonas. 1986.

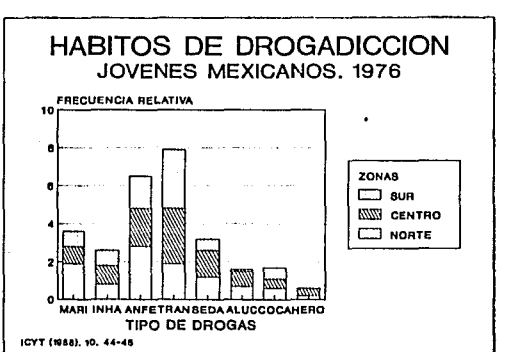


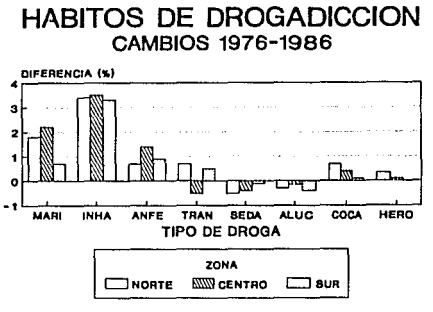
FIGURA I.8. Barras encimadas, tendencias de drogadicción. Las diferentes densidades distinguen las zonas. 1976.

Atendiendo por un momento sólo a las barras inferiores de cada gráfica se distinguen los hábitos de drogadicción de los estudiantes durante 1976. Las barras superiores corresponden a 1986. En la mayoría de los casos se registró un incremento en los porcentajes de consumo, por lo que se deduce que más que encontrar una solución al problema, éste empeoró al paso de diez años.

Por otro lado, se observa que la mariguana era más consumida en el Norte que en cualquier otra región; a lo largo de los dos períodos, en el centro se consumían básicamente inhalantes, anfetaminas y tranquilizantes; y, en la zona sur no se consumía heroína pero los tranquilizantes registraron porcentajes muy altos.

En resumen, los hábitos de consumo de drogas varía de una zona a otra.

Si fuera necesario enfatizar aún más las *diferencias en el consumo* de cada droga de 1976 a 1986,la figura I. 9, donde a través de barras se ilustran los cambios, es de singular utilidad. A partir del eje horizontal en cero, hacia arriba se grafican las *diferencias positivas* de 1986 con respecto a 1976, y, de la misma línea hacia abajo, están las *diferencias negativas.*



ICYT (1988). 10. 44-48

FIGURA I.. 9.Diagrama de barras que ilustra las diferencias en los porcentajes de las tendencias de drogadicción por región entre 1976 y 1986.

En este sentido, es fácil *confirmar* que los problemas de drogadicción entre los jóvenes estudiantes, más que disminuir han aumentado . Además se *ratifica* el incremento que se refiere al consumo de inhalantes y mariguana. Por otro lado, en el norte y centro del país, aunque la diferencia es mínima, ha logrado disminuir el porcentaje de adictos a sedantes y alucinógenos principalmente.

D. Ventajas y Desventajas.-

Al percibir un diagrama de barras encimadas se puede *detectar inmediatamente los totales para cada categoría* de la primer variable pero se pierde un poco el detalle para la segunda.

Otra ventaja radica en que la combinación de las dos últimas técnicas, tolera la exhibición de tres variables. Tal es el caso de la fig. I.10 donde además de distinguir las tres zonas, es posible *comparar* las proporciones de la muestra que consumían las diferentes drogas en 1976 y una década posterior.

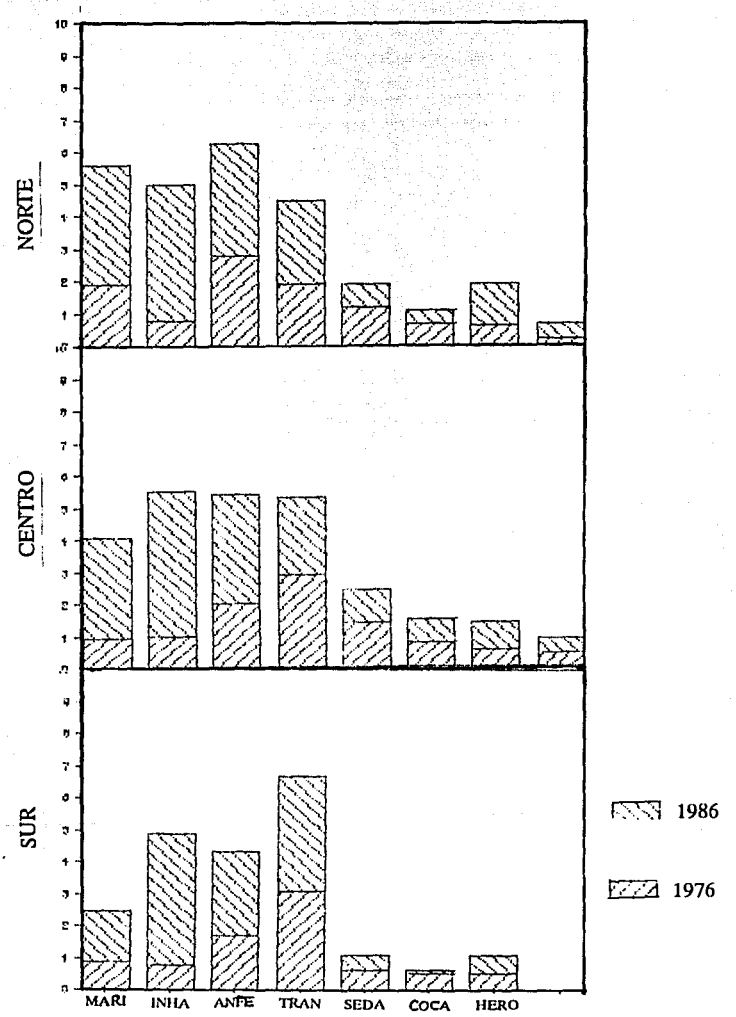


FIGURA I.10 Combinación de grupos de barras con barars encimadas. Se logra representar 3 variables.

I.1.3 HISTOGRAMA DUAL.

A. Descripción de la Técnica:-

Si se tiene que una de las variables es dicotómica, una alternativa de graficación es el Histograma Dual. Como el nombre lo indica, sólo se trata de una versión modificada a partir de un histograma. Los elementos de percepción para el lector son la longitud de las líneas y el área encerrada por las barras.

B. Desarrollo de la Técnica.-

Por su origen, el diseño de este diagrama surge de la elaboración de dos histogramas "apareados", uno por cada valor que toma la variable dicotómica. A partir de un eje vertical, el lado izquierdo representa una categoría, y el lado derecho, la otra.

C. Aplicaciones de la Técnica.-

Considerando nuevamente los datos sobre las tendencias de drogadicción en los jóvenes, la variable que se refiere al año , puede ser tomada como la variable dicotómica pues, oscila entre dos valores: 1976 o 1986. De esta manera, a partir del eje vertical hacia la izquierda, se tienen las barras que corresponden a 1976 y a la derecha se grafican los datos relativos a 1986. Al observar la figura I.11 se pueden confirmar las deducciones obtenidas anteriormente en cuanto al incremento en el consumo de drogas durante la década de 1976 a 1986, sin embargo, la interpretación sólo puede hacerse de manera global, pues no se distinguen otros factores.

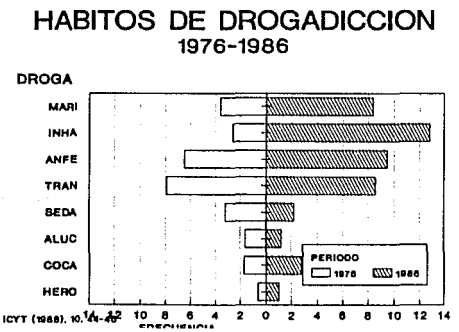


FIGURA I.11. Histograma dual. Tendencias de drogadicción 1976-1986.

D. Ventajas y Desventajas.-

Por este medio se pueden distinguir, además de la distribución de cada categoría a lo largo de los intervalos de la otra variable, una posible *simetría* entre ambas categorías. El hecho de no localizarse del mismo lado del eje de apoyo, la comparación de las barras puede verse complicada sobre todo si las diferencias entre una y otra son demasiado pequeñas

I.1.4 DOBLE HISTOGRAMA DUAL

A. Descripción de la Técnica.-

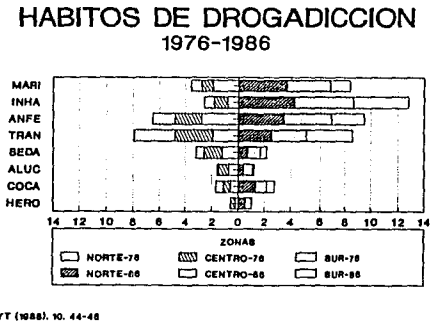
Esta técnica es una opción sencilla para *representar tres variables donde* alguna de ellas sea dicotómica. La idea básica es empalmar dos histogramas duales.

B. Desarrollo de la Técnica.-

Su diseño consiste en dos histogramas duales graficados uno sobre el otro. Se sugiere que la textura asignada a los valores de la variable que esta sobrepuesta sea más oscura que la de abajo para notar la diferencia.

C. Aplicaciones de la Técnica.-

Si se toma como base la idea principal, esta puede aprovecharse no sólo para mostrar el comportamiento de dos variables duales con respecto a una tercera. Es decir, la restricción puede reducirse a que *sólo una de las variables sea dicotómica* con lo que se llega a obtener una gráfica combinada de dos diagramas de barras encimadas colocados en lados opuestos sobre un sólo eje común. De esta manera, se logra el esquema de la Fig. I.12 donde se han combinado nuevamente las variables año, tipo de droga y región.



ICYT (1988). 10. 44-46

FIGURA 1.12. Doble histograma dual. Tendencias de drogadicción, año, tipo de droga y región.

D. Ventajas y Desventajas.-

Por medio de esta técnica se consigue una gráfica similar al histograma dual, con la diferencia que las barras están seccionadas marcando las categorías de una tercer variable. La comparación entre las distribuciones de ambas variables es más clara y sencilla. Se recomienda, al igual que en algunas de las técnicas anteriores, emplear densidades de dibujo muy distintas una de otra para evitar confusiones.

I.1.5 HISTOGRAMA TRIDIMENSIONAL

A. Descripción de la Técnica.-

El histograma en tres dimensiones tiene las mismas características que el histograma bidimensional con la diferencia de que una dimensión más dá entrada a otra variable. De esta manera, se observa la *distribución de dos variables en forma simultánea.*

B. Desarrollo de la Técnica.-

A lo largo del eje X, se hacen particiones generalmente iguales que representan la amplitud de los intervalos de la primer variable. En forma análoga, sobre el eje Y, se trazan los intervalos que corresponden a la segunda variable. Sobre el eje Z, se registran las frecuencias, dando lugar así a un diagrama de barras en tres dimensiones.

C. Aplicaciones de la Técnica.-

En la figura I.13 está el histograma tridimensional con la suma total de los índices de drogadicción para 1976 y 1986 divididos en ocho intervalos.

De nueva cuenta, la superioridad de los porcentajes que corresponden a 1986 con respecto a los de 1976 se hace evidente por medio de este diagrama.

Las barras de la fig I.13 están divididas en dos grupos, lo que habla de una gran concentración entre los porcentajes de los primeros intervalos por un lado, pero también proporciones altas para ambos períodos por el otro.

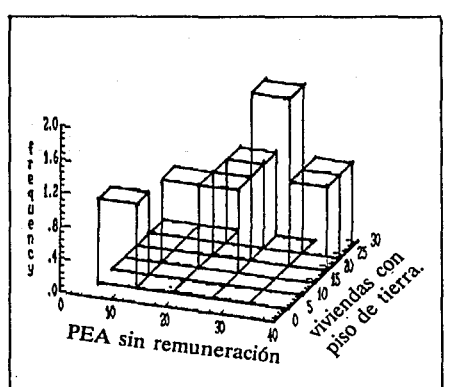


FIGURA I.13 Histograma tridimensional. Suma de los índices de drogadicción 1976-1986

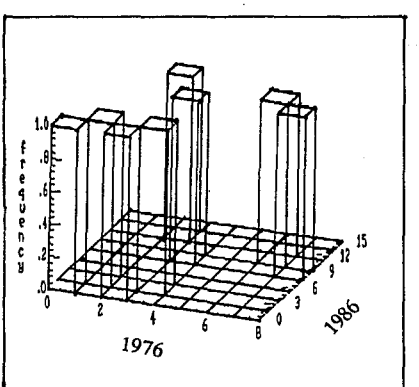


FIGURA I.14 Histograma tridimensional. Tasa de PEA que no reciben ingresos vs. tasa de viviendas con piso de tierra vs. frecuencia.

Para el conjunto de datos sobre el nivel de vida de la población del estado de Aguascalientes durante 1980, se tiene el histograma tridimensional de la figura I.14. Este incluye las variables que corresponden a los índices de población que no recibe ingresos y el porcentaje de viviendas con piso de tierra. Los datos tienden a concentrarse entre los valores 20-40 de la primer variable y 10-20 de la segunda. Además, se nota la relación entre ambas, es decir, a altos índices de PEA sin remuneración corresponde también un alto porcentaje de viviendas con piso de tierra.

D Ventajas y Desventajas.

Se puede considerar una técnica con ventajas desde el punto de vista de que representa a dos variables en forma simultánea, sin embargo, si las categorías no están claramente definidas, el diagrama es confuso y se dificulta su interpretación.

I.2 DIAGRAMA DE TALLOS Y HOJAS.

A. Descripción de la Técnica.-

Con el propósito de representar distribuciones y valores numéricos de manera simultánea, John Tukey (1977) ideó los Diagramas de tallos y Hojas que no son más que la combinación de los aspectos visuales de un histograma con la información que posée una tabla de distribución de frecuencias.

La idea fundamental es substituir las cantidades de cada intervalo de la tabla por símbolos de manera tal que las barras estén formadas por todos los valores que toma la variable.

B. Desarrollo de la Técnica.-

El principio básico es similar al de las tablas de conteo. Por un lado se tiene un renglón para cada clase y en el otro, mediante símbolos se representa la frecuencia en cada una de ellas. De esta manera, además de conocer los valores para los intervalos se puede percibir, siguiendo la figura formada por los símbolos, la distribución de los datos.

El siguiente es un ejemplo de diagramas de tallos y hojas para variables cualitativas

Estado Civil		Frecuencia
Soltero	///// ///	8
Unión Libre	///	3
Casado	///// /	6
Divorciado	//	2
Viudo	/	1

Cada renglón es un tallo y cada símbolo en un tallo es una hoja.

El desarrollo de la técnica es particularmente interesante en el caso de las variables numéricas.

A manera de ejemplo, considérese el siguiente conjunto de datos

{ 4, 0, 2, 1, 5, 2, 0, 3, 2, 3, 5, 0, 2, 2, 1 }

El diagrama es:

0	|	///
1	|	//
2	|	/////
3	|	//
4	|	/
5	|	//

Si se emplea como hoja un símbolo idéntico al dato que representa, la imagen reflejada es aún más significativa

0	000
1	11
2	22222
3	33
4	4
5	55

Cuando los datos constan de más de una cifra, puede ser confuso su empleo como símbolos

30	30
31	31,31,31
32	32,32,32
33	33,33
34	34

Por lo que en general es preferible emplear como tallo la primera parte de la cifra y la restante toma el papel de las hojas.

3	0
3	1,1,1
3	2,2,2
3	3,3,
3	4

De esta manera, al observar el siguiente diagrama , no habría confusión alguna

21	1,2,2,3,8
22	3,7,7,7
23	2,5,7,8
24	1,2,3
25	1,4

para interpretar que se tienen 18 datos, donde el menor y el mayor son 211 y 254 en ese orden o que el dato que mas se repite es 227. Cuando un tallo consta de demasiadas hojas, este puede seccionarse en dos partes. Al primer renglón del tallo se asignan las hojas con valores entre 0 y 4, por consiguiente, en la segunda parte se localizan aquellas cuyos valores oscilan entre 5 y 9.

3	8
4	0,1,1,2,4
4	5,7,8,9
5	0,0,1,1,2,4
5	5,6,7,7,7,8

El uso de la "coma" entre los valores es una ayuda para distinguir la longitud de cada una de las hojas, sin embargo, pueden suprimirse si el esquema es por si mismo claro o se hacen algunas advertencias.

Si se tiene un largo conjunto de datos que provoquen un árbol con demasiados tallos, puede recurrirse al principio de las tablas de frecuencias, es decir, agrupar los datos en intervalos.

En el espacio que separa los tallos de las hojas, se encuentran las marcas del renglón donde se localiza la mediana (M) y los cuartiles inferior y superior (H).

C. Aplicaciones de la Técnica.-

En la tabla 4 se enlistan los registros diarios de los niveles de contaminación ambiental metropolitana referentes a los 28 días del mes de febrero de 1989, proporcionados por la Secretaría de Desarrollo Urbano y Ecología (SEDUE).

Esta Institución ha dividido el área metropolitana en cinco zonas, a saber Noroeste (NO), Noreste (NE), Centro (CE), Sureste (SE),y Suroeste (SO) razón por la cuál cada día cuenta con cinco registros.Es decir, cada elemento tiene cinco variables. Además, por la naturaleza de la información, se puede ver que se trata de variables discretas.

La figura I.15 ilustra mediante diagramas de tallos y hojas la contaminación por zona registrada durante febrero de 1989. (tabla 4) Para cuestiones de interpretación, los diagramas *a, c* y *e* tiene tallos en la escala de las decenas y cada hoja representa a las unidades. En la gráfica *b*, cada raíz ocupa 2 renglones y, en la gráfica *d*, la escala es diferente, el dígito localizado en el tallo pertenece a las centenas y las hojas son las decenas, con lo que se pierde la información exacta de las unidades. De esta manera, del dato menor en la zona suroeste, sabemos que pertenece al intervalo [50 - 59] en tanto elmayor oscila entre [210 - 219]. (En la parte inferior del diagrama seaclara que son 59 y 214 respectivamente). Además, cada tallo está seccionado en cinco renglones.

Si se observan los valores máximo y mínimo para cada diagrama, no es difícil percatarse de que el noreste fue la zona menos contaminada y, en el otro extremo se localizó el suroeste llegando incluso a superar los 200 puntos que el índice marca como ligeramente molesto para la población en general.

Dirigiendo la atención a las marcas de los cuartiles (H)también pueden encontrarse aspectos interesantes. En la zona centro (c) el 50% de los datos se concentró entre los valores [71 - 98]. En el sureste (e), el rango intercuartílico es mayor [70 a 104] lo que dá una idea de mayor dispersión en estos términos, la zona Noroeste tiene datos aún más dispersos. Por otro lado, los cuartiles de la zona Noreste son menores a los cuartiles correspondientes de otras zonas por lo que persiste la idea de menor contaminación. Por su parte, en el suroeste, los datos se concentraron en niveles superiores a los cien puntos.

En el centro de la ciudad de México, los datos registrados están muy concentrados por lo que se puede pensar que no hubo varios días excesivamente contaminados y otros tantos sin contaminación. En general, el índice se mantenía sin grandes variaciones.

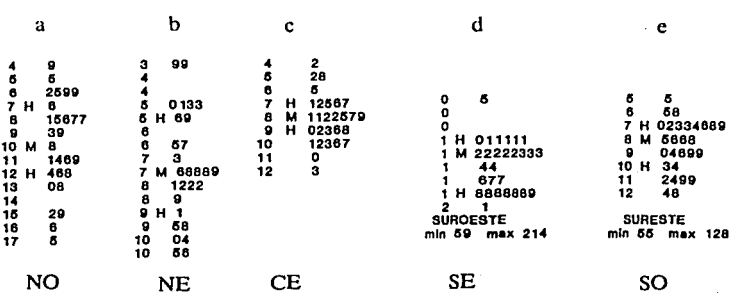


FIGURA 1.15. Diagrama de tallos y hojas. Datos de IMECA. Febrero de 1989

D. Ventajas y Desventajas.-

La longitud del rango de valores es ilustrado con claridad por los diagramas de tallos y hojas (sin olvidar las advertencias anteriores).

La tendencia de concentración de los datos se percibe con ayuda de las etiquetas de los cuartiles y el tamaño del rango.

Al igual que en los histogramas, *se detectan posibles simetrías* en la distribución de los datos.

Los valores que permanecen alejados del resto del conjunto se detectan rápidamente, como es el caso de la figura I.15 d donde todos los valores son superiores a los 100 puntos a excepción de un dato con 59 puntos.

Por la idea original de combinar las características de un histograma con la información numérica de la tabla de frecuencias, el diagrama de tallos y hojas toma ventaja sobre ambos.

El almacenamiento de datos, tarea que no siempre cubren de manera adecuada los diagramas de barras, encuentra solución en los tallos y hojas.

Una ventaja más tiene que ver con el número de dígitos empleados no sólo para almacenar los datos sino también para proporcionar de manera simultánea una representación gráfica del comportamiento de datos.

En la tabla 1 se emplearon 100 dígitos para presentar los registros sin embargo, en la figura I.16 sólo se requirieron 64.

CALIFICACIONES
ECONOMIA POLITICA

```
3    3
3    559
4    112
4    578
5 H  0234
5    5579
6    0014
6 M  55566789
7    13344
7 H  6778
8    014
8    5589
9    144
9    8

min 33   max 98
```

FIGURA I.16. Registro de calificaciones de los estudiantes de economía política.

Los elementos necesarios para la elaboración del diagrama no significan traba alguna. Lápiz y papel son suficientes en el diseño,aunque, en la actualidad los programas de computadora son muy eficaces y representan un ahorro sobre todo en lo que se refiere al factor tiempo.

La factibilidad de elaborar un diagrama dual de tallos y hojas, amplía el provecho de la técnica, pues, a las ventajas anteriores se agrega la disposición a representar dos conjuntos de datos en forma simultánea permitiendo así, la *comparación de las distribuciones* y observar la presencia o ausencia de simetría entre ambas. (fig. I.17)

IMECA
FEBRERO 1989

SURESTE		CENTRO
	4	2
5	5	28
58	6	5
02334689	H 7 H	12567
5668	M 8 M	1122579
04699	9 H	02368
34	H10	12367
2499	11	0
48	12	3

SEDUE

FIGURA I.17. Diagrama doble de tallos y hojas. Sureste-centro. febrero 1989.

En la figura I.17 se compara el diagrama de tallos y hojas para las zonas Centro y Sureste de la tabla 4.

La localización de las medianas es la misma pero la distribución de los datos de la zona centro tiende a ser simétrica, lo que no se cumple en el sureste, y, en este sentido, tampoco se puede hablar de simetría entre ambas distribuciones.

E. Revisión Actualizada.-

Una alternativa para simplificar aún más la representación de datos con varias cifras la sugiere Tuckey (1977) al usar símbolos para indicar las unidades.

Para ejemplificar simboliza $250- de las siguientes maneras:

Si se tiene una escala de $100- quedaría

2 ! 5

pero, si es de $10-, deberá aparecer un asterisco que lo indique

2* ! 5

y si se tiene valores de $1- habrá que aumentar un asterisco más

2** ! 5

Con respecto a los árboles que emplean más de un renglón por tallo, Tuckey distingue el primero del segundo (si son dos renglones) mediante un asterisco y un punto.

3.			8
4*		H	01124
4.			5789
5*		M	001124
5.		H	567778

y si son 5 renglones, se auxilia además con letras

1*			1
t			2333
f		H	445555
s			66677
.		M	88
2*			001
t			3333
f		H	445
s			6667
.			899

el asterisco es asignado para valores 0 y 1; "t" para 2 y 3; "f" para 4 y 5; "s" para 6 y 7 y . para 8 y 9.

La razón de las letras tienen que ver con la gramática del idioma inglés: "t"(two, three),"f"(four,five), "s"(six,seven). Chambers (1983) usa los dígitos entre paréntesis en lugar de las letras: (2,3), (4,5), (6,7).

Tuckey (1977) además, hace mención de algunas variaciones útiles en los diagramas de tallos y hojas.

. Si a un lado del diagrama se anota entre paréntesis el número de hojas por cada tallo, no

habrá necesidad de contar el número de símbolos porque incluye, en la parte final, un dígito que representa al total de datos.

Jaime Curts (1987), por su parte, anota en un extremo la frecuencia acumulada pero, para resaltar la posición de la mediana,contabiliza de arriba hacia abajo y en sentido contrario teniendo como tope el renglón donde esta se localiza.

Frecuencia acumulada			
5		1	00112
9		2	5678
14	M	3	02468
9		4	12345
4		5	3799

I.3 DIAGRAMAS DE CAJA.

A. Descripción de la Técnica.-

Para conocer la estructura o el comportamiento de los datos es importante tener presente los valores máximo y mínimo así como los cuartiles y la mediana.

Aunque esta información puede obtenerse por medio del conteo en el diagrama de tallos y hojas correspondiente, su representación gráfica es de singular utilidad. Los diagramas de caja resaltan adecuadamente estas características de medición, y mediante un rectángulo y dos líneas auxiliares describen el conjunto de datos.

B. Desarrollo de la Técnica.-

Como resultado del diagrama de tallos y hojas, se conoce el total de datos; la posición de la mediana y la posición de los cuartiles.

El objetivo de los diagramas de caja es ilustrar gráficamente estos aspectos por medio de un rectángulo.

Los bordes superior e inferior del rectángulo representan los cuartiles mayor y menor respectivamente. El segmento que divide al rectángulo dá cuenta de la posición de la mediana. Las líneas exteriores que parten de los bordes de la caja muestran la distancia entre el valor de los cuartiles y los datos calculados como mínimo y máximo de tal manera que la longitud que existe de extremo a extremo, representa una aproximación del rango.

Los valores extremos son calculados de la siguiente manera:

$$\text{mín} = Q(.25) - (1.5)(IQR)$$

$$\text{máx} = Q(.75) + (1.5)(IQR)$$

donde:
Q(.25) es el cuartil inferior,
Q(.75) es el cuartil superior,
IQR es el rango intercuartílico (Q(.75) - Q(.25))

Si algún punto queda fuera de los límites mín y máx, se dice que es un punto aislado, también se conoce como punto discrepante o aberrante.

Si la mediana divide el rectángulo en dos partes iguales, se refleja una *distribución simétrica en los datos* acentuándose aún más cuando las líneas exteriores son de igual tamaño.

La dispersión de los datos está dada por la longitud de la caja. El anchó del rectángulo no tiene significado particular.

C. Aplicaciones de la Técnica.-

La figura I.18 es el diagrama de caja que corresponde a los datos de la tabla 1. La posición de la mediana y la similaridad entre las longitudes de los segmentos externos, parecen ser elementos suficientes para deducir que la muestra de los 50 estudiantes, tiende a ser simétrica. Por otro lado, el borde izquierdo de la caja indica que el 25% de los alumnos obtuvo una calificación inferior a los 53 puntos mientras igual proporción logró superar los 77 puntos. la mediana de los datos es igual a 66.

Para el caso de los índices sobre las tendencias de drogadicción (tabla 3) , la caja que corresponde al período de 1976 se encuentra en la fig I.19.

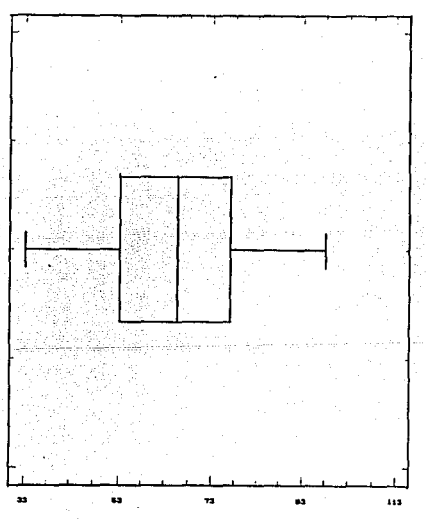


FIGURA I.18. Diagrama de caja

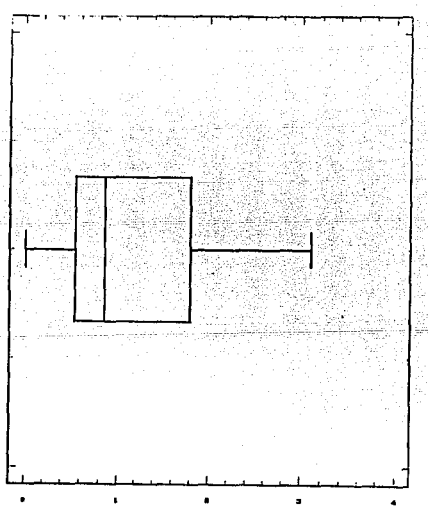
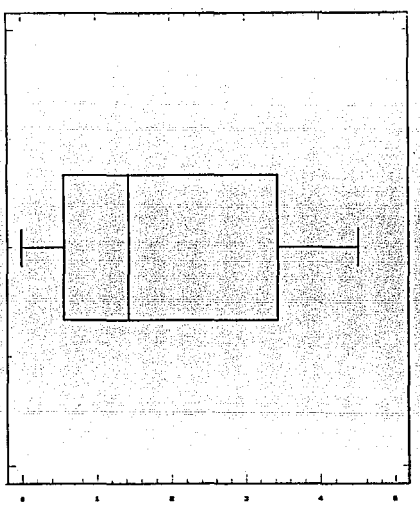


FIG. I.19. Diagrama de caja. Tendencias de drogadicción en 1976

La mediana de los índices es .85 aproximadamente y se localiza muy cerca del cuartil inferior por lo que se *deduce* que la *distribución de los datos no es simétrica*. Es decir, el 50% de los datos se concentran en un intervalo de valores inferiores al .85. El tamaño de dicho intervalos es menor con respecto al 50% restante cuyos valores se encuentrán más disapresos. El 25% de los índices superan el 1.8.

La caja para los índices de 1986 está en la figura I.20. En ella también se *refleja la asimetría de los datos*. Además, por el tamaño de la caja, se puede hablar de una *mayor dispersión* en la distribución de los índices.Por otro lado, el 25% de los datos supera el 3.5%. No se registran puntos extremos.

Con respecto a los datos de contaminación del área metropolitana durante noviembre de 1989 (tabla 5), se puede decir que la mitad se *concentraron* entre los 85 y 135 puntos; la *distribución es simétrica con respecto a la mediana*. Se registraron dos índices extremos.(Fig. I.21)



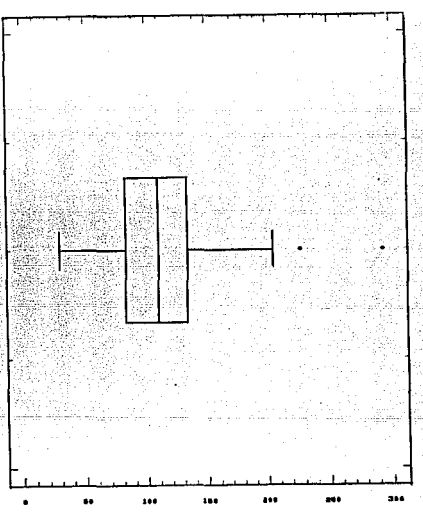


FIGURA I.20 Diagrama de caja. Tendencias de drogadicción 1986

FIG. I.21 IMECA. Noviembre 1981

En cuanto a los datos sobre el nivel de vida en Aguascalientes (tabla 7), el diagrama de caja se encuentra en la fig. I.22. En él se pueden distinguir sólo aspectos generales del comportamiento de las variables, por ejemplo, los valores de cinco de las variables son lejanos a los valores de las restantes. La mediana de los indicadores es 24 y el 75% de los índices son menores al 40, aunque más del 25% de ellos son mayores al 14.

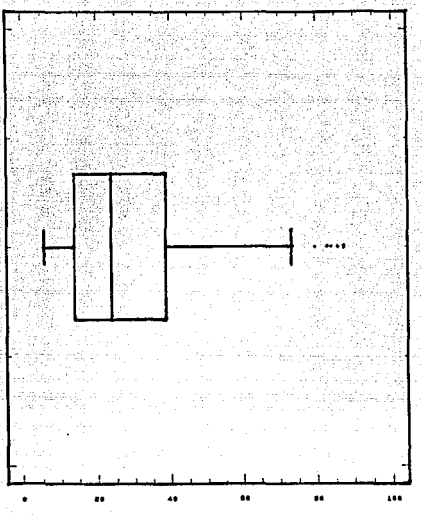


FIGURA I.22. Diagrama de caja para los indicadores del nivel de vida del edo. de Aguascalientes

D. Ventajas y Desventajas.-

Los diagramas de caja son útiles en situaciones donde no es necesario o no es posible graficar todos los aspectos de la distribución, por ejemplo, para la comparación de varias distribuciones. Esto es, *las cajas proporcionan una idea general e inmediata sobre la distribución de los datos* (sin importar el tamaño de la muestra), sin embargo, se pierde detalle en la información, por lo cual, en todo caso, sería recomendable hacer una combinación con otra técnica.

E. Revisión Actualizada.-

I.3.1 DIAGRAMA DE CAJAS MULTIPLES

La observación de las características de las variables puede ampliarse al desglosar e ilustrar la información de cada variable bajo un cierto criterio. Por ejemplo, en fig. I.23 se tiene una gráfica de las tendencias de drogadicción para cada período. En ella se observa con mayor claridad las diferencias que se establecieron apartir de las figuras I.19 y I.20

En la figura I.24 se encuentra la misma información pero clasificada por zonas. La idea de que en el sur los índices de drogadicción son los más bajos dentro de la República mexicana queda *confirmada.* La mediana de los datos de la región Norte es la más alta por lo que se plantea que en esta zona el problema es más grave.

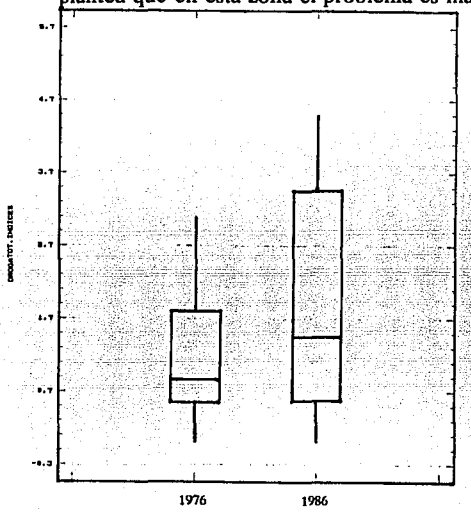


FIGURA I.23. Cajas múltiples.
Tendencias de drogadicción 1976-1986

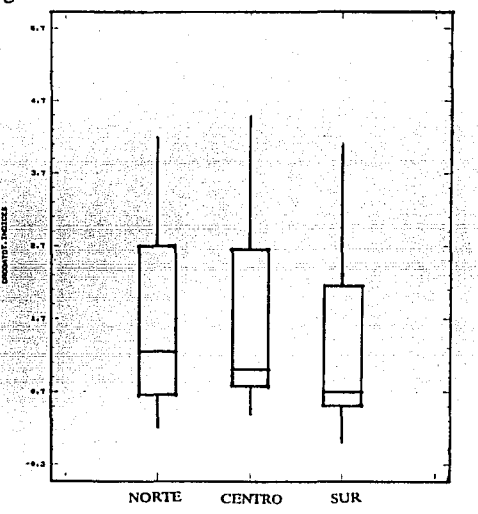


FIGURA I.24. Tendencias de drogadicción por zona

Los datos del IMECA en el área metropolitana para cada una de las zonas están en la figura I.25. Es evidente que la zona más contaminada durante febrero de 1989 fue el Suroeste, además de registrar índices muy dispersos (la longitud de la caja es mayor) en comparación con las demás zonas.

Por otro lado, se tiene que, en efecto, la región Noreste resultó ser la menos afectada y el Centro la zona con índices más homogéneos entre sí que el resto.

De esta manera, se *confirman* todas las observaciones que surgieron del diagrama de tallos y hojas (fig I.15).

Otra aplicación interesante de esta técnica es la *comparación del comportamiento* de los índices de polución a partir de febrero de 1989. en la Cd. de México. La figura I.26 ilustra la cajas para la zona centro en los períodos: del 1º al 28 de febrero de 1989, del 1º al 30 de noviembre de 1989 y del 1º al 28 de febrero de 1990 en ese orden.

Dado que los 3 casos, se tienen datos en términos de índices y, como en la elaboración de la caja no interviene el tamaño de la muestra, las cajas son comparables.

El incremento en los niveles de contaminación es muy claro pese al establecimiento del programa "Hoy no circula" (El programa entró en vigencia el 20 de noviembre de 1989). Al observar las distintas cajas en una sola escala, se facilita de manera importante la comparación y por lo tanto la obtención de conclusiones.

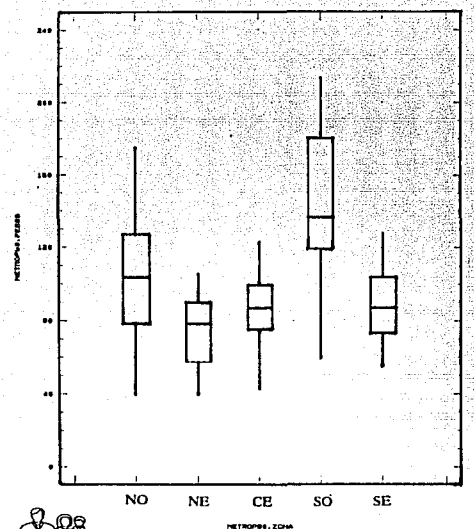


FIGURA I.25. IMECA febrero 1989

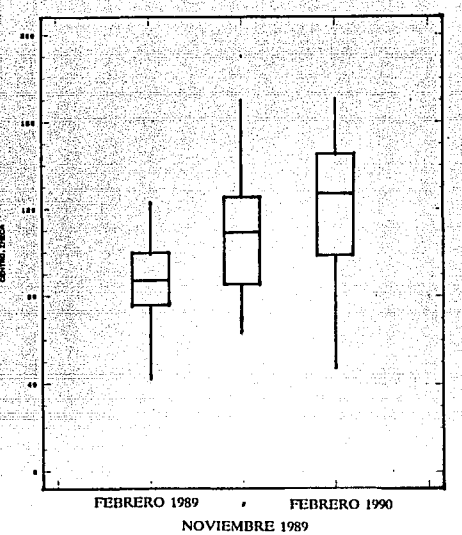


FIGURA I.26. IMECA zona centro en tres periodos

Para el caso de los índices del nivel de vida en Aguascalientes, en la figura I.27 se tiene el diagrama de cajas múltiples donde es posible observar el comportamiento de las variables por cada municipio.

Los índices están muy dispersos en San José de Gracia y en Cosío (son las cajas mas largas). En la mayoría de los casos, una variable toma valores lejanos al resto, sin embargo, por este medio no es posible su identificación. La capital del estado, manifiesta índices inferiores a los demás.

De acuerdo con el tamaño, la localización del punto lejano y la posición de las cajas, Asientos parece tener un nivel de vida semejante al de Tepezalá.

I.3.2 DIAGRAMA DE CAJAS "CORTADAS"

Una versión modificada de la técnica anterior, la constituyen las cajas 'cortadas', las cuales, en esencia conservan las características de los diagramas de caja, la diferencia se establecen que a cada caja se le adhiere una marca que corresponde al ancho de un intervalo de confianza para la mediana, mientras el ancho de la caja es proporcional a la raíz cuadrada del número de observaciones en el conjunto de datos. Los limites superior e inferior (izquierdo y derecho si la caja es horizontal) se calculan mediante la fórmula:

$M \pm 1.57(IQR / \sqrt{n})$ donde:

M, IQR, y n son la mediana, el rango intercuartílico y el número de observaciones respectivamente para cada conjunto de datos.

De esta manera, en la fig. I.28 es evidente que el intervalo de confianza para la mediana en la zona Suroeste es mayor que en las demás zonas por contar con los datos mas dispersos. Y, en la figura I.29, dicha característica le corresponde también a la caja mayor, la cual representa a los niveles de contaminación en el centro de la Ciudad durante febrero de 1990.

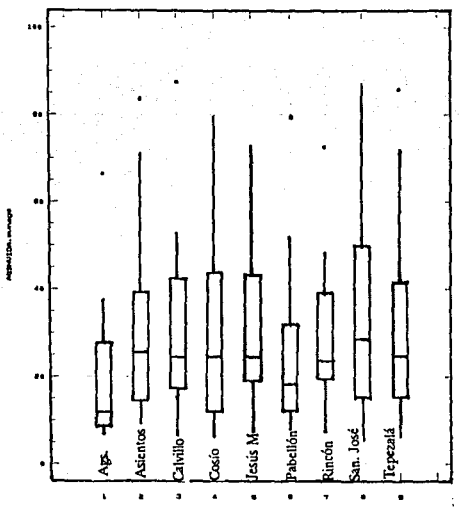


FIGURA I.27. Cajas múltiples. Nivel de vida de los municipios de Aguascalientes

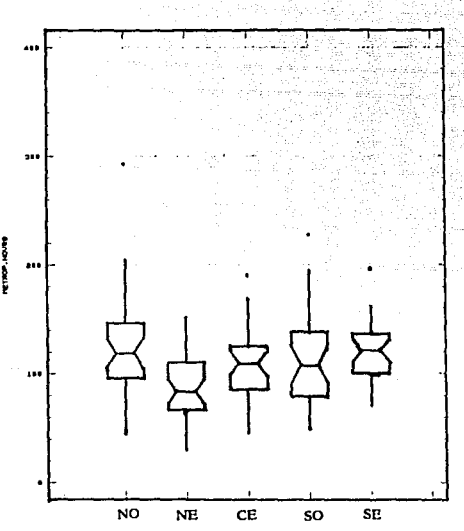


FIGURA I.28.Cajas "cortadas". IMECA noviembre 1989 por zonas

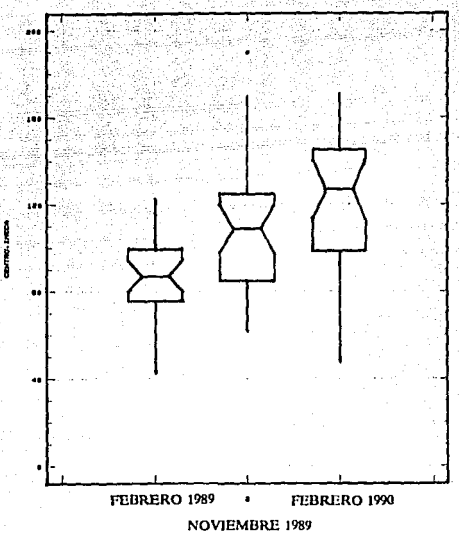


FIGURA I.29. Cajas "cortadas". Zona Centro en tres periodos

COMPARACIONES

Al comparar el diagrama de caja con el histograma del mismo conjunto de datos, y el diagrama de tallos y hojas correspondiente, se distinguen las diferentes capacidades de cada técnica, es decir, mientras las cajas provocan una idea general de la distribución de los datos; los histogramas ilustran la frecuencia en que se presentan las observaciones por intervalo, en tanto, por medio de tallos y hojas se tiene la capacidad de reproducir el conjunto original de datos. De aquí la importancia de seleccionar la técnica adecuada para cada objetivo sin olvidar que son técnicas complementarias.

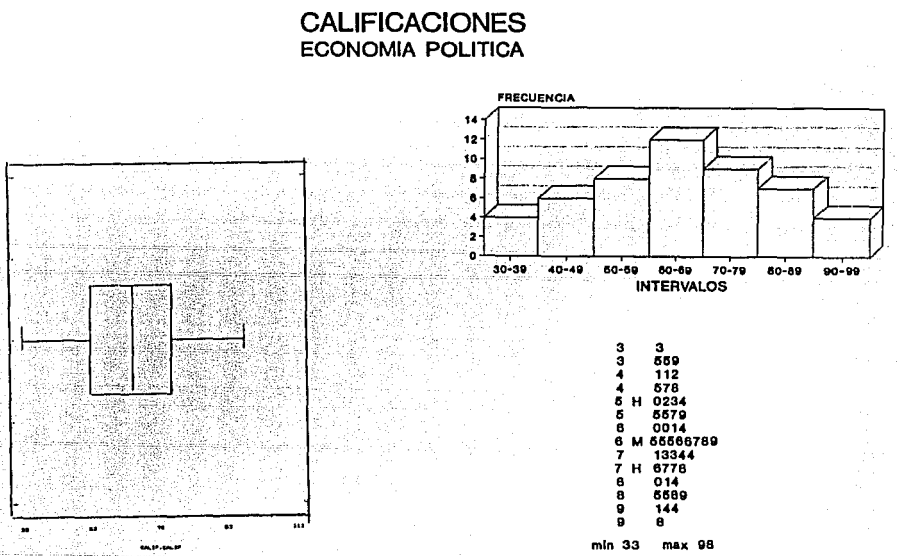


FIGURA I.30. Comparación de tres técnicas al mismo conjunto de datos: caja, histograma, tallos y hojas

REFERENCIAS
1.-ANSCOMBE, F.J. 1973 ."Graphs in Statistical Analysis" The American Statistician. Vol. 27-1
2.-CHAMBERS, KLEINER. 1982 "Graphical Techniques for multivariate data and for clus- tering" Handbook of Statistics. Vol.2
3.-CHAMBERS,J. 1983 "Graphical Methods for data Analysis"
4.-COX,D.R. Journal Royal Statistical Society. 1978" Some re- marks on the role in Statistics of graphical - methods. Appl. Statistics Vol. 27-1
5.-CURTS, J.1987 "Introducción al Análisis exploratorio de datos multidimensionale". Ciencias.
6.-MC GILL, TUKEY. 1978 "Variations of box plots" The American Statistician.. Vol.32
7.-TUKEY, J.1977 "Exploratory Data Analysis".
8.-VALLEMAN, HOAGLIN. 1981 "Aplications, basics and computing of exploratory data analysis."

CAPITULO
II

II DIAGRAMAS PARA LA RELACION ENTRE VARIABLES.

Hablar del comportamiento simultáneo de dos variables con fundamento meramente numérico puede ser riesgoso si no se cuenta con el apoyo gráfico.

Los diagramas de dispersión y de cuantiles son una valiosa herramienta en el análisis de la relación entre dos variables. A partir de algunas modificaciones, se logra una serie de esquemas que amplían la utilidad de estas técnicas.

II.1 DIAGRAMAS DE DISPERSION.

A. Descripción de la Técnica.-

Elaborar un diagrama de dispersión tiene que ver con la búsqueda de una simple relación entre variables.

Cuando el valor de una variable depende de el valor que tome la otra (independiente) se dice que existe una relación de dependencia. La variable independiente también se conoce como factor y la dependiente es la respuesta.

En un diagrama de dispersión, el eje de las abscisas se asigna al factor y el eje de las ordenadas corresponde a la respuesta.

B. Desarrollo de la Técnica.-

La elaboración de un diagrama de dispersión individual consiste en graficar un plano cartesiano de dos dimensiones, recordando que si existe una relación de dependencia,la variable factor debe localizarse sobre las abscisas y la respuesta sobre las ordenadas. En caso contrario, el orden de las variables es indistinto.

El diagrama de dispersión contiene a todas las parejas de la forma (x_i , y_i) con $i = 1,2,3,...,n$ donde n es el tamaño de la muestra.

En la figura II.1 se tiene el diagrama de dispersión que forman el número de hombres contra el número de mujeres por grupo de edad durante 1980 para la población de Guanajuato.

C Aplicación de la Técnica.-

En el ejemplo de la población de Guanajuato por grupo de edad y sexo durante 1980, es clara la relación entre las variables, es decir, el comportamiento para ambos sexos tiende a ser directamente proporcional aunque , al trazar una recta que parte del origen con pendiente uno, ($x = y$), se observa que en algunos intervalos, la diferencia entre el número de hombres y mujeres es más notoria que entre otros.

Los puntos que quedan por arriba de la recta, corresponde a aquellos intervalos de edad donde el número de mujeres es superior que el número de hombres; y, los puntos que se encuentran por debajo de la recta tienen interpretación contraria.

Al asignar una etiqueta a cada grupo de edad, se obtiene la figura II.2 donde ahora se identifica con claridad que durante 1980, la población masculina con 14 años o menos, diferia poco numéricamente con la femenina en los mismos grupos de edad, pero, a partir de los 15 años en adelante, el número de mujeres era mayor al de sus compañeros notándose una tendencia a igualar proporciones conforme se avanza en los grupos de edad.

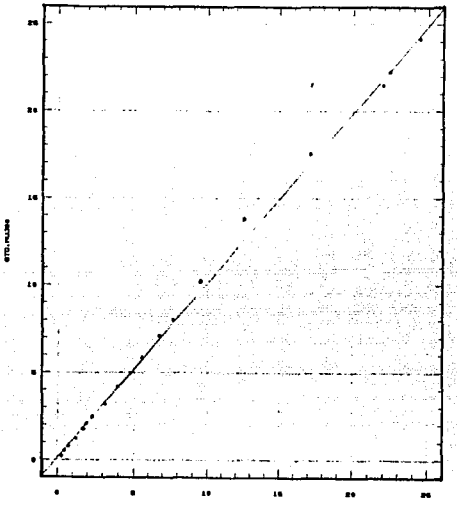


FIGURA II.1. Diagrama de dispersión. Censo de Guanajuato 1980.

EDAD
GUANAJUATO
0 - 4
5 - 9
10 - 14
15 - 19
20 - 24
25 - 29
30 - 34
35 - 39
40 - 44
45 - 49
50 - 54
55 - 59
60 - 64
65 - 69
70 - 74
75 - 79
80 - 84
85 y más
NO ESPECIFICADA

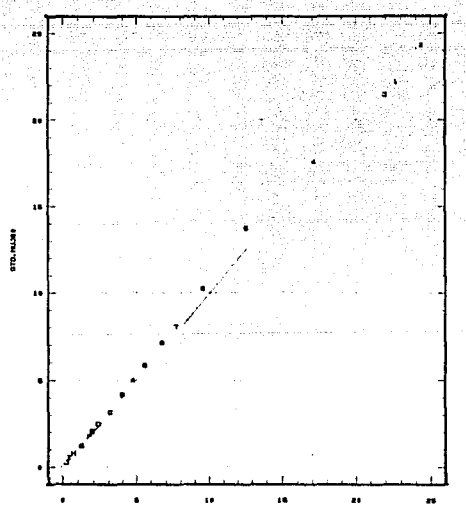


FIGURA II.2. Diagrama de dispersión con etiquetas para las variables. Hombres 1980 vs. Mujeres 1980. Gto.

La situación demográfica por grupo de edad y sexo durante 1970 se ilustra en la figura II.3 donde se distingue una relación semejante a la anteriormente descrita, con la salvedad de que en el intervalo de 5 a 9 años, el número de hombres era proporcionalmente menor al del grupo femenino.

El diagrama de dispersión considerando los datos de la contaminación, se muestran en la figura II.4 ahí se localizan los puntos que corresponden a los índices de la zona centro durante febrero de 1990 (eje X) contra los del mismo mes, un año antes (eje Y).

Dado que la mayoría de los puntos permanecen en la parte inferior a la recta identidad, se plantea que la suposición obtenida en la figura I.26 (a partir del diagrama de cajas múltiples) con respecto a la superioridad de los niveles de polución en la zona centro durante febrero de 1990, es cierta.

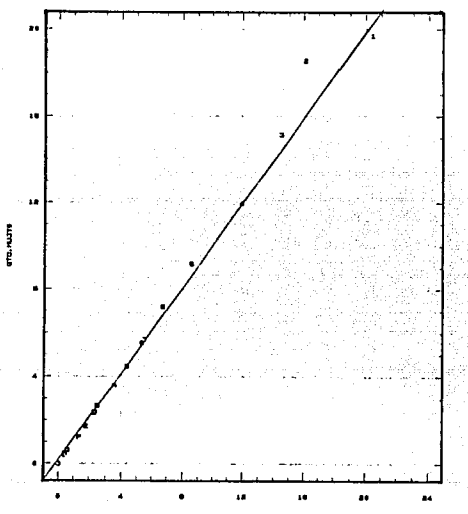


FIGURA II.3. Diagrama de dispersión con etiquetas para cada grupo de edad Guanajuato 1970.

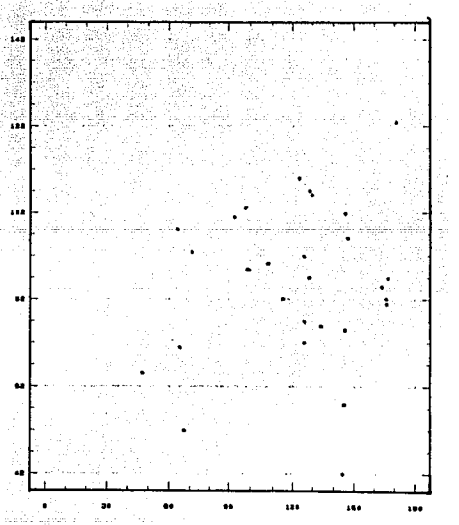


FIGURA II.4. Diagrama de dispersión. Zonas centro en 1989 y 1990.

La figura II.5 contiene el diagrama de dispersión con los respectivos índices en el área metropolitana que se registraron en febrero de 1989 y 1990.

En la gráfica se observa una fuerte concentración de datos entre los intervalos de 40 a 160 para 1989 y de 40 a 200 un año después.

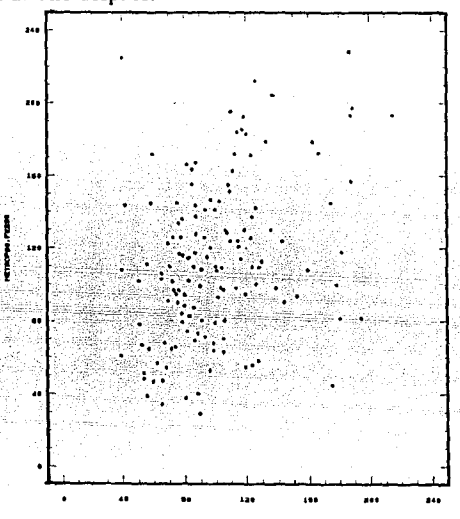


FIGURA II.5. IMECA. Febrero 1989 vs. febrero 1990

Chambers (1983), hace una combinación de diagramas de dispersión y gráficas de caja para cada variable. De esta manera, además de percibir la relación, se observan las tendencias de distribución individual de las variables. (figura II.6)

D Ventajas y desventajas.-

Por medio de esta técnica, es sencillo obtener una idea visual sobre la existencia o ausencia de la relación entre el comportamiento de dos variables y el sentido de ésta.

Es posible detectar puntos aislados que no se distinguen a través de cálculos numéricos y que pueden distorsionar los resultados provocando conclusiones equivocadas.

La idea básica del diagrama de dispersión permite hacer algunas modificaciones para conseguir gráficas aún más significativas y ampliar el número de variables a representar.

Cuando el número de datos a graficar es muy grande, el diagrama de dispersión muestra puntos encimados con lo que se pierde el detalle de cada observación pero se mantiene la idea general de la relación.

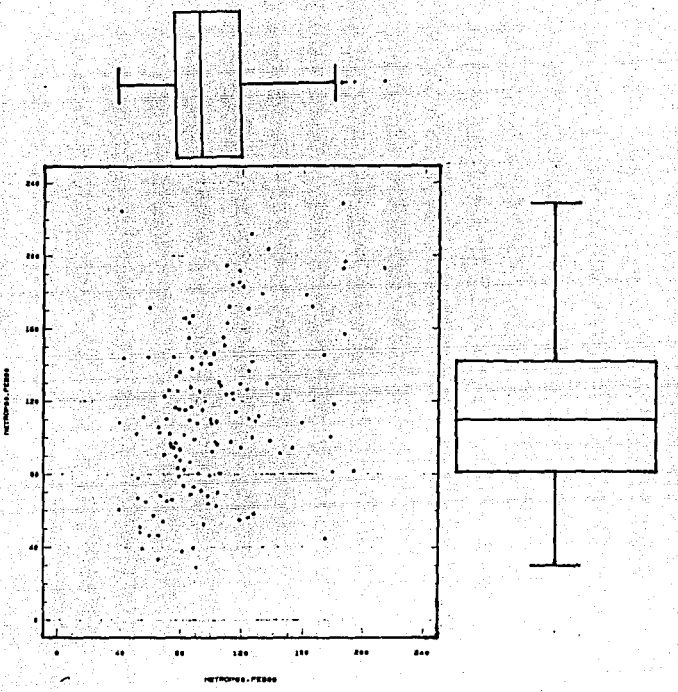


FIGURA II.6.Diagrama de dispersión con cajas.

E Revisión Actualizada.-

II.1.1 DIAGRAMAS DE DISPERSION CON SIMBOLOS.

A Descripción de la Técnica.-

La creatividad de los investigadores y sus diferentes necesidades ha generado una serie de esquemas útiles que conservan la idea de los diagramas de dispersión. Entre estas innovaciones se encuentra el diagrama de dispersión con símbolos.

B Desarrollo de la Técnica.-

Se tiene nuevamente un plano cartesiano que contiene las parejas de la forma (x_i , y_i) para $i = 1, 2, 3, \ldots, n$ donde n es el total de datos. Una tercer variable es involucrada en este

diagrama por medio de un símbolo, esto es, se asigna un símbolo a cada punto del plano cartesiano que corresponde a los distintos valores que esta toma.

Los símbolos pueden tener la forma de un figura geométrica de tal manera que su área está en función de los valores a representar.

Así, al interpretar la gráfica, se deben tomar en cuenta tres aspectos:

-la posición del símbolo con respecto al eje X,

-la posición del símbolo con respecto al eje Y,

-el área o tamaño del símbolo.

El área de una figura geométrica es una variable continúa por lo cual, ésta técnica no tiene problema alguno para representar variables también continúas. Sin embargo, el impacto visual de las áreas puede resultar ambiguo en el caso de valores similares. Esto es, la comparación no es tan inmediata.

La asignación de símbolos a valores discretos es muy simple, a cada intervalo corresponde un símbolo distinto. Si una variable continua se clasifica en intervalos, puede tener el mismo tratamiento que las discretas aunque con esta medida se pierde el detalle del valor individual.

C Aplicación de la Técnica.-

En la figura II.7 se tiene el diagrama de dispersión que incluye los datos del IMECA registrados en la zona Noroeste y Sureste y Centro durante febrero de 1990.

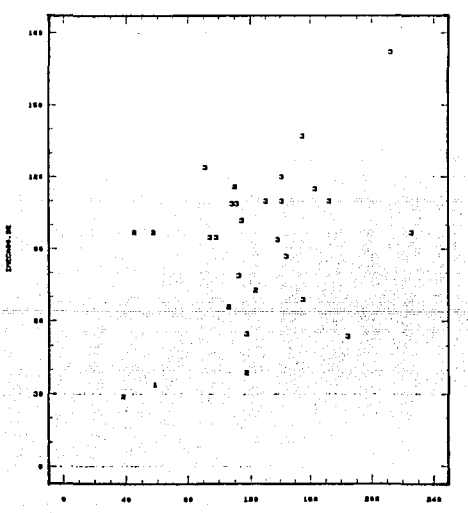


FIGURA II.7. Sureste vs. noroeste. Febrero 1990. La tercer variable (centro) está representada por los símbolos

La primera y segunda variable están graficadas a lo largo de los ejes coordenados y los valores que corresponden a los niveles de polución en el centro quedan representados con dígitos.

Cada dígito corresponde a cada intervalo de la clasificación que elaboró la Secretaría de Desarrollo Urbano y Ecología. Esto es:

En esta gráfica, es clara la mayoría para el símbolo 3. Este hecho, provoca la idea de que en general, en la zona Centro, los niveles de contaminación oscilaban entre los 101 y 200 puntos. Por otro lado, los datos tienden a concentrarse entre los valores 80-200 para el noroeste y 30-150 en la parte este del sur de la ciudad quedando fuera solo algunos puntos aislados.

D. Ventajas y desventajas.-

Una ventaja de la técnica es que se logra la representación de tres variables. Sin embargo, la comparación entre áreas puede ser confusa. Además, si se dá el caso de varios puntos sobrepuestos, la observación se complica aún más.

II.1.2 DIAGRAMA DE DISPERSION EN TRES DIMENSIONES

A. Descripción de la Técnica.-

Por medio del diagrama de dispersión en tres dimensiones, también se logra comparar el comportamiento de tres variables simultáneamente.

B. Desarrollo de la Técnica.-

De nueva cuenta se tiene un plano cartesiano que contiene a los puntos de la forma (x_i, y_i) con $i = 1, 2, 3, \ldots, n$; $n =$ tamaño de la muestra. La tercer variable se representa por medio de la altura del punto a partir del plano. Esto es, a cada punto se le proporciona una altura que corresponde al valor que toma la tercer variable. De esta manera, se dice que se tiene un espacio euclidiano y los puntos tienen la forma (x_i, y_i, z_i) para $i = 1, 2, 3, \ldots, n$.

C. Aplicaciones de la Técnica.-

En la figura II.8, se encuentra el diagrama de dispersión en tres dimensiones para los datos del IMECA de las zonas Centro, Sureste, Noroeste en febrero de 1990. Es decir, esta es otra forma de representar los datos del diagrama II.7 de donde se obtienen las mismas conclusiones.

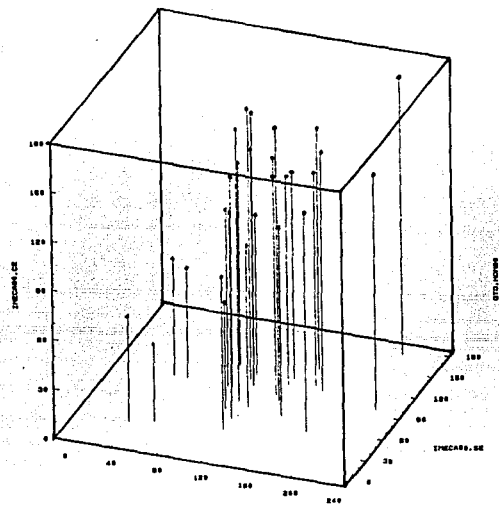
Con respecto al ejemplo del censo de la población del Edo. de Guanajuato, en las figuras II.9 y II.10 están los diagramas de dispersión que muestran el comportamiento por sexo y grupo de edad a través de tres décadas.

Al observar únicamente el plano (x,y) de la figura II.9 se compara el comportamiento de la población masculina entre 1960 y 1970. Se distingue una tendencia lineal con pendiente menor que uno , lo que significa que el número de hombres que había en 1970 era proporcionalmente menor al que reportó el censo de una década anterior aunque en 1980 tuvo un ligero incremento.

La situación en el caso del sexo femenino era diferente, pues, el incremento de 1960 a 1970 fue proporcional pero, para 1980 se registraron incrementos en los intervalos de 5 a 24 años. (figura II.10)

Por otro lado, en los dos primeros períodos, la población en ambos sexos tendía a disminuir

a partir de un máximo registrado en el intervalo de 0 a 4 años, pero, en 1980, este máximo se registró en un grupo de edad posterior (5-9 años), es decir, el número de nacimientos en 1980 disminuyó con respecto a 2 décadas anteriores.



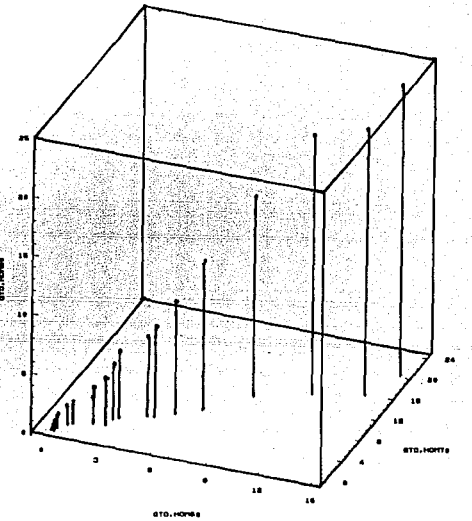


FIGURA II.8. Diagramas de dispresión tridimensional. FIGURA II.9. Población masculina en tres decadas. Gto. Noroeste vs. soreste vs. centro. Febrero 1990.

En la figura II.11 se tiene el diagrama de dispersión en tres dimensiones para los datos que corresponden a los municipios Calvillo, Aguascalientes y Asientos. En esta gráfica se distinguen claramente la formación de tres grupos de variables, uno de ellos contiene sólo un elemento. Más adelante (IV) se verá que en cada grupo se encuentran las variables que se correlacionan entre sí.

D Ventajas y Desventajas.-

Por medio de esta técnica, se analiza el comportamiento de *tres variables en forma simultánea.*

La observación en el espacio tridimensional puede ser confusa.

Si el número de variables es grande, se tendrían que elaborar varios diagramas, lo que dificulta la observación general del desarrollo de las variables.

Si hay muchos puntos, se tiene el riesgo de que queden puntos sobrepuestos con lo que se pierden detalles en la gráfica.

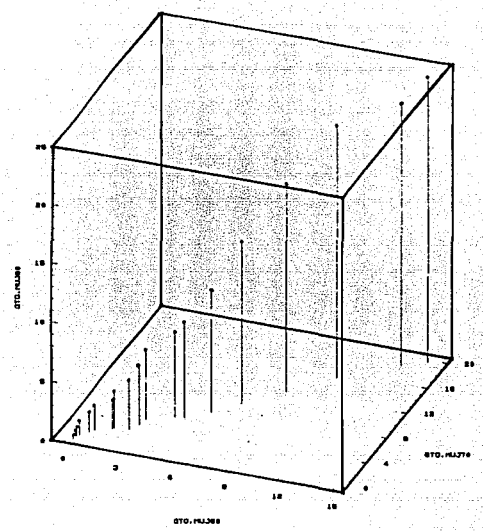


FIGURA II.10. Población femenina en Guanajuato a través de tres decadas.

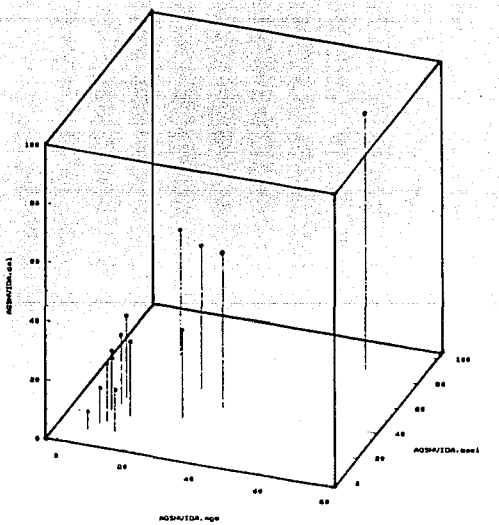


FIGURA II.11. Diagrama de dispersión en tres dimensiones Aguascalientes vs. Asientos vs. Calvillo.

II.1.3 DIAGRAMAS MULTIPLES DE DISPERSION.

A. Descripción de la Técnica.-

El arreglo escalonado de varios diagramas de dispersión provée al analista de una visión general de la relación entre variables de tal manera que se puede hacer una selección inicial de acuerdo al interés del estudio.

B. Desarrollo de la Técnica.-

Los diagramas múltiples de dispersión se caracterizan por ser un arreglo de diagramas de dispersión donde cualquier par de gráficas adyacentes tienen un eje común . En la figura II.12 esto significa que buscando los "cruces" en el último renglón, se tiene la variable que corresponde a la población femenil durante 1960 contra el número de hombres en 1980, 1970, 1960 y mujeres en 1980 y 1970 en ese orden para el censo de Guanajuato (tabla 7).

C. Aplicación de la Técnica.-

La figura II.12 contiene el diagrama múltiple de dispersión para los resultados del censo de población por sexo y grupo de edad en Guanajuato. este esquema incluye todas las combinaciones posibles entre variables. de hecho, la figura II.1 forma parte de este diagrama (1,3).

Para el ejemplo de los índices de contaminación ambiental en febrero de 1990, la figura II.13 muestra el arreglo triangular de todas las combinaciones posibles que forman un diagrama de dispersión.De esta manera, se observan las relaciones entre las diferentes zonas; por un lado, la gráfica (3,2) habla de que el Centro y Suroeste son las regiones que durante ese período tendían a tener cierto grado de similitud, lo cual no sucedía entre cualquier otro par de zonas. Por otra parte, es claro que tanto el Norte, tanto en su área este y oeste como el Sureste,registraron índices inferiores a los correspondientes a las zonas Suroeste y Centro de la Ciudad.

El arreglo completo de los diagramas de dispersión está dado en la figura II.14. Las gráficas del triángulo superior son las mismas que las que forman el triángulo inferior, con la diferencia de que los ejes son inversos.

La figura II.15 muestra el panorama general del comportamiento de todas las variables del nivel de vida en los nueve municipios de Aguascalientes. Así entonces, al observar el conjunto total, se puede confirmar que los municipios con mayor relación entre sí son Tepezalá y Asientos (2,1) tal como se había notado a partir de las cajas múltiples.(fig. I.27). Además, Cosío refleja cierta relación con ambos.

D. Ventajas y Desventajas.-

La utilidad del arreglo escalonado radica en la sencillez de cada uno de los diagramas de dispersión tanto en su elaboración como en su interpretación. Además, envuelven directamente a las variables originales por lo que no se tiene dificultad para comprender transformación alguna.

La desventaja del método está en lo problemático que resulta inferir, desde las gráficas, patrones para más de dos variables. Por otra parte, el número de esquemas a elaborar es impráctico para cuestiones con una elevada cantidad de variables.

Este último punto puede encontrar solución parcial si se decide graficar solo algunos de los pares de variables. La selección adecuada depende de los objetivos a cubrir y las características

mismas del problema. Al escalonar las gráficas simbólicas, es decir, combinar dos métodos, se consigue aumentar la cantidad de variables ilustradas.

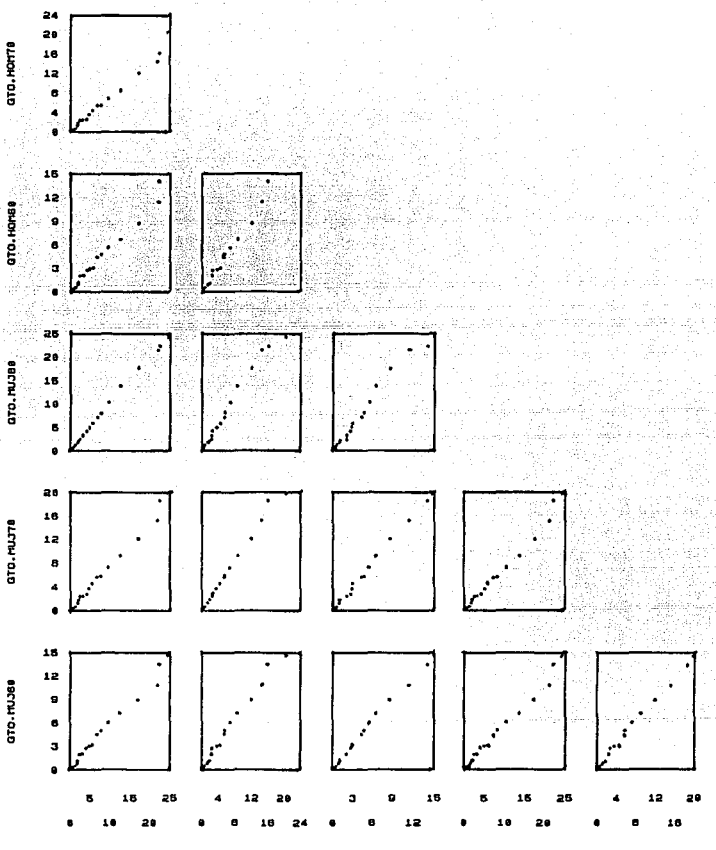


FIGURA II.12. Diagrama múltiple de dispersión. Población de Guanjuato.

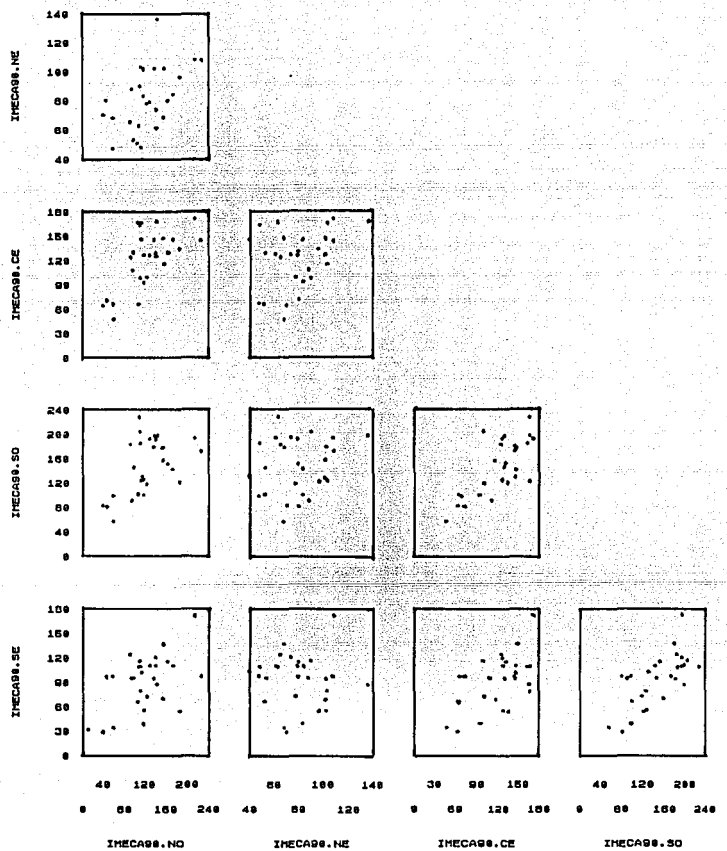


FIGURA II.13. Diagrama multiple de dispersión. IMECA febrero 1990.

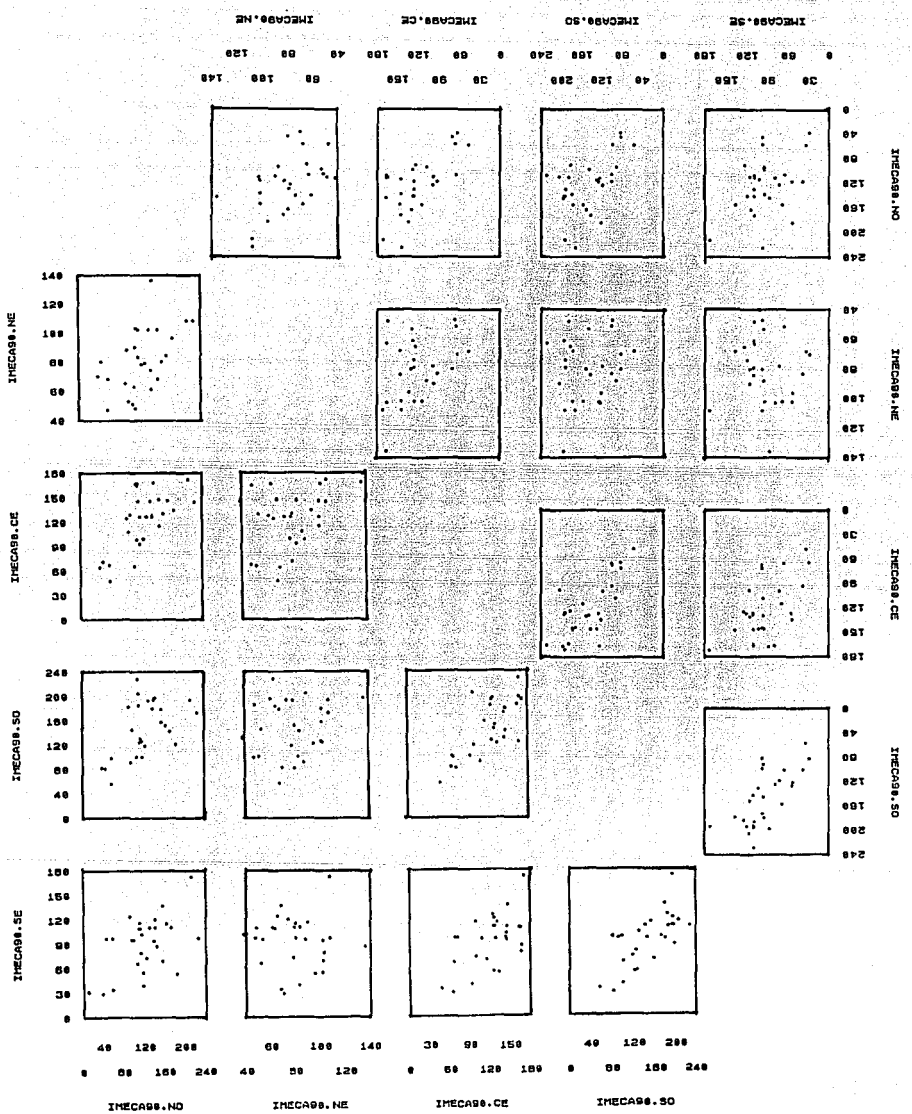


FIGURA II.14. Diagrama multiple de dispersión Completo.

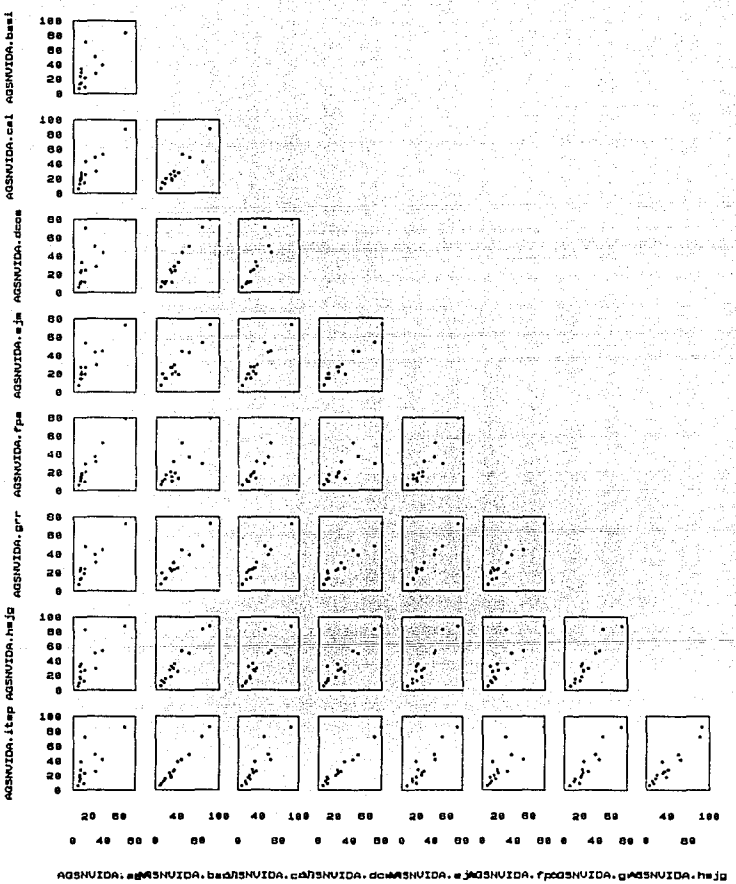


FIGURA II.15. Diagrama múltiple de dispersión. Municipios de Aguascalientes

II.1.4 PARTICIONES.

A Descripción de la Técnica.-

Por medio del diagrama múltiple de dispersión se logra un panorama general de la relación entre variables. La técnica de particiones tiene una tarea opuesta, es decir, hace un "acercamiento" a un diagrama de dispersión. Se incluyen tres variables que pueden ser discretas o continuas.

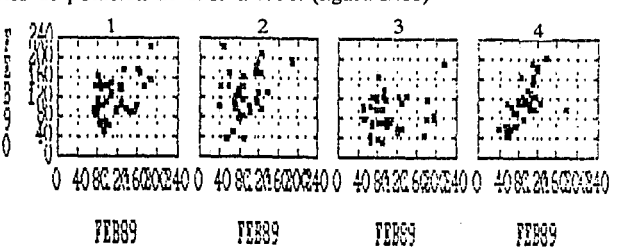
B Desarrollo de la Técnica.-

La técnica está basada en la partición de las n observaciones en subconjuntos de acuerdo a los valores de una tercer variable y, entonces, por cada subconjunto se hace una gráfica separada, de esta manera, lo que originalmente era un diagrama de dispersión para dos variables puede ser aprovechado como un grupo de esquemas para tres variables.

Por ejemplo, en la figura II.16 se tiene el diagrama de dispersión para los datos de contaminación en febrero 89 contra los correspondientes a febrero 90 dividido en cuatro particiones, uno por cada semana del mes. La observación es más clara pues, cada diagrama tiene menos puntos. Para el caso de variables continuas, se hacen las particiones de acuerdo a los intervalos o bandas de valores de manera similar al procedimiento en los histogramas.

C Aplicaciones de la Técnica.-

Durante la segunda y la cuarta semana de febrero, fue más evidente el incremento de los niveles de polución de 1989 a 1990. (figura II.16)



PARTICIONES
1.- Del 1 al 7 de febreo
2.- Del 8 al 14 de febrero
3.- Del 15 al 21 de febrero
4.- Del 23 al 28 de febrero

FIGURA II.16. Particiones.

El la figura II.17 se encuentra el diagrama de dispersión Noroeste vs. Centro de acuerdo a las particiones en los datos de la zona Suroeste.

La figura II.18 contiene la gráfica de particiones para las variables del nivel de vida en los municipios de Asientos y Tepezalá en base a los datos de Cosío.

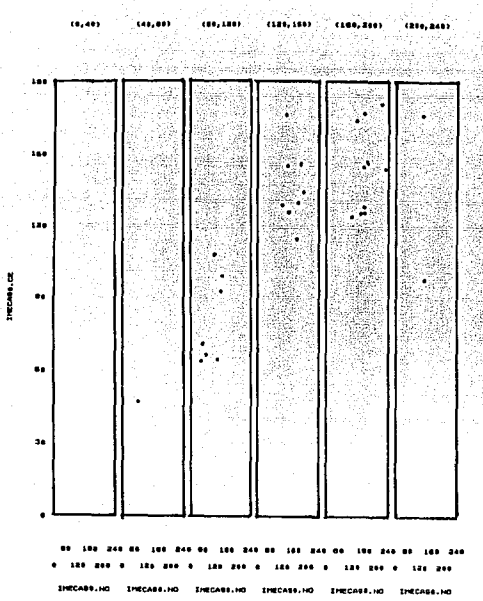


FIGURA II.17. Centro vs. noroeste en particiones de la variable suroeste.

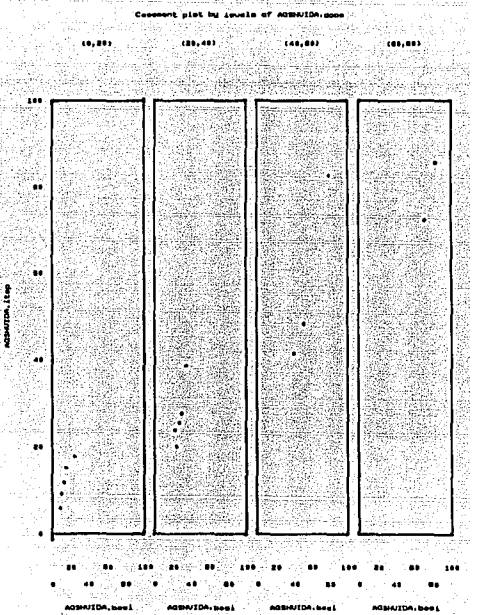


FIGURA II.18. Particiones. Asientos vs. Tepezalá vs Cosío

D. Ventajas y Desventajas.-

Si un diagrama de dispersión tiene varios puntos sobrepuestos, la técnica de particiones es un buen recurso para observar la estructura con mayor detalle.

Es importante seleccionar un número adecuado de particiones pues, si es pequeño, puede no ser suficiente pero, por otro lado, un número grande provoca muy pocos puntos en cada diagrama.

Las escalas para los ejes deben ser idénticas en todos los diagramas componentes.

II.1.5 VENTANAS MULTIPLES.

A. Descripción de la Técnica.-

Con el mismo principio de las particiones, surge una nueva técnica que permite la representación de cuatro variables.

B. Desarrollo de la Técnica.-

Haciendo particiones para ambas variables, se tendrían tantos diagramas como el producto de intervalos de cada variable. En cada recuadro se tiene ahora un pequeño diagrama de dispersión con menor numero de observaciones, lo que facilita su análisis.

Si a este nuevo conjunto de diagramas se le aumenta en los márgenes superior y derecho, las gráficas correspondientes a otras dos variables, se tiene entonces, la representación de cuatro variables al mismo tiempo.

C. Aplicaciones de la Ténica.-

La figura II.19 es la gráfica de ventanas para los datos IMECA de febrero de 1990. Las variables Suroeste y Sureste fueron divididas en tres intervalos cada una por lo que generan nueve ventanas. Algunas de ellas permanecen vacías porque no existen puntos que satisfagan ambas condiciones, es decir, no hubo un día, a lo largo del período, que registrara entre 0 y 60 puntos para la zona sureste y de 160 a 240 en la parte oeste del sur de la ciudad.

El renglón adicional de la parte superior contiene los diagramas de dispersión para las variables Centro vs. Noroeste divididas de acuerdo a las particiones del Suroeste. De manera análoga, la columna adicional de la derecha forma el despliegue de las particiones del diagrama Centro vs. Noroeste conforme los intervalos de la variable Sureste. La gráfica de la esquina superior derecha es la superposición de todas las ventanas. Es la gráfica completa de Centro contra Noroeste para las zonas Suroeste y Sureste.

En la figura II.20 está el diagrama de ventanas múltiples para cuatro municipios del Estado de Aguascalientes.

D Ventajas y Desventajas.-

Si se tiene particular interés en una variable, es conveniente elaborar diagramas a través de los cuales se pueda observar más de cerca el comportamiento de las unidades en cada una de las categorías que comprende.

El diagrama tiene capacidad para graficar cuatro variables. El problema esta en la selección de las dos variables que serán fraccionadas y cual de las dos variables restantes será asignada al eje x y cual al eje y.

Si el número de puntos es pequeño, pueden quedar muy pocos datos en cada ventana por lo que el diagrama sería poco informativo.

Este diagrama es particularmente útil para detectar variables que se relacionan por "partes".

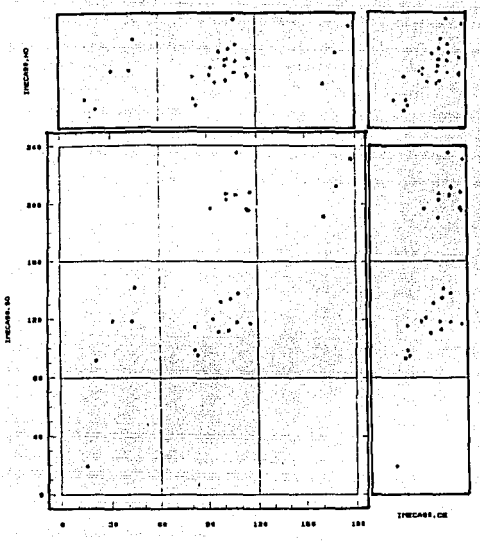


FIGURA II.19. Ventanas multiples. Se logra la representación de cuatro variables.

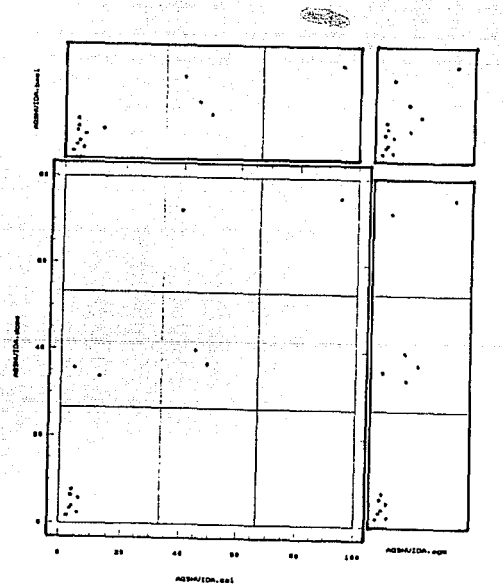


FIGURA II.20. Ventanas multiples. Asientos vs. Cosío vs. Calvillo

II.2 DIAGRAMAS DE CUANTILES.

A. Descripción de la Técnica.-

Los cuartiles y percentiles son medidas asociadas a la mediana puesto que se basan también en la posición que ocupan en una serie de observaciones.

Hay 99 percentiles que dividen a un conjunto de observaciones en 100 partes iguales.

Los cuartiles son aquellos que dividen el conjunto de datos en cuatro partes iguales de acuerdo a su posición, esto es, el cuartil inferior delimita el 25% de las observaciones de la misma manera en que el cuartil superior marca el 75% y la mediana, por supuesto indica la posición del valor que se encuentra justo a la mitad de los datos.

B. Desarrollo de la Técnica.-

Se tiene un conjunto de datos discretos ordenados de menor a mayor Y_j para $i = 1,2,3,...,n$; p representa cualquier fracción entre 0 y 1.

$$C(p) = Y_i \text{ cuando p es una de las fracciones de } p_i \text{ donde } p_i = \frac{i - .5}{n}$$

Así, los cuantiles $C(p_i)$ de los datos, son justo los valores de las observaciones ordenadas.

En el ejemplo de los estudiantes de Economía Política (tabla 3), se tiene que los datos ordenados son:

33 35 35 39 41 41 42 45 47 48
50 52 53 54 55 55 57 59 60 60
61 64 65 65 65 66 66 67 68 69
71 73 73 74 74 76 77 77 78 80
81 84 85 85 88 89 91 94 94 98

como n = 50,

$p_1 = (1-.5)/50 = .01$ y se forma el punto P_1 (.01 , 33)

$p_2 = (2-.5)/50 = .03$ y se forma el punto P_2 (.03 , 35)

esto es, en la gráfica participan p_i contra $C(p_i)$ localizando p_i en el eje horizontal y por consiguiente $C(p_i)$ en las ordenadas. (fig.II.20).

C(.25) es el valor antes del cual se localiza una cuarta parte de las observaciones, C(.75) es el valor antes del cual quedarán tres cuartas partes de los datos y C(.5) es idéntico a la mediana. Los tres cuantiles coinciden con los percentiles 25,75 y 50 respectivamente.

Para datos continuos se hace necesaria una interpolación para conectar puntos consecutivos con segmentos rectilíneos.

Si p es la fracción f del camino entre p_1 y p_{i+1}, entonces, C(p) se define como:

$$C(p) = (1-f)C(p_i) + fC(p_{i+1})$$

C.Aplicaciones de la Técnica.-

La figura II.21 es el diagrama de cuantiles para la población masculina durante 1980 en Guanajuato. Contiene un total de 19 puntos que corresponden al número de habitantes censados por grupo de edad.

El comportamiento de los datos es decreciente pero, a partir del primer cuantil, el descenso es más lento. Es decir, después de que la población masculina llegó a 95,482 personas, (de 25 a 29 años) ésta continuó disminuyendo en los siguientes grupos de edad con decrementos más suaves.

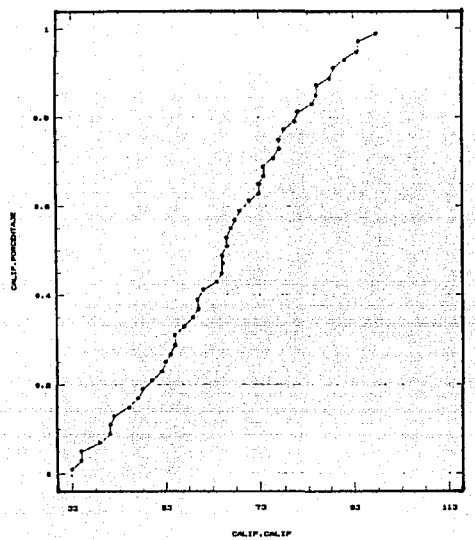


FIGURA II.20. Diagrama de cuantiles

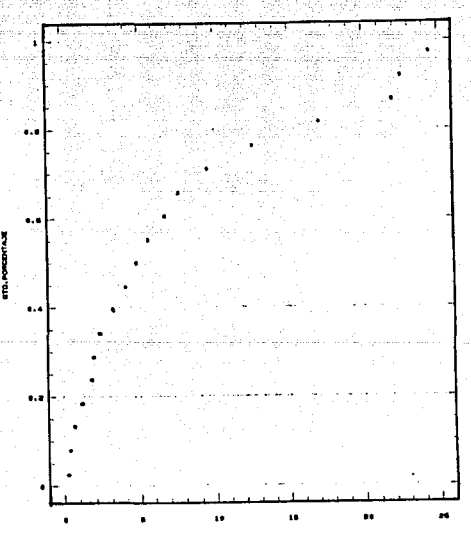


FIGURA II.21. Población masculina de Guanjuato. 1980.

Para el caso del número de mujeres Guanajuatenses en 1980, el comportamiento es similar. (figura II.22).

La gráfica cuantil-cuantil es un método efectivo para hacer una comparación detallada de las distribuciones de dos conjuntos de datos. Se construye graficando los cuantiles de una distribución contra los correspondientes cuantiles de la otra.

Sean X_i y Y_j, i,j = 1, 2, 3, . . ., n dos conjuntos de datos, entonces, el diagrama cuantil-cuantil es un gráfica Qy(p) contra Qx(p) para un rango de p valores donde p = (i-.5)/n. Si las dos distribuciones son iguales, todos los puntos se localizan exactamente sobre la línea y = x.

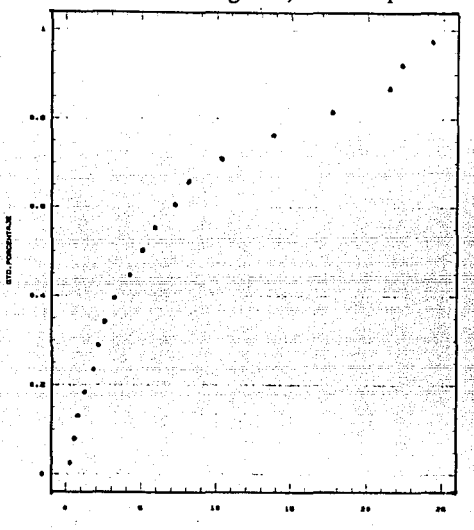


FIGURA II.22. Población femenina de Guanajuato de 1980.

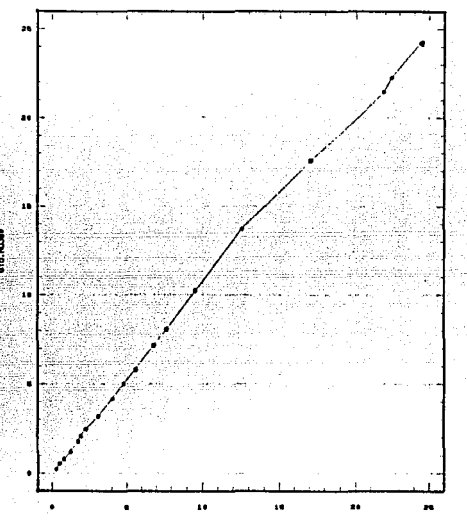


FIGURA II.23. Diagrama cuantil-cuantil. Población en Guanajuato 1980.

La comparación entre las distribuciones de la población masculina contra la femenina de Guanajuato en 1980 se hace en el diagrama cuantil-cuantil de la figura II.23 donde se demuestra la diferencia de comportamientos.

La figura II.24 muestra el despliegue de la gráfica cuantil-cuantil de los datos del índice metropolitano de calidad del aire registrados en el Noroeste y Suroeste durante febrero de 1989. La parte oeste del sur registró niveles superiores de polución.

Al cabo de un año, la situación comparativa entre ambas zonas mostró algunos cambios significativos.(fig. II.25).

Para hacer una comparación entre los índices de febrero 89 contra los de un año después en toda el área metropolitana, la figura II.26 es muy significativa. A pesar de todas las medidas tomadas para mejorar la calidad del aire en la Ciudad de México, los niveles de contaminación mostraron un ascenso.

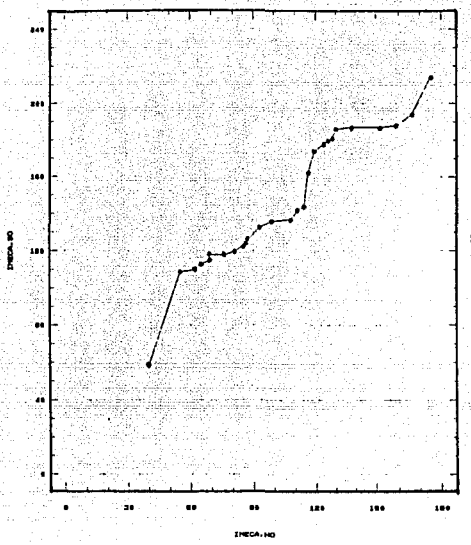


FIGURA II.24. Noroeste vs. suroeste. Febrero 1989

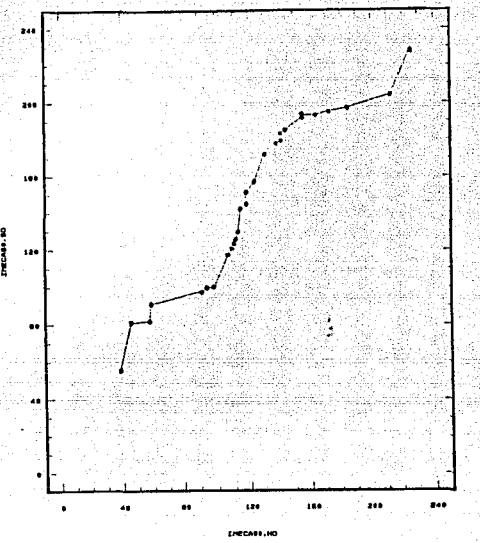


FIGURA II.25. Noroeste vs. suroeste. Febreo 1990.

D Ventajas y desventajas.-

La gráfica de cuantiles es por demás sencilla de construir.
No se hacen selecciones arbitrarias en cuanto al valor de los parámetros.
No se tienen suposiciones a cerca de la distribución de los datos.
Por medio de esta técnica, todos los puntos son graficados en una posición diferente aún si se duplican algunos de ellos. El número de puntos que pueden ser graficados sin encimar, está limitado sólo por la resolución de la graficadora.

E Revisión actualizada.-

La observación de los datos se dificulta cuando se tiene una concentración de puntos en un pequeño espacio de la gráfica. En este sentido, Chambers (1983) sugiere graficar los logaritmos correspondientes, es decir,

$$X_i^* = \log X_i \quad y \quad Y_i^* = \log Y_i$$

con lo que se obtiene una gráfica más uniforme de tal manera que no sería difícil expresar la relación entre las variables por medio de la ecuación de una recta:

$$Y_i^* = kX_i^* + c$$

si k = 1 y c = 0 se puede asegurar que ambas distribuciones son idénticas.
La figura II.27 contiene el diagrama de cuantiles para los logaritmos de los índices de contaminación en febrero 89 y 90.

La gráfica generada es similar a la figura II.26, con la diferencia de que por medio de los logaritmos, la observación de los puntos es más clara pues disminuye el número de puntos sobrepuestos.

La comparación de la distribución de dos muestras no se limita a los casos en que ambas cuentan con el mismo numero de datos. Esto es, si se tiene Y con i = 1,...,m y X con j = 1,...,n con m < n, es común tomar el valor de m e interpolar un valor correspondiente para n.

La interpolación puede formularse de la siguiente manera:

se busca un valor v que satisfaga la ecuación

$$\frac{v-.5}{n} = \frac{i-.5}{m} \qquad \text{de aquí que}$$

$$v = n\,(i - .5)/m + .5$$

Si v es un entero, los elementos de la gráfica serán Y_i vs. X_v. En otros casos, se toma j como entero y θ como la parte fraccionaria positiva, es decir, $v = j + \theta$ con $\theta \in (0.1)$. Entonces, el cuantil interpolado $Q_x\frac{(1-.5)}{m}$ es calculado como $(1 - \theta)X_j = \theta_{(j+1)}$

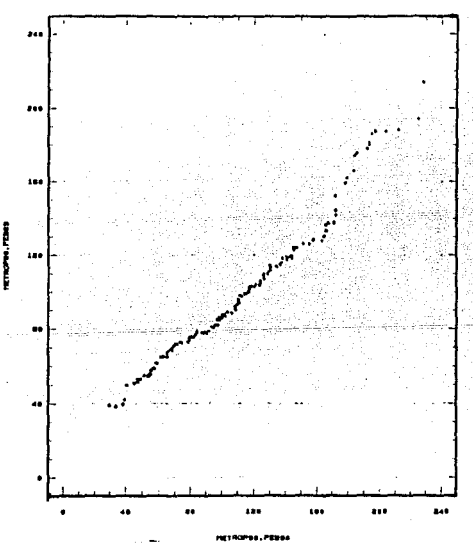


FIGURA II.26. Febrero 1989 vs. febrero 1990.

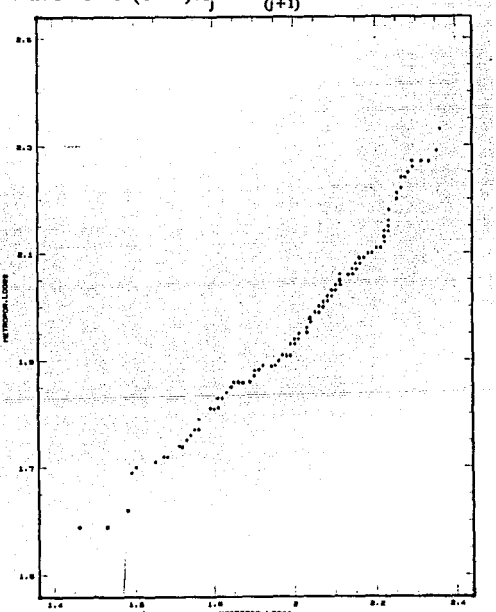


FIGURA II.27. Logaritmos. Febrero 1989 vs. 1990.

CAPITULO
III

REFERENCIAS
1.-CHAMBERS,J. (1983) Graphical Methods for data Analysis. USA
2.-CHAMBERS,J. and Kleiner B. (1982) Graphical Techiques for multivariate data and for clustering. Handbook of Statics. Vol 2. 209-244

III DIAGRAMAS SIMBOLICOS

Las gráficas simbólicas usan los valores de los datos como parámetros, de tal manera que sus variaciones causan la apariencia de un símbolo. El objetivo es obtener un símbolo con una forma distintiva para cada observación.

Si el método está bien diseñado, se espera inferir propiedades de los datos a través de la observación de los símbolos.

III.1 DIAGRAMAS DE SOLES Y ESTRELLAS

A. Descripción de la Técnica.-

Los diagramas de soles y estrellas muestran las características de los datos de manera individual. Es decir, se hace una pequeña ilustración *por elemento con respecto a las diversas variables* mediante rayos que parten de un mismo origen.

B. Desarrollo de la Ténica.-

Dadas p variables. cada diagrama tendrá p rayos (uno por variable) manteniendo igual espacio entre uno y otro sobre un círculo o, si se prefiere, la mitad del círculo. A partir de un rayo arbitrario, el ángulo entre este y el j-ésimo es de longitud

$$\theta_j = 2\pi\,(j-1)/p \qquad j = 1, 2, \ldots, p$$

para el círculo completo y

$$\theta_j = \pi\,(j-1)/p - 1 \qquad j = 1, 2, \ldots, p$$

para medio círculo.

Localizando el centro del símbolo en el origen de un plano cartesiano y escalando los valores de las variables al rango [0,1], el punto correspondiente a la j-ésima variable sobre el i-ésimo elemento (x_{ij}), tiene coordenadas polares

$$P_{ij} = (x_{ij} \cos\theta_j , x_{ij} \operatorname{sen}\theta_j)$$

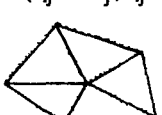


Figura III.1. Diagrama de estrella.

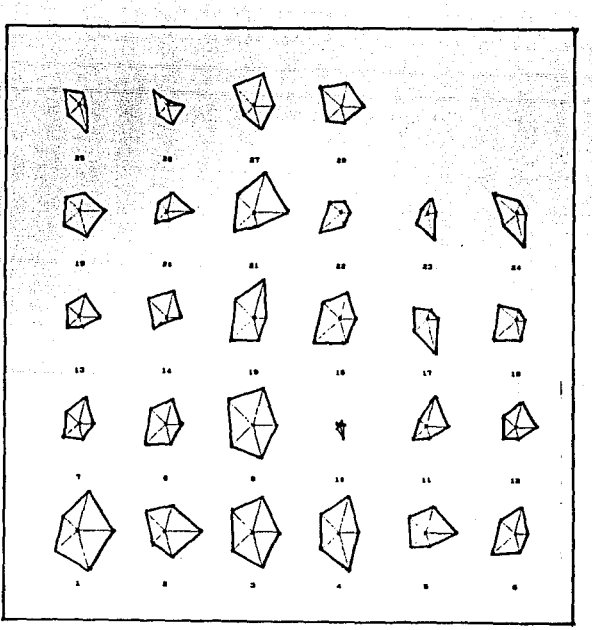
El diseño de los soles está basado en rayos de igual magnitud. Sobre este principio, los extremos del polígono formado son los que dan cuenta de la magnitud de cada variable.

En las estrellas, se prescinde de la base de rayos iguales. Simplemente se toma como símbolo el polígono obtenido.

C. Aplicaciones de la Técnica.-

Las figuras II.2 y II.3 representan las estrellas y los soles, en ese orden, del Indice Metropolitano de Calidad del Aire (IMECA). Más que tratar de describir las zonas, se intenta distinguir el comportamiento de la contaminación ambiental por día. Cada rayo simboliza a cada zona de la metrópoli, por lo que es un tanto complicado obtener conclusiones por zona

En base a las imágenes percibidas y, dado que cada símbolo es un día del mes, se puede decir que el 10 de febrero fue el día menos contaminado durante ese mes en 1989; mientras los días primero, tres y nueve registraron niveles elevados en las cinco zonas. La diversidad de formas tiene origen en la diferencia de la calidad del aire en los distintos rumbos de la ciudad de México durante el mismo día. Así, por ejemplo, el 24 de febrero fue un día con poca contaminación para los habitantes del noroeste y noreste, sin embargo se padeció de la contaminación en el centro y en la zona sur.



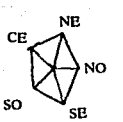


FIGURA III.2. Diagrama de estrellas. IMECA febrero 1989.

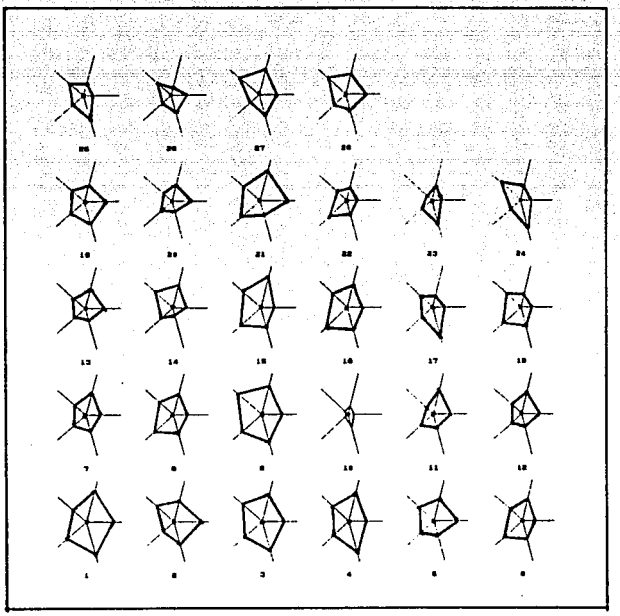


FIGURA III.3. Diagrama de soles. IMECA febrero 1989.

En la elaboración del diagrama de estrellas para los datos de tendencias de drogadicción (tabla 3), se asignó, a partir del rayo horizontal derecho y en sentido contrario a las manecillas del reloj, el siguiente orden de variables: zona norte 1976, 1986; zona centro 1976, 1986; zona sur 1976, 1986.

En la figura III.4 se tiene el diagrama de estrellas para los datos de la tabla 3. Este consta de ocho elementos, uno por cada enervante y cada símbolo cuenta con seis rayos que corresponden a las tres zonas en los dos perfodos tal como se describe en párrafo anterior.

Los incrementos en los porcentajes de consumo son notorios sobre todo en cuanto a inhalantes se refiere. Pero en el caso de sedantes y alucinógenos, la situación es contraria tal como se había observado ya por medio de diagramas de barras. *La semejanza entre los símbolos* correspondientes a estas dos drogas reflejan una semejanza también en el *comportamiento de los datos.*

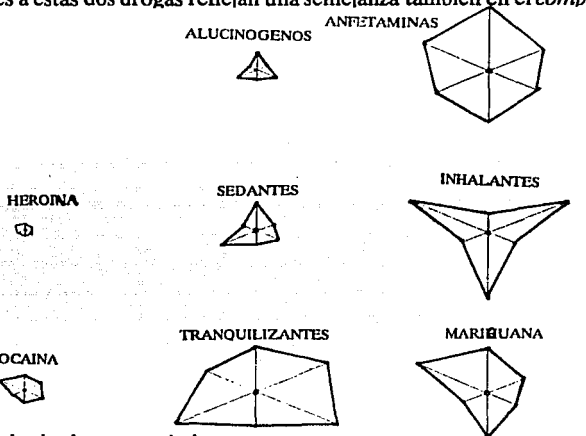


FIGURA III.4. Tendencias de consumo de drogas.

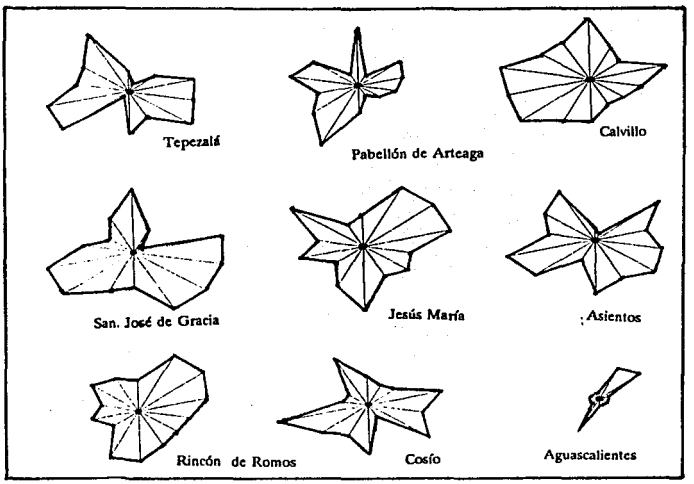
La figura III.5 contiene los símbolos de estrellas que corresponden a los nueve municipios del estado de Aguascalientes. Observando con detalle cada estrella, se puede obtener una idea general sobre la semejanza entre los niveles de vida de los diferentes lugares e *intuir los grupos* que se pueden formar. Esto es, a manera de hipótesis, los símbolos de Cosío, Asientos y Tepezalá podrían integrar un conjunto. San José de Gracia, Pabellón de Arteaga y Calvillo son elementos del segundo cúmulo, mientras Rincón de Romos y Jesús María forman un tercer conjunto en tanto Aguascalientes presenta condiciones de vida completamente distintos a los demás.

D. Ventajas y Desventajas.-

Debido a la diversidad de formas logradas, es sencillo de interpretar. En este sentido, el círculo completo es de una forma mas compacta y tiende a generar más símbolos distintos. De esta manera, proporcionan la imagen individual del conjunto de datos y es posible comparar la magnitud de diferentes variables. En cuanto a su diseño, los diagramas de soles y estrellas son de elaboración sencilla, sin embargo, puede ser tedioso cuando el número de datos es demasiado grande, y confuso si la cantidad de variables es excesiva. Es decir, son técnicas ilustrativas si el número de variables es moderado, de manera que se puedan distinguir una

de otra. Por otro lado, al normalizar las variables, se pierde la información sobre la localización y escala de cada variable lo cual puede significar un problema.

FIGURA III.5. Municipios de Aguascalientes.



III.2 DIAGRAMAS DE CARAS

A. Descripción de la Técnica.-

Con el objetivo de estudiar datos multidimensionales y, dadas las limitaciones y dificultades que implicaba la representación gráfica, Herman Chernoff (1971) ideó una técnica con la que logró plasmar datos con dieciocho variables.

Las características que distinguen o asemejan la cara de una persona con otra, son las mismas que rendirán cuentas de cada una de las 18 dimensiones. Es decir, la forma de la cara, el tamaño y la posición de los ojos, etc. son las que describen cada variable sobre un elemento.

B. Desarrollo de la Técnica.-

La elaboración de una cara gira alrededor de cinco aspectos básicos: la forma de la cara, la boca, , la nariz, los ojos y las cejas. Aunque, cabe mencionar que mas adelante (Bruckner 1978) aumento como característica importante lo referente a las orejas.

A partir de un punto central O, se dibuja un rayo hacia el punto P en una esquina y otro hacia P' en el otro extremo, de tal manera que OP y OP' son simétricos con respecto a un eje vertical que pasa por O. Fig. III.6 (a).

Se marcan los puntos S e I en los extremos superior e inferior de la cara tal que OS y OI son verticales y tienen igual longitud. Fig. III.6 (b).

La parte superior de la cara es la elipse determinada por PSP' y una excentricidad, análogamente, la parte inferior se determina por P'IP y otra excentricidad. Fig. III.6 (c).

La nariz se forma con un triángulo centrado en el punto O cuya longitud esta controlada por una variable. Fig. III.6 (d).

La boca es un arco de circunferencia centrado en el eje vertical y pasa por Pm. Sus variables son posición, curvatura y ancho. Fig. III.6 (e).

Los ojos son elipses orientadas simétricamente sobre el eje vertical. La posición ,separación, inclinación, excentricidad y tamaño pueden variar. Fig. III.6 (f).

Las cejas son segmentos lineales localizados simétricamente con respecto a la línea vertical a través de O. Se determinan por posición, inclinación y tamaño. Fig. III.6 (g).

Las pupilas se localizan en la misma distancia horizontal desde el centro de los ojos. Las orejas están representadas por círculos tangentes a P y P' cuyo radio también depende del valor que tome una variable. Fig. III.6 (h).

Con la intención de evitar diagramas deformes o irregulares es conveniente hacer una normalización de los datos. Fig. III.7.

Los valores máximo y mínimo en una dimensión formarán el rango de la misma. Los valores tomados serán mapeados linealmente desde el rango de los datos al rango de las características faciales. Es decir, la correspondiente parte de la cara se obtiene por interpolación de los extremos.

C. Aplicaciones de la Técnica.-

Los datos de la tabla 6 han sido reproducidos del ejemplo desarrollado por Lawrence (1978).

Se tienen diez grupos de compañías aceiteras Norteaméricanas:

Nombre	Compañías
ARCO	Atlantic Richfield, Richfield Oil, Sinclair, B.B. Barber, Barber Oil Exploration, Royal Gorge Company.
UNION	Union, Pure Oil, Pure Transportation Company
GETTY	Getty, Skelly
MOBIL	Mobil, Magnolia Petroleum
TEXACO	Texaco, Texaco Seaboard
CHEVRON	Chevron, California Company, Standard Oil of Texas.
GULF	Gulf, British American Oil
AMOCO	Amoco, Midwest Oil, Standolind, Pan American
SHELL	Shell, Shell P/L Corporation
EXXON	Exxon, Humble, Exxon Pipeline Company.

Las catorce variables atribuidas a cada grupo son las siguientes:

1.-Prima neta.	Total neto en dolares pagados.
2.-Exceso $ / renta	Promedio total de dolares pagados sobre la segunda oferta más alta.
3.-Superficie neta	Total de acres rentados (millones)
4.-Rentas ganadas	Número de rentas ganadas
5.-Prom. Propiedades	Porcentaje Promedio de propiedades de rentas
6.-Pct. prod.	Porcentaje de rentas encontradas últimamente ganando la compañía
7.-Prom. años	Promedio de años entre la venta y la primer producción
8.- Prod. gas	Producción neta de gas
9.- Prod. Liq.	Producción neta de líquido
10.- Pago real	Pago real neto al gobierno
11.- Tiempo real	Número real de años de producción
12.- Regresión	Cuadrado del coeficiente de correlación desde regresión lineal múltiple de producción real sobre datos ordenados, producción retrazada y años de producción.
13.- Prod. anual	Producción anual real por acre ($)
14.- Pct. Rentas	Porcentaje de rentas pagadas terminadas.

Cada uno de estos atributos está relacionado con una característica facial. La asignación de las variables con los rasgos de la cara se enlista a continuación:

Número	Característica facial
1	Amplitud de la cara
2	Longitud de la ceja
3	Altura de la cara
4	Separación de los ojos
5	Posición de las pupilas
6	Longitud de la nariz
7	Ancho de la nariz
8	Diámetro de las orejas
9	Nivel de las orejas
10	Longitud de la boca
11	Inclinación de los ojos
12	Curvatura de la boca
13	Nivel de la boca
14	Nivel de los ojos
15	Altura de las cejas

En la figura III.8 destacan los diagramas que corresponden a las compañías Shell y Exxon lo que induce a pensar que son las empresas con mayor éxito comercial, pues, de acuerdo con las variables son las que tienen mayor número de propiedades y sus niveles de producción son superiores a las del resto además, el promedio de años entre la primer producción y su venta es menor.

En el otro extremo se encuentran las compañías Getty, Arco y Amoco, a las que de acuerdo al diagrama el negocio no reditúa tanto como a las anteriores.

La asignación efectuada por el autor no parece tan aleatoria pues, aparentemente, las caritas "sonrientes" pueden relacionarse con el éxito, mientras las otras reflejan dificultades.

D. Ventajas y desventajas.-

La mayoría de las técnicas de graficación están limitados a tres dimensiones como máximo, en cambio, las caras de Chernoff permiten ilustrar en un plano, datos multivariados.

Pueden percibirse las características de un individuo (elemento) por medio de una simple exploración a la cara respectiva.

La comparación de diversas caras hace posible detectar patrones mas generales. La relación entre los rasgos y las diversas variables depende de la asignación aleatoria o deliberada del investigador. De acuerdo con el experimento desarrollado por Huff y Black (1978), la correspondencia que existe entre características y variables es subjetiva, pues, teniendo el mismo conjunto de datos, las diferentes asignaciones generan caras también distintas. Fig.III.9.

Lo anterior no seria tan problemático de no ser porque algunos rasgos son más significativos que otros. es decir, en general la gente puede marcar diferencias o semejanzas centrando su atención sólo en algunas características.

Sin embargo, Bruckner (1978) y Chernoff (1975) sostienen que esta subjetividad mas que desventaja puede ser una gran ventaja pues realza la habilidad del usuario para detectar y comprender fenómenos importantes y ademas, le obliga a ser mas cuidadoso en sus afirmaciones.

Por otro lado, en cuanto a la obtención de caras no muy reales, Chernoff opina que la falta de realismo se ve compensado por la habilidad de caricaturizar. Y, las caricaturas, después de todo, significan una ventaja por ser un medio mnemotécnico para obtener y comunicar conclusiones.

Observar formas y tamaños de caras es más complejo que percibir longitudes de rayos, por lo que esta técnica no es muy común sin embargo, debe tenerse en cuenta como una alternativa más de graficación.

E. Revisión Actualizada.-

Bernhard Flury (1988) ideó, en base a la versión original de Chernoff, duplicar la cantidad de variables a representar en una cara dejando de lado la simetría, es decir, en la parte facial izquierda es posible graficar dieciocho variables y, otras tantas del lado derecho, sumando así, treinta y seis características representadas. Fig. III.10.

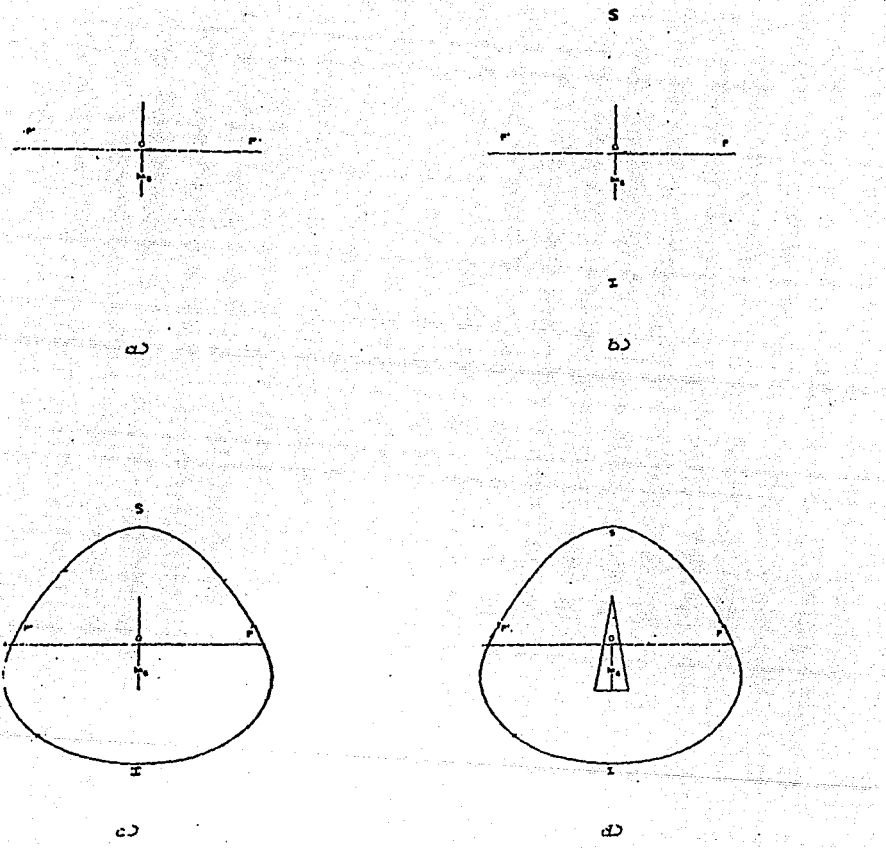


FIGURA III.6. Desarrollo de la técnica de diagramas de caras.

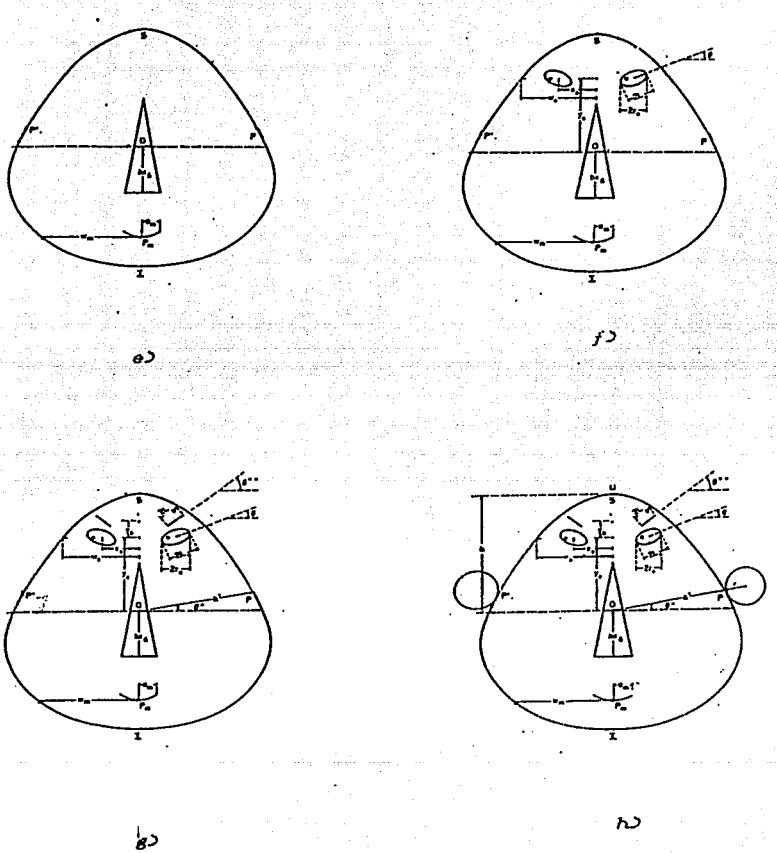


FIGURA III.6. Desarrollo de la técnica de diagramas de caras.

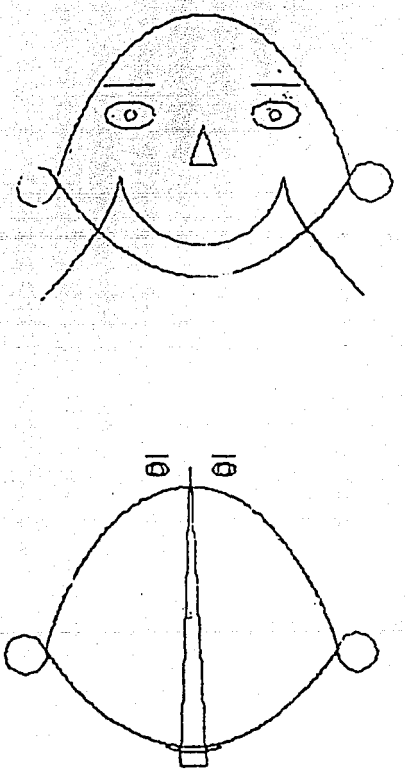


FIGURA III.7. Diagramas de caras deformes, resultado de la representción de variables no estandarizadas.

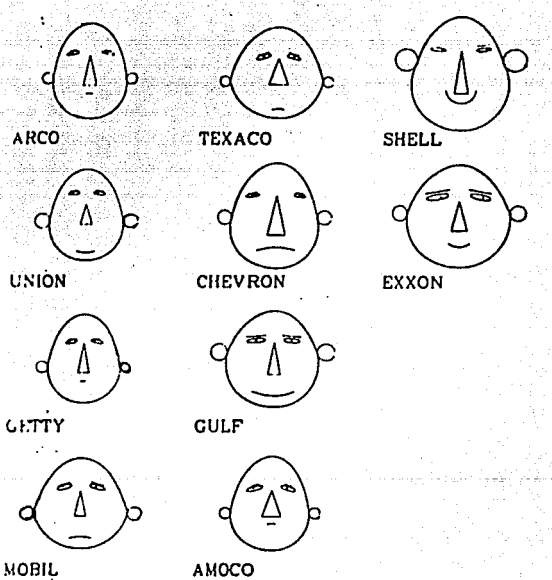


FIGURA III.8. Diagramas de caras para las diez compañías aceiteras

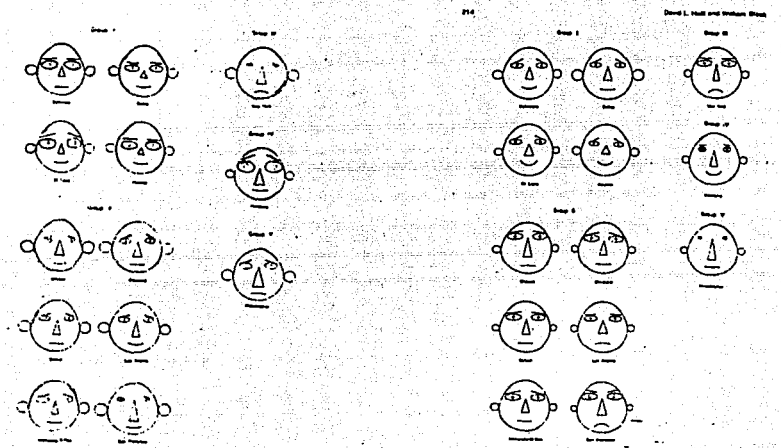
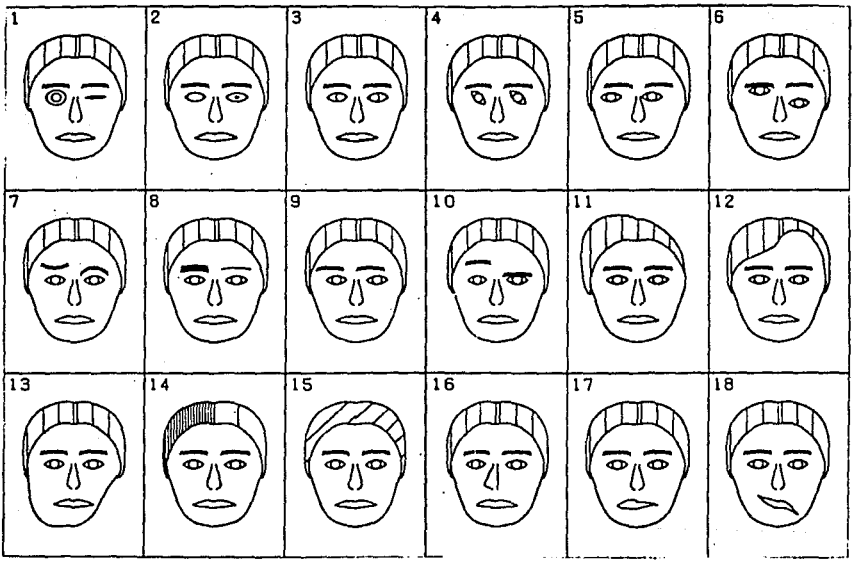


FIGURA III.9. La diferencia correspondencia entre variables y características faciales, genera caras distintas aún trtándose de los mismos datos.



Fuente: Flury, B. (1988)

FIGURA III.10. Caras asimetricas, se logra representar 36 variables

III.3 OTROS DIAGRAMAS SIMBOLICOS

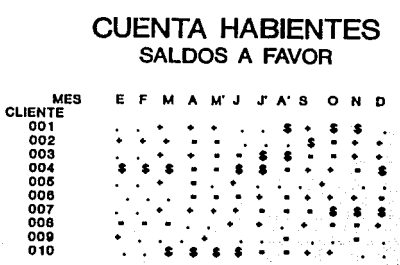
MATRIZ SIMBOLICA

Esta técnica consiste en asignar un pequeño símbolo a cada elemento de la matriz de acuerdo al intervalo al que pertenece el valor que este toma.

Así, por ejemplo, los caracteres '.', '+', '=', '$' forman cuatro niveles de clasificación. '.' representa valores menores que los representados por '+', éste, a su vez simboliza valores menores que '=' y el carácter '$' se emplea para los datos con valores superiores.

Esta técnica no es efectiva cuando el tamaño de la matriz es grande.

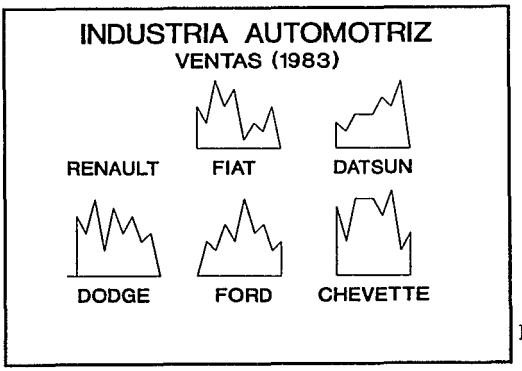
Los puntos alejados requieren de alguna notación especial.



Fuente: Chambers (1982)

PERFILES

Cada elemento se representa por tantas barras como variables (grupos de barras). Cada barra tiene un peso proporcional al valor de la variable. El perfil se refiere a la parte superior de las barras; algunas veces, los perfiles se muestran como una línea continua.

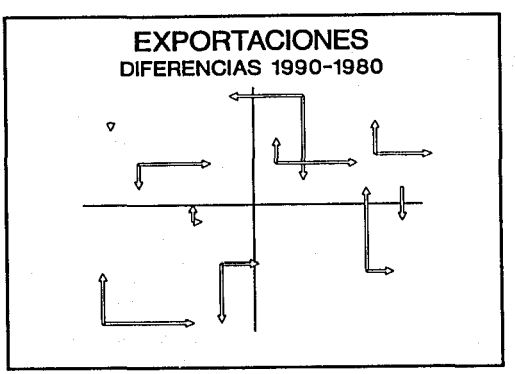


Fuente: Chambers (1982)

GRAFICA DE FLECHAS

Se tiene un plano cartesiano; las dos primeras variables se localizan en forma ordinaria. A partir del punto generado, la tercer variable, en esta observación se representa como una línea horizontal con longitud proporcional al valor que toma. Si el valor es positivo, la línea tiene dirección al este y al oeste si el valor es negativo.

De manera análoga, una cuarta variable interviene con dirección norte-sur según sea el caso. Al usar las direcciones noreste-suroeste, el diagrama permite graficar dos variables más. Esta técnica es efectiva sólo cuando se manejan a lo más seis variables.

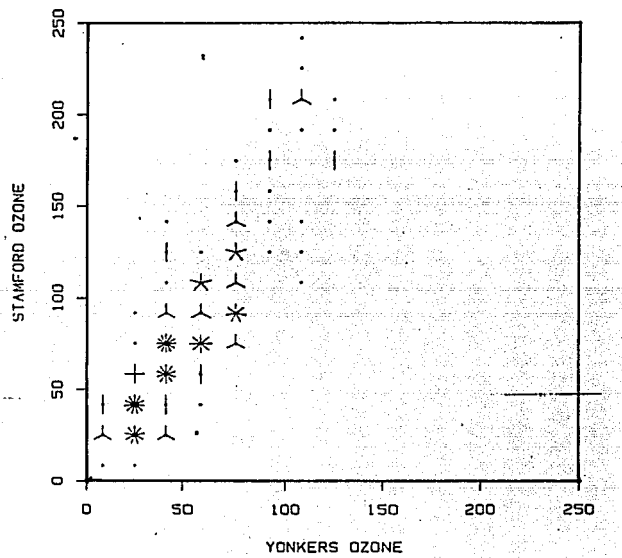


Fuente: Everitt (1978)

GIRASOLES

Los símbolos llamados girasoles fueron desarrollados para resolver el problema de puntos sobrepuestos. La idea es encerrar los puntos en celdas cuadradas y contar el número de elementos que pertenecen a cada celda. Un punto significa una observación, un punto con dos segmentos representa dos observaciones, un punto con tres líneas simboliza tres observaciones, etc.

Por medio de esta técnica se percibe fácilmente la densidad de las variables.



Fuente: Chambers (1982)

CURVAS DE ANDREWS

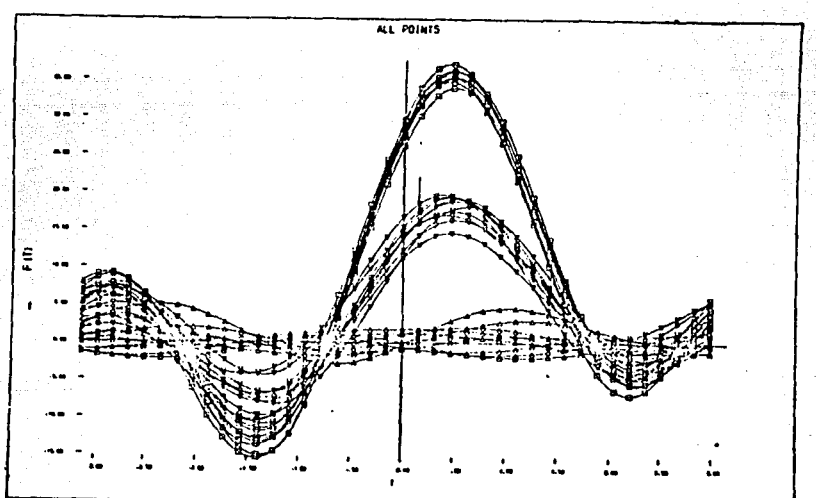
La técnica que propone Andrews (1972) para graficar datos multivariados parte de la idea de que cada una de las n observaciones x de dimensión p define una función

$$f_x(t) = x_1/2 + x_2 \operatorname{sen}(t) + x_3 \cos(t) + x_4 \operatorname{sen}(2t) + x_5 \cos(2t) \ldots$$

para $-\pi \leq t \leq \pi$

De esta manera, las *n* observaciones aparecerán como un conjunto de líneas dibujadas a través de la gráfica. La representación de esta función preserva la distancia euclidiana por lo cual, los puntos que permanecen juntos en el espacio p-dimensional original, quedan representados por curvas cercanas entre sí para todos los valores de t.

Esta propiedad permite usar una gráfica para la identificación de cúmulos, puntos lejanos y otras características de la distribución de los datos.



Fuente: Everitt (1978)

CAPITULO
IV

IV DESCRIPCION UTILIZANDO TECNICAS MULTIVARIADAS.

Construir una imágen mental o el modelo de una configuración de los datos en dos o tres dimensiones no implica mayor dificultad.

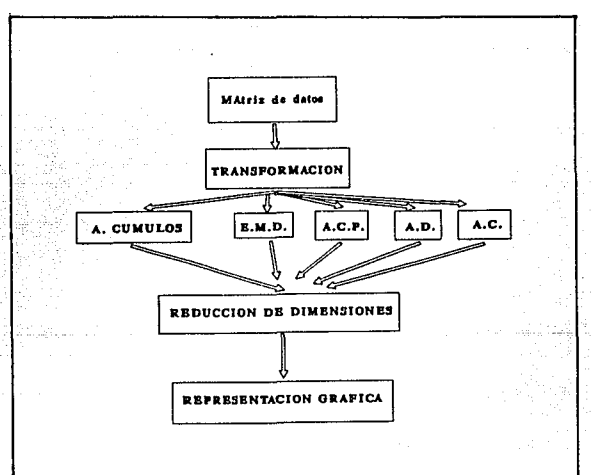
Sin embargo, pese a que todos los problemas reales son por sí mismos multivariados, su representación gráfica en más de tres dimensiones, sólo es concebida en la imaginación. Por lo que la descripción de problemas que involucran n elementos en un espacio p-dimensional requiere de aplicaciones matemáticas y computacionales para lograr plasmar sus características en un plano con el objetivo de visualizarlas y simplificar su entendimiento e intrepretación.

Por medio de algunas técnicas de Análisis Multivariado puede obtenerse un menor número de variables que representen a las originales mostrando que un subconjunto de ellas es adecuado para los propósitos específicos del análisis.

Gran parte del trabajo sobre Análisis Multivariado tiene fundamento en procedimientos formales de inferencias o de teorías de distribución.

En este trabajo se emplearon algunos resultados teóricos para rescatar la parte de la gráfica (como herramienta descriptiva) que ellas tienen para representar a individuos o variables.

Las técnicas que se emplean en este capítulo pueden mostrarse de manera general en el siguiente esquema:



IV.1 ANALISIS DE CUMULOS

A. Descripción de la Técnica.-

La descripción de un grupo de elementos o individuos remite de manera inmediata a la serie de características que cada uno posee.

Y, es justamente, a partir de los atributos individuales que el observador puede formarse un juicio de similaridad o diferencia entre dos o más elementos.

El análisis de cúmulos tiene como propósito principal identificar y agrupar objetos similares. El resultado esperado es una serie de conjuntos homogéneos en su interior y lo más heterogéneos posible entre si. Es decir, los sujetos que pertenecen al mismo conjunto son demasiado parecidos pero los elementos de dos cúmulos distintos son por demás diferentes.

Hablar en términos como *"parecidos", "diferentes","similares"* puede resultar arbitrario. Para contrarestar la ambigüedad, los diferentes algoritmos de análisis de cúmulos toman como medida de similaridad la distancia entre puntos.

De esta manera, los objetos similares están asociados con puntos cercanos entre sí. Así mismo, los objetos diferentes permanecen alejados uno del otro.

Una vez formados los cúmulos puede considerarse cada uno de ellos como un solo individuo y, en lugar de hablar de la información de una población total, referirse a las características de pequeños subgrupos, logrando así reducir las dimensiones del problema.

Cuando el estudio tiene propósitos meramente descriptivos, no hay suposiciones a cerca de la forma de la población. En otros casos, se supone un modelo donde cada observación en la muestra puede tener origen en una de las diferentes distribuciones.

Bajo los supuestos de que el análisis requiere de la agrupación de observaciones y de que es posible formar los cúmulos, se aplica la técnica a una matriz D_{ij} de distancias o similaridades[1] derivada de los datos originales.

Los objetivos pueden ser obtener únicamente la partición del conjunto de datos o lograr la jerarquización de los cúmulos.

B. Desarrollo de la Técnica.-

IV.1.1 CUMULOS NO JERARQUICOS.-

Las técnicas no jerárquicas pueden enfocarse de las siguientes formas:

1.- Centros móviles.- Estos métodos son considerados los más apropiados para muestras grandes.

Algoritmo.-
Se supone que se desea particionar un conjunto **I** de n individuos caracterizados por p-variables.
Se tiene una matriz D ($n \times p$) con distancia Euclidiana.
El número máximo de grupos a formar esta dado por kc ($kc < n$).

0) Se determina el grupo provisional de kc centros.
(selección seudoaleatoria sin reemplazamiento de kc en n.)
Los kc centros son $C_1^0,\ C_2^0,\ C_3^0,\ \ldots,\ C_{kc}^0$

1.- Apéndice B

Se crea una partición P^0 de I en k_c cúmulos
$I_1^0, I_2^0, \ldots, I_{kc}^0$

Un individuo i pertenece a I_k^0 si i esta más cerca de C_k^0 que de otros centros.

i) A partir de los centros de gravedad de los cúmulos $I_1^0, I_2^0, \ldots, I_{kc}^0$
se determinan kc nuevos centros $C_1', C_2', \ldots, C_{kc}'$
Estos nuevos centros crean una nueva partición P' construida de acuerdo a las mismas reglas de P^0.
P' esta formada por $I_1', I_2', \ldots, I_{kc}'$

ii) A partir de los centros de gravedad de los cúmulos
$I_1', I_2', \ldots, I_{kc}'$ se determinan k_0 centros nuevos $C_1', C_2', \ldots, C_{kc}'$

El algoritmo termina cuando dos iteraciones sucesivas conducen a la misma partición o con un criterio conveniente de selección como el porcentaje de varianza o el numero de iteraciones.
La varianza dentro de cada grupo debe ir disminuyendo en cada iteración.
Suponiendo que cada individuo tiene un peso relativo p_i tal que
Todos los algoritmos de cúmulos no jerárquicos son similares al anterior, sólo difieren en pequeños aspectos.

2.-Selección secuencial.- Se escoge sólo un centro u origen del cúmulo y todos los objetos cercanos son incluidos. Una vez formado el primer cúmulo, se busca el centro para un nuevo cúmulo formado de manera análoga al anterior. Se continua de la misma manera hasta que todos tengan un grupo asignado.
Cuando un objeto ha sido incluido en un cúmulo, no se considera para los cúmulos subsecuentes.

3.-Cúmulos dinámicos.- Los cúmulos no se caracterizan por un centro de gravedad pero sí por el número de individuos que constituyen un grupo.

4.-K-medias.- También empieza con selección seudo aleatoria pero, para calcular los nuevos centros, se modifican las posiciones de los centros antes que todos los individuos sean reasignados. Cada reasignación de los individuos permite una modificación de la posición de los centros correspondientes. (Un objeto puede cambiar de cúmulo en cada reasignación).

C Aplicación de la Técnica.-

La fig.IV.1 contiene los resultados emitidos por el paquete SYSTAT al aplicar el algoritmo k-medias a los datos de la tabla 4.
Las variables consideradas son las zonas en que se divide el área metropolitana: Noroeste (NO), Noreste (NE), Centro (CE), Suroeste (SO) y Sureste (SE).
El despliegue muestra cinco cúmulos; el primero consta de un solo elemento (el día 10 de febrero) quien parece no tener similitud con los otros días del mes. El cúmulo 3 es el que contiene mayor numero de elementos y si se observa la columna correspondiente a la media (MEAN) no es difícil percatarse de que el grupo 4 contiene a los días con mayores índices de contaminación en el período.
De hecho, esta situación ya se percibía con los diagramas de soles y estrellas en las figuras

VA	BETWEEN SS	DF	WITHIN SS	DF	F-RATIO	PROB
NO	21693.430	4	10219.240	23	12.206	.000
NE	7466.060	4	3012.036	23	14.253	.000
CE	3039.029	4	5989.828	23	2.917	.043
SO	32172.870	4	2636.569	23	70.165	.000
SE	5567.091	4	4765.027	23	6.718	.001

CLUSTER NUMBER: 1

MEMBERS			STATISTICS			
CASE	DISTANCE	VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
10	.00	NO	49.00	49.00	49.00	.00
		NE	39.00	39.00	39.00	.00
		CE	42.00	42.00	42.00	.00
		SO	59.00	59.00	59.00	.00
		SE	73.00	73.00	73.00	.00

CLUSTER NUMBER: 2

MEMBERS			STATISTICS			
CASE	DISTANCE	VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
6	4.42	NO	65.00	80.00	93.00	9.92
8	4.10	NE	51.00	75.00	105.00	18.50
15	14.72	CE	71.00	85.50	98.00	8.69
16	13.40	SO	178.00	190.33	214.00	11.76
18	12.64	SE	73.00	86.50	96.00	6.92
22	15.83					

CLUSTER NUMBER: 3

MEMBERS			STATISTICS			
CASE	DISTANCE	VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
2	23.59	NO	85.00	121.10	166.00	23.73
5	15.77	NE	65.00	76.18	91.00	7.64
7	16.11	CE	58.00	84.45	106.00	16.03
11	13.68	SO	109.00	128.18	144.00	10.77
12	8.01	SE	55.00	77.55	99.00	13.65
13	9.11					
14	20.19					
19	13.53					
20	12.04					
27	16.56					
28	9.32					

CLUSTER NUMBER: 4

MEMBERS			STATISTICS			
CASE	DISTANCE	VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
1	14.07	NO	108.00	141.20	175.00	23.68
3	5.55	NE	95.00	100.60	106.00	3.98
4	17.46	CE	82.00	100.20	120.00	13.91
9	12.93	SO	162.00	177.80	187.00	9.17
21	22.41	SE	76.00	111.80	123.00	18.68

CLUSTER NUMBER: 5

MEMBERS			STATISTICS			
CASE	DISTANCE	VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
17	7.45	NO	55.00	72.20	99.00	15.12
23	14.76	NE	39.00	51.40	59.00	6.89
24	13.38	CE	52.00	82.20	110.00	18.67
25	7.90	SO	110.00	116.80	123.00	4.71
26	16.38	SE	85.00	105.00	119.00	11.68

FIGURA IV.1. Datos IMECA agrupados en 5 cúmulos

	BETWEEN SS	DF	WITHIN SS	DF	F-RATIO	PROB
NO	25250.370	6	6662.333	21	13.265	.000
NE	7475.879	6	2049.250	21	5.715	.002
CE	4369.091	6	4659.750	21	3.292	.019
SO	32592.050	6	2217.417	21	51.444	.000
SE	6914.232	6	3417.884	21	7.080	.000

CLUSTER NUMBER: 1

MEMBERS

CASE	DISTANCE
10	.00

STATISTICS

VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
NO	49.00	49.00	49.00	.00
NE	39.00	39.00	39.00	.00
CE	42.00	42.00	42.00	.00
SO	59.00	59.00	59.00	.00
SE	73.00	73.00	73.00	.00

CLUSTER NUMBER: 2

MEMBERS

CASE	DISTANCE
6	4.42
8	4.10
15	14.72
16	13.40
18	12.64
22	15.63

STATISTICS

VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
NO	65.00	80.00	93.00	9.92
NE	51.00	75.00	105.00	18.50
CE	71.00	85.50	98.00	8.69
SO	178.00	190.33	214.00	11.76
SE	73.00	86.50	96.00	6.92

CLUSTER NUMBER: 3

MEMBERS

CASE	DISTANCE
7	11.35
11	10.67
12	5.96
13	5.86
14	16.66
20	11.15

STATISTICS

VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
NO	86.00	108.33	128.00	16.05
NE	67.00	78.50	89.00	7.04
CE	58.00	74.50	101.00	13.38
SO	115.00	125.33	136.00	7.48
SE	55.00	68.17	79.00	7.29

LUSTER NUMBER: 4

MEMBERS

CASE	DISTANCE
4	.00

STATISTICS

VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
NO	108.00	108.00	108.00	.00
NE	98.00	98.00	98.00	.00
CE	96.00	96.00	96.00	.00
SO	162.00	162.00	162.00	.00
SE	124.00	124.00	124.00	.00

LUSTER NUMBER: 5

MEMBERS

CASE	DISTANCE
17	7.45
23	14.76
24	13.38
25	7.90
26	16.38

STATISTICS

VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
NO	55.00	72.20	99.00	15.12
NE	39.00	51.40	59.00	6.89
CE	82.00	82.20	110.00	10.67
SO	110.00	116.80	123.00	4.71
SE	85.00	105.00	119.00	11.68

LUSTER NUMBER: 6

MEMBERS

CASE	DISTANCE
1	13.09
3	8.60
9	14.59
21	19.06

STATISTICS

VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
NO	126.00	149.50	175.00	18.87
NE	93.00	101.25	106.00	4.21
CE	82.00	101.25	123.00	15.37
SO	174.00	181.75	187.00	5.21
SE	76.00	108.75	128.00	19.74

LUSTER NUMBER: 7

MEMBERS

CASE	DISTANCE
2	14.36
5	11.42
19	11.46
27	17.09
28	8.38

STATISTICS

VARIABLE	MINIMUM	MEAN	MAXIMUM	ST.DEV.
NO	111.00	136.60	166.00	19.79
NE	65.00	77.80	91.00	8.28
CE	81.00	96.40	106.00	9.44
SO	109.00	131.60	144.00	12.91
SE	74.00	88.80	99.00	10.68

FIGURA IV.2. Datos IMECA agrupados en 7 cúmulos

III.2 y III.3, pues, justamente los días 1, 3, 4, 9 y 21 (agrupados en el cuarto cúmulo) generaron gráficas similares entre si y mayores que el resto, además, la estrella del día 10, en efecto es la menor y no tiene "parecido" con las otras.

En la figura IV.2 el mismo conjunto de datos se ha seccionado en siete cúmulos. El resultado sólo difiere en que a partir de un grupo se obtuvieron otros, pero, sustancialmente los elementos conservan su lugar en cuanto a sus vecinos, es decir, no se distingue ningún día que haya saltado bruscamente de un conjunto para aparecer en otro. De hecho, los cúmulos 1, 2 y 5 no sufrieron alteración alguna, esto habla de la estrecha similaridad entre sus integrantes; el cúmulo 3 se dividió en 2 grupos (3 y 7) y, finalmente el cuarto cúmulo sólo arrojo al elemento 4 a formar otro conjunto (4 y 6).

El despliegue gráfico de esta clasificación está dado en la figura IV.3. En ella, los elementos de los siete cúmulos se distinguen entre sí por los diferentes símbolos.

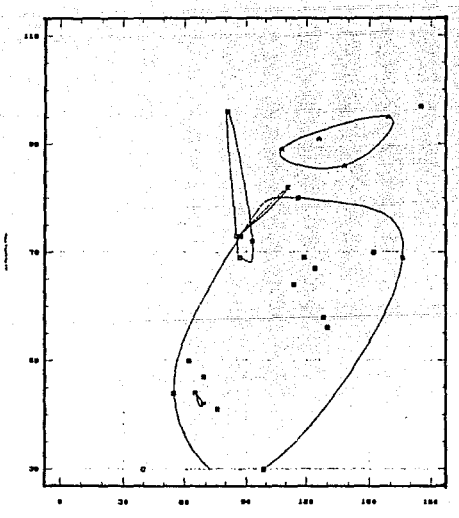


FIGURA IV.3 Datos IMECA. Siete cúmulos.

El resultado del Análisis de cúmulos para los datos de Tendencias de drogadicción se

encuentran en la figura IV.4. Se distinguen cuatro cúmulos que son graficados en un plano cartesiano.

El diagrama que ilustra los municipios del Edo. de Aguascalientes en un plano cartesiano, acumulados en cinco grupos, se encuentra en la figura IV.5. En ella se distinguen tres grupos unitarios.

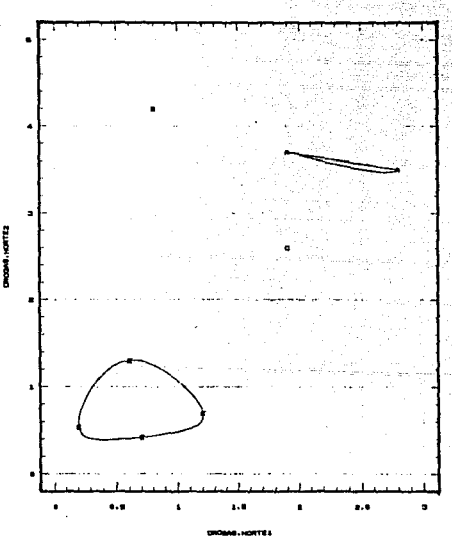


FIGURA IV.4. Drogas consumidas por los jóvenes mexicanos

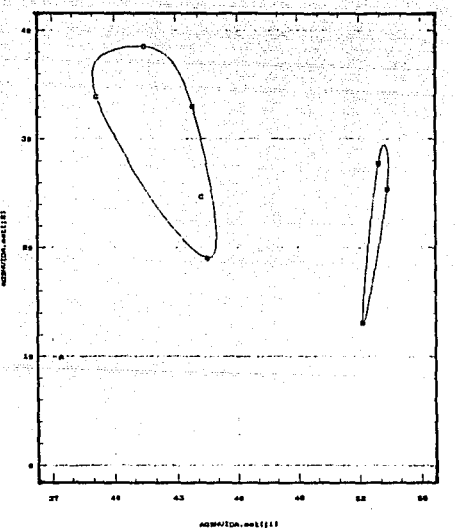


FIGURA IV.5. Municipios del Edo. de Aguascalientes en 5 cúmulos.

D. Ventajas y Desventajas.-

En el algoritmo para acumular alrededor de centros móviles, el resultado depende directamente de la selección inicial del conjunto de centros en el paso cero. Es decir, a cada conjunto diferente de centros corresponde un resultado distinto. El problema es encontrar la solución óptima.

Para contrarrestar el problema, Lebart (1984) propone una pequeña modificación en la técnica llamada "cúmulos estables". Consiste en desarrollar diferentes particiones con distintos conjuntos de centros cada vez y escogiendo como cúmulos estables los grupos de individuos que son siempre asignados al mismo cúmulo en cada una de las particiones.

Otra desventaja radica en la posibilidad de obtener resultados equivocados por suponer ciertas algunas características sobre la estructura de los datos.

IV.1.2 CUMULOS JERARQUICOS.

B. Desarrollo de la Técnica.-

Al aplicar las técnicas de clasificación jerárquica, es necesario tomar como cierto que:

-Algunas medidas de distancia o similaridad pueden ser definidas entre pares de objetos a clasificar.

-Existen reglas para calcular distancias entre cúmulos separados.

El uso común que se le da a la técnica es en el sentido aglomerativo. Se parte de un estado inicial de n individuos y se van fusionando en cada paso de manera tal que el estado final es aquel en que todos los individuos forman un solo grupo.

En algunos problemas puede ser útil partir del grupo acumulado y mediante particiones sucesivas, obtener los cúmulos individuales se conoce como (método divisivo).

Algoritmo general.-

i) Se tiene una matriz de distancia D(i,j) derivada de los datos originales.
ii) Se localiza los puntos más cercanos y se acumulan en un nuevo punto
iii) Se calcula la distancia entre el nuevo punto y los sobrantes.
iv) regresar al paso i con $n = n-1$
El algoritmo termina cuando se tiene un solo punto (el acumulado).

A diferencia de los métodos de optimización, en los cúmulos jerárquicos, si un objeto es localizado en un grupo, no puede asignarse a otro.

Los algoritmos comunes son Encadenamiento simple (SINGLE LINKAGE) y Encadenamiento Completo (COMPLETE LINKAGE).

1.- ENCADENAMIENTO SIMPLE.-

A partir de la idea del Algoritmo general, el método de encadenamiento simple agrupa a los puntos que están más cercanos unos a otros de la siguiente manera:

Primero, ordena ascendentemente las $a = n(n-1)/2$ distancias entre los puntos.

i) Sean $C_1, C_2, \ldots, C_n$ los cúmulos con un elemento cada uno llamado $X_i \in C_i$.

ii) Sin perdida de generalidad, sea $d(r_1 s_1) = \text{mín}(r, s)$ tal que $Xr_1 Xs_1$ están cerca, entonces, estos dos puntos son agrupados en un cúmulo, de manera que ahora se tiene n-1 cúmulos donde $Cr_1 + Cs_1$ forman un nuevo cúmulo.

iii) Sean $d(r_2\ s_2)$ la siguiente distancia menor.

Si r_2 o s_2 es igual a r_1 o s_1 los nuevos n-2 cúmulos son $C_{r1} + C_{s1}$, $C_{r2} + C_{s2}$ y los restantes cúmulos unitarios.

Si $r_1 = r_2$ y $s_1 \neq s_2$, los nuevos n-2 cúmulos son $C_{r1} + C_{s1}$, C_{s2} y los demás cúmulos unitarios.

iv) El proceso continua como en iii hasta determinar las $a = n(n-1)/2$ distancias.

En el i-ésimo estado, sea $d_{r\,s}$ la i-ésima menor distancia.

Entonces, el cúmulo que contiene r_i se une con el que contiene $a_{s\,i}$. Pero si r_i y s_i ya están en el mismo cúmulo, entonces no se forma nuevo grupo.

v) El proceso puede detenerse antes de que todos los cúmulos hayan estado juntos en un solo grupo, si se determina un criterio para las distancias. Esto es, terminar cuando las distancias entre cúmulos son mayores que una cierta d_0 arbitraria.

Sean $C_1^*, C_2^*, \ldots, C_g^*$ los cúmulos restantes de tal manera que $d_0' > d_0$.

Entonces, dos cúmulos C_j, C_k serán unidos en el umbral d_0 si al menos una distancia (o ligue singular) existe entre r y s con

$X_r \in C_j$, $X_s \in C_k$ y $d_0 < d_{rs} < d_0'$

Cuando se ha establecido un ligue entre objetos, este no se puede romper.

2.- ENCADENAMIENTO COMPLETO

La técnica es similar a la anterior, la diferencia esta en que las distancias entre dos cúmulos se define como la mayor distancia entre pares de elementos en cada cúmulo.

Algoritmo.-

a) Sean $C_1, C_2, \ldots, C_n$ los cúmulos individuales conteniendo $X_1, X_2, \ldots, X_n$ respectivamente.

b) Asumiendo que $d_{12} = \text{mín}\,(d_{ij})$ para todo i , j y sean $C_2^*, C_3^*, \ldots, C_n^*$ los nuevos cúmulos después de haber agrupado los pares C_1, C_2 *en* C_2^*

c) Se define una nueva matriz de distancias $D^* = (d_{i\,j}^*)$ pero ahora de dimensión

(n-1) x (n-1) con $d_{2j}^* = \text{máx}\,(d_{ij}, d_{2j})$ para $j = 3, 4, \ldots, n$; $d_{ij}^* = d_{ij}$ para i,j = 3 , 4, ... , n.

Encontrando mín d_{ij}^* i,j = 2, ... , n ; entonces, seguir como en b.

Se continúa hasta que todas las distancias entre cúmulos son mayores que un valor inicial arbitrario d_0.

Cuando se ha terminado, las distancias dentro del cúmulo son menores que d_0, es decir,

máx $i, j \in C$ $d_{ij} \leq d_0$ para cualquier cúmulo C. Así, este método tiende a producir cúmulos compactos sin cambio de efecto.

C. Aplicaciones de la Técnica.-

El dendrograma o diagrama de árbol constituye un resumen gráfico de la información que contiene la matriz de distancias o similaridades.

La distancia entre dos individuos en un dendrograma puede ser tomada como la fusión del nivel en el cual aparecen ambos por primera vez en el mismo cúmulo.

En la fig.IV.6 se muestra el dendrograma que corresponde a los datos IMECA de la tabla 4 y, en términos del párrafo anterior, los días 8 y 6 de febrero de 1990 registraron índices por demás similares entre si en todas las zonas del área metropolitana, situación semejante a la que se observa entre los elementos 12 y 13.

Como lo indica el algoritmo general, el resultado final acumula el total de datos en un solo

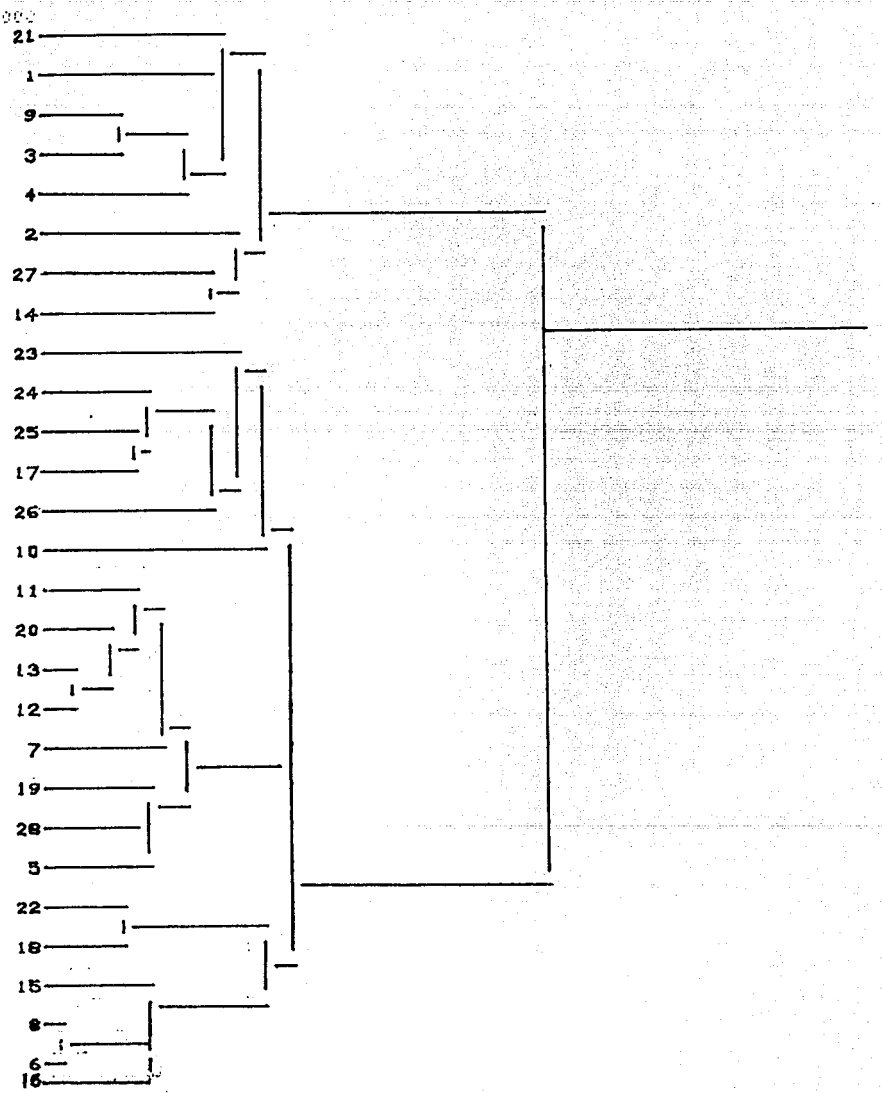


FIGURA IV.6. Datos IMECA Dendrograma

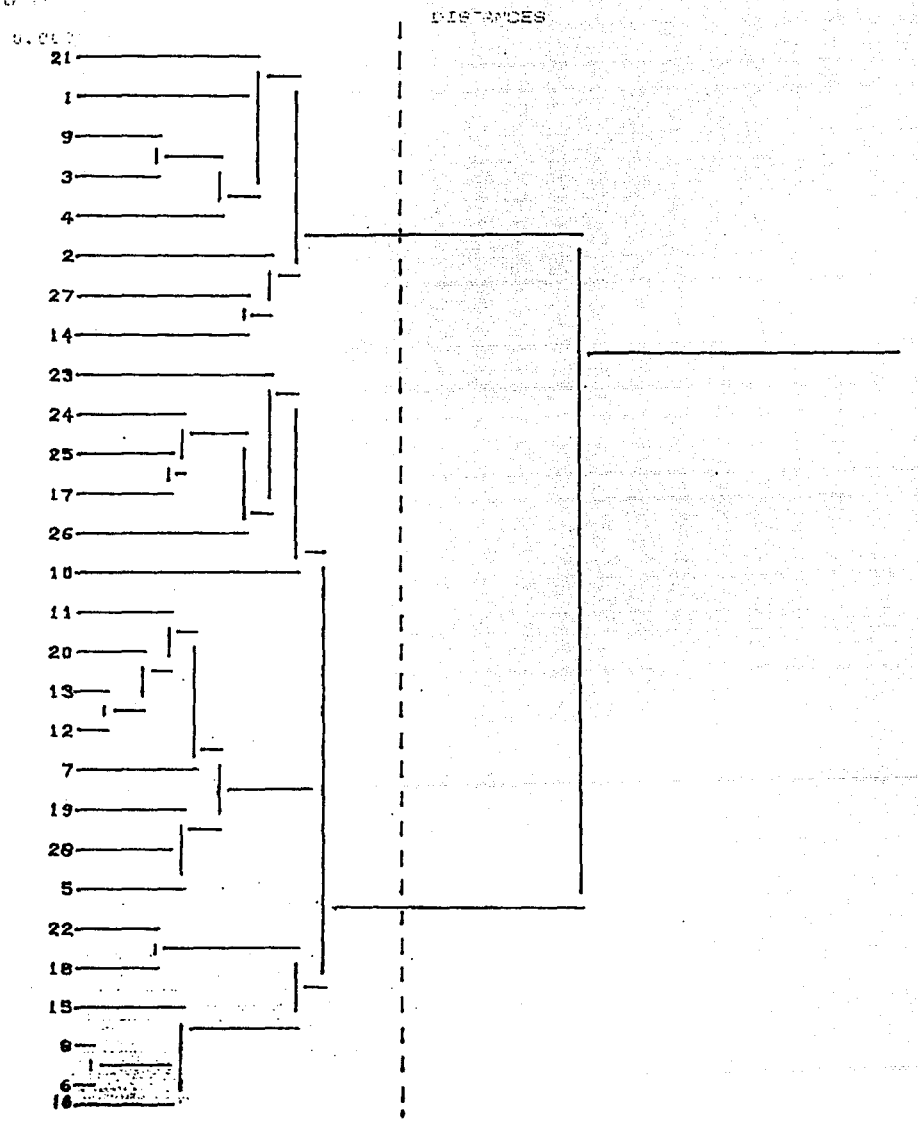


FIGURA IV.7. Datos IMECA Dendrograma seccionado formando 2 cúmulos.

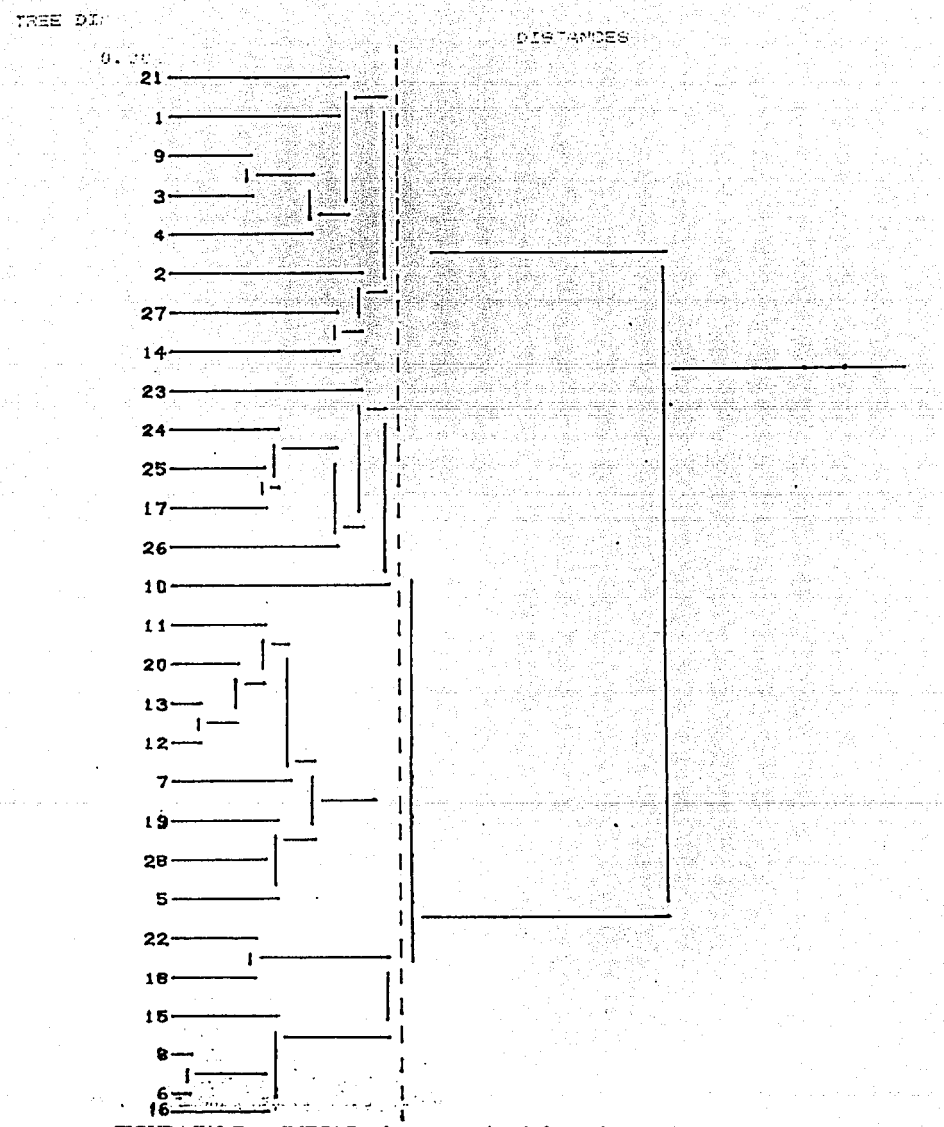


FIGURA IV.8. Datos IMECA Dendrograma seccionado formando 4 cúmulos.

grupo. Si se trazara una recta vertical de manera que cortara el árbol de izquierda a derecha, se tendrían dos cúmulos en el primer corte (fig.IV.7) y cuatro grupos en el segundo corte (fig.IV.8). De esta manera se puede proceder para obtener los cúmulos individuales partiendo de un solo grupo acumulado.

El resultado graficado por el árbol puede confirmarse mediante el despliegue de la fig. IV.1 (el grupo de la parte inferior del dendrograma es quizá el más claro y corresponde al segundo cúmulo de la fig IV.1), y con un poco más de dificultad visual, en los diagramas de soles y estrellas.

Para el caso de los datos sobre tendencias de drogadicción en la República Mexicana, el dendrograma correspondiente se encuentra en la figura IV.9. El consumo de heroína y alucinógenos es el más parecido entre sí. Por otro lado, se distinguen con claridad dos grupos: Anfetaminas, mariguana, tranquilizantes y heroína; en uno y heroína, alucinógenos, cocaína y sedantes en el otro.

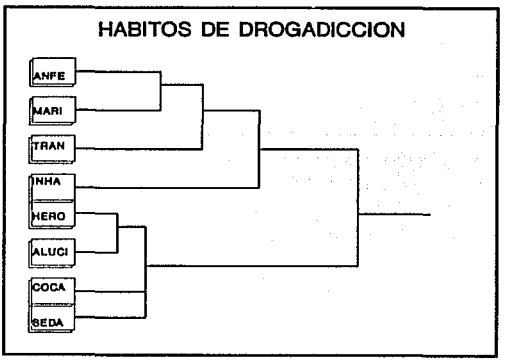


FIGURA IV.9. Dendrograma. Jerarquización por tipo de drogas.

Al desarrollar el Análisis de Cúmulos a los mismos datos de drogadicción pero con respecto a las zonas, se tiene la figura IV.10.

En este caso, es evidente también la formación de dos grupos.Los datos de las tres zonas durante 1976 son más parecidas entre sí que con cualquiera de 1986; de la misma manera, los datos de 1986 forman otro grupo. Se puede deducir además, que en 1976, las tendencias de drogadicción de la zona sur eran similares a las del centro, pero la situación cambió para la década posterior, pues, quienes registraron tendencias similares fueron el norte y el centro.

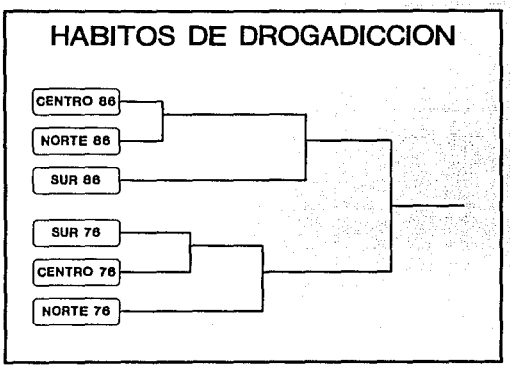


FIGURA IV.10. Dendrogramas para los datos sobre tendencias de drogadicción por zonas y período.

En cuanto a los datos del nivel de vida de Aguascalientes durante 1980, la figura IV.11 refleja la situación de similaridad entre los nueve municipios del Edo. Al cortar el dendrograma con una línea vertical, se distinguen con claridad los cinco cúmulos de la figura IV.5 pero, ahora se puede determinar que los cúmulos unitarios estaban formados por la capital del Edo., Pabellón de Arteaga y San José de Gracia.

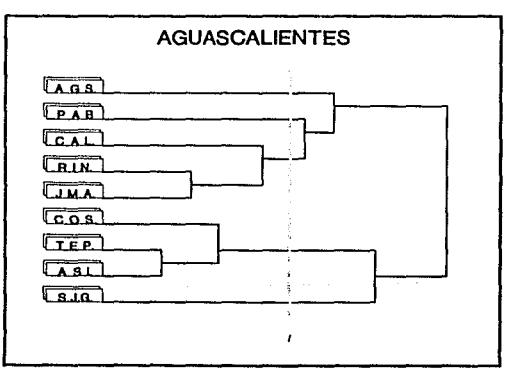


FIGURA IV.11. Municipios de Aguascalientes.

El primer grupo formado de izquierda a derecha está integrado por los elementos que representan a Asientos y Tepezalá respectivamente, esto significa que la similitud entre los niveles de vida en ambos lugares es más fuerte que entre cualquier otro. En ese orden, el siguiente cúmulo lo integran Jesús María y Rincón de Romos. Un paso después, Cosío se incorpora al primer grupo, y más adelante, Calvillo se une a Jesús María y Rincón de Romos. Es importante aclarar que el orden de las uniones no corresponde al orden jerárquico de los índices, es decir, la primer unión

se dá porque la similitud más estrecha se dió entre Tepezalá y Asientos pero esto no significa que ambos registren los índices del nivel de vida más bajos.

Por medio del dendrograma, se puede *rectificar* o *ratificar las hipótesis* que se habían planteado intuitivamente por medio de otras técnicas. De esta manera, se confirma que los índices del nivel de vida en Asientos, Tepezalá y Cosío son semejantes, pero los de Calvillo se parecen más a los indicadores de Jesús María y Rincón de Romos que a San José de Gracia y Pabellón de Arteaga, como se había supuesto a través de los diagramas de estrellas, y además, las condiciones de vida en estos dos últimos lugares y la capital del estado son por demás diferentes a los otros poblados.

En la figura IV.12 se tiene el dendrograma que corresponde a las variables. El grupo unitario está formado por la variable B6, otro conjunto está formado por las variables B1, B5 y B10 lo que significa que existe una relación entre las tasa de la población que recibía, en 1980, ingresos menores a $3611.00 con la población de 15 años o más con la primaria incompleta y con los índices de vivienda sin tubería de drenaje.

Por otro lado, la tasa de PEA que labora desde menos de una hora hasta 32 horas a la semana se relaciona con la tasa de analfabetismo de la población de 10 años o más. La tasa de viviendas sin energía eléctrica se asocia con las viviendas de un solo cuarto y con piso de tierra.

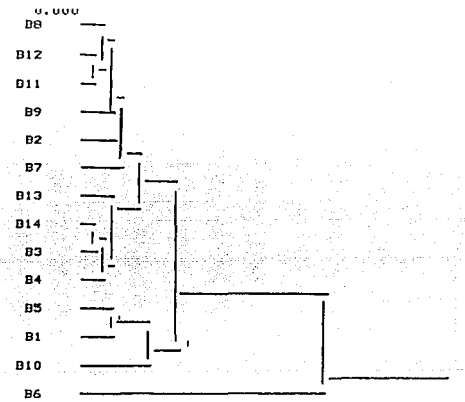


FIGURA IV.12 Variables del nivel de vida del Edo. de Aguascalientes.

D. Ventajas y Desventajas.-

Al tomar como ciertas algunas suposiciones sobre la estructura de los datos, se puede llegar a un resultado erróneo (de la misma manera que en las técnicas no jerárquicas) pues, es posible que se este tratando de imponer una jerarquía en grupos que no la tienen.

Desde el punto de vista de Mardia (1979), "El uso de métodos jerárquicos involucra perdida de información" debido al reemplazo de la matriz de distancias originales por una nueva matriz de distancias que satisfacen la desigualdad ultramétrica*.

La selección de métodos jerárquicos o no jerárquicos, depende del enfoque del problema.

En cuanto al número de cúmulos a formar, aun no se tiene un criterio definido.

Las técnicas de cúmulos pueden ser útiles como métodos de reducción de dimensiones.

* Apéndice C.

Los métodos jerárquicos permiten la construcción de un diagrama de árbol o dendograma.

La mayor dificultad de las técnicas jerárquicas radica en la selección de la medida de distancias.

La representación de un dendrograma de matrices de similaridades o de distancias es mucho más útil si los datos están fuertemente marcados en los cúmulos jerarquizados.

E. Revisión Actualizada

La forma de los dendrogramas no es única, esto es, el diseño puede modificarse de acuerdo a las necesidades. Las siguientes figuras muestran diferentes presentaciones del árbol de los municipios de Aguascalientes.

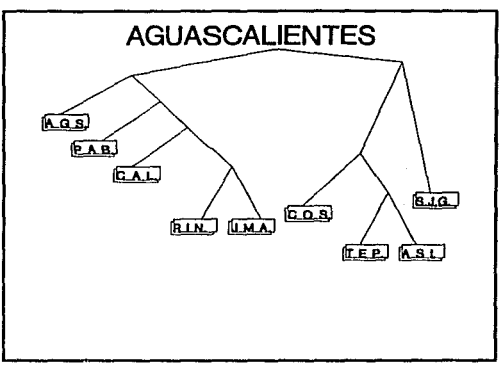


FIGURA IV.13 Dendrogramas para los municipios de Aguascalientes

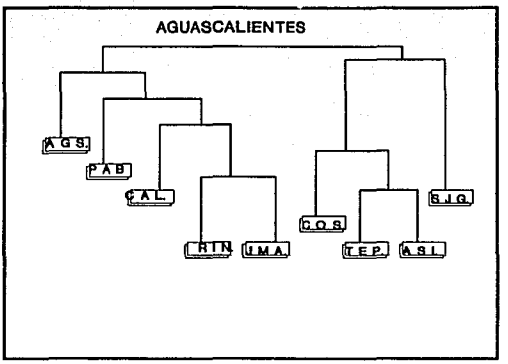


FIGURA IV.14 Dendrogramas para los municipios de Aguascalientes

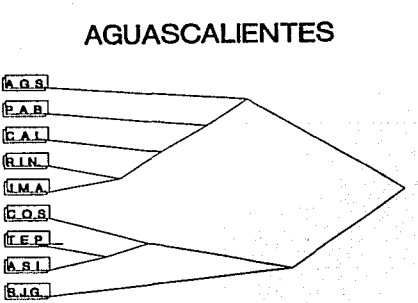


FIGURA IV.15 Dendrogramas para los municipios de Aguascalientes

Los dendrogramas no son la única alternativa de graficación para las distancias entre individuos o grupos. Chambers (1982) presenta una serie de diagramas que representa esta información.

MATRIZ DE DISTANCIA SOMBREADA.-

Es una técnica útil para representar la distancia entre objetos individuales.

Los objetos se reacomodan de tal manera que los objetos del mismo cúmulo son adyacentes unos a otros. Mientras las distancias incrementan, decrecen los niveles de sombreado. de esta manera, se tiene una serie de triángulos bajo cada estrecho y bien delimitado cúmulo.

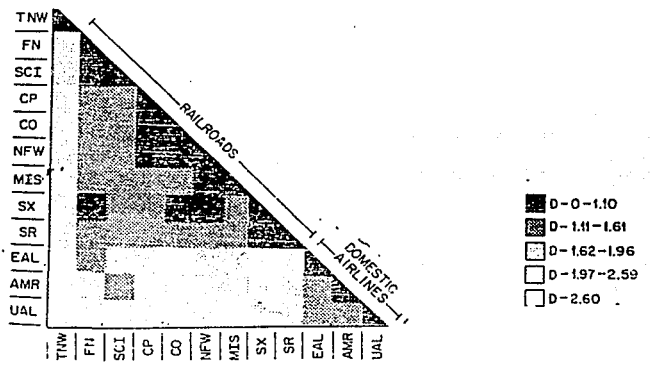


FIGURA IV.16. Matriz de distancias.

DISTANCIA ENTRE CENTROS

Fowlks (1976) propone cortar el dendrograma en un cierto nivel creando con esto un deseado número de cúmulos. Los cúmulos resultantes son representados por círculos con diámetros iguales a los diámetros de los cúmulos. La interpretación del diámetro del círculo depende del algoritmo y la métrica usada. Para el caso de la distancia euclidiana, el diámetro es igual a la máxima distancia entre cualesquiera dos objetos dentro del cúmulo. Finalmente, los círculos son conectados por lineas horizontales y verticales cuyas longitudes son iguales a las distancias entre los cúmulos correspondientes.

Por ejemplo, se tiene el dendrograma de la fig.IV.17. Se hace el corte a la altura del nivel 2 generando 6 cúmulos, 4 de ellos son unitarios. La distancia entre los centros de los dos círculos es 2.07, es decir, la distancia entre los cúmulos representados por los dos círculos. La distancia entre TNW y la mitad de la línea que conecta los dos círculos es 2.43, o sea, la distancia entre TNW y la unión de los 2 círculos.

Aparte de los diámetros de los círculos, sólo importan las longitudes de las rectas horizontales y verticales.

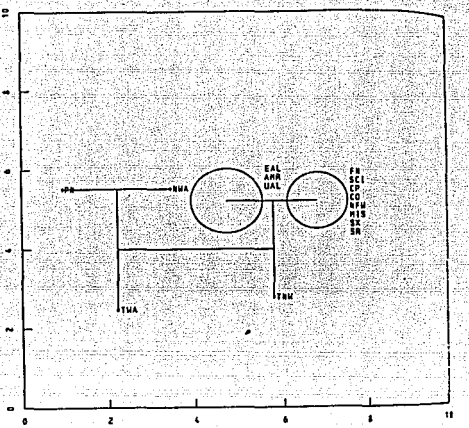


FIGURA IV.17. Distancias entre centros.

DISTANCIAS ENTRE CENTROS Y OBJETOS.

Dunn y Landwehr (1980) proponen un diagrama que involucra las distancias entre los centros de los cúmulos y los objetos.

Para cada grupo se gráfica la distancia entre su centro y todos los objetos que están cerca del centro, indicando cuales objetos pertenecen al cúmulo. Si los objetos en el cúmulo están cerca del centro, el grupo puede considerarse como estrecho. de lo contrario, se dice que el cúmulo es amplio. Por este medio se detectan los objetos lejanos del centro en un cúmulo relativamente estrecho. si el objeto permanece cerca del centro de otro cúmulo, puede considerarse una revisión y la posibilidad de asignarse a otro cúmulo.

OBJETOS INDIVIDUALES Y VARIABLES

El objetivo es representar el comportamiento de un conjunto dado de cúmulos y sus objetos para cada variable individual.

Se elaboran diagramas separados para cada variable donde los valores de cada objeto son graficados contra el número de cúmulo que contiene a este objeto. por este medio, se observa la localización de los diferentes cúmulos para cada variable y su dispersión dentro del cúmulo. Una variación a la técnica consiste en substraer el centro de cada cúmulo y después graficar. Con esto se puede comparar la dispersión de cada cúmulo y detectar puntos fuera, pero se pierde información sobre los niveles de los cúmulos.

Con el objetivo de observar la relación de los centros de los cúmulos, Dunn y Landwehr (1980) sugieren graficar las diferencias entre un centro y todos los centros de las demás variables; para observar la relación entre un objeto y el centro del cúmulo, proponen graficar las diferencias entre el objeto y el centro del cúmulo.

Si las variables están dadas en diferentes unidades, se deben estandarizar las diferencias por medio de la división entre una medida de la escala dentro del cúmulo.

REFERENCIAS.-
1.- Everit B.S. (1978). Graphical Techiques for multivariate data.
2.- Mardia Mutivariate Analysis
3.- Gnanadesikan (1977). Methods for Statiscal Data Analysis of multivariate observations.
4.- Hair, J. Multivariate data Analysis: With readings
5.- Chambers (1982). Graphical Techniques for multivariate data Analysis and for clustering.
6.- Fowlkes (1976). Graphical Techniques for displaying multidimensional clusters.
7.- Dunnand Landwehr (1980). Analysis clustering effects across time. JASA. 75, 8-15
8.- Lebart et al (1984) Multivariate descriptived statistical Analysis. New York.

IV.2 ESCALAMIENTO MULTIDIMENSIONAL

A. Descripción de la Técnica.

La técnica de Escalamiento Multidimensional (EM) es un conjunto de procedimientos que permiten desplegar, mediante gráficas, la interelación entre distancias o similaridades e inferir las dimensiones importantes a partir de las mismas.

Los algoritmos pueden ser métricos o no métricos.

El escalamiento multidimensional no métrico busca la representación de coordenadas de matrices de disimilaridades o distancias usando únicamente su propiedad ordinal. El mejor diagrama es aquel que preserva el ordenamiento de la relación entre el rango de los datos originales y las distancias entre puntos.

Los métodos métricos emplean la escala numérica ya sea discreta o continua para medir las características de los individuos.

En general, el propósito de la técnica es proporcionar una gráfica útil para dar una interpretación significativa de los datos obteniendo información acerca de las relaciones entre los objetos.

Un ejemplo de aplicación real de las técnicas de Escalamiento Multidimensional son los problemas de mercadotecnia donde se logra, mediante este proceso, determinar la competitividad entre productos y deducir el criterio que la gente usa cuando juzga un conjunto de objetos (productos).

B. Desarrollo de la Técnica.-

ALGORITMO NO METRICO (Shepard - Kruskal)

Se tiene la matriz D de disimilaridades generada al medir alguna característica entre los n individuos.

Por ejemplo, si n = 4, se tiene

$$D = \begin{matrix} 0 & & & \\ \delta_{21} & 0 & & \\ \delta_{31} & \delta_{32} & 0 & \\ \delta_{41} & \delta_{42} & \delta_{43} & 0 \end{matrix}$$

Debido a la simetría, D tiene n(n-1)/2 elementos que es posible ordenar de manera ascendente.

Para el ejemplo, se supone la siguiente relación:

$\delta_{21} < \delta_{41} < \delta_{42} < \delta_{31} < \delta_{32} < \delta_{43}$

El objetivo principal es representar los n individuos en un espacio euclidiano.

Se hace una propuesta inicial sobre la localización de los puntos (individuos) en el plano cartesiano.

En base a lo anterior, se calculan las n(n-1)/2 distancias entre los n puntos.

Las distancias se ordenan de manera ascendente como las disimilaridades.

El orden de las distancias debe corresponder, en índices, al de las disimilitudes. Esto es, en el ejemplo: $\delta_{21} < \delta_{41} < \delta_{42} < \delta_{31} < \delta_{32} < \delta_{43}$

En caso de que el orden indicado no se cumpla, se construye una nueva propuesta y se

calculan las distancias d^*_{ij} cuya naturaleza depende de los d_{ij} de tal manera que se aproximen al orden establecido por los índices de las disimilaridades.

La falta de ajuste entre la representación propuesta y la relación original de las disimilaridades, se mide con el STRESS cuya expresión es la siguiente:

$$S=\frac{\sqrt{min\sum(d_{ij}-d^*_{ij})^2}}{\sum d^*_{ij}}$$

La idea, entonces, es iniciar un proceso iterativo cuya finalidad es la de encontrara una expresión de distancias ordenadas en forma creciente tan 'cercana' como sea posible a la indicada por las disimilitudes. La estrategia principal para lograr un 'buen' ajuste es obtener un valor mínimo para el STRESS. Una regla propuesta por Kruskal para el juicio de tolerabilidad de S es: S 20% muy pobre, S = 10% regular, S 5% bueno, S = 0 perfecto.

ALGORITMO METRICO (SOLUCION CLASICA)

Se tiene una matriz simétrica D(*nxn*) de distancias d_{ij}

1.- A partir de D se construye la matriz A.

$$A = (-1/2\, drs)$$

2.- Se obtiene la matriz B con $b_{rs} = a_{rs} - \bar{a}_{r.} - \bar{a}_{.s} + \bar{a}_{..}$

3.- Se encuentran los k eigenvalores mayores (previa elección de k) , k < n.

$\lambda_1 , \lambda_2 , ..., \lambda_k$ de B con sus correspondientes eigenvectores $X = (x_1 , x_2 , ... , x_k)$ el cual es normalizado por $x_i x_i = \lambda_i \quad i = 1, 2, 3, \ldots , k.$ (Debe suponerse que todos los primeros k eigenvalores son positivos.)

Si los primeros k eigenvalores de B son positivos, entonces, las distancias entre puntos de esta configuración sera aproximadamente D. Esta configuración es llamada la solución clásica de MDS en k dimensiones. Es una solución métrica.

C. Aplicación de la Técnica.-

Al describir la técnica se habló del objetivo del Escalamiento Multidimensional: desplegar graficamente la interelación entre distancias o similaridades. También se afirmó que "el mejor diagrama es aquel que preserva el ordenamiento de la relación entre el rango de los datos originales y las distancias entre puntos". En este tono parece un buen ejemplo el intento de reproducir un mapa. La fig IV.18 es el resultado de aplicar la técnica de escalamiento multidimensional a una matriz diagonal cuyos datos son las distancias entre las capitales de algunos de los estados de la República Mexicana (tabla 10). Después de 50 iteraciones, la representación obtenida tiene 3.3% de error.

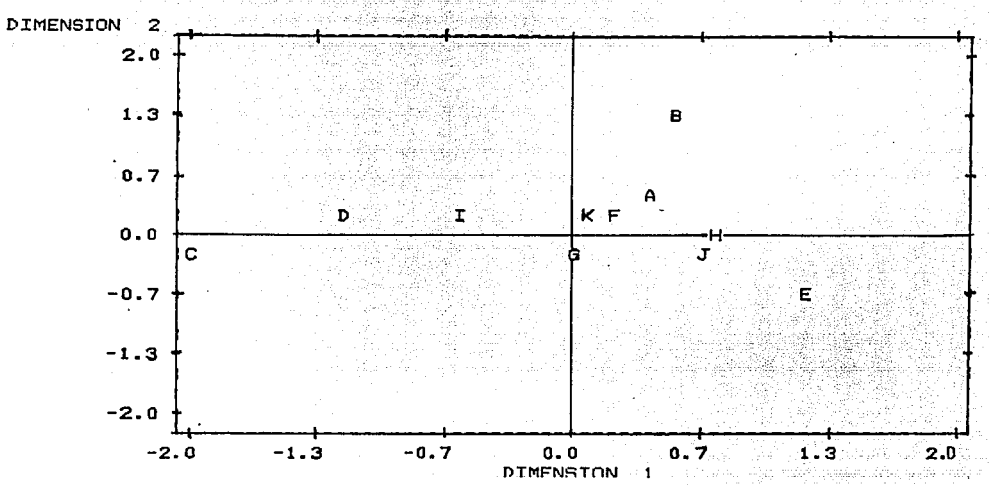


FIGURA IV.18. Escalamiento Multidimensional. Mapa.

Para los datos de contaminación en febrero de 1989, el resultado del escalamiento multidimensional tiene un stress del 2.4%. En el se refleja una similitud entre las zonas Suroeste (SO) y Noreste (NE).

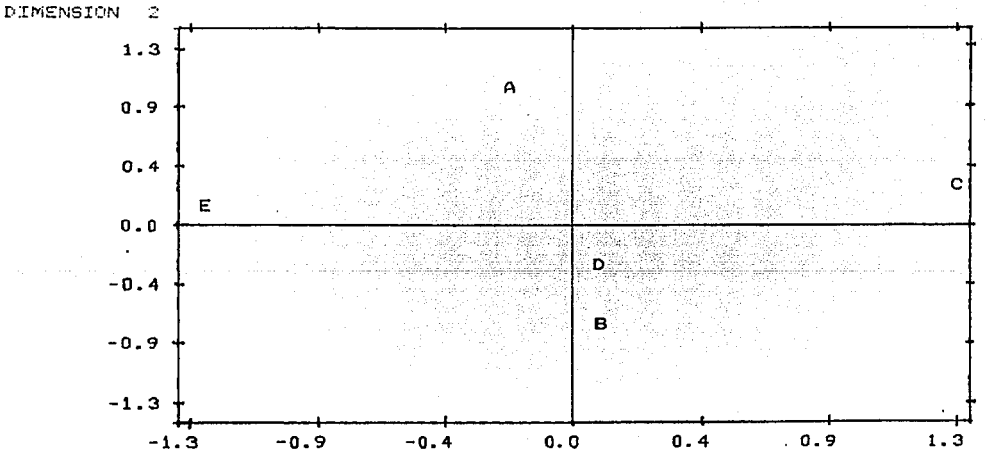


FIGURA IV.19. Escalamiento Multidimensional. IMECA por zonas en febrero 1989.

Para el ejemplo de las tendencias de drogadicción, en la primera iteración se logró la representación perfecta (S=0). De esta manera, se comprueba la estrecha similitud entre las zonas Centro y Sur en 1976, por un lado y, el Norte y Centro en 1986 por el otro, tal como se había distinguido en el dendrograma correspondiente (fig. IV.10)

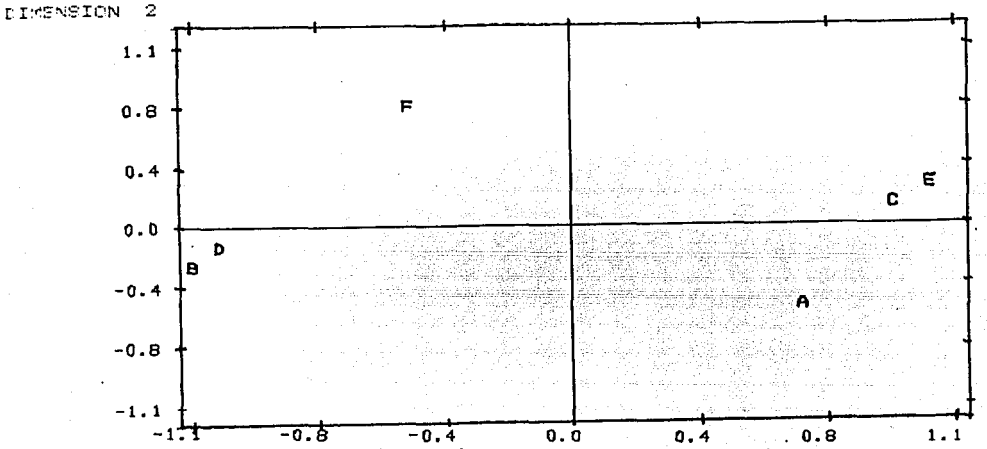


FIGURA IV.20. Tendencias de drogadicción.

D. Ventajas y desventajas.-

Por cálculos computacionales es más simple el método clásico que el no métrico. Quizá debido a la fortaleza que le da las transformaciones monótonas de la función distancia. Sin embargo, ambos métodos parecen dar respuestas similares. El método no métrico es invariante bajo rotaciones, traslaciones y expansión o contracción uniforme de la mejor configuración. La mejor configuración proporcionada por el método no métrico se obtiene a partir de una configuración arbitraria. Esta puede ser obtenida por la solución clásica.

REFERENCIAS

1.- Anderson, Al (1986) Multidimensional Scaling in product development. The fascination of statistics. New York 103-109

2.- Carroll, J.D. (1980). "Multidimensional scaling". Ann Rev. Psychol. 31 607 - 49

3.- Gnanadesikan (1977). Methods for Statiscal Data Analysis of multivariate observations. Ney York: J. Willey

4.- Mardia, K.V. (1979). Multivariate Analysis. London. Ney York: Academic.

5.- Hair, J. (1987). Multivariate data Analysis: with reading: Anderson, R.E. Ney York:MAcmillan

IV.3 ANALISIS DISCRIMINANTE

A. Descripción de la Técnica.-

Análisis discriminante es la técnica estadística apropiada para problemas que involucran varias variables métricas independientes y una sola variable categórica dependiente. A diferencia del análisis de cúmulos, Análisis discriminante busca la descripción y clasificación de elementos teniendo los grupos pre-establecidos por medio de una variable categórica. Un ejemplo claro es el problema que plantea distinguir entre aquellos individuos que son buenos sujetos de crédito de los que no lo son (variable categórica) con base a sus ingresos, datos familiares, antigüedad en el trabajo, edad, etc. (variables métricas). Esta técnica identifica las características donde existe la mayor diferencia entre grupos: deriva un coeficiente discriminante (peso) para cada variable que reflejan estas diferencias y asigna cada individuo a un grupo.

B. Desarrollo de la Técnica.-

El enfoque que se dá en este trabajo al Análisis Discriminante no asume ninguna forma particular para la distribución de las muestras pues sólo se busca una regla discriminante entre ellas.

Sea $\Psi = [Y_{ij}]$ la matriz de datos con n renglones (individuos u observaciones) y p columnas (variables). La media de la variable j es

$$\overline{Yi} = \frac{1}{n}\sum_{i=j}^{n} y_{ij}$$

Los n renglones de Y son particionados a priori en q grupos. El grupo k esta caracterizado por un conjunto I de n valores de índice i con

$$\sum_{k=1}^{q} n_k = n$$

Sea $\overline{Y}_{kj}$ la media de la variable j en el grupo k

$$\overline{Y}_{kj} = \frac{i}{n_k}\sum_{i \in I_k} Y_{ij}$$

Para cualquier variable j se tiene que

$$\overline{Y}_j = \sum_{k=1}^{q} \frac{n_k}{n} \overline{Y}_{kj}$$

La covarianza total entre dos variables j y j' es

$$Cov\,(j,j') = \frac{i}{n}\sum_{k=1}^{q} \left[\sum_{i \in I_k} (Y_{ij} - Y_j)\,(Y_{ij} - Y_{j'}) \right] \frac{i}{n_k} \sum_{i \in I_k} Y_{ij}$$

Por otro lado

$$((Y_{ij}) - Y_j) = (Y_{ij} - Y_{kj}) + (Y_{kj} - Y_j) \qquad \text{y}$$

$$\sum_{i \in I_K} (Y_{ij} - Y_{Kj})\,(Y_{Kj'} - Y_{j'}) = (Y_{Kj'} - Y_{j'}) \sum_{i \in I_K} (Y_{ij} - Y_{Kj}) = 0$$

Factorizando en cuatro términos, donde dos de los cuales son cero y por la definición de $\overline{Y}_{kj}$ se tiene:

$$\sum_{i \in I_K} (Y_{ij} - Y_{Kj})(Y_{Kj'} - Y_{j'}) = (Y_{Kj'} - Y_{j'}) \sum_{i \in I_K} (Y_{ij} - Y_{Kj}) = 0$$

De manera análoga

$$\sum_{i \in I_K} (Y_{Kj} - Y_j)(Y_{ij'} - Y_{Kj'}) = 0$$

Con lo que se obtiene la ecuación del análisis de varianza

$$Cov(j, j') = \frac{1}{n} \sum_{K=1}^{q} \sum_{i \in I_K} (Y_{ij} - Y_{Kj})(Y_{ij'} - Y_{Kj'}) + \sum \frac{nK}{n} (Y_{Kj} - Y_j)(Y_{Kj'} - Y_{j'})$$

En forma matricial queda T = W + B

donde

T = covarianza total
W = covarianza dentro del grupo
B = covarianza entre grupos

Sea a_i el valor de una combinación lineal de las p variables centradas para el individuo i

$$a_{(i)} = \sum_{j=1}^{p} a_j (Y_{ij} - Y_i)$$

La varianza de la variable derivada a_i tiene el siguiente valor donde a_i esta centrada

$$Var(a) = \frac{1}{n} \sum_{i=1}^{n} a_i^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{p} a_j (Y_{ij} - Y_i) \right]^2$$

$$Var(a) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{j'=1}^{p'} a_j a_{j'} (Y_{ij} - Y_i)(Y_{ij'} - Y_{j'})$$

cambiando el orden de la sumatoria, se obtiene

$$Var(a) = \sum_{j=1}^{p} \sum_{j'=1}^{p} a_j a_{j'} Cov(j, j') = a'Ta$$

donde a es el vector cuyos p componentes son

$$a_p\, a_2 \ldots, a_p$$

La varianza de la combinación lineal a es particionada en varianza dentro y entre grupos.

$$\mathbf{a'Ta = a'Wa + a'Ba.}$$

Entonces, el problema puede ser formulado así: Entre todas las combinaciones lineales de las variables hay que encontrar aquellas que tienen una varianza máxima entre grupos (para maximizar las diferencias entre grupos) y mínima dentro de los grupos (para minimizar la dispersión dentro del grupo). Estas combinaciones lineales son las funciones lineales discriminantes.

Se busca entonces, el vector a tal que

$$\frac{a'Ba}{a'Wa} \quad sea\ máxima\ o \quad \frac{a'Wa}{a'Ba} \quad sea\ mínima$$

A manera de resumen, dado un problema de análisis discriminante,la computadora deriva una combinación lineal de las variables independientes. El resultado es una serie de valores discriminantes para cada individuo en cada grupo. Los valores son calculados bajo el criterio de maximizar la varianza entre grupos y minimizarla interiormente.

La técnica de Análisis discriminante es aún más significativa por su **representación gráfica.** Los valores discriminantes se localizan en un plano cartesiano con lo que se consigue el despliegue de los n individuos con p características en un espacio de dimensión dos. Dada la naturaleza de técnica de clasificación, en el diagrama se distinguen los grupos.

C. Aplicaciones de la técnica.-

Ejemplo 1.-

Una inmobiliaria cuenta con tres ofertas de vivienda para las diferentes posibilidades económicas de sus clientes:

-El primer paquete consiste en un departamento de $42m^2$ con una recámara.

-El segundo incluye un departamento de dos recámaras y $60m^2$ de construcción.

-La tercera oferta es para el cliente que busque comodidad y desée vivir en una zona residencial.

Para asegurar que un cliente tiene capacidad de solventar los gastos que implica la adquisición, la inmobiliaria hace una investigación sobre cada uno de los solicitantes en cuanto a los siguientes tres aspectos:

a)Antigüedad en el trabajo.

b)Ingresos semanales (x $ 100,000)

c)Saldo a favor en la cuenta de cheques (x $1'000 000) Para tomar la decisión adecuada sobre el tipo de vivienda a ofrecer, la inmobiliaria toma como base el resultado de encuestas que en su momento se hicieron a clientes que no han presentado problemas en los pagos. Se eligió una muestra de 6 personas por grupo (tabla 11). La función discriminante (combinación lineal de las variables independientes) resultante es:

$$D = .85x + .13x - 1.56x$$

la representación gráfica se encuentra en la figura IV.21 donde no es difícil percatarse de los tres grupos.

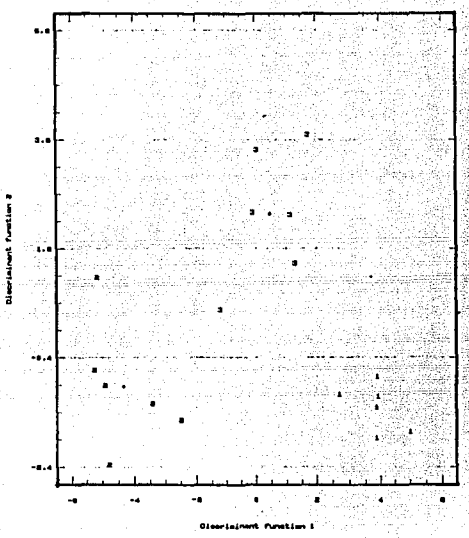


FIGURA IV.21. Análisis discriminante.

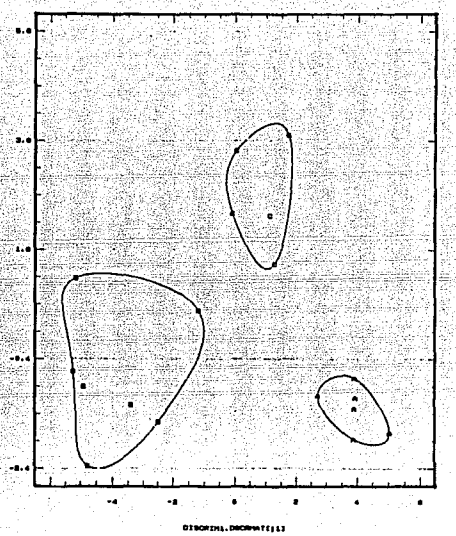


FIGURA IV.22. Cúmulos en los resultados de A. Discriminante.

El resultado de aplicar Análisis de cúmulos a los valores generados por la función discriminante, es el diagrama de la figura IV.22

Ejemplo 2.- Una institución educativa, ha mostrado un serio interés al problema de la enseñanza de las matemáticas en el nivel medio superior. A través de un examen de exploración ha podido detectarse que los alumnos pueden clasificarse en tres grupos:

a) Aquellos que no muestran habilidades numéricas.

b) Los que tienen habilidades pero no pueden resolver problemas que involucran raciocinio.

c) Los alumnos que logran conjugar las habilidades numéricas y el razonamiento para resolver problemas con mayor grado de dificultad.

La estrategia de atención que se sigue en cada grupo es diferente por lo que es importante ubicar a cada alumno de nuevo ingreso en el grupo adecuado. Para hacer la selección, se toma

como base el resultado obtenido por 21 alumnos (7 de cada grupo) en las tres fases de la evaluación (tabla 12). La función discriminante que se obtiene es:

$$D = .77x_1 + .29x_2 + .12x_3$$

La representación gráfica en el plano cartesiano está en la figura IV.23.

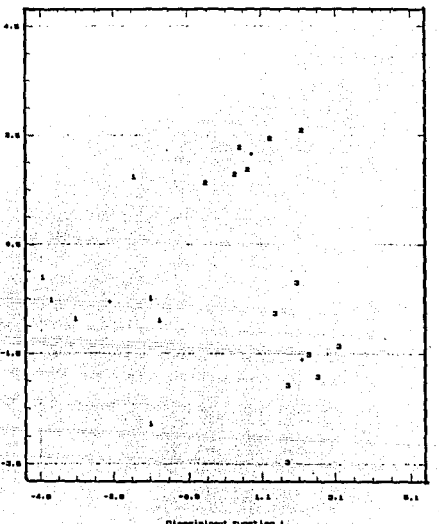


FIGURA IV.23. Análisis discriminante. Estudiantes de matemáticas.

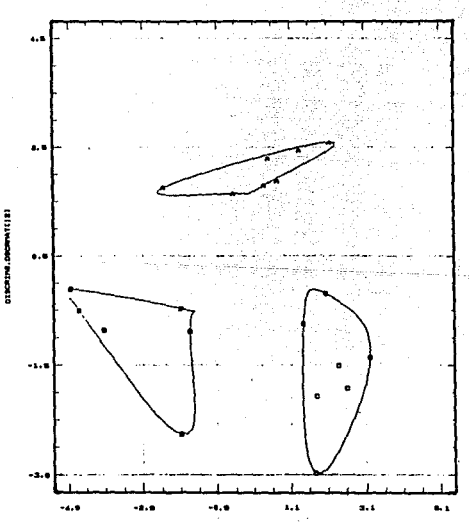


FIGURA IV.24. Cúmulos

IV.4 ANALISIS DE COMPONENTES PRINCIPALES (A.C.P.)

A. Descripción de la Técnica.-

El Análisis de Componentes Principales es una técnica que permite describir un conjunto de datos por medio de la representación gráfica de las observaciones. Dada la naturaleza multivariada de los datos (n individuos o elementos con p variables) se hacen necesarios una serie de cálculos algebraicos para "resumir" los puntos en un pequeño número de combinaciones lineales de los datos originales. Una vez determinadas las combinaciones lineales (son los componentes principales) estos forman los ejes coordenados en la gráfica y entonces, todas las observaciones son transformadas para quedar plasmadas en términos de los nuevos ejes sin perder información ni distorsionar la relación entre las variables originales.

B. Desarrollo de la técnica.-

Aplicando esta técnica, se pretende introducir un nuevo conjunto de combinaciones lineales ortogonales llamadas Componentes Principales, tal que la varianza de los puntos dados, con respecto a estas coordenadas derivadas están en orden decreciente de magnitud. Así, el primer componente principal es aquel cuya proyección de los puntos dados tiene máxima varianza entre todas las posibles combinaciones lineales, el segundo componente principal tiene máxima varianza después del primero y ademas, es ortogonal a el.

Las coordenadas originales $x_1, x_2, x_3, \ldots, x_n$ son transformadas por un cambio de origen a la media de la muestra $x_1, x_2, x_3, \ldots, x_n$, seguidas de una rotación rígida.

Considerese:

a) La matriz de datos
b) El vector media
c) La matriz de covarianza

de la siguiente manera :

a) La matriz de datos

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

n observaciones en un espacio p-dimensional.

Toda la información del individuo i con respecto a las p variables:

$$X_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \cdot \\ \cdot \\ \cdot \\ x_{in} \end{pmatrix}$$

Toda la información de la variable j sobre todos los individuos:

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \cdot \\ \cdot \\ \cdot \\ x_{nj} \end{pmatrix}$$

b) El vector media

$$\mu = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{pmatrix} \qquad donde \qquad x_i = \frac{\sum_{i=1}^{n} x_{i1}}{n}$$

c) La matriz se covarianza S (pxp)

$$S = \begin{bmatrix} x_1^2 & x_{12} & \cdots & x_{1p} \\ x_{12} & x_2^2 & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{1p} & x_{2p} & \cdots & x_p^2 \end{bmatrix}$$

La matriz de covarianza tiene validez en el análisis cuando las variables están dadas en las mismas unidades de medida. Si las variables pertenecen a diferentes unidades (edad en anos, peso en kg., estatura en metros, etc.), las combinaciones lineales tienen poco significado, por lo que se deben estandarizar las variables y usar la matriz de correlaciones.

Si se estandariza cada coordenada dividiendo entre la desviación estandard, entonces, la matriz de covarianza de las variables estandarizadas es justamente la matriz de correlación de las variables originales.

Los componentes principales están expresados en términos de combinaciones lineales de las variables observadas.

Así, el primer componente de las n observaciones en un espacio p-dimensional es

$$Y_1 = a_{11}x_1 + a_{21}x_2 + \ldots + a_{p1}x_p$$

cuyos coeficientes a_{11} son los elementos del vector característico a_1 asociado con la mayor

raíz característica λ_1 de la matriz de covarianza S y si se cumple la restricción $a_1'a = 1$ (el vector es unitario), la raíz característica λ_1 es interpretada como la varianza de Y_1 .

El segundo componente principal es la combinación lineal

$$Y_2 = a_{12}x_1 + a_{22}x_2 + \ldots + a_{p2}x_p$$

cuyos coeficientes serán escogidos bajo las siguientes restricciones

a) $a_2'a_2 = 1$ para asegurar la unicidad de los coeficientes.

b) $a_1'a_2 = 0$ es decir, $a_1 \perp a_2$.

Los coeficientes del segundo componente son elementos del vector característico correspondiente a la segunda raíz característica.

Los siguientes componentes principales son encontrados en su momento, a partir de los vectores característicos.

El uso y la importancia de cada componente sera medido por la proporción del total de varianza atribuible a el.

$$\frac{\lambda_j}{\sum\lambda} = \frac{\lambda_j}{trS}$$

donde $\lambda 1$ = la varianza del j-ésimo componente y tr S es la varianza total del sistema.

El signo y la magnitud de a los coeficientes indican la dirección y la importancia de la contribución de la i-ésima respuesta en el j-ésimo componente.

C. Aplicaciones de la Técnica.-

Los resultados de aplicar componentes principales a los datos de contaminación en el área metropolitana durante febrero de 1989 (tabla 4), son los siguientes:

La varianza explicada por los dos primeros componentes es del 79.95%.
La representación de las zonas Noroeste (A), noreste (B), centro (C), suroeste (D) y sureste (E) en un plano cartesiano se encuentra en la figura IV.24

La correlación entre las variables noreste, centro y sureste se refleja en este diagrama. Por otro lado, la figura IV.25 contiene la representación bidimensional de los 28 días del mes de febrero de 1989. A partir de este esquema pueden intuirse algunos cúmulos.

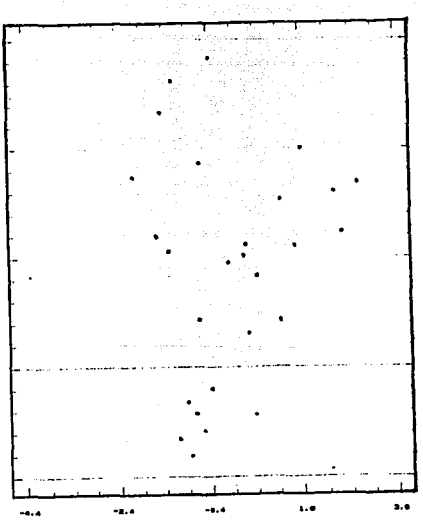


FIGURA IV.24. A.C.P.
IMECA, Febrero 1989 por día.

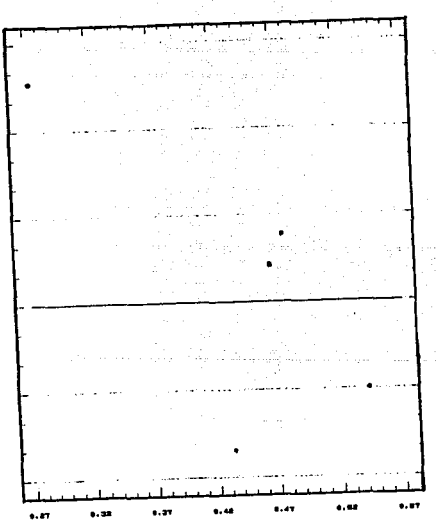


FIGURA IV.25. A.C.P. IMECA, Febrero 1989 por zona.

Para aclarar la idea de acumulamiento, en la figura IV.26 se tiene el resultado de aplicar Análisis de Cúmulos a la información obtenida por medio de A.C.P.

El Análisis de componentes principales también se aplicó a los datos sobre las tendencias de drogadicción de los jóvenes de la República mexicana.

Los dos primeros componentes principales explican el 93.48% de la varianza.

En la figura IV.27. se tiene el plano cartesiano con la localización de las tres zonas en ambos períodos.

Los puntos con etiquetas D y B (Norte y Centro 1986) permanecen juntos de la misma manera en que el dendrograma (fig.IV.10) los había declarado como elementos similares. La situación es la misma con los puntos C y E (Centro y Sur 1976).

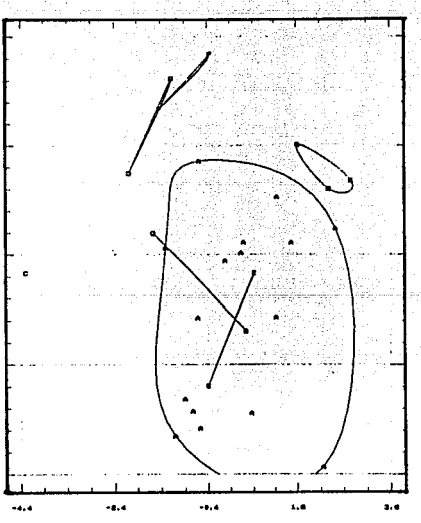


FIGURA IV.26. Acumulamiento del resultado de A.C.P

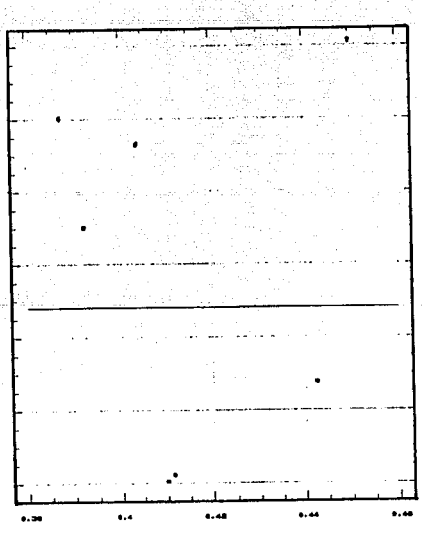


FIGURA IV.27. Tendencias de drogadicción por zonas.

La representación que corresponde a los tipos de drogas está en la figura IV.28 . En ella se pueden determinar 'a ojo' los elementos que son similares entre sí. Aunque, para tener una idea más clara y precisa, en la figura IV29 se encuentran armados los grupos. En base a los esquemas, se determina, de nueva cuenta, que el consumo de sedantes, alucinógenos, heroína y cocaína es similar entre sí. (Se observó con anterioridad por medio de otras técnicas y se ratificó con el dedrograma de la fig IV.9).

Además, conjugando las tres zonas y los dos períodos, la mariguana y las anfetaminas han tenido similar demanda entre los jóvenes estudiantes. Los puntos alejados corresponden a los tranquilizantes y a los inhalantes que, por este medio reflejan poco parecido a las demás drogas aunque, el dendrograma correspondiente indica que finalmente tienden a formar parte del grupo de la mariguana y las anfetaminas.

D. Ventajas y desventajas.-

La técnica logra transformar el conjunto original de datos en un conjunto de puntos que son una aproximación a la configuración geométrica real pero están sobre una base diferente de menor dimensión con lo que es posible observar su representación gráfica. Por medio de Análisis de Componentes Principales, es posible identificar cúmulos y "puntos fuera". Se debe tener especial cuidado al hacer las interpretaciones y no perder de vista que las distancias entre individuos son explicadas en términos de patrones similares de respuesta a las variables (dos individuos cercanos son "parecidos") y la distancia entre variables debe entenderse en términos de correlación.

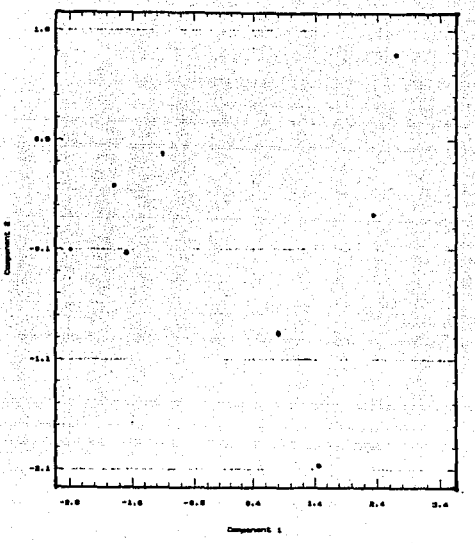


FIGURA IV.28. Tendencias de drogadicción por droga.

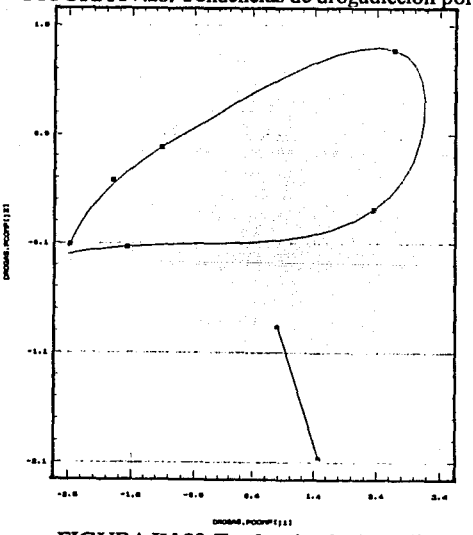


FIGURA IV.29. Tendencias de drogadicción. Cúmulos.

E. Revisión Actualizada.-

Gabriel propone otra técnica de graficación para datos multivariados llamada BIPLOT. Este diagrama permite observar en forma simultánea y en un plano cartesiano, el despliegue de la relación entre individuos en términos de distancias y entre variables como covarianzas y correlaciones. La varianza de las variables también tiene representación gráfica lo que hace posible la observación de los puntos individuales y sus diferencias. En la figura IV.30 se tiene el BIPLOT correspondiente a los datos del nivel de vida de Aguascalientes. La longitud de cada vector es proporcional a la varianza respectiva. Los valores cercanos corresponden a las variables correlacionadas. La localización de los puntos dá cuenta de la relación entre individuos (municipios para el ejemplo) y su tendencia de agrupación.

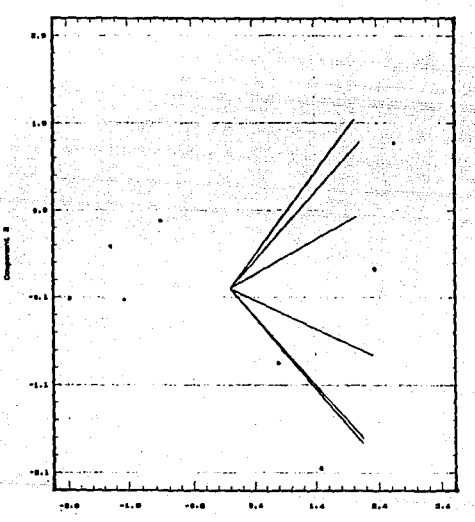


FIGURA IV.30 BIPLOT

REFERENCIAS.-

1.- Everit B.S. (1978). Graphical Techiques for multivariate data. Ney York: North Holland.

2.- Hair, J. (1978) Multivariate data Analysis: With readings. Ney York. Macmillan.

3- Lebart (1984) Multivariate descriptived statistical Analysis. New York.

4.- Mardia K.V. (1979) Mutivariate Analysis. Ney York: Academic.

3.- Gnanadesikan (1977). Methods for Statiscal Data Analysis of multivariate observations.

IV.5 ANALISIS DE CORRESPONDENCIAS

A. Descripción de la Técnica.-

El Análisis de Componentes Principales (A.C.P.) y el Análisis de Correspondencias (A.C.) tienen en común la idea de reducir la dimensión del espacio, es decir, intentan representar por medio de una gráfica, una matriz rectangular de datos en términos de un pequeño número de combinaciones lineales de los datos originales sin perder información ni distorsionar la relación entre las variables.

La diferencia entre ambos, radica en el tipo de datos a analizar: En A C P, las columnas de la matriz de datos son generalmente un conjunto de variables o medidas tomadas sobre una muestra relativamente homogénea de objetos o individuos que representan los renglones de la matriz. En tanto que Análisis de Correspondencias es una técnica apropiada para tablas de contingencia.

Análisis de Correspondencias es una técnica que gráfica simultáneamente la nube de renglones y la nube de columnas de la matriz de datos.

B Desarrollo de la Técnica.-

Sea **X** (nxp) la matriz de observaciones:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

donde **X** representa el número de individuos que tienen las características i y j

El interés de la técnica no está enfocado al análisis de las frecuencias absolutas sino en las distribuciones de los puntos, es decir, lo importante es analizar la frecuencia condicional de pertenecer a la columna j dado que pertenece al renglón i o viceversa.

$$\text{Sea } x_{i\bullet} = \sum_{j=1}^{p} x_{ij} \qquad \text{el total de personas en el renglón } i$$

$$x_{\bullet j} = \sum_{i=1}^{n} x_{ij} \qquad \text{el total de personas en la columna } j$$

$$x = \sum_{ij} x_{ij} \qquad \text{la muestra total}$$

En análisis de Correspondencias se definen dos nubes de puntos: la que forman los n renglones en R^p y la formada por las p columnas en R^n. El objetivo es poner en correspondencia ambas nubes en un espacio multidimensional y representarlas gráficamente de manera simultánea en un plano cartesiano.

El *j-ésimo* componente del *i-ésimo* vector en R^p es

$$\frac{x_{ij}}{x_{i\bullet}} \qquad j=1,2,3,...,p$$

El *i-ésimo* componente del *j-ésimo* vector en R^n es

$$\frac{x_{ij}}{x_{\bullet j}} \qquad j=1,2,3,...,p$$

Las frecuencias relativas son:

$$f_{ij} = \frac{x_{ij}}{x} \quad ; \quad \sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij} = 1$$

$$f_{i\bullet} = \sum_{j=1}^{p} f_{ij} = \frac{x_{i\bullet}}{x} \quad ; \quad \sum_{i=1}^{n} f_{i\bullet} = 1$$

$$f_{\bullet j} = \sum_{j=1}^{p} f_{ij} = \frac{x_{\bullet j}}{x} \quad ; \quad \sum_{i=1}^{p} f_{\bullet j} = 1$$

De esta manera, en R^p, se tienen n renglones-puntos ($i=1,2,3,...,n$) donde las coordenadas del punto i son $\frac{f_{ij}}{f_{i\bullet}}$ (en frecuencia relativa); a cada punto se le da un peso que es proporcional a su frecuencia, entonces $f_{i\bullet}$ es el peso asignado al punto i.

En R^n se tienen p columnas-puntos ($j=1,2,3,...,p$) donde las coordenadas del punto j son $\frac{f_{ij}}{f_{\bullet j}}$ y el peso asignado a este punto es $f_{\bullet j}$.

La distancia entre los puntos i e i' en R^n es

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{\bullet j}} \left[\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right]^2$$

análogamente, la distancia entre dos puntos j y j' en R^p es

$$d^2(j,j') = \sum_{j=1}^{n} \frac{1}{f_{i\bullet}} \left[\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right]^2$$

El objetivo entonces, es encontrar el eigenvector u de la matriz **X'NXM** correspondiente al eigenvalor mayor λ; donde :

X es la matriz de datos,

N es la matriz diagonal (nxn) cuyos elementos son los pesos de los *n-puntos*,

M es la matriz de distancia simétrica positiva.

C. Aplicación de la Técnica.-

Un ejemplo de aplicación del Análisis de Correspondencias es el que Guadalupe Irizar y Delfino Vargas (1987) desarrollaron en el estudio sobre la evaluación sensorial de panes elaborados con mezclas de harina de trigo y girasol.

El experimento consistió en elaborar 10 panes de 100g. Nueve de ello contenían una mezcla de harina de trigo con 5%, 10% y 15% de harina de girasol. El décimo pan fue llamado testigo y contenía 100% harían de trigo.

- *9 panes con mezcla de harina de girasol*
  - *Criollo Atotonilco*: 5%, 10%, 15%
  - *Rio Verde*: 5%, 10%, 15%
  - *CIANOC-2*: 5%, 10%, 15%
- *1 testigo 100% trigo*

Se diseñó un cuestionario en el que 15 jueces (catadores) evaluaron cuatro características del pan:

- *color*: *A Blanco*, *B Crema*, *C café*
- *aroma*: *A Excelente*, *B Agradable*, *C regular*, *D malo*
- *textura*: *A muy suave*, *B suave*, *C semisuave*, *D duro*
- *sabor*: *A Excelente*, *B bueno*, *C regular*, *D malo*, *E desagradable*

Antes de desarrollar Análisis de Correspondencias, los autores sometieron la información

a un Análisis de Componentes Principales con el objeto de seleccionar las observaciones que pertenecen a una cierta región de consenso y a los catadores más consistentes en sus juicios de esta manera se generó la tabla de contingencia que aparece en el apéndice A (tabla 13).

El Análisis de Correspondencias se realizó a partir de los datos para determinar las características de las mezclas aceptables para los jueces.

Los resultados que obtuvieron los investigadores se enlistan a continuación:

La inercia total es de 0.7902 y las tres inercias principales son 0.2505 (31.7%), 0.1582 (19.78%) y 0.1394 (17.6%). Dado que la suma de las dos primeras inercias acumula más de la mitad de la inercia total (varianza), se les consideró como los dos primeros ejes principales.

La figura IV.31 es la representación bidimensional simultánea de las características (columnas) y las mezclas (renglones). Cada característica tiende a dibujar una parábola en el plano.

Es claro que las características favorables se localizan a la izquierda: color blanco, aroma excelente, textura muy suave y sabor excelente; mientras las desfavorables permanecen del lado derecho.

Las mezclas al 10%. y 15% de harina de girasol tipo criollo atotonilco, y al 15% del tipo CIANOC-2 recibieron mejores calificaciones que las demás mezclas, aunque no tan buenas como el testigo. En el otro extremo, las mezclas al 10% y 15% del tipo CIANOC-2 resultan malas.

En general, para obtener buenos resultados con una mezcla de harina de girasol tipo Criollo Atotonilco, ésta debe ser al 10% o al 15%, de otra manera, la calidad del pan es 'regular'; en tanto, con el tipo Río Verde, lo más que se logra con una mezcla al 15% es una textura suave; en las demás proporciones, pasa de regular a malo. Y, finalmente, con el tipo CIANOC-2 se tiene éxito sólo en la mezcla al 15%, las dos restantes son malas.

D. Ventajas y Desventajas.-

El Análisis de Correspondencias es una técnica de gran valor para el análisis de tablas de contingencia. Por medio de esta técnica se obtiene una representación gráfica simultánea del espacio de los individuos y el de sus características. Es posible observar los datos multivariados en un plano cartesiano. La interpretación de la gráfica requiere de especial cuidado; los ejes principales no tiene necesariamente una interpretación específica.

REFERENCIAS*

1.- Lebart (1984), Multivariate descriptive Statistical Analysis. New York.

2.- Greenacre, M. and Trevor, H. (1987). The geometric Interoretation of Correspondence Analysis. JASA Vol 82-398, 437-447

3.- Irizar G., Vargas D. (1987). Análisis de Correspondencia como método para la evaluación sensorial de panes elaborados con mezclas trigo-girasol. Agricultura Técnica en México Vol. 13 153-159

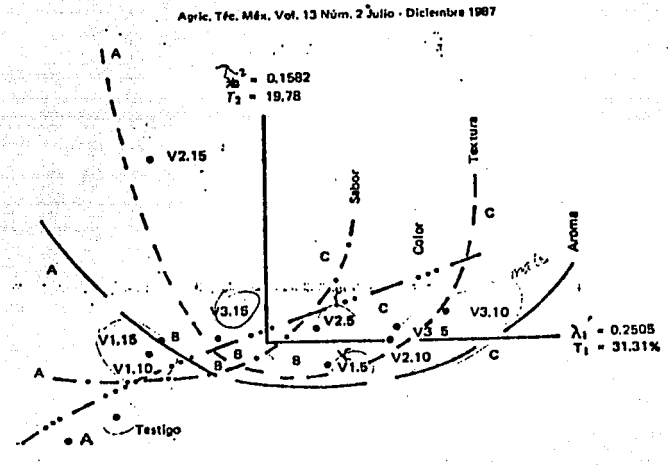


FIGURA IV.31. Análisis de Correspondencias.
Representaión gráfica de los resultados del estudio sobre evaluación sensorial de panes

CAPITULO
V

V CANASTA BASICA

La Secretaría de Comercio y Fomento Industrial (SECOFI) y el Instituto Nacional del Consumidor (INC) publican periódicamente una lista de los precios que alcanzan los productos de la canasta básica en los diferentes almacenes de autoservicio del Distrito federal con el objeto de dar, a las amas de casa, un elemento de comparación al hacer sus compras.

Las tablas 14 y 15 contienen los datos que estas instituciones emitieron el 16 de marzo de 1990 y el primero de abril de 1991 respectivamente. Cada columna contiene el valor de los diferentes productos en una sola tienda y, cada renglón incluye el costo de un producto de la misma marca e igual contenido, salvo los casos del arroz y el frijol cuyas marcas varían.

Los datos marcados con asterisco no fueron proporcionados de manera oficial por no encontrarse el articulo referido en la tienda correspondiente en el momento del muestreo, por lo que los valores fueron estimados por medio de un índice simple relativo de precios considerando el precio oficial de marzo de 1990 como el periodo base.

A partir de esta información, puede obtenerse una serie de conclusiones interesantes más alla de identificar el precio más barato en los diversos supermnercados, por ejemplo:

-Comparar las proporciones del ingreso familiar que absorvían los elementos de la canasta básica.

-Observar el comportamiento de los precios entre diferentes abarrotes.

-Estudiar las diferencias entre los precios de un sólo producto entre un período y otro.

-Analizar la actitud de las tiendas de autoservicio en cuanto a los precios que ofrecen de un período a otro.

-Determinar la diferencia en costo total entre adquirir la canasta básica en una tienda o en otra.

-Observar la similaridad entre almacenes en términos de precios al público.

V.1 LOS PRODUCTOS

Para empezar con una revisión global de los productos, los diagramas de estrellas son un buen recurso (fig V.1 y V.2). Cada estrella representa a un renglón de la tabla respectiva, esto es, la dimensión de los rayos dá cuenta del precio que cada producto alcanza en cada uno de los almacenes involucrados. En este sentido, se tiene que un kilogramo de carne molida de res requiere del porcentaje mayor sobre el gasto familiar en ambos perfodos (es la estrella mayor) en contraste con la pasta para sopa, la harina de trigo y la lata de leche clavel que generan los elemntos mas pequenos del conjunto.

La 'uniformidad' de los rayos de cada estrella reflejan que las diferencias de precios entre una tienda y otra no son excesivas,es decir, los datos no son muy dispersos.

En el diagrama de cajas múltiples (figuras V.3 y V.4), el tamaño relativo de las cajas confirma que los valores de cada producto en las diferentes tiendas son poco dispersos; aunque algunos de ellos muestran mayor variación que otros como es el caso de las tres presentaciones de café,la leche para lactantes, el arroz y la carne en 1990. La situación en cuanto a dispersión de datos en 1991, cambió en el sentido de que los precios tendieron a homogeneizarse.

Por este medio es más sencillo además observar la relación entre los precios de los

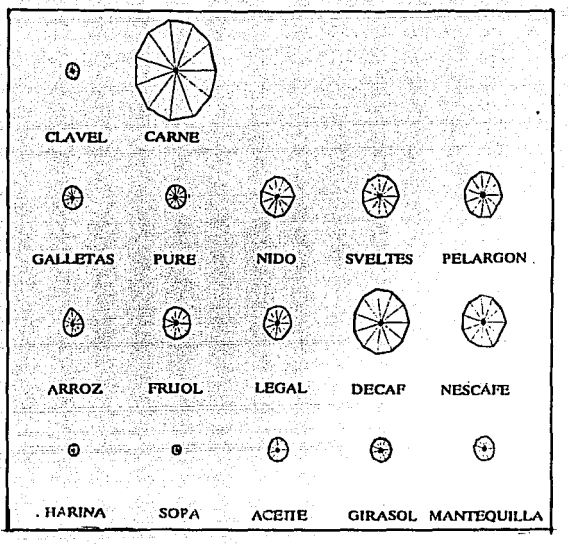
diferentes productos.

La caja mas alta es la que representa a los precios de 1kd. de carne de res y las que permanecen cercana a la base del diagrama son equivalentes a las estrellas menores (harina de trigo, sopa cora y leche clavel).

La posición relativa de las cajas es la misma en ambos perfodos excepto para el caso del arroz, frijol y el puré de tomate. Esta situación provoca la idea de una aparente equivalencia de los precios entre ambos perfodos, y la reducción del costo del arroz, frijol y puré de tomate no solo con respecto al valor del resto de los productos, sino también en relación al precio en que se ofrecían un año anterior.

Las figuras V.5, V.6 y V.7 son los diagramas cuantil- cuantil para los precios 90-91 del arroz, frijol y puré de tomate respectivamente. En ellos se refleja con claridad que de marzo de 1990 a abril de 1991, el costo de 1kg. de arroz y el de 1kg. de frijol se redujo notablemente. Para el caso del bote de puré de tomate, el importe permaneció casi sin cambios.

a) Marzo 1990



b) Abril 1991

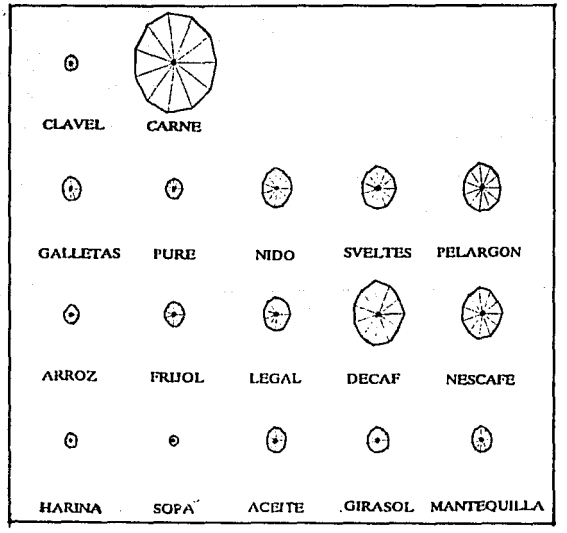


FIGURA V.1. Diagramas de estrellas. Cada símbolo representa el precio del producto en los diferentes almacenes. El rayo horizontalderecho es asignado a la primer columna de la tabla respectiva continuando la asignación en sentido contrario a las manecillas del reloj.

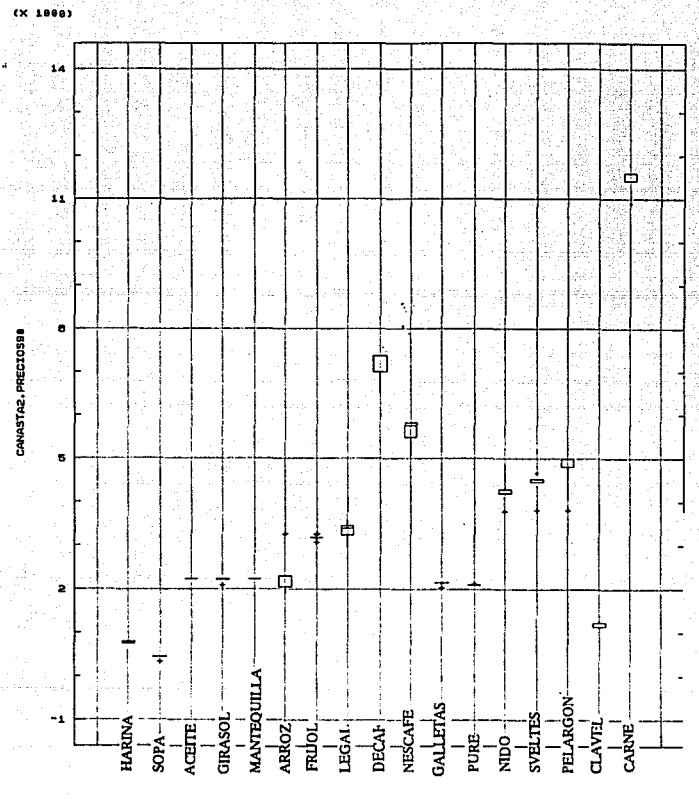


FIGURA V.3. Cajas multiples. Precios de los productos 1990.

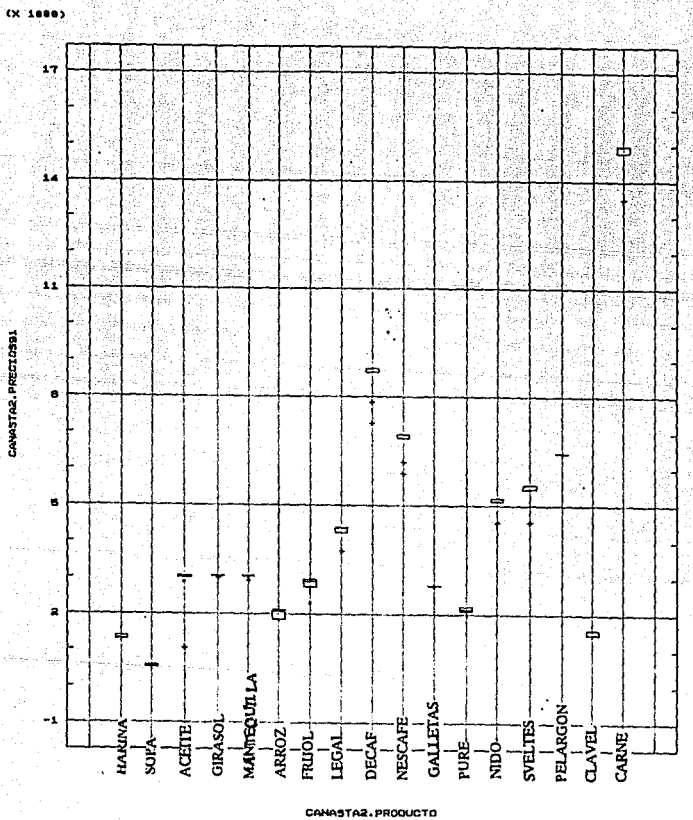


FIGURA V.4. Cajas multiples. Precio de los productos en abril 1991.

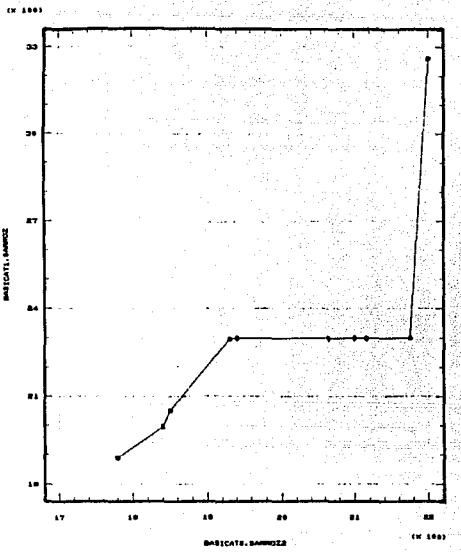


FIGURA V.5. Cuantil-cuantil. Arroz 1990-1991.

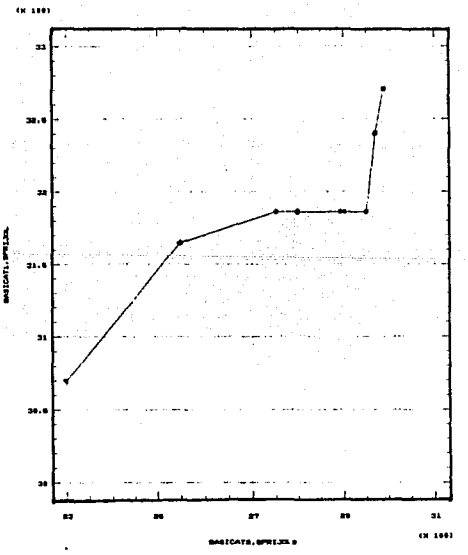


FIGURA V.6. Cuantil-cuantil. Frijol 1990-1991.

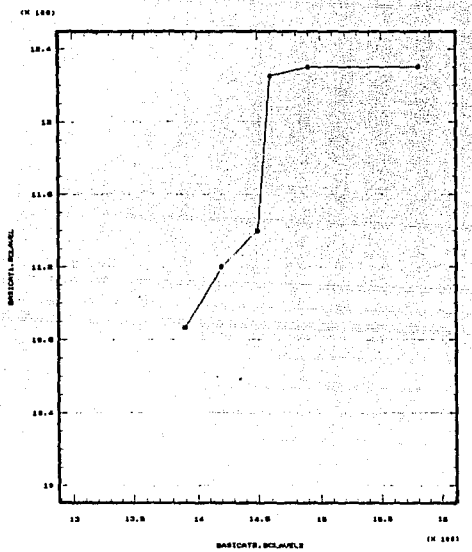


FIGURA V.7. Cuantil-cuantil. Leche Clavel 1990-1991.

V.2 LAS TIENDAS.

En las figuras V.8 y V.9 se tienen los diagramas de estrella que corresponden a los precios que ofrecían los diferentes almacenes de autoservicio. Cada símbolo es una tienda y la longitud de cada rayo representa el precio de cada uno de los 17 productos de la canasta básica.

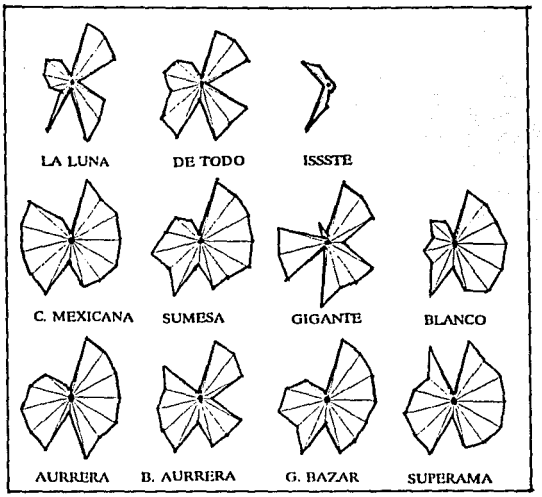


FIGURA V.8. Tiendas 1990.

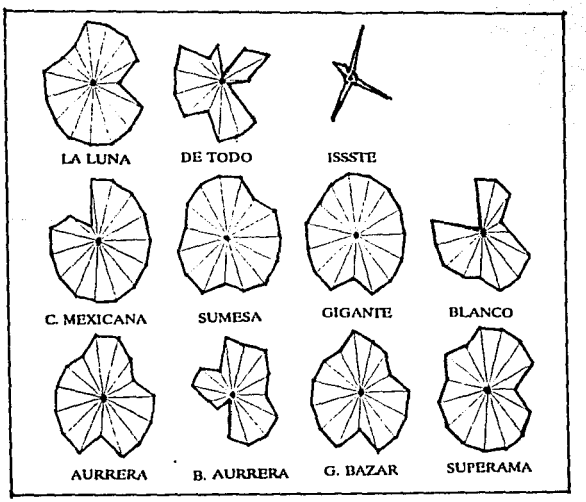


FIGURA V.9. Tiendas 1991.

Los símbolos semejantes reflejan la idea de tiendas similares en cuanto al comportamiento de los precios que ofrecen. así, entonces, en 1990 (fig V.8), las condiciones de compra en 'Comercial Mexicana' no eran muy diferentes a las ofrecidas por 'Gran Bazar'. pero, en 1991 (fig V.9), 'Gran Bazar" presentó precios similares a los de 'Sumesa'. En cuanto a la dispersión de los precios por tienda, los almacenes del sector público (ISSSTE) presentaron los datos más 'compactos' que el resto en ambos perfodos (fig V.10 y V.11).

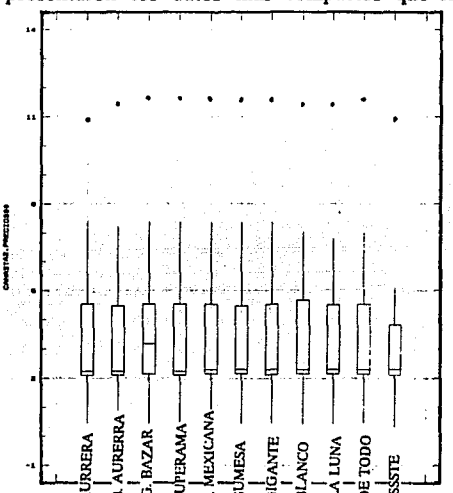


FIGURA V.10. Cajas multiples. Tiendas 1990.

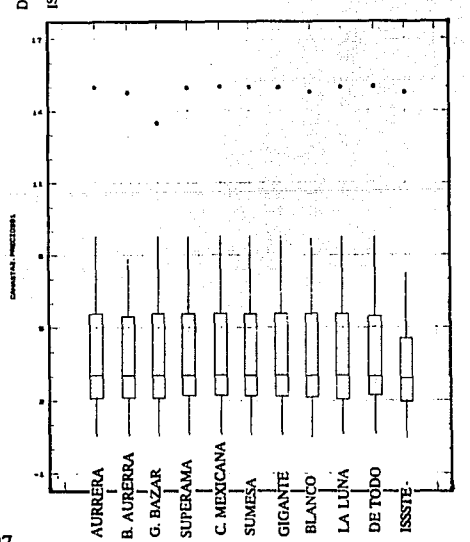


FIGURA V.11. Cajas multiples. Tiendas 1991.

La posición de la mediana habla de una clara asimetría de los datos. En marzo de 1990, los datos posicionados entre el primer cuartil y la mediana son de valores sumamente parecidos (poco dispersos) en contraste con los que rebasan el 59% de los datos. En abril de 1991, además de que el tamaño de las cajas tiende a igualarse, la posición de la mediana es idéntica por lo que se deduce que las condiciones de compra entre los diferentes almacenes se homogeneizó para el último período.

La idea de tiendas con precios 'parecidos' se refleja también en el diagrama de dispersión múltiple para cada período (figuras V.12 y V.13). de la misma manera en que en las cajas múltiples se distingue un punto aislado, en los diagramas de dispersión, la idea persiste.

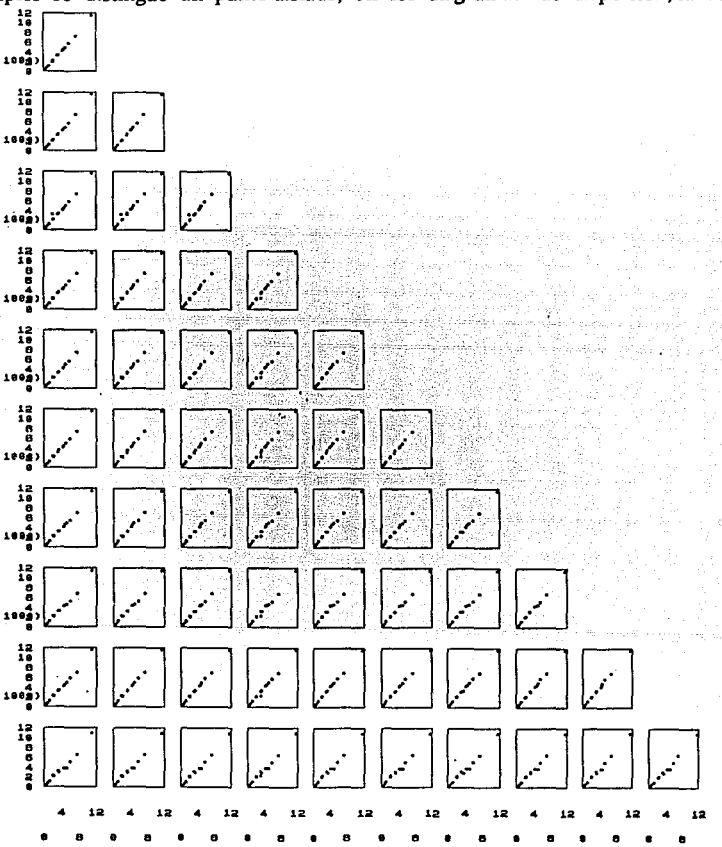


FIGURA V.12. Diagramas multiples de dispersión tiendas 1990.

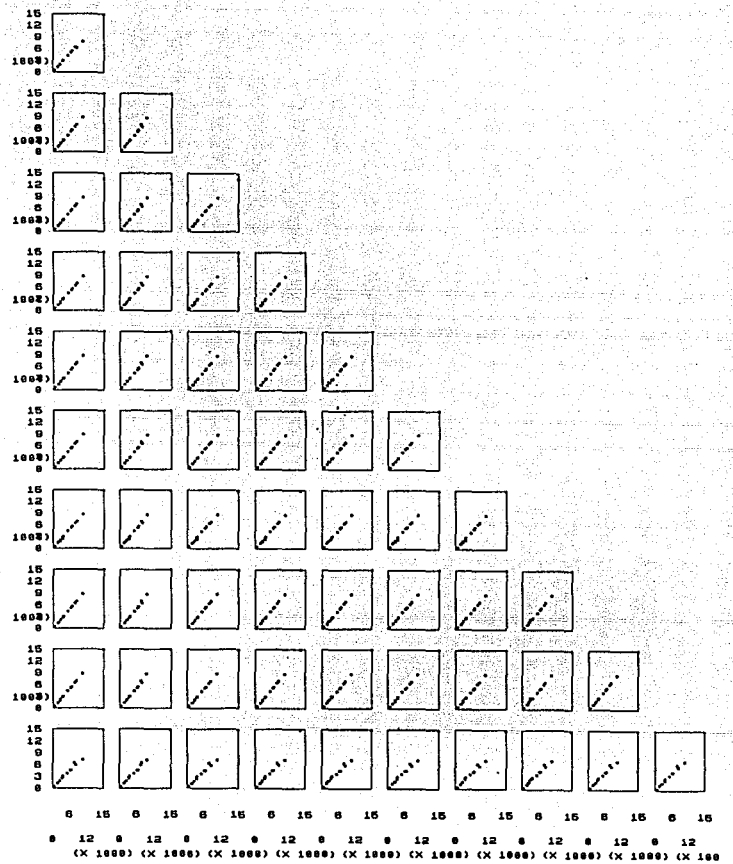


FIGURA V.13. Diagramas multiples de dispersión. Tiendas 1991.

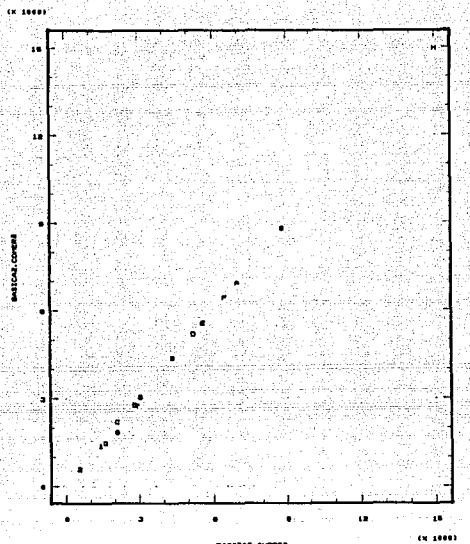


FIGURA V.14. Aurrerá vs Comercial Mexicana.

Al 'ampliar' el diagrama de dispersión de los almacenes Aurrera contra Comercial Mexicana, y asignar además una etiqueta a cada producto, (fig V.14) se distingue que el valor aislado corresponde al precio de la carne de res.

Pese a la apariencia similar entre las tiendas declarada por las gráficas de caja y los diagramas múltiples de dsipersión, algunas tiendas son más 'parecidas ' entre sí que con otras. El dendrograma de la figura V.15 indica que en 1990, los almacenes con mayor similaridad fueron 'Gigante' y 'Comercial Mexicana'. En un nivel posterior, 'Sumesa' se integró a los dos almacenes anteriores. Por otro lado, 'Bodega Aurrera' y 'Superama' también mostraron condiciones semejantes entre sí aunque la relación no era tan fuerte como la que se dió entre 'Gigante' y 'Comercial Mexicana'.

La tienda del sector público, ISSSTE, no tenía semejanza con algún otro autoservicio.

En 1991, se presentaron algunos cambios (fig.V.16), pues, 'La luna' mostró políticas de precios similares a las de 'Aurrera' y 'Gigante' y no a las de 'Blanco' como en el período anterior. La tienda ISSSTE, Bodega Aurrera y el almacén 'Gran Bazar' permanecieron completamente ajenos a sus competidores.

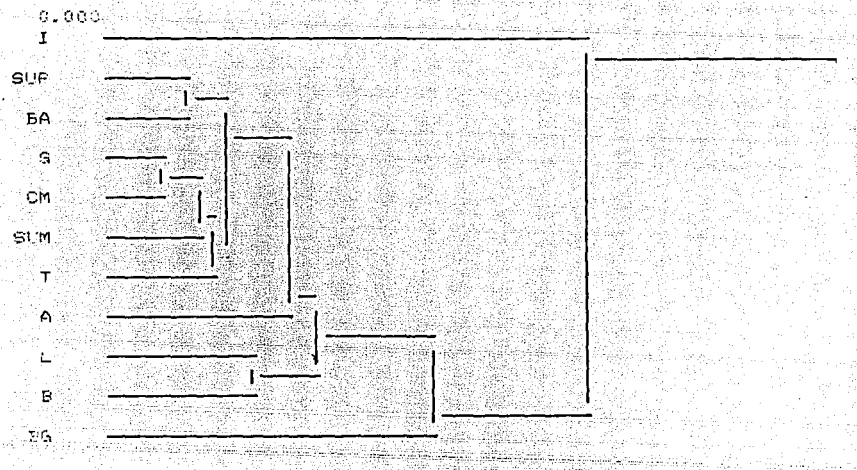


FIGURA V.15. Dendrograma. Tiendas 1990.

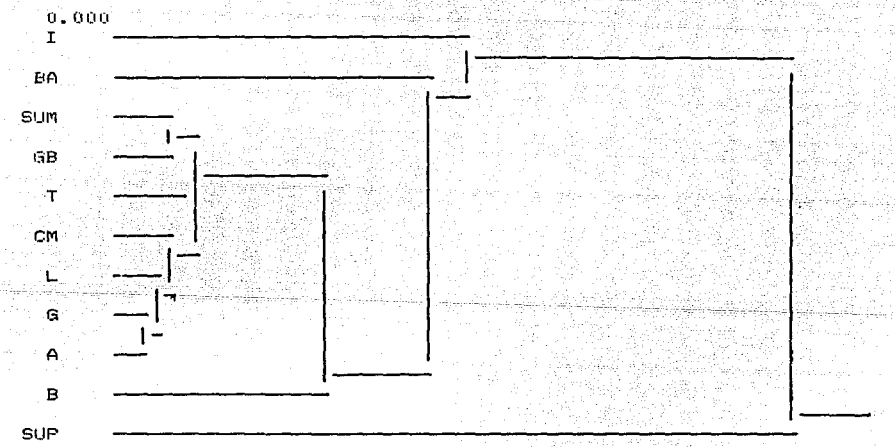


FIGURA V.16. Dendrograma. Tiendas 1991.

Para observar el comportamiento individual de un sólo almacén durante los dos perfodos, se tiene el diagrama cuantil-cuantil para el caso de 'Aurrera' (fig.V.17), 'Comercial Mexicana' (fig.v.18) y para la tienda ISSSTE (fig. V.19).

La curva que grafican las dos primeras es más suave que la generada por los precios del almacén de los trabajadores del Edo. esto indica que el aumento de precios entre 1990 y 1991 fue más brusco.

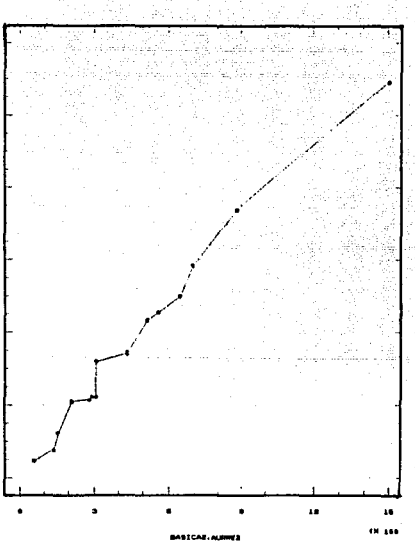


FIGURA V.17. Cuantil-cuantil.
Aurrera 1990 vs Aurrera 1991.

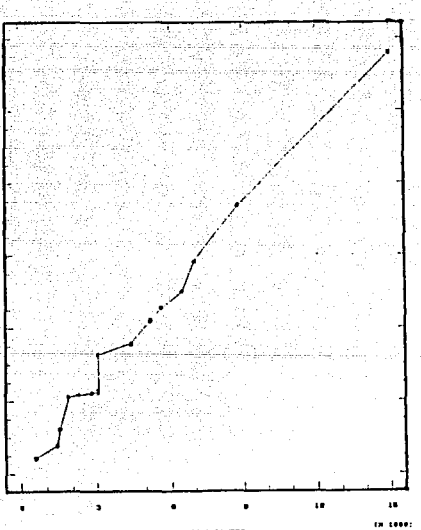


FIGURA V.18. Cuantil-cuantil.
C. Mexicana 1990 vs C. Mexicana1991.

El resultado de hacer un Análisis de Componentes Principales confirman varias observaciones que se han venido haciendo.

En la figura V.20 se tiene la representación gráfica, en un plano cartesiano formado por los dos primeros componentes principales, de las once tiendas en cuestión durante 1990.

Puntos cercanos representan correlación entre variables, esto es, los precios que ofrece Comercial mexicana están relacionados con los de Gigante y Sumesa tal como lo mostró el dendograma (fig. V.15). Por otro lado, la tienda ISSSTE está representado por un punto aislado por no tener relación con el resto de los elementos.

La posición de los puntos parece indicar que a la derecha de la gráfica se encuentran los almacenes con precios más altos en tanto a la izquierda se localizan aquellos que cuentan con precios más accesibles.

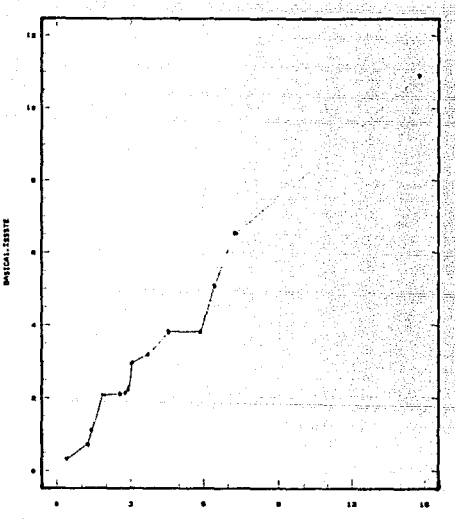


FIGURA V.19. Cuantil-cuantil. ISSSTE 1990 vs. ISSSTE 1991.

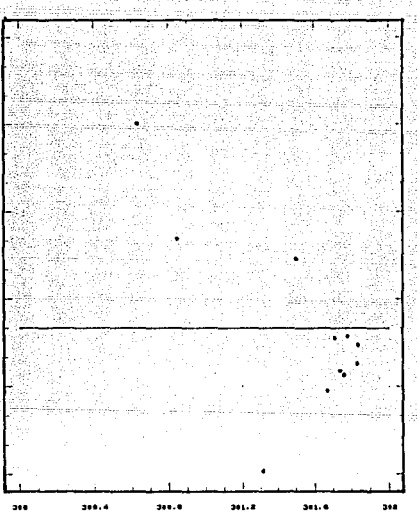


FIGURA V.20. Análisis de componetes pricipales. Tiendas 1990.

En la figura V.21 se tiene la localización de los productos a través de los dos primeros componentes principales.

Después de aplicar un análisis de cúmulos, los cinco grupos formados son los siguientes:

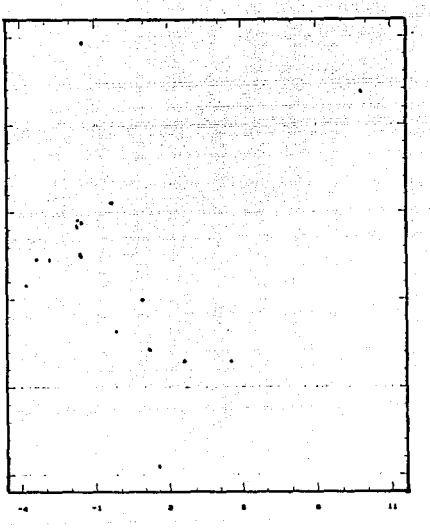


FIGURA V.21. Componentes principales. Productos 1990.

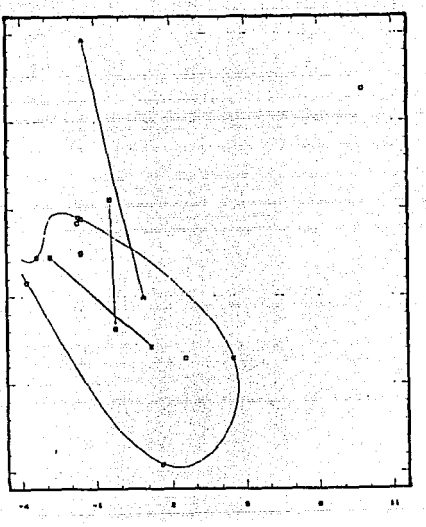


FIGURA V.22. Productos 1990. 5 cumulos

La distribución de los productos a lo largo del plano cartesiano respeta un orden cresciente, es decir, empieza por los productos cuyo importe es inferior como son la sopa, la harina y la lata de leche clavel (además forman un cúmulo); y termina el producto que siempre apareció como el más oneroso: el kg. de carne de res (cúmulo unitario).

En cuanto a la situación durante abril de 1991, el análisis de Componentes Principales reafirma la tendencia, por parte de las tiendas a homogeneizarse, pues mapean puntos sobrepuestos con la excepción para la tienda ISSSTE, Bodega Aurrera y Gran Bazar.

La representación gráfica de los productos en 1991 se encuentran en la figura V.23. La posición que presentan es diferente a la generada en 1990. Esto es, la relación que se daba entre los precios de los productos en 1990 se modificó en el siguiente período. Al aplicar análisis de Cúmulos(fig. V.24) parece reflejarse que más productos entran dentro de la clasificación de ser económicos (primer grupo).

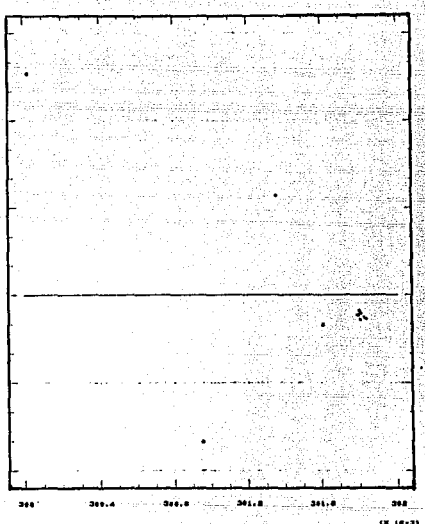


FIGURA V.23. Análisis de componentes principales. Tiendas 1991.

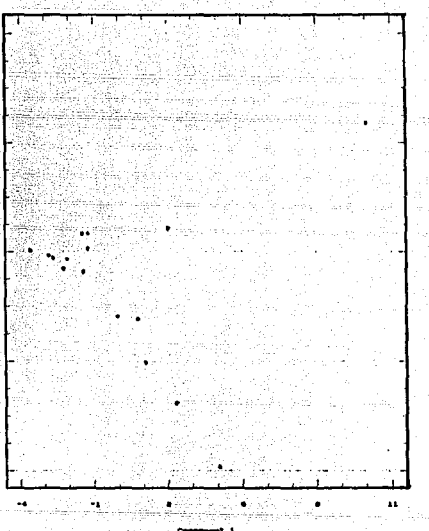


FIGURA V.24. Análisis de componentes principales. 1991.

CONCLUSIONES

Las representaciones gráficas son una valiosa herramienta en el Análisis de datos; por medio de la imagen visual de los datos, se distinguen aspectos importantes sobre su comportamiento que con frecuencia permanecen ocultos a través de los cálculos numéricos.

Las técnicas de graficación son un dispositivo que previene al investigador sobre la valides de los supuestos planteados y lo provee de información que le permite elegir el tratamiento adecuado a los datos.

Las técnicas de graficación permiten observar la representación de los datos en un plano cartesiano aún tratándose de problemas multivariados, con lo que la interpretación y obtención de conclusiones es más clara e inmediata que la información obtenida a través de la tabla de datos o de los cálculos numéricos.

No existe una técnica que funcione como llave maestra. Es decir, no se puede esperara toda la información necesaria a través de un solo diagrama; generalmente se requiere de la combinación de esquemas para obtener mejores resultados. En este sentido, se dice que las técnicas de graficación integran un LABORATORIO para los datos.

La selección de las técnicas a aplicar está en función de la naturaleza de los datos y los objetivos del estudio.

Las técnicas de graficación tienen un gran campo de aplicación en diferentes área, tantos como el número de problemas multivariados que pueden plantearse.

La computadora y la existencia de paquetes estadísticos hacen de las técnicas de graficación un recurso aún mas accesible. La representación de un conjunto de datos no se limita a los diagramas ya establecidos. La creatividad e ingenio del investigador para combinar las técnicas o para generar nuevos esquemas enriquece el tema de **"Técnicas de graficación para datos multivariados"**

APENDICES

APENDICE A.

LOS DATOS

TABLA 1

Registros de las calificaciones individuales de cincuenta estudiantes que se examinaron en Economía Política. Escala discreta. Datos ficticios utilizados para ejemplos.

60 33 85 52 65 77 84 65 57 74
71 81 35 50 35 64 74 47 68 54
80 41 61 91 55 73 59 53 45 77
41 78 55 48 69 85 67 39 76 60
94 66 98 66 73 42 65 94 89 88

Se tiene una sola variable que es la calificación obtenida por cada estudiante en una escala de 0 a 100 puntos. En total son 50 observaciones.

TABLA 2

El Hospital de Nutrición detectó en 1983 al primer caso de SIDA en México en un paciente aún con vida. En 1984, el Hospital Médico "La Raza" detectó el SIDA en una autopsia y, a partir de esa fecha a noviembre de 1988, había atendido a 376 personas con esta enfermedad.

El siguiente cuadro incluye datos de los primeros 220 pacientes con SIDA que atendió el Hospital 'La Raza'.

Se hicieron 3 clasificaciones: una por grupo de edad y sexo y otra por grupo de fuente de contagio. En total se tiene 3 variables categóricas: el grupo de edad, el sexo y el grupo de fuente de contagio

Grupo de edad	Masculino	Femenino
-15 años	2	2
15 - 24	29	2
25 - 44	142	5
45 - 64	37	-
65 o más	1	-
TOTAL	211	9

Grupo	Número de casos
Homosexuales (masculinos)	119
Bisexuales (masculinos)	53
Heterosexuales	19
Hemotransfundidos	10
Hemofílicos	2
Se ignora	17
	220

Fuente:López, Lucía (1989). La lucha por el tratamiento.ICYT. 11. 39-41

TABLA 3

El Instituto Mexicano de Psiquiatría realizó una investigación epidemiológica sobre las tendencias de drogadicción de estudiantes de enseñanza media y media superior en todos los estados de la República Mexicana en el período que abarca de 1976 a 1986. Los resultados de la investigación generaron la siguiente tabla de contigencias.

	REGION NORTE		REGION CENTRO		REGION SUR	
	1976	1986	1976	1986	1976	1986
	* 3247	2568	5643	6751	1010	596
	%	%	%	%	%	%
DROGAS						
Mariguana	1.9	3.7	.9	3.1	.8	1.6
Inhalantes	.8	4.2	1.0	4.5	.8	4.1
Anfetaminas	2.8	3.5	2.0	3.4	1.7	2.6
Tranquilizantes	1.9	2.6	2.9	2.4	3.1	3.6
Sedantes	1.2	.7	1.4	1.0	.6	.5
Alucinógenos	.7	.4	.8	.7	.1	.1
Cocaína	.6	1.3	.5	.9	.6	.6
Heroína	.2	.5	.4	.5	0.0	0.0
Totales	10.1	16.9	9.9	16.5	7.7	13.1

El problema consta de 3 variables:

-El tipo de droga con 8 categorías (una por droga)
-El período con 2 categorúas (1976-1986)
-La Región en sus 3 categorías (NORTE, CENTRO y SUR).

Los estados que contemplan cada región se enlistan a continuación:

REGION NORTE: Baja California Norte y Sur, Sinaloa, Sonora, Coahuila, Chihuahua, Tamaulipas y Nuevo León.
REGION CENTRO: Durango, San Luis Potosí, Nayarit, Aguascalientes, Jalisco, Michoacán, Guanajuato, Hidalgo, Estado de México, Distrito Federal, Puebla, Veracruz y Guerrero.
REGION SUR :Campeche, Tabasco, Yucatán, Chiapas y Oaxaca.

* población encuestada.

Fuente: Girón, Elvia (1988). La Huida Mágica. Información Científica y Tecnológica. 10. 44-46.

Para cuestión de interpretación de la tabla, se puede seguir el siguiente ejemplo:
En 1976, el 1.9% de los 3247 estudiantes encuestados en la región Norte, declaró adicción a la Mariguana.

IMECA

La Secretaría de Desarrollo Urbano y Ecología (SEDUE), es el organismo encargado de monitorear los niveles de Contaminación ambiental en el área metropolitana, generando así el Indice Metropolitano de Calidad del Aire (IMECA).

Esta institución ha dividido el área metropolitana en 5 zonas, a saber Noroeste (NO), Noreste (NE), Centro (CE), Sureste (SU), y Suroeste (SO) (ver mapa anexo en la siguiente página).

La clasificación por puntos, de los niveles de contaminación proporcionada por la fuente es la siguiente:

0-50 Situación muy favorable para la realización de todo tipo de actividades físicas.
51-100 Situación favorable para todo tipo de actividades.
101-200 Aumento de molestias menores en personas sensibles.
201-300 Aumento de molestias e intolerancia relativa al ejercicio en personas con padecimientos respiratorios y cardiovasculares: aparición de ligeras molestias en la población en general.
301-500 Aparición de diversos sintomas e intolerancia al ejercicio en la población sana.

Fuente: SEDUE

En la tabla 4, 5 y 6 se encuentran los datos relativos a febrero de 1989, noviembre de 1989 y febrero de 1990 respectivamente.

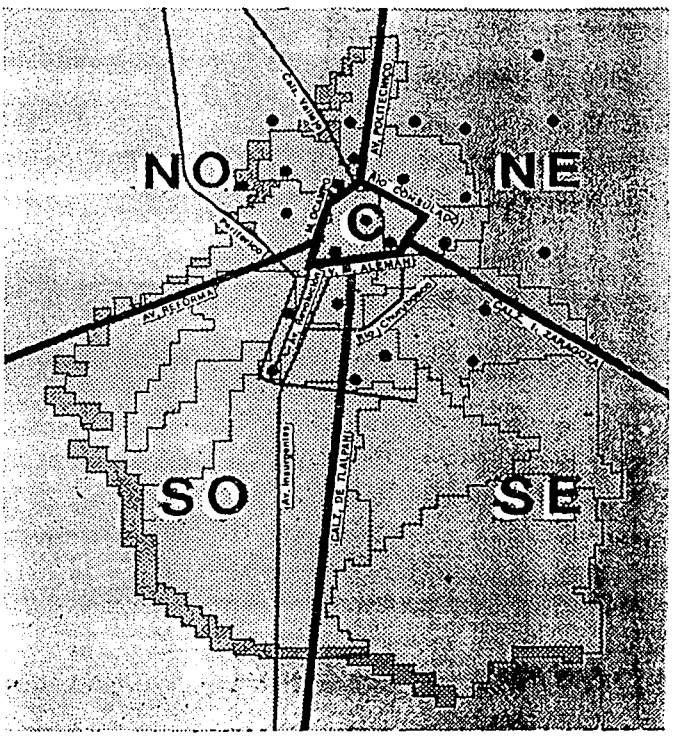
Cada tabla tiene tantos registros (elementos) como días en el mes. Cada día consta de 5 variables, una por cada zona.

De esta manera, los elementos de la tabla son de la forma

(NO, NE, CE, SO, SE)

Así el 1 de febrero de 1989, por ejemplo, forma el elememto (159, 104, 82, 186, 128).

IMECA



Fuente: Secretaría de Desarrollo Urbano y Ecología.

TABLA 4

Registros diarios de los niveles de contaminación alcanzados en cada una de las cinco zonas del área metropolitana durante febrero de 1989.

Indice Metropolitano de Calidad del Aire (IMECA)					
Día	Noroeste	Noreste	Centro	Suroeste	Sureste
1	159	104	82	186	128
2	166	78	102	126	99
3	138	95	107	174	119
4	108	98	96	162	124
5	152	79	90	144	74
6	85	82	82	187	86
7	87	82	75	133	79
8	87	78	87	188	86
9	126	100	123	187	112
10	40	39	42	59	73
11	116	89	58	136	72
12	119	78	72	118	70
13	114	73	76	115	68
14	86	82	101	126	55
15	81	105	98	194	90
16	93	81	92	214	96
17	76	50	77	123	119
18	69	51	89	181	88
19	130	65	81	142	99
20	128	67	65	124	65
21	175	106	93	180	76
22	65	53	71	178	73
23	62	59	52	113	104
24	69	56	110	120	114
25	55	53	85	118	103
26	99	39	87	110	85
27	111	91	106	109	94
28	124	76	103	137	78

TABLA 5

Registros diarios de los niveles de contaminación alcanzados en cada una de las cinco zonas del área metropolitana durante noviembre de 1989.

Indice Metropolitano de Calidad del Aire noviembre 1989					
Día	Noroeste	Noreste	Centro	Suroeste	Sureste
1	142	83	108	116	147
2	44	29	81	50	138
3	70	56	75	96	117
4	118	110	104	95	135
5	119	84	80	73	122
6	96	110	125	89	137
7	91	82	142	106	137
8	141	80	106	108	131
9	205	114	141	11	123
10	164	120	110	115	158
11	122	60	85	69	89
12	115	144	85	69	89
13	88	56	83	60	118
14	293	134	190	106	92
15	152	128	170	138	107
16	11	92	124	181	195
17	88	78	134	196	152
18	159	64	124	138	114
19	59	58	62	79	75
20	104	68	102	191	70
21	65	44	62	140	110
22	176	113	110	227	158
23	175	152	135	141	129
24	146	94	114	129	120
25	134	71	102	79	99
26	115	100	101	82	98
27	120	99	110	71	100
28	114	89	112	105	109
29	102	67	130	137	122
30	128	83	101	129	162

Fuente: Secretaría de Desarrollo Urbano y Ecología.

TABLA 6

Registros diarios de los niveles de contaminación alcanzados en cada una de las zonas del área metropolitana durante el mes de febrero de 1990.

Indice Metropolitano de Calidad del Aire febrero 1990					
Día	Noroeste	Noreste	Centro	Suroeste	Sureste
1	108	63	166	228	109
2	172	84	146	142	110
3	98	53	129	145	95
4	155	68	147	178	137
5	94	88	108	91	95
6	155	102	115	157	69
7	138	102	145	179	94
8	144	136	167	197	87
9	212	108	171	193	172
10	225	108	144	172	97
11	114	40	145	130	102
12	130	79	126	192	110
13	184	96	134	121	54
14	117	83	93	100	39
15	38	70	64	82	29
16	141	74	126	193	120
17	117	102	126	126	55
18	123	78	99	118	73
19	112	103	166	124	79
20	58	68	47	56	34
21	45	80	71	81	97
22	106	51	65	100	66
23	57	47	67	98	97
24	91	65	124	183	124
25	11	48	164	185	109
26	141	61	128	195	110
27	163	80	130	151	115
28	110	90	98	204	116

Fuente: Secretaría de Desarrollo Urbano y Ecología.

TABLA 7

POBLACION POR GRUPO DE EDAD Y SEXO EN 1960, 1970 Y 1980 EN EL ESTADO DE GUANAJUATO.

	1960		1970		1980	
EDAD	HOMBRES	MUJERES	HOMBRES	MUJERES	HOMBRES	MUJERES
GUANAJUATO	867,219	868,278	1,139,123	1,131,247	1,484,934	1,521,176
0 - 4	149,192	145,745	204,379	197,122	224,548	222,559
5 - 9	139,936	134,451	194,886	185,618	244,442	242,370
10 - 14	114,323	107,897	161,052	151,067	219,266	214,616
15 - 19	86,868	89,004	119,520	119,920	170,653	175,737
20 - 24	66,283	71,649	85,718	91,717	125,712	137,628
25 - 29	55,884	60,329	67,405	71,858	95,482	102,640
30 - 34	47,731	50,112	55,334	57,031	76,755	80,535
35 - 39	44,021	44,989	53,424	55,516	67,536	71,547
40 - 44	29,978	31,325	43,385	44,659	55,720	58,464
45 - 49	28,408	30,202	35,575	36,073	47,446	50,133
50 - 54	27,099	27,419	25,335	26,952	39,819	41,610
55 - 59	20,796	20,343	23,447	24,089	30,938	31,686
60 - 64	20,266	19,450	23,717	23,631	23,022	24,922
65 - 69	10,774	10,950	17,990	17,439	19,240	20,673
70 - 74	9,160	9,065	13,659	12,455	17,597	17,809
75 - 79	5,216	4,856	6,284	8,384	12,419	12,265
80 - 84	3,657	3,851	4,407	4,950	7,428	7,846
85 y más	3,766	3,517	3,606	4,226	4,342	5,464
NO ESPECIFICADA	3,854	3,114			2,569	2,672

FUENTE: DIRECCION GENERAL DE ESTADISTICA. VIII, IX, Y X CENSOS GENERALES DE POBLACION, 1960, 1970 Y 1980, ESTADO DE GUANAJUATO, MEXICO, D.F.

Esta es una tabla de contigencia con 3 variables categóricas: Grupo de edad, sexo y período.

En el censo de 1960 se registraron 149,192 varones entre 0 y 4 años.

TABLA 8

AGUSCALIENTES: INDICADORES DEL NIVEL DE VIDA
LA POBLACION POR MUNICIPIO 1980.

Municipio	INDICADORES													
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14
NACIONAL	25.34	19.69	15.14	13.90	27.89	62.79	27.20	26.45	28.44	42.84	21.78	29.95	6.49	15.11
ESTADO	40.01	14.54	9.57	11.71	32.64	70.39	28.94	12.41	11.55	26.21	14.17	17.50	6.52	14.06
AGUASCALIENTES	37.40	9.96	8.36	9.98	27.61	66.38	28.78	8.64	8.56	15.05	9.95	14.73	6.46	13.99
ASIENTOS	39.19	33.90	12.57	14.55	50.40	83.65	27.21	23.52	23.84	71.37	28.76	21.55	7.12	8.88
CALVILLO	52.92	27.72	12.32	20.02	48.71	87.28	29.73	19.88	17.23	42.44	23.52	25.37	6.10	14.39
COSIO	43.75	33.00	9.51	12.06	50.55	79.48	28.49	22.33	11.16	70.60	24.21	24.84	5.99	11.85
JESUS MARIA	44.44	19.05	14.75	14.58	43.19	73.18	29.62	26.89	19.56	53.52	21.94	26.94	7.26	19.85
PABELLON DE ARTEAGA	52.14	13.14	12.03	17.20	36.94	79.26	31.82	15.15	10.27	29.34	19.16	20.29	6.44	9.81
RINCON DE ROMOS	44.61	24.72	12.41	13.63	39.16	72.67	30.41	22.32	21.55	48.20	23.25	24.14	7.21	19.30
SAN JOSE DE GRACIA	53.33	25.42	10.01	15.29	50.02	87.28	29.56	17.67	32.27	82.78	35.46	27.25	5.43	12.46
TEPEZALA	41.58	38.53	12.06	15.30	48.34	85.64	25.67	20.21	17.73	72.23	27.50	23.63	6.00	9.11

INGRESOS

B1 TASA DE PEA QUE RECIBE INGRESOS MENORES A $3,611.00.
B2 TASA DE PEA QUE NO RECIBE INGRESOS

EDUCACION

B3 TASA DE ANALFABETISMIO DE LA POBLACION DE 10 AÑOS Y MAS
B4 TASA DE POBLACION DE 15 AÑOS Y MAS SIN INSTRUCCION.
B5 TASA DE POBLACION DE 15 AÑOS Y MAS CON PRIMARIA INCOMPLETA.
B6 TASA DE POBLACION DE 18 AÑOS Y MAS SIN ENSEÑANZA MEDIA
B7 TASA DE POBLACION DE 6 A 14 AÑOS QUE NO ASISTE A LA ESCUELA.

VIVIENDA.

B8 TASA DE VIVIENDAS CON PISO DE TIERRA.
B9 TASA DE VIVIENDAS SIN AGUA ENTUBADA.
B10 TASA DE VIVIENDAS SIN TUBERIA DE DRENAJES.
B11 TASA DE VIVIENDAS SIN ENERGIA ELECTRICA.
B12 TASA DE VIVIENDAS DE UN SOLO CUARTO.

SALUD

B13 TASA BRUTA DE MORTALIDAD (POR CADA 1000 HABITANTES).

EMPLEO

B14 TASA DE PEA QUE LABORA DESDE MENOS DE UNA HORA HASTA 32 HORAS A LA SEMANA.

FUENTE: INSTITUTO NACIONAL DE ESTADISTICA, GEOGRAFIA E INFORMATICA.
X CENSO GENERAL DE POBLACION Y VIVIENDA, 1980.

Este problema contempla 14 variables (los inidicadores del nivel de vida) para 9 elementos (los municipios del edo. de Aguascalientes).

Por su naturaleza de indicadores, las variables son contínuas en un rango de valores de cero a cien.

TABLA 9

Compañía	VARIABLE													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ARCO	0.56	1.10	0.78	306.00	49.00	10.00	4.50	0.38	66.00	62.00	0.11	174.00	0.84	2.80
UNION	0.53	1.20	0.49	203.00	47.00	4.00	4.20	1.22	103.00	99.00	0.19	527.00	0.98	8.50
GETTY	0.54	1.00	0.32	197.00	31.00	11.00	4.00	0.67	51.00	57.00	0.11	160.00	0.38	2.80
MOBIL	1.21	2.80	0.50	211.00	50.00	8.00	3.90	1.04	68.00	78.00	0.06	339.00	0.81	4.00
TEXACO	1.16	2.70	0.56	176.00	66.00	8.00	7.80	0.31	56.00	50.00	0.04	277.00	0.91	2.50
CHEVRON	0.84	1.20	1.16	378.00	70.00	13.00	5.80	0.70	197.00	141.00	0.17	355.00	0.50	1.60
GULF	1.01	2.20	0.67	219.00	65.00	11.00	4.10	1.53	338.00	235.00	0.23	481.00	0.83	2.90
AMOCO	0.66	1.30	0.66	258.00	53.00	8.00	7.30	0.45	37.00	44.00	0.07	213.00	0.31	2.70
SHELL	0.97	1.70	1.59	336.00	95.00	13.00	3.60	1.90	430.00	378.00	0.39	656.00	0.38	2.70
EXXON	1.44	2.90	1.02	250.00	84.00	8.00	5.20	0.99	276.00	199.00	0.14	609.00	0.58	4.30

Las 10 compañías, son los elememtos a describir:

A cada compañía se le atribuyen las sigientes 14 variables:

1.- Prima Neta
2.- Exceso $ sobre renta
3.- Superficie Neta
4.- Rentas ganadas
5.- Promedio de Propiedades
6.- Porcentaje de Producción
7.- Promedio de años
8.- Promedioneto de gas
9.- Producción neta de líquido
10.- Pago real
11.- Tiempo real
12.- Coeficiente de correlación
13.- Producción anual real
14.- % de rentas pagadas

TABLA 10

	Cuernavaca	Chilpancingo	Guadalajara	Guanajuato	Jalapa	D.F	Pachuca	Puebla	Querétaro	Tlaxcala
Cuernavaca										
Chilpancingo	192									
Guadalajara	663	835								
Guanajuato	448	620	303							
Jalapa	368	554	887	671						
D.F	85	287	579	363	308					
Pachuca	177	379	587	371	365	92				
Puebla	179	365	704	489	189	125	221			
Querétaro	302	474	361	146	526	218	226	343		
Tlaxcala	198	396	692	477	188	113	190	31	331	
Toluca	150	282	553	338	374	66	157	191	192	179

Distancia en Km. entre diferentes capitales de Estado en la República Mexicana.

Por ejemplo, entre Chilpancingo Y Cuernavaca se recorren 192 Km

Los datos de las tablas 11 y12 son ficticios y se emplearon para desarrollar los ejemplos de Análisis Discriminante.

TABLA 11

Datos recopilados por una inmobiliaria sobre los clientes que han merecido crédito para la compra de un departamento.

Grupo	Antigüedad en el trabajo	Ingresos semanales (x $100 0000)	Saldo a favor cta. cheques (x $1,000 000)	
A	6	7	4	
A	7	5	1	
A	9	10	11	
A	8	8	6	
A	8	9	6	
A	10	9	8	
B	11	13	30	
B	15	16	42	
B	22	20	50	
B	17	16	45	
B	12	11	38	
B	13	14	35	
C	18	16	25	
C	24	22	30	
C	20	21	32	Se tiene 18 elementos
C	19	20	27	(individuos); una variable
C	22	25	34	categórica (el grupo) y tres
C	17	16	32	variables métricas.

TABLA 12

Calificaciones obtenidas por 21 alumnos en un examen de matemáticas.

Grupo	Habilidad	Raciocinio	Ambos	Gpo.	Habilidad	Raciocinio	Ambos
A	5	3	1	B	10	5	5
A	4	7	3	B	9	5	4
A	5	5	3	B	8	4	4
A	3	2	2	C	8	8	5
A	5	4	3	C	9	7	6
A	2	2	0	C	10	8	8
A	2	3	0	C	9	9	7
B	8	5	3	C	8	9	6
B	7	4	3	C	9	8	7
B	9	5	4	C	8	8	8
B	8	6	3				

En este problema particular participan 21 elementos evaluados por 3 variables métricas en una escala discreta de cero a diez. Las variables categóricas es el grupo al que pertenece cada alumno.

TABLA 13

Número de votos emitidos para cada característica en la evaluación sensorial de panes elaborados con mezclas de trigo - girasol.

		CARACTERISTICAS														
1	2	COLOR			SABOR				AROMA			TEXTURA				
MEZCLA	VARIDAD	A	B	C	A	B	C	D	A	B	C	A	B	C	D	TOTAL
5	1	0	3	1	0	4	0	0	0	1	3	0	3	1	0	17
10	1	1	2	1	1	2	1	0	1	3	0	0	3	1	0	16
15	1	0	2	1	1	1	0	1	1	2	0	0	2	0	1	13
5	2	0	2	1	0	1	2	0	0	2	1	0	2	1	0	14
10	2	0	1	2	0	2	1	0	0	0	3	0	2	0	1	14
15	2	0	1	1	0	1	1	0	1	1	0	1	1	0	0	10
5	3	0	1	2	0	2	1	0	0	1	2	0	1	2	0	15
10	3	0	1	2	0	1	1	1	0	1	2	0	0	3	0	15
15	3	0	2	1	0	3	0	0	1	2	0	0	2	1	0	15
0	4	1	1	0	0	2	0	0	0	2	0	2	0	0	0	12
TOTAL		2	16	12	2	19	7	2		15	11	3	16	9	2	141

1 Mezcla de (% de girasol)
2 variedad (1=Criollo Atotonilco; 2=Criollo Río Verde; 3=CIANOC; 4=Trigo comercial)

TABLA 14

CANASTA BASICA MARZO 1990

	AURRERA	BODEGA	G. BAZAR	SUPERAMA	COMERCIAL	SUMESA	GIGANTE	BLANCO	LUNA	DE TODO	ISSSTE
HARINA	785	745	785	785	785	785	749	785	740	739	740
SOPA	446	432	446	446	446	446	446	432	440	445	330
ACEITE	2230	2230	2230	2230	2230	2230	2230	2230	2230	2230	2230
GIRASOL	2230	2225	2230	2230	2230	2230	2230	2211	2230	2211	2090
MANTECA	2230	2230	2230	2230	2230	2230	2230	2230	2230	2230	2230
ARROZ	2050	1890	3260	1995	2299	2299	2299	2299	2299	2299	2299
FRIJOL	3186	3240	3186	3165	3270	3186	3070	3186	3186	3186	3186
LEGAL	3425	3360	3425	3450	3562	3356	3562	3178	3240	3399	2960
DECAF	7372	7230	7372	7367	7372	7372	7372	7023	6800	6990	6500
NESCAFE	5815	5755	5815	5625	5815	5625	5849	5487	5200	5810	5080
GAMESA	2150	2150	2150	2155	2155	2155	2030	2155	2150	2150	2140
PURE	2095	2100	2095	2100	2100	2100	2125	2100	2090	2095	2100
NIDO	4340	4200	4200	4250	4200	4200	4200	4200	4340	4290	3780
SVELTES	4527	4460	4527	4525	4527	4459	4520	4668	4500	4525	3810
PELARGON	4990	4810	4990	4810	4990	4990	4990	4990	4600	4940	3810
CLAVEL	1230	1140	1230	1230	1230	1087	1230	1230	1230	1285	1120
CARNE	10900	11425	11600	11600	11600	11600	11600	11425	11425	11600	10900

Para cada uno de los 17 productos de la canasta básica, se tienen como variables, los precios en que se ofrecían en 11 almacenes de autoservicio.

Las unidades de medida son pesos mexicanos.

Por ejemplo, en marzo de 1990, laharina de trigo costaba $785 en los almacenes Aurrerá

Fuente: Instituto Nacional del Consumidor.

TABLA 15

CANASTA BASICA ABRIL 1991

	AURRERA	BODEGA	G. BAZAR	SUPERAMA	COMERCIAL	SUMESA	GIGANTE	BLANCO	LUNA	DE TODO	ISSSTE
HARINA	1390	1320	1390	1305	1390	1390	1390	1305	1300	1300	1230
SOPA	530	550	530	590	590	590	590	490	580	585	440
ACEITE	3017*	3040	3017*	3050	3050	3017*	3050	3017*	3050	2990	2890
GIRASOL	3050	3040	3041*	3050	3050	3050	3050	3041*	3050	2990	3041*
MANTECA	3050	3040	3050	3050	3050	3050	3050	3030*	3050	2990	2920
ARROZ	2100	1940	2100	2115	1850	2200	2175	1780	2065	1930	1840
FRIJOL	2895	2800*	2895	2990	2750	2950	2900	2300	2800*	2970	2550
LEGAL	4370	4230	4370	4250	4390	4290*	4395	4395	4400	4360	3740
DECAF	8790	7850	8790	8780	8790	8790	8790	8690	8790	8750	7260
NESCAFE	6950	6190	6950	6940	6950	6950	6950	6850	6950	6890	5870
GALLETAS	2790	2780	2790	2790	2790	2790	2790	2788*	2790	2785	2788*
PURE	2095	2120	2095	2220	2210	2136*	2125	2136*	2210	2210	1940
NIDO	5200	5090	5200	5200	5200	5117*	5200	5200	5200	5150	4530
SVELTES	5570	5457*	5570	5570	5570	5270	5570	5570	5570	5457	4550
PELARGON	6440	6439*	6440	6439*	6440	6440	6440	6440	6440	6430	6439*
CLAVEL	1580	1460	1580	1580	1490	1450	1580	1580	1580	1390	1420
CARNE	14990	14803*	13500	14990	14990	14990	14990	14803*	14990	14990	14803*

Fuente: Instituto Nacional del Consumidor.

APENDICE.-B Matrices

Definición.-Una matriz *mxn* sobre un campo *f* es un arreglo rectangular de escalares consistente en *i* renglones ($1 \leq i \leq m$) y *j* columnas ($1 \leq j \leq m$) brevemente denotado por

$$A = (\alpha_{ij}).$$

Si m = n, entonces A recibe el nombre de matriz cuadrada de orden n(ó m).

Definición.-Dos matrices A(mxn) y B(pxq) son iguales si:

a) $m = p, n = q$

b) $A = (\alpha_{ij}) = (\beta_{ij}) = B \Leftrightarrow \alpha_{ij} = \beta_{ij} \quad \begin{cases} 1 \leq i \leq m \\ 1 \leq j \leq n \end{cases}$

Si $\alpha_{ij} = 0 \; \forall_{i,j}$ se dice que se tiene una matriz nula

$\hat{O} = (oij)$

Definición.-La suma de dos matrices $A = \alpha_{ij}$ $B = \beta_{ij}$ con el mismo número de renglones y columnas se define como otra matriz $S = (\delta_{ij})$

$$S = A + B = (\alpha_{ij} + \beta_{ij}) = \delta_{ij}$$

Definición.-El producto de una matriz por un escalar λ se define por

$$\lambda A = \lambda(\alpha_{ij}) = \lambda\alpha_{ij}$$

La suma de matrices y el producto de una matriz por un escalar λ satisfacen las siguientes propiedades:

a) Cerradura

b) $A + (B + C) = (A + B) + C$

c) $A + B = B + A$

d) $(\lambda + \mu)A = \lambda A + \mu A$

e) $\rho(A + B) = \rho A + \rho B$

f) $\hat{O} + A = A$

g) $1.A = A$; $(-1)A = -A = -\alpha_{ij}$

h) $O \cdot A = \hat{O}$

i) $\rho(\mu A) = \rho\mu A = (\rho\mu)A$

El conjunto de matrices definidas sobre un campo *f* que satisfacen las propiedades anteriores, es considerado un espacio vectorial de dimensión (m • n)

Definición.- Sean A(nxm) y B(mxp) dos matrices tales que el número de columnas de A coincide con el número de renglones de B, entonces, el producto AB = P será una matriz con el mismo número de renglones que A y con igual número de columnas que B, es decir, P(nxp). Además los elementos de P, están dados por la relación

Consecuencias del producto de matrices:

1) $\gamma(AB) = (\gamma A)B$

2) $(A + B)C = AC + BC$; $C(A + B) = CA + CB$

3) $A(BC) = (AB)C$

Definición.- A una matriz de la forma $\begin{pmatrix} \alpha & 0 & \dots & 0 \\ 0 & \beta & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \gamma \end{pmatrix}$ se le denomina matriz diagonal.

En el caso de que $\alpha = \beta = \gamma = 1$ se dice que es una matriz idéntica o unitaria I y tiene la propiedad de que: $AI = A = IA$

$$I = \delta_{ij} \qquad con \quad \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

Si se cumple AB = BA, se dice que AyB son matrices conmutables entre si.En este caso, dado un numero natural p se tendrá que

$$(AB)^P = A^P B^P$$

Definición.-Dada la matriz A(mxn), a la matriz que se obtiene de A al cambiar los renglones por las columnas A'(nxm) se le llama la matriz transpuesta de A:

$$A^T = A' \; n \times m$$

Propiedades:

$$(\alpha A + \beta B)^T = \alpha A^T + \beta B^T$$
$$(AB)^T = B^T A^T$$

Si A es una matriz cuadrada tal que $A^T = A$, entonces se dice que A es una matriz simétrica.Si $A^T = -A$, entonces A es antisimétrica.(Antihermiteana en C) (complejos).

A es simétrica si $a_{ij} = a_{ji} \qquad \forall_{i,j}$

Si las matrices simétricas A,B son conmutables entonces:

$$(AB)^T = B^T A^T = BA = AB$$

es decir, su producto también será simétrico.

Definición.-A una matriz cuadrada A. se le llama invertible si existe una matriz X tal que:

$$XA = I = AX \Rightarrow X = A^{-1}$$

X recibe el nombre de inversa de A y es única.

Propiedades:

$$(A^{-1}) = A \qquad A^{-2} = A^{-1}A^{-1}; \dots; A^{-n} = A^{-1}A^{-1} \dots A^{-1}$$
$$(AB)^{-1} = B^{-1}A^{-1} \qquad (AA^{-1})^T = E = (A^{-1}A)^T \qquad (A^{-1})^T = (A^T)^{-1}$$

Definición.-Dada una matriz cuadrada A, si se cumple que $A^T = A^{-1}$,es decir si su matriz transpuesta es igual a su inversa, entonces, A es una matriz ortogonal.

Corolario.- La inversa de una matriz ortogonal es a su vez ortogonal.

Consecuencias: El producto de 2 matrices ortogonales, es a su vez una matriz ortogonal, es decir:

Sean A,B dos matrices invertibles tales que: $A^T = A^{-1}$;

$$B^T = B^{-1} \Rightarrow (AB)^T = B^T A^T = B^{-1}A^{-1} = (AB)^{-1}$$

Definición.-El rango de una matriz A es igual al número máximo de vectores linealmente independientes en A, es decir se trata de la dimensión del subespacio $V^m \leq V^{mn}$ generado por m vectores.

Definición.-Por una operación elemental se entiende una operación que transforma a una base de un subespacio vectorial en otra base del mismo subespacio.Estas operaciones son:
-Intercambio de dos vectores.
-Producto de un vector x un escalar nulo (deformación).
-Suma de un vector con la deformación de otro vector.

Definición.-Una matriz elemental es cualquier matriz cuadrada que se obtiene de aplicar al menos una operación elemental a los vectores de una matriz idéntica.

Teorema.-Una matriz a la que se apliquen una o mas operaciones elementales no altera su rango.

Definición.-Se dice que una matriz A(*mxn*) es no singular (regular) si rango (A) = máx(m,n) en caso contrario será singular.

Definición.-Una matriz cuadrada cuyo determinante es igual a cero, se llama singular. En caso contrario es no singular o regular.

Definición.-Una matriz A(mxn) se llama reducida por filas si:
a) El primer elemento no nulo de cada fila no nula de A es igual a 1.
b) Cada columna de R que tiene el primer elemento no nulo de alguna fila tiene el primer elemento no nulo de alguna fila tiene todos sus otros elementos igual a cero.
EJEMPLO: La matriz Identidad.

Definición.-Una matriz A(mxn) se llama matriz escalonada reducida por filas si:
a) A es reducida por filas.
b) Toda fila nula de A esta debajo de todas las filas que tienen elementos no nulos.
c)Si las filas 1,...,r son las filas no nulas de R y si el primer elemento no nulo de la fila i esta en la columna K_i, i = 1,...r entonces $K_1 < K_2 < ...Kr$.

Corolario.-Una matriz A es escalonada si A^t se expresa en forma escalonada.

Definición.-Una matriz $A = (\alpha_{ij})$ tiene forma triangular si sucede que:

1) $\alpha_{ij} = 0 \quad \forall\, i > j \qquad i = 1,2,...,m;\ j = 1,...n$
2) $\alpha_{ij} = 0 \quad \forall\, i < j \qquad i \le i \le m;\ \ 1 \le j \le m$

A tendrá forma diagonal si se tiene que $\alpha_{ij}=0$ para $i \neq j$

Definición.- Dos matrices A, B cuadradas de orden n, sobre el mismo campo *f* son similares si existe una matriz no singular X tal que $A = XBX^{-1}$

Teorema.- $A = XBX^{-1}$ y $B = YCY^{-1} = A = (XY)C(XY)^{-1}$.

APENDICE.C. Distancias y Similaridades.

Definición.- Una función real no negativa S(x,y), de pares de puntos de un conjunto E; se dice que es una medida de similaridad para E si se cumple:

1) $0 \le S(x,y) \le 1$
2) S(x,y)=1
3) S(x1y)=S(y,x)

Una matriz S(nxm) es llamada matriz de similaridades si es simétrica Sij=Sij.

$$S = \begin{pmatrix} 1 & a & a \\ a & 1 & c \\ b & c & 1 \end{pmatrix} \qquad y \qquad S_{ij} \le S_{ii}$$

$$S_{ij} = 1 \qquad \forall\, i = 1,2,\ldots\ n$$

Sij= Es el fndice de similaridad entre las características del individuo i con las del individuo j.

El concepto de similaridad tiene que ver con la idea intuitiva de "se parece a".

Como el coeficiente de correlación toma valores negativos, no es considerado una medida de similaridad.

En variables cualitativas la similaridad se relaciona con la presencia o ausencia de p atributos en un objeto.
Sean P y Q dos objetos donde:

P=(X1,X2,...Xp) y Q=(y1,y2,...yp)
Xi=0 ó 1 , Yi=0 ó 1.

dependiendo de la ausencia (0) o presencia (1) del atributo i-ésimo.

La medida de similaridad mas simple entre PyQ es: $s_1(P,Q) = \frac{a}{p}$

donde $a = \sum x_i y_i$

otra alternativa es: $S_2(P,Q) = \frac{a+d}{p}$

donde $d = \sum (1-x_i)(1-y_i)$

En S_1 y S_2 todos los atributos tienen igual peso pero, en algunas aplicaciones es preferible diferenciar los pesos de los atributos.

Sea X(nxp) la matriz de datos Xi=0 ó 1
n individuos y p atributos.

La matriz de similaridades correspondiente a X sería:

$$S_1 = \frac{xx'}{p}$$

$$S_2 = \{xx' + (J-x)(J-x)'\} \qquad donde \quad J = 11'$$

3.- Distancia Mahalanobis.-

$$d_{ij} = (x_i x_j)^{-1} W^{-1} (x_i - x_j)$$

donde W es la matriz de varianza-covarianza
y x'_i y x'_j son los vectores (1xp) de registros individuales i y j . Es invariante bajo transformaciones lineales no singulares y no es afectada por problemas de escalamiento.

4.- Distancia City Bllock.-
Si 2 individuos son especificados por 2 variables cuyas escalas de valores son las mismas, deben tener la misma distancia, además son 2 unidades aparte sobre cada variable o 1 unidad aparte sobre 1 variable y 3 sobre otra.

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

5.- Distancias Minkowski $\quad d_{ij} = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right]^{\frac{1}{r}}$

6.- Distancia Sup-norm $\quad d_{ij}^{(\infty)} = \sup_{k=1,\dots p} \{ |x_{ik} - x_{jk}| \}$

DISTANCIAS PARA DATOS CUALITATIVOS.-

$$A = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{g1} & x_{g2} & \cdots & x_{gp} \end{pmatrix}$$

$$\sum_{i=1}^{p} x_{ri} = 1 \qquad r = 1, \dots, g$$

x_r denota la proporción de una población de tamaño n_r situada en cada una de los p categorías.

1.- Distancia euclidiana.- La distancia euclidiana puede ser apoyada si las proporciones son meramente cantidades mediadas en las que no se ha considerado algún modelo de variación estocásticas. (¿alguna función de distribución?)

$$d_{rs}^2 = \sum_{i=1}^{p} (x_{ri} - x_{si})^2$$

$$\therefore \quad d_{rs} = \left[\sum_{i=1}^{p} \left((x_{ri} - x_{si})^2 \right) \right]^{\frac{1}{2}}$$

2.- Distancias Mahalanobis.- Suponiendo que x_r para r=1... g representa las proporciones basadas en una muestra de tamaño nr desde una distribución multinominal con parámetros
Entonces x_r tiene media $a = (a_1,...,a_p)'$ y matriz de covarianza $a(\bar{x}_r = a)$

$$\sum_r = n_r^{-1}\sum \qquad donde \qquad \sum = \theta_{ij}$$

$$\theta_{ij} = \begin{cases} a_i(1-a_i) & i = j \\ -a_i a_j & i \neq j \end{cases}$$

esto es:

$$\sum = diag(a) - aa'$$

Como x_r permanece sobre un hiperplano, $\sum$ es singular

$$g - \sum^{-1} = \sum^{-} = diag\left(a_1^{-1},...,a_p^{-1}\right)$$

La distancia generalizada Mahalanobis entre xr y xs se define como $\frac{n_r n_s}{(n_r + n_s)}$ veces

$$\sum_{i=1}^{p} \frac{(x_{ri} - x_{si})^2}{a_i}$$

Para reducir errores en problemas con variables en diferentes unidades, se procede mediante la estandarización

$$\sum_{i=1}^{p} \frac{(x_{ri} - x_{si})^2}{a_i}$$

3.- Distancia Bhattacharyya.- $Vr = \left(x_{r1}^{\frac{1}{2}},...x_{rp}^{\frac{1}{2}}\right) \qquad r = 1,... g$

los vectores Vr son puntos en una esfera unitaria en con centro en el origen.

El coseno del ángulo entre Vr y Vs $\qquad \cos B_{rs} = \sum_{i=1}^{p} v_{ri} v_{si} = \sum_{i=1}^{p} (x_{ri} x_{si})^{\frac{1}{2}}$

Así, el ángulo Brs es la distancia entre Vr y Vs.

La distancia euclidiana es:

$$D_{2:rs} = \left[\sum_{i=1}^{p}\left(x_{ri}^{\frac{1}{2}} - x_{si}^{\frac{1}{2}}\right)^2\right]^{\frac{1}{2}}$$

Los medidas son conectadas por la ecuación

$$D_{2:rs}^2 = 4sen^2\left(\tfrac{1}{2}B_{rs}\right)$$

Si el problema involucra variables mixtas (cualitativas y cuantitativas).Gower (1971) propuso el siguiente coeficiente. de similaridad entre los puntos i,j.

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

donde

$W_k = 1$ si k es cualitativa

y $W_k = \frac{1}{R_x}$ si k es cuantitativa

R_k es el rango de la k-ésima variable.

Sij es semidefinida. positiva pero si en lugar de R se usa la desviación estandard de la muestra, S puede no serlo.

DISTANCIAS ENTRE DOS PUNTOS.

Definición.- Una función real no negativa d(R,Q) de pares de puntos de un conjunto E, se dice que es una función de distancia para E si:

1) d(P,Q) >=o
2) d(P,Q) = d(P,Q).
3)d (P,Q) =d(Q,R)
4) d(P,P) =O
5) d(P,Q) =O <=> P=Q
6) d(P,Q) =d (P,R) + d(R,Q) (Desig.del *).

La matriz D(NXM) es llamada matriz de distancias si es simétrica Dij. = Dji: y Dii=O.

$$D\begin{pmatrix} 0 & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{12} & 0 & \cdots & \cdots & x_{2n} \\ \vdots & & & & \vdots \\ x_{1n} & & & & 0 \end{pmatrix}$$

1.- Distancia euclidiana.- Sea X(nxp) una matriz de datos con reglones $(x_1, x_2 \cdots x_n)$ Entonces, la d euclidiana entre los puntos x_i y x_j es d_{ij}, donde

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = \|x_i - x_j\|^2$$

2.- Distancia Karl Pearson (Disttancia estandarizada).- Cuando las varianzas no son conmensurables, lo mejor es estandarizar.

$$d_{ij}^2 = \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{S_k^2}$$

donde S_k^2 es la varianza de la K-ésima variable

Otra manera de estandarizar es reemplazar S_k por el rango $R_k = \max_{i,j} |x_{ik} - x_{jk}|$

Esta distancia es invariante ante el cambio de escala

DISTANCIA ULTRAMETRICA.-

En el análisis de cúmulos, el concepto de jerarquía está ligado a una clase de distancias entre objetos llamados Distancias utramétricas.

Sea E un conjunto con distancias positivas que satisface:

$$d(P,Q) = 0 \quad \Leftrightarrow \quad P = Q$$
$$d(P,Q) = d(P,Q)$$
$$d(P,Q) \leq d(P,Q) + d(R,Q) \quad \Delta$$

la distancia es utramétrica si satisface

$$d(P,Q) \leq \max\{d(p,R), d(P,Q)\}$$

MEDIDAS DE DISTANCIA Y SIMILARIDAD ENTRE GRUPOS.

Cuando se busca la distancia o similaridad entre grupos se puede recurrir a tomar el promedio de similaridad o distancias entre pares de individuos (uno desde cada grupo)

Otra alternativa es medir la similaridad o distancia entre cúmulos usando una medida entre individuos a partir de los grupos.

Otras medidas son:

1) Distancias Mahalanobis $d_{G_1G_2} = (\bar{x}_{G_1} - \bar{x}_{G_2})^1 W^{-1} (\bar{x}_{G_1} - \bar{x}_{G_2})$

donde W es la matriz de varianza-covarianza entre grupos y son los vectores (1xp) media de G1 y G2 respectivamente.

2) Jardín y Sibson (1971)

$$d_{G1\,G2} = \tfrac{1}{2} log_2 \left\{ \frac{|\tfrac{1}{2} W_{G1} + W_{G2}| + \tfrac{1}{4} (\bar{X}_{G1} - \bar{X}_2)' (\bar{X}_{G1} - X_{G2})}{|W_{G1}|^{1/2} \; |W_{G2}|^{1/2}} \right\}$$

donde $W_{G_{1,2}}$ son matrices de varianza-covarianza de G_1 y G_2

$|W_{G_1}|$ es el determinate de la matriz W_{G_1} . Cuando $W_{G_1} = W_{G_2}$ se reduce a Mahalanobis.

3) Distancia Generalizada para variables discretas

$$d_{G1G2} = (P_{G1} - P_{G2})^1 S^{-1} (P_{G1} - P_{G2})$$

donde: S es la matriz de varianza-covarianza común y P_{Gi} es el vector de elementos en los que P_{Gijk} da por grupo Gi la proporción e individuos faltando en cada categoría K de la variable j.