UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE CIENCIAS

METODOS MULTIVARIADOS APLICADOS A LOS DATOS DEL CENSO

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

ADRIAN AVALOS PADILLA

MEXICO, D. F. MARZO 1995

UNIVERSIDAD NACIONAL
AVENIDA DE
MEXICO

M. EN C. VIRGINIA ABRIN BATULE
Jefe de la División de Estudios Profesionales
Facultad de Ciencias
P r e s e n t e

Los abajo firmantes, comunicamos a Usted, que habiendo revisado el trabajo de Tesis que realiz(ó)ron EL pasante(s) ADRIAN AVALOS PADILLA

con número de cuenta 8723957-7 con el Título: "METODOS MULTIVARIADOS APLICADOS A LOS DATOS DEL CENSO"

Otorgamos nuestro Voto Aprobatorio y consideramos que a la brevedad deberá presentar su Examen Profesional para obtener el título de ACTUARIO

GRADO	NOMBRE(S)	APELLIDOS COMPLETOS	FIRMA
M en C.	LETICIA GRACIA	MEDRANO VALDEMAR	
Director de Tesis M en C.	GUILLERMO	GOMEZ ALCARAZ	
M en C.	GUSTAVO	MARQUEZ FLORES	
M en C.	JAVIER	GARCIA GARCIA	
Suplente ACT.	AURORA	VALDEZ MICHEL	

Suplente

DEDICATORIA

A Dios.

Por permitirme llegar a esta meta, estar conmigo en todo momento y darme lo mejor de la vida.

A mi Madre y a mi Padre.

Por su dedicación y entrega. Y por que sin ellos este logro nunca hubiera sido alcanzado.

A Roxanita y a la Osa.

Por ser dos personas muy importantes, a las que dedico mis mejores esfuerzos.

A la Universidad Nacional Autónoma de México.

Por ser la Casa de Estudios por excelencia y brindarme las mejores oportunidades para mi desarrollo personal.

Agradezco a la M. en C. Leticia Gracia Medrano por haber aceptado dirigir este trabajo y por dedicar su tiempo para revisarlo y corregirlo; así como su valiosa asesoría y orientación.

ÍNDICE

INTRODUCCIÓN 1

CAPÍTULO 1
MÉTODOS GRÁFICOS 8

1.1 Un primer enfoque de los datos 9
1.2 Descripción por polígonos 11
1.3 Descripción por barras 13
1.4 Graficación por líneas 13
1.5 Otras técnicas 14
1.5.1 Boxplot 16
1.6 Aplicación a los datos del censo 20
1.6.1 Gráficas de polígonos 20
1.6.2 Gráficas de barras 26
1.6.3 Gráficas de líneas 29
1.6.4 Caritas de Chernoff 34
1.6.5 Draftsman plot 38
1.6.6 Steam and leaf 40
1.6.7 Box plot 42
1.7 Conclusiones 44

CAPÍTULO 2
COMPONENTES PRINCIPALES 46

2.1 Objetivos 47
2.2 Desarrollo 48
2.3 Propiedades 51
2.4 Aplicación a los datos del censo 53
2.5 Conclusiones 57

INDICE

INTRODUCCIÓN 1

CAPÍTULO 1
MÉTODOS GRÁFICOS 8

1.1 Un primer enfoque de los datos 9
1.2 Descripción por polígonos 11
1.3 Descripción por barras 13
1.4 Graficación por líneas 13
1.5 Otras técnicas 14
1.5.1 Boxplot 16
1.6 Aplicación a los datos del censo 20
1.6.1 Gráficas de polígonos 20
1.6.2 Gráficas de barras 26
1.6.3 Gráficas de líneas 29
1.6.4 Caritas de Chernoff 34
1.6.5 Draftsman plot 38
1.6.6 Steam and leaf 40
1.6.7 Box plot 42
1.7 Conclusiones 44

CAPÍTULO 2
COMPONENTES PRINCIPALES 46

2.1 Objetivos 47
2.2 Desarrollo 48
2.3 Propiedades 51
2.4 Aplicación a los datos del censo 53
2.5 Conclusiones 57

CAPÍTULO 3
ESCALAMIENTO MULTIDIMENSIONAL 62

3.1 Introducción 63
3.2 Descripción 65
3.2.1 Escalamiento clásico 66
3.2.2 Escalamiento no métrico 67
3.3 Aplicación a los datos del censo 69

CAPÍTULO 4
ANÁLISIS DE CONGLOMERADOS 89

4.1 Objetivo 90
4.2 Métodos jerárquicos 92
4.3 Métodos no jerárquicos 93
4.4 Aplicación a los datos del censo 96

CAPÍTULO 5
ANÁLISIS DE CORRESPONDENCIA 99

5.1 Introducción 100
5.2 Desarrollo 102
5.3 El problema dual 104
5.4 Inercia 106
5.5 Aplicación a los datos del censo 108

CONCLUSIONES 125

APÉNDICE 128

BIBLIOGRAFÍA 136

CAPÍTULO 3
ESCALAMIENTO MULTIDIMENSIONAL 62

3.1 Introducción 63
3.2 Descripción 65
3.2.1 Escalamiento clásico 66
3.2.2 Escalamiento no métrico 67
3.3 Aplicación a los datos del censo 69

CAPÍTULO 4
ANÁLISIS DE CONGLOMERADOS 89

4.1 Objetivo 90
4.2 Métodos jerárquicos 92
4.3 Métodos no jerárquicos 93
4.4 Aplicación a los datos del censo 96

CAPÍTULO 5
ANÁLISIS DE CORRESPONDENCIA 99

5.1 Introducción 100
5.2 Desarrollo 102
5.3 El problema dual 104
5.4 Inercia 106
5.5 Aplicación a los datos del censo 108

CONCLUSIONES 125

APÉNDICE 128

BIBLIOGRAFÍA 136

INTRODUCCIÓN.

El análisis multivariado es, en la actualidad, una de las herramientas más utilizadas para obtener conclusiones a partir de un conjunto de datos multidimensional y de gran magnitud. El auge de ésta rama de la Estadística se ha provocado, en gran medida, por el desarrollo de herramientas computacionales. Logrando con ello una rápida generación de resultados los cuales serán analizados y servirán para hacer inferencias u obtener conclusiones acerca del conjunto de datos tratados.

El objetivo de este trabajo es proporcionar una breve explicación y ejemplificación de algunas de las técnicas más utilizadas del análisis multivariado, y para su análisis se eligió un conjunto de datos muy especial: Los datos del Censo de México de 1990

De esta manera, se consideró que al utilizar datos reales se puede conseguir una mejor comprensión del modo de operación y utilidad de las técnicas que se describen. Obviamente este trabajo no pretende ser un libro que trate detalladamente los métodos del análisis multivariado ya que cada una de las técnicas maneja conceptos y resultados que por sí mismos podrían ocupar un libro completo. Lo que se da es una breve explicación acerca del objetivo de cada método, que engloba los siguientes aspectos:

- ¿ Qué es cada una de las técnicas?
- ¿ Para qué sirve?

- ¿ Cuándo se utiliza?
- ¿ Qué tipo de datos maneja?
- ¿ Qué conclusiones se pueden obtener a partir de cada una de ellas?

Además, se explica brevemente cómo funcionan los algoritmos que se utilizan, lo cual se hace a un nivel general. Sin embargo, se hace referencia a algunos artículos y bibliografía que puede servir a quienes estén interesados en obtener información más detallada de cada método.

Por otra parte, al manejar datos reales de nuestro país, se obtienen conclusiones que pueden decirnos mucho acerca de nuestro entorno. Esto hace que el trabajo sea más interesante ya que es más ilustrativo analizar datos que pueden proporcionarnos información relevante y que, en cierta medida, nos involucra en lugar de analizar datos generados artificialmente.

Así, el formato de cada capítulo es el siguiente:

a) Una breve descripción de la técnica y del algoritmo utilizado para la ejemplificación con los datos censales.

b) Aplicación a los datos del Censo con ayuda de paquetes estadísticos tales como SYSTAT5, CSS, STAT-GRAPHICS y SPLUS. Los cuales generan la salida de cada uno de los algoritmos y realizan todos los cálculos que se involucren.

c) A partir de lo anterior, se analizan los resultados obtenidos y en base a ellos se obtienen conclusiones y/o inferencias acerca del comportamiento de los datos.

DATOS UTILIZADOS.

Para la ejemplificación de cada una de las técnicas descritas en los capítulos siguientes, se utilizaron los datos del Censo de 1990 para la República Mexicana . Dichos datos involucran aspectos de población, vivienda, económicos, culturales y sociales, entre otros.

Todos los datos recopilados durante el censo se publicaron en una serie de libros y en un CD (disco compacto conocido como códice90) que mostraban a detalle la información obtenida. De esta manera, la información se presentó desde unidades básicas de información conocidas como AGEB (área geoestadística básica), la cual puede ser vista como un conjunto de manzanas que engloba a cierto número de viviendas, hasta unidades mayores, tales como los estados del país (pasando por localidades, municipios o delegaciones, etc.) y finalmente obtener los datos para el país en general.

El presente estudio analiza algunos datos para cada uno de los estados de la República, los cuales son vistos como individuos a los que se miden ciertas características (variables) y en base a ello se aplican las técnicas y se obtienen conclusiones. Con lo anterior se pudo vislumbrar un comportamiento general de los estados con respecto a las variables seleccionadas.

A continuación se describe el manejo de los datos censales para llegar a la selección de las variables presentadas y su transformación para poder utilizarlas como elementos de entrada para cada una de las técnicas presentadas.

TRATAMIENTO DE LOS DATOS DEL CÓDICE90.

Debido a la gran cantidad de información presentada en el censo, era imposible obtener los datos a analizar a partir de los libro publicados por el INEGI (Instituto Nacional de Estadística Geografía e Informática), razón por la cual se obtuvieron de una publicación de los mismos datos por parte del INEGI en un CD, el cual es conocido como ***CÓDICE90*** y que contiene toda la información censal, además de otras cuestiones como cartografía, tablas de datos por grupos quinquenales de edad, sexo y otras subdivisiones.

El códice90 contiene una cantidad elevada de tablas, cada una con determinado número de variables que hacen referencia a población, migración e inmigración, escolaridad, fecundidad, economía, nivel de ingresos, viviendas y religión entre otros aspectos. La información guardada en él se encuentra en un formato particular, y no puede ser copiada directamente a un diskette para su manejo por separado.

Sin embargo, para realizar el presente trabajo se necesitaba utilizar la información por separado y analizar secciones del códice90, para poder seleccionar grupos de variables de una manera más fácil. Para ello se recurrió a una utilería que presenta el paquete mencionado, que consiste en una opción de exportar los datos a un archivo en formato ASCII. De esta manera, se obtuvo una copia en varios diskettes de las principales tablas del censo.

Las tablas mencionadas anteriormente contenían una serie de datos desglosados al máximo. Lo que primeramente se decidió fue compactar o reducir la información presentada en grupos quinquenales en datos generales por estado. Así, de las tablas que se extraerían los datos a analizar se seleccionaron únicamente los renglones que contenían la información completa que correspondía a cada uno de los estados. Cabe señalar que para este caso se tomó al Distrito Federal como uno más de los estados de la República Mexicana, aunque oficialmente no se considera como tal, pero se hizo para simplificar las expresiones al referirse a los individuos en estudio.

Una vez que los archivos contenían renglones que correspondían a cada uno de los estados, se procedió a hacer una selección de variables para formar las matrices a las que posteriormente se aplicarían los métodos del análisis multivariado.

Los datos presentados por el INEGI están en cifras totales, razón por la cual no se podía hacer un análisis comparativo entre los estados de la República; es decir, no se puede comparar el total de personas que asisten a la escuela primaria en el Distrito Federal con el que asiste en Quintana Roo, por ejemplo, ya que el D.F. posee un mayor número de habitantes que Quintana Roo. Debido a esta situación, se optó por presentar los resultados en porcentajes, ya que resulta mucho más ilustrativo el porcentaje de personas que poseen alguna característica; de esta manera, se pueden hacer comparaciones, como por ejemplo, que en el estado de Chiapas casi el 50 % de la población gana menos de un salario mínimo y que en el Distrito Federal casi el 50 % de la población gana más de un salario mínimo.

Así, se aplicaron los métodos descritos a lo largo de este trabajo a matrices que contenían como individuos u observaciones a los estados y que a cada uno se asociaba un valor en porcentaje, generalmente, para cada una de las variables seleccionadas.

Para poder llegar a estas matrices finales se requirió de bastante tiempo debido a la gran cantidad de variables que se manejaron y a las dificultades que se encontraron cuando esta tarea se llevaba a cabo. Los pasos que se siguieron para conseguir lo anterior fueron:

1) Copiar los datos del CD a diskettes con la utilería que proporciona el paquete. Estos archivos quedaron con formato ASCII.

2) Emigrar archivo por archivo a formato Lotus para poder hacer operaciones de eliminación de columnas, renglones, obtener porcentajes, etc. de una manera más fácil y rápida, razón por la cual se eligió Lotus 1-2-3 para manipular la información.

3) Cada uno de los archivos obtenidos de la manera anterior se manejó mediante la hoja de cálculo Lotus para llegar a la estructura individuos-variables. En este proceso se eliminó toda aquella información que no iba a ser utilizada, como los datos referentes a grupos quinquenales de edad, datos desplegados por sexo, etc. Es decir, se eliminó el desglose de los datos para llegar a totales por estado.

4) Una vez que se pudo manejar la información en un formato adecuado, se procedió a seleccionar las variables que se utilizarían.

5) Como se iban a manejar diferentes tablas, éstas se construyeron combinando variables de diferentes archivos, los cuales debían estar todos en el mismo formato, en este caso en formato Lotus.

6) Una vez hecho lo anterior se obtuvieron porcentajes para cada entrada de las tablas construidas. Estas fueron las tablas que se utilizaron para aplicar los métodos multivariados que se describen en los capítulos siguientes. Cabe señalar que el único método para el cual no se requirió de obtener porcentajes fué el de **Análisis de Correspondencia**, en el cual se manejan tablas utilizando los totales presentados por INEGI, aunque en este caso, también se realizó una simplificación de la información para llegar al formato de datos que ésta técnica requiere.

CAPITULO 1

MÉTODOS GRÁFICOS

I.- MÉTODOS GRÁFICOS

1.1 UN PRIMER ENFOQUE DE LOS DATOS

Cuando se tiene una serie de individuos (entiéndase objetos, observaciones, y en este caso los estados de la República Mexicana) de los cuales se posee información acerca de determinadas características (a las cuales se conoce como variables) atribuibles a cada uno de ellos, después de realizar una cuantificación y tabulación de los datos obtenidos se tendrá una representación del siguiente tipo:

#Individuo	X_1	X_2	.....	X_p
1	x_{11}	x_{12}	.....	x_{1p}
2	x_{22}	x_{22}	.....	x_{2p}
...	.....	.....	.....	.....
...	.....	.....	.....	.....
...	.....	.....	.....	.....
n	x_{n1}	x_{n2}	.....	x_{np}

En este caso se tienen **p** variables, **n** individuos u observaciones y x_{ij} será el valor que toma la variable j en la i-ésima observación.

La tabla anterior, que corresponde a una matriz de n x p, contiene la situación que guarda cada uno de los individuos respecto a las variables en

cuestión. Sin embargo, en una representación así muchas veces resulta difícil tener una idea general de cada uno de los individuos con respecto a los demás, y con respecto a las variables de estudio. Esto se da porque al estar comparando números únicamente lo podemos hacer en pequeños grupos (2 o 3 a la vez), lo que ocasiona que se complique un enfoque global de los datos.

Existen diferentes métodos gráficos que pretenden facilitar la tarea anterior, algunos cumplen mejor su objetivo que otros; aunque todos tienen los mismos principios y filosofía: proporcionar una representación visual que describa los datos de tal manera que permita, al ojo humano, enfocarlos globalmente y describir su comportamiento en una representación más fácil de interpretar en una inspección inicial.

Para ello, a cada observación se le asocia una figura (polígonos, estrellas, rostros, curvas senoidales, etc.) que, de acuerdo a sus características, represente a las variables en cuestión.

Por ejemplo, supongamos que tenemos en estudio la variable DENSIDAD DE POBLACIÓN y representamos cada estado con un círculo de radio igual al valor de esa variable. Con dar un vistazo general a la gráfica de círculos podemos identificar al D.F. como el círculo más grande y a Quintana Roo como el círculo más pequeño. Es claro que tratándose de una sola variable puede resultar más fácil remitirse a la tabla inicial que realizar una gráfica como la descrita. Sin embargo, en la mayoría de los casos reales se maneja una gran cantidad de variables, lo que hace más difícil remitirse a la tabla de datos que a una representación gráfica, por lo que casi siempre es muy útil contar con una visualización de este tipo ya que tiene ciertas ventajas:

a) Es fácil detectar y comprender fenómenos importantes.

b) Sirve como regla mnemotécnica para recordar las características principales de las observaciones.

c) Se pueden extraer conclusiones más fácilmente.

d) Con base en ello, se pueden hacer aproximaciones con cálculos informales o llegar a determinadas conclusiones iniciales.

Algunas de las representaciones más usuales son: poligonales, graficación por barras, por líneas, las curvas de Andrews, las caritas de Chernoff, gráficas de caja y gráficas de tronco y hoja.

A continuación se describe brevemente cada una de ellas, aunque para una mayor información referirse a la revista Chernoff [4], Krzanowski [12] o Marriott [13].

1.2.-DESCRIPCION POR POLÍGONOS.

Consiste en asociar a cada observación una figura poligonal construida con base en p aristas:

A partir del punto **O** se mide hacia la variable 1 el valor que ésta presentó en el individuo i. Se procede igual con las demás variables para el mismo

individuo. Al final se unen los valores obtenidos, generándose un polígono de la forma de la figura 1.1.

Al final se tendrán n polígonos como el anterior, que representarán a las observaciones en una configuración más fácil de interpretar. Como es de esperarse, los datos que se proporcionen deben estar dentro de una escala definida, ya que no resulta eficiente comparar datos que tengan magnitudes pronunciadamente diferentes. Una comparación entre un conjunto de datos, de los cuales uno está dado en porcentaje y el otro en totales proporciona una gráfica poco representativa, ya que las aristas deben estandarizarse y manejar la misma escala de medida. Así, si de un lado se tienen datos que oscilan entre 0.2 y 0.9 y del otro datos comprendidos entre 500 y 999, por decir algo, la escala de las aristas debería estandarizarse y aceptar un máximo de 999, lo cual impediría que se apreciaran correctamente las magnitudes del primer conjunto de datos. Este fue uno de los inconvenientes que se evitó transformando los datos que presenta el INEGI como totales a porcentajes por estado.

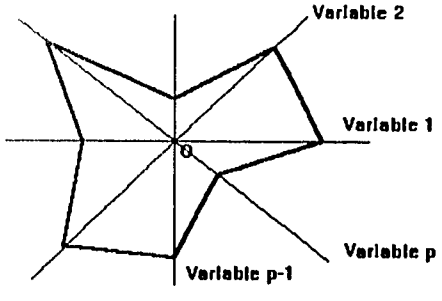


FIGURA 1.1 Construcción de un polígono

Como ya se ha mencionado, a medida que crece el número de variables la representación anterior resulta más útil para el ojo humano al permitirle hacer distinciones entre los individuos.

1.3.-DESCRIPCION POR BARRAS

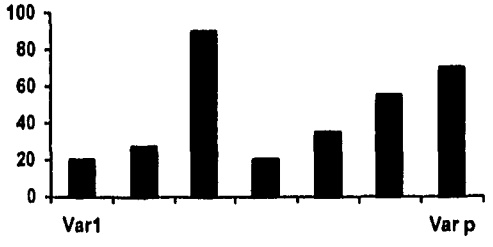


FIGURA 1.2 Graficación por medio de barras.

En este caso se obtiene una gráfica de barras para cada estado. Se construyen n gráficas con p barras cada una. La altura de cada barra será el valor observado en la variable correspondiente para el individuo en cuestión. La configuración resultante constará de n gráficas del tipo de la figura 1.2.

Generalmente el ancho de las gráficas será uniforme, aunque pueden construirse configuraciones en las cuales dicho valor varíe de acuerdo a determinados parámetros.

1.4.-GRAFICACIÓN POR LINEAS

La graficación por líneas es muy similar a la anterior. En este tipo de representaciones los puntos más altos de cada barra se unen por medio de una linea poligonal obteniéndose una salida como la representada en la figura 1.3.

En esta gráfica cada vértice corresponde a alguna variable. Así, se tendrán n gráficas como la que se presenta a continuación.

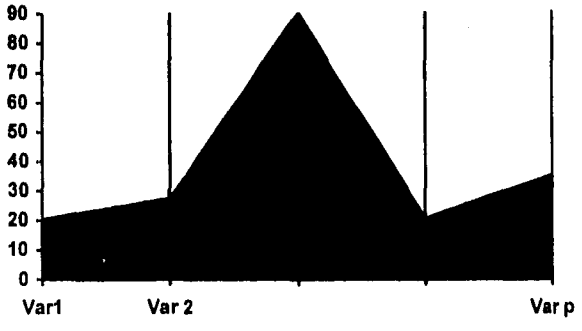


FIGURA 1.3 Graficación por líneas.

1.5.-OTRAS TÉCNICAS

Existen otras técnicas de diferente estilo, como las llamadas **CURVAS DE ANDREWS**, en las cuales a cada individuo se le asocia la función:

$$f(x) = X_1(2)^{1/2} + X_2\text{sen}(t) + X_3\cos(t) + X_4\text{sen}(2t) + X_5\cos(2t) + \cdots\cdots$$

En este caso, cada X_i corresponde al valor de la variable **X** para el i-ésimo individuo. Así, en una sola gráfica se tendrán n curvas senoidales que representarán a la población en estudio. La salida será del estilo de la figura 1.4.

Obviamente, en dicha gráfica se debe distinguir cuál curva corresponde a cada individuo. Esta técnica es muy valiosa cuando se pretender hacer agrupaciones entre los individuos.

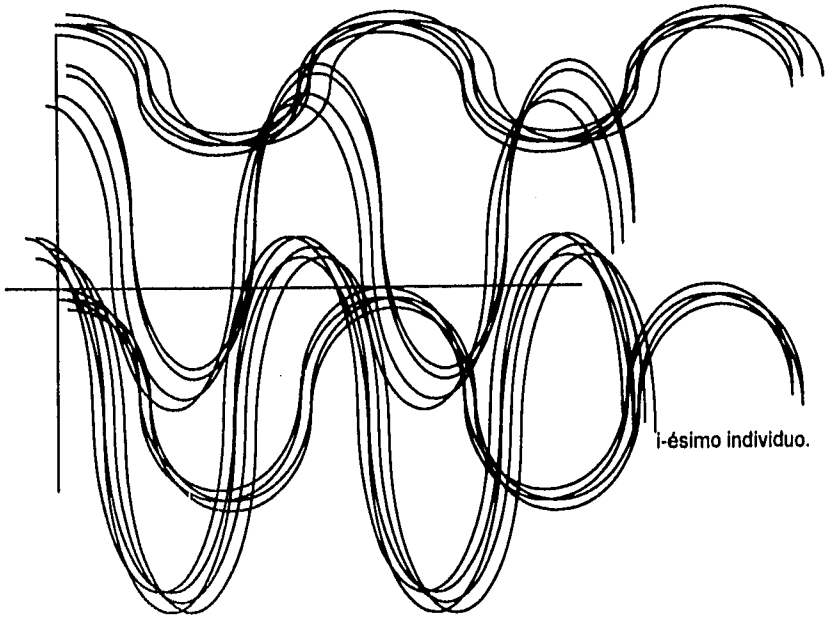


Figura 1.4: Ejemplificación de una gráfica de Curvas de Andrews, en la cual se pueden observar claramente 4 grupos de individuos..

Existe otra técnica conocida como las **CARITAS DE CHERNOFF**, propuestas por Chernoff, que consiste en asociar un rostro a cada observación, en el cual los rasgos estarán determinados por cada una de las variables. Por ejemplo, la variable 1 puede estar asociada con la forma de la cara, la variable 2 con la boca, etc. Así se obtendrán n caritas que representan la situación de los individuos.

También se encuentran otras representaciones como son las **DRAFTSMAN PLOT**, las cuales consiste en graficar todos los pares de

variables entre sí, proporcionando una visión respecto al comportamiento de las variables. Esto último se puede aplicar en análisis de regresión para ver la relación entre variables. Más adelante se ejemplifica éste tipo de gráficas.

Existe también, una representación numérica conocida como **STEAM AND LEAF,** traduciendo literalmente "steam and leaf", se tiene que es "tronco y hoja". Este tipo de gráfica se aplica a un conjunto de datos colectados, ya sea para un solo individuo con diferentes variables o para varios individuos con una sola variable. Así, se tiene que las gráficas de "tronco y hoja" trabajan con vectores de nx1 (o de 1xn). Hoaglin, Mosteller y Tukey [10] dan una descripción bastante completa acerca de ellas.

Lo que se pretende es resumir un conjunto de datos en una descripción gráfica sin pérdida de información y en caso de que la haya ésta es mínima. Su construcción es como sigue: Si se tienen datos que contienen n dígitos, se hace una lista ordenada con los n-1 primeros dígitos, sin que se repita ninguna de éstas cifras; de esta lista ordenada se hace una separación hacia el lado derecho, donde se escribirán los dígitos restantes para completar cada una de las cifras observadas. En caso de que haya cifras repetidas, o que coincidan en los primeros n-1 dígitos se escribirán en el mismo renglón. Más adelante se da un ejemplo de esta representación, aplicándola a datos obtenidos del *códice 90.*

1.5.1.- BOXPLOT.

Una gráfica de boxplot consiste en un recuadro que representa de una manera diferente el comportamiento de los datos numéricos.

De este tipo de gráficas se pueden ver las siguientes características: localización, dispersión y detección de observaciones discrepantes, entre otras. Es decir, muestran aspectos importantes del comportamiento de un conjunto de datos.

Para construirlas se inicia obteniendo los siguientes datos:

1) Mediana. Que corresponde a la observación $\mathbf{X}_{((n+1)/2)}$ si n es impar o al promedio de las observaciones $\mathbf{X}_k$ y $\mathbf{X}_{k+1}$ si n es par. Donde k=n/2.

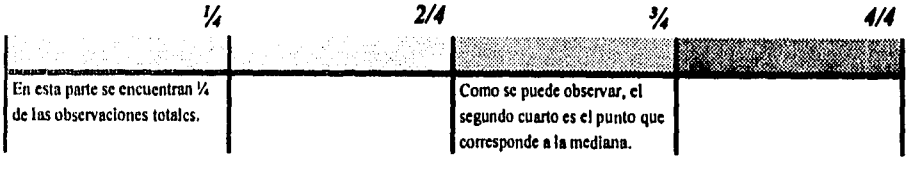
2) Límite inferior.

3) Límite superior.

4) Cuarto inferior. Denotado por F_L.

5) Cuarto Superior.Denotado por F_U.

Los datos 2, 3, 4 y 5 se definen a continuación: Los cuartos inferior y superior se obtienen al ordenar las observaciones en forma ascendente y "dividir" el conjunto de observaciones en cuatro partes iguales, como se muestra a continuación:



Para calcular el primero y tercer cuartos cuando **n** no sea divisible entre cuatro, se toma el promedio de las observaciones $\mathbf{X}_{[n/4]}$ con $\mathbf{X}_{([n/4]+1)}$, y $\mathbf{X}_{[3n/4]}$ con $\mathbf{X}_{([n/4]+1)}$, respectivamente.

Al rango comprendido entre el cuarto inferior y el cuarto superior se le conoce como "amplitud entre cuartos", que está muy relacionado con lo que se conoce como rango intercuartil. Basándose en dicho rango se puede construir un intervalo para la detección de observaciones discrepantes. Para ello se procede de la siguiente manera:

$$L_I = F_L - 3/2d_F$$

$$\text{y} \quad L_S = F_U + 3/2d_F$$

Donde $d_F = F_U - F_L$.

Además, L_I es el valor del límite inferior y L_S el del límite superior mencionados en los puntos 2 y 3.

Para construir las "cajas" o recuadros se dibuja un cuadrado con esquina superior izquierda en F_U y esquina inferior derecha en F_L. A partir de este cuadro se trazan líneas hacia el último punto que no es menor al límite inferior y otro hacia el último punto que no excede al límite superior respectivamente y finalmente se traza una línea que cruza el cuadro en el punto correspondiente a la mediana. Esta construcción puede ser de manera horizontal o vertical.

Visto en un recuadro queda como la figura 1.5, donde se nota el por qué se llama gráfica de boxplot, ya que da apariencia de ser una caja. Y su traducción literal sería "gráficas de caja".

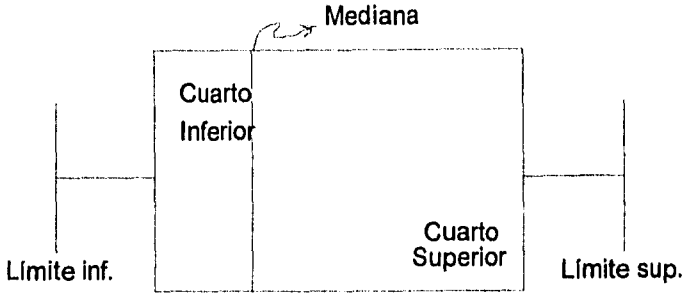


Figura 1.5: Representación de una gráfica boxplot.

La longitud de la caja muestra la dispersión que hay basándose en F_U y F_L. Un ejemplo real de este tipo de gráficas se dará en la figura 1.12.

De esta manera, independientemente de los objetivos que se tengan al comenzar el análisis de los datos, es muy útil tener una representación que nos permita guardar en la mente la situación de las observaciones de una manera más fácil ya que, por ejemplo, en análisis de conglomerados, factores y componentes principales puede ayudar el hecho de contar con un agrupamiento previo de los individuos.

Además, si tenemos asociado un esquema gráfico a un problema estadístico, estaremos en condiciones de iniciar un análisis más completo puesto que se tiene una visión "panorámica" del problema.

1.6.-APLICACIÓN A LOS DATOS DEL CENSO

Para aplicar éstas técnicas a los datos del censo las observaciones serán los 32 estados de la República Mexicana. Se despliegan 4 tipos de representaciones: poligonal, gráfica de líneas, gráfica de barras y las caritas de Chernoff. Las variables estarán dadas como sigue:

a) Diversas variables representativas. Descritas en la tabla 1.1.

b) Variables de ingreso. Descritas en la tabla 1.2.

c) Variables que indican el sector de actividad. Detalladas en la tabla 1.3.

d) Variables que indican la actividad desempeñada. Descritas en la tabla 1.4.

A continuación se presentan los resultados obtenidos para cada grupo de variables de acuerdo a las técnicas señaladas. Además, a las variables de la tabla 1.1 se aplicó la gráfica "Draftsman Plot", la cual se analiza al final.

1.6.1.-GRÁFICAS DE POLÍGONOS

Las variables que aparecen en la tabla 1.1 son:

VARIABLE 1: Porcentaje de la población económicamente activa mayor de 12 años.
VARIABLE 2: Porcentaje de desocupados de la PEA.

VARIABLE 3: Porcentaje de "Otro tipo de inactivos" de la PEA.
VARIABLE 4: Porcentaje de la PEA que gana entre 0 y 1 salario mínimo.
VARIABLE 5: Porcentaje de la PEA que gana entre 1 y 3 salarios mínimos.
VARIABLE 6: Porcentaje de la PEA que gana más de 3 salarios mínimos.
VARIABLE 7: Porcentaje de la población universitaria respecto a la población mayor de 18 años.
VARIABLE 8: Porcentaje de la población que cuenta con energía eléctrica en su casa.
VARIABLE 9: Porcentaje de alfabetas respecto a la población mayor de 15 años.
VARIABLE 10: Porcentaje de la población que no tiene estudios de primaria respecto a la población mayor de 6 años.

Este grupo de variables será analizada por medio de la técnica de polígonos. La matriz obtenida al tabular los datos es de dimensión 32 x 10, por lo que se obtendrán 32 figuras . En este caso se utilizó el paquete Stat-Graphics para obtener la salida de la figura 1.6.

Al analizar dicha gráfica, se observa que los estados 7 y 20, que corresponden a Chiapas y Oaxaca presentan los valores del lado derecho del polígono muy pequeños, es decir desde la variable 5 hasta la variable 9 se presentan valores bajos para dichos individuos. Estas variables corresponden a nivel de ingresos, educación y disposición de servicios de energía eléctrica. Lo anterior contrasta con los casos 9 y 19, que corresponden al D.F.. y a Nuevo León, en cuyos casos los valores de las mismas variables son los más altos.

Es notable que el caso 7 (Chiapas) presenta un pico muy pronunciado respecto a la variable 10, en Oaxaca la misma variable presenta igualmente un

TABLA 1.1
VARIABLES DESCRIPTIVAS

	ENTIDAD FEDERATIVA	VARIABLE 1 % PEA > 12	VARIABLE 2 % PEA DESOCUP.	VARIABLE3 %OTROS INAC. DE PEI	VARIABLE 4 (0,1) S.M.	VARIABLE 5 (1,3) S.M.	VARIABLE 6 (3, +) S.M.	VARIABLE 7 % POB UNIV..	VARIABLE 8 % CON ELECTRIC..
	ESTADOS UNIDOS MEXICANOS	43.03635	2.742227	9.095101	26.94856	51.41121	17.36632	8.154809	87.00775
1	AGUASCALIENTES	44.88849	2.177417	8.951781	18.85809	61.09245	16.87048	8.068601	95.02520
2	BAJA CALIFORNIA	49.40941	2.234459	10.49570	10.15684	54.84030	30.30659	9.168417	89.50981
3	BAJA CALIFORNIA SUR	47.28892	2.111830	8.423482	16.10891	59.47763	20.46261	8.198038	88.89878
4	CAMPECHE	42.81913	1.909065	7.991772	33.43712	47.98543	13.12282	5.939056	84.97793
5	COAHUILA	43.31410	3.153402	9.719744	18.59084	59.29832	16.52447	10.25412	94.75671
6	COLIMA	45.93913	1.890537	9.015220	14.39456	58.87663	23.39706	7.964969	94.17978
7	CHIAPAS	42.91418	2.299983	7.174752	58.94359	29.12595	7.705591	3.409608	65.07960
8	CHIHUAHUA	46.22185	3.004952	12.70888	15.23748	55.95485	23.68257	7.921441	86.76831
9	DISTRITO FEDERAL	47.62848	2.582101	6.194831	20.72734	55.12763	21.11084	16.20107	99.24238
10	DURANGO	39.44845	3.533114	12.19253	29.65632	52.46332	14.08883	6.449390	86.26610
11	GUANAJUATO	40.45233	3.108328	11.98604	25.92704	50.36615	18.00234	4.549131	87.49570
12	GUERRERO	37.59201	3.953760	12.10395	38.31599	43.72175	11.58895	4.723563	77.36944
13	HIDALGO	40.50875	2.995963	8.773975	39.51694	45.33026	10.46714	4.838851	77.39133
14	JALISCO	43.85707	2.203010	9.446006	19.62030	53.87142	22.45870	8.042764	92.13250
15	MEXICO	43.41790	2.957201	7.385821	20.63491	57.95448	17.90553	8.101117	93.56295
16	MICHOACAN	39.11530	3.073507	11.37177	29.11871	45.71076	15.97435	5.358443	86.86913
17	MORELOS	42.99925	3.183876	10.38975	18.28526	59.15913	19.41801	7.786207	96.04284
18	NAYARIT	42.60700	2.133325	10.08675	21.65064	54.18884	18.42618	6.694331	91.32828
19	NUEVO LEON	45.94298	2.622182	7.648881	15.91279	58.22843	22.30255	12.55352	96.45067
20	OAXACA	39.24155	2.776202	10.32886	53.07958	34.53496	8.211267	3.514798	76.17273
21	PUEBLA	40.35604	2.356889	8.134660	38.80686	45.05125	12.30407	7.027951	84.51690
22	QUERETARO	43.05826	3.094339	9.990226	23.98146	52.28966	20.09245	7.886668	84.34548
23	QUINTANA ROO	51.21153	1.350469	7.087828	21.96764	45.85758	24.90532	6.573370	84.60901
24	SAN LUIS POTOSI	40.52218	2.379001	10.16689	36.57034	45.89256	12.21343	6.899758	72.01787
25	SINALOA	44.15237	2.005542	10.28622	15.47302	58.65911	20.53396	9.557039	91.00962
26	SONORA	44.64217	2.567372	9.839298	12.10129	59.31868	24.56693	8.669344	90.76028
27	TABASCO	41.08160	3.117981	7.696367	36.36526	42.67119	15.49764	5.918224	84.54671
28	TAMAULIPAS	44.09556	3.590889	11.40556	23.42385	55.79504	16.39982	9.253758	84.09916
29	TLAXCALA	39.70240	3.579555	8.027376	31.80678	54.04330	10.90285	6.303524	94.34987
30	VERACRUZ	41.82094	2.797733	8.227675	36.59878	47.51381	12.40952	5.877336	72.82007
31	YUCATAN	43.60256	1.512598	6.702433	38.94613	46.78632	11.44997	5.915988	90.80429
32	ZACATECAS	36.00075	4.002164	12.89349	38.67818	45.14599	11.00292	4.922935	86.69838

TABLA 1.1 (CONTINUACIÓN)
VARIABLES DESCRIPTIVAS

	ENTIDAD FEDERATIVA	VARIABLE 9 % ALFABETAS >15	VARIABLE 10 % SIN PRIMARIA >=6
	ESTADOS UNIDOS MEXICANOS	87.38823	14.00790
1	AGUASCALIENTES	92.84916	10.72876
2	BAJA CALIFORNIA	95.05645	8.580399
3	BAJA CALIFORNIA SUR	94.23655	8.919416
4	CAMPECHE	84.38860	16.46757
5	COAHUILA DE ZARAGOZA	94.40733	8.971977
6	COLIMA	90.58562	12.07846
7	CHIAPAS	69.60788	28.16729
8	CHIHUAHUA	93.71077	9.658746
9	DISTRITO FEDERAL	95.87304	6.686970
10	DURANGO	92.92465	10.79164
11	GUANAJUATO	83.20347	17.96562
12	GUERRERO	72.97493	24.72764
13	HIDALGO	79.14353	17.83173
14	JALISCO	90.95039	12.24416
15	MEXICO	90.84121	11.05321
16	MICHOACAN	82.35511	17.96009
17	MORELOS	87.98331	14.01486
18	NAYARIT	88.30774	13.23831
19	NUEVO LEON	95.22778	7.795570
20	OAXACA	72.32314	23.60764
21	PUEBLA	80.59984	18.07698
22	QUERETARO	84.44835	16.79992
23	QUINTANA ROO	87.40465	13.60646
24	SAN LUIS POTOSI	84.88791	15.20474
25	SINALOA	89.94179	11.41149
26	SONORA	94.12195	8.334347
27	TABASCO	87.13855	12.82818
28	TAMAULIPAS	93.00933	9.978433
29	TLAXCALA	88.79205	12.15861
30	VERACRUZ	81.61882	18.56202
31	YUCATAN	83.98834	15.87681
32	ZACATECAS	90.01658	12.81056

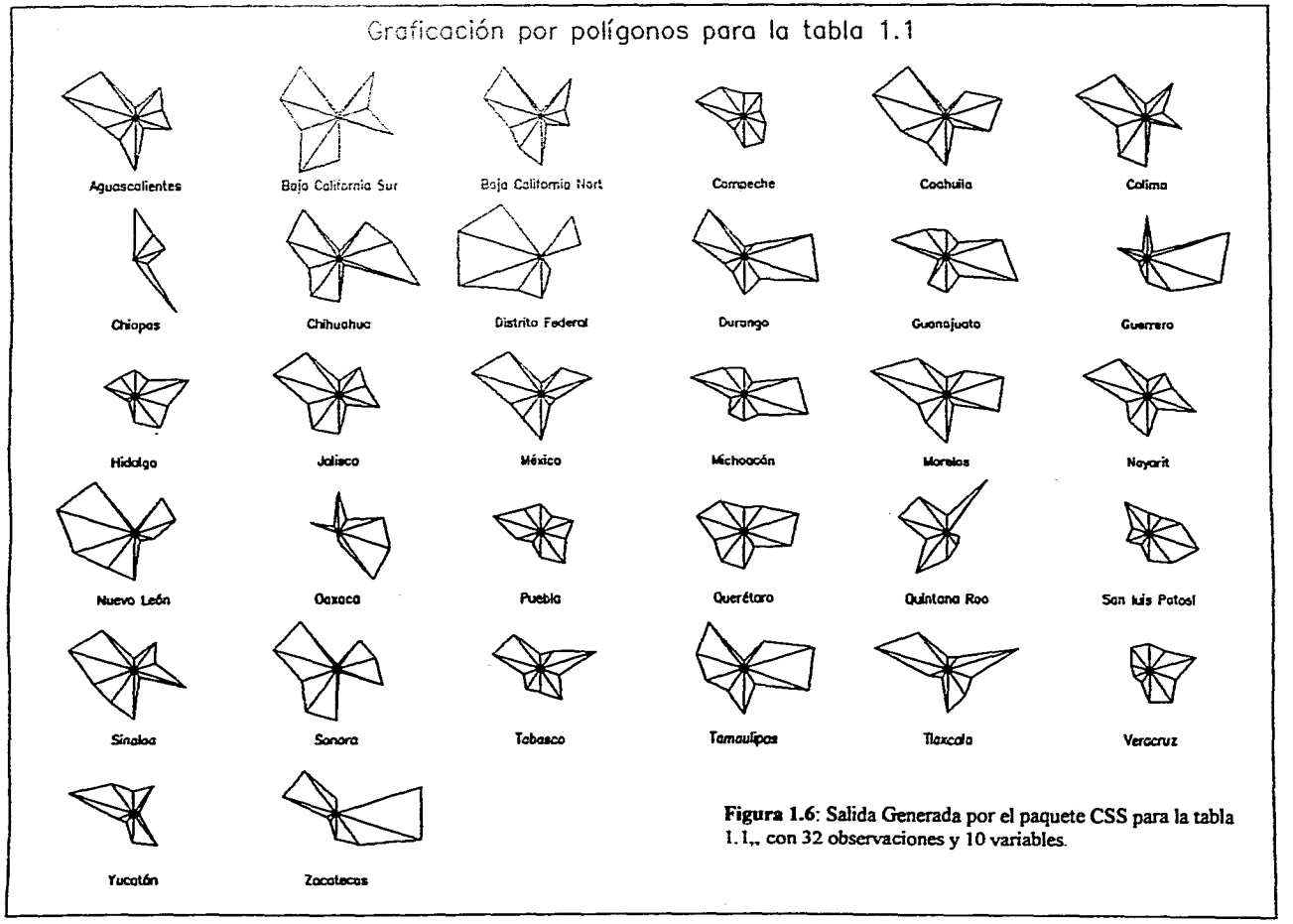


Figura 1.6: Salida Generada por el paquete CSS para la tabla 1.1,, con 32 observaciones y 10 variables.

valor alto. Lo anterior indica que el nivel de educación de dichos estados es sumamente bajo.

Los estados de Baja California N., Chihuahua, Jalisco y Sonora presentan similitudes respecto a las variables 5 a 8, que corresponden a nivel de ingresos y educación, así como a disposición de energía eléctrica. Los casos correspondientes a Zacatecas, Guerrero y Durango se asemejan en las variables 1 y 2, que corresponden a porcentajes de la PEA. Lo que parecería indicar similitudes en aspectos laborales, lo cual se podría justificar debido a que ambos estados tienen características muy parecidas en situación geográfica dentro del país, además de que sus características climáticas son semejantes.

De la gráfica se puede concluir que Chiapas y Oaxaca son estados con escasos servicios de energía eléctrica, alto grado de analfabetismo y en donde un gran porcentaje de la población gana apenas el salario mínimo.

Zacatecas presenta una situación muy extraña, ya que a pesar de haber un porcentaje bajo de PEA existe un alto índice de desempleados. Caso contrario es Quintana Roo, donde un valor alto de la variable 6 (porcentaje de la PEA que gana más de 3 salarios mínimos) y uno bajo de la variable 2 (porcentaje de PEA desocupada) indican que no existen condiciones graves de desempleo.

De esta manera podemos distinguir 3 tipos de individuos: los que se parecen a Chiapas y Oaxaca (con bajo nivel de desarrollo), los que se asemejan al D.F. y a Nuevo León (con condiciones más favorables de vida), los parecidos a Zacatecas con altos índices de desempleo y como caso particular Quintana Roo, con pocos problemas de desempleo.

1.6.2.- GRÁFICAS DE BARRAS

En la tabla 1.2 se muestran las siguientes variables:

VARIABLE 1: Porcentaje de la PEA ocupada que no percibe ingresos.
VARIABLE 2:: Porcentaje de la PEA ocupada que gana menos de la mitad de un salario mínimo.
VARIABLE 3:: Porcentaje de la PEA ocupada que gana desde medio hasta menos de un salario mínimo.
VARIABLE 4:: Porcentaje de la PEA ocupada que gana un salario mínimo.
VARIABLE 5:: Porcentaje de la PEA ocupada que gana más de uno y hasta dos salarios mínimos.
VARIABLE 6:: Porcentaje de la PEA ocupada que gana entre 2 y 3 salarios mínimos.
VARIABLE 7:: Porcentaje de la PEA ocupada que gana de 3 a 5 salarios mínimos.
VARIABLE 8:: Porcentaje de la PEA ocupada que gana entre 5 y 10 salarios mínimos.
VARIABLE 9:: Porcentaje de la PEA ocupada que gana más de 10 salarios mínimos.

Para este grupo de variables se utilizó un gráfica de barras (ver figura 1.7). Y la información extraída al analizarla es la siguiente:

Respecto a las variables de la tabla 1.2, salta a la vista la escasez de recursos económicos por parte de Chiapas y Oaxaca, donde la mayoría percibe ingresos mínimos. Un gran porcentaje no recibe siquiera un salario mínimo

TABLA 1.2
VARIABLES DE INGRESO
PORCENTAJES DE LA POBLACION OCUPADA DE ACUERDO AL NÚMERO DE SALARIOS MÍNIMOS QUE GANAN.

	ENTIDAD FEDERATIVA	VARIABLE 1 0	VARIABLE 2 (0, ½)	VARIABLE 3 (½, 1)	VARIABLE 4 1	VARIABLE 5 (1,2]	VARIABLE 6 (2,3)	VARIABLE 7 [3,5]	VARIABLE 8 (5,10)	VARIABLE 9 > 10
	ESTADOS UNIDOS MEXICANOS	7.221707	6.657148	12.64811	4.21600	36.27637	15.13483	9.757307	5.094607	2.514406
1	AGUASCALIENTES	4.037859	4.421161	9.934311	4.64765	43.66915	17.42330	9.515221	5.015421	2.339839
2	BAJA CALIFORNIA	1.287068	4.515350	3.605313	7.49110	29.80082	25.03948	16.35892	9.810405	4.137259
3	BAJA CALIFORNIA SUR2	.698442	3.148020	9.735994	5.26454	37.94848	21.52914	12.53661	5.637242	2.288761
4	CAMPECHE	8.369615	8.414286	16.49253	1.60684	34.63392	13.35151	7.824220	3.731089	1.567510
5	COAHUILA DE ZARAGOZA	2.463640	4.606893	11.19343	3.26870	42.29798	17.00033	10.43443	5.530183	2.559859
6	COLIMA	3.708587	3.614936	6.942925	1.28114	36.00551	22.87112	14.27394	6.721908	2.401216
7	CHIAPAS	19.00383	18.85901	21.02453	0.56195	21.13377	7.992188	4.124758	2.265971	1.314860
8	CHIHUAHUA	5.519725	4.546889	4.695252	4.75617	37.57268	18.38196	12.80804	7.544819	3.329711
9	DISTRITO FEDERAL	1.054628	3.838801	15.06856	7.65354	39.74335	15.38428	10.97948	6.645643	3.485709
10	DURANGO	11.68296	6.077892	11.64284	2.52825	37.99784	14.46548	8.273270	4.028219	1.787344
11	GUANAJUATO	7.957598	6.344548	11.22922	3.95666	35.39421	14.97194	10.48293	5.211423	2.307990
12	GUERRERO	14.70605	10.50518	12.70623	3.98525	29.49301	14.22873	7.071621	3.122818	1.394512
13	HIDALGO	8.935467	8.298551	22.02264	2.60279	34.17877	11.15149	6.388183	2.892472	1.188490
14	JALISCO	5.359380	5.148976	8.618582	4.93367	35.87447	17.99695	13.10537	6.219796	3.133526
15	MEXICO	3.680387	5.037022	11.15357	7.63935	42.09462	15.85986	9.782955	5.130836	2.991741
16	MICHOACAN	11.72061	7.583590	9.436545	3.77968	30.85977	14.85099	9.158927	4.602673	2.212758
17	MORELOS	5.170844	3.956286	8.861598	2.96534	41.21404	17.94509	11.72389	5.327867	2.366250
18	NAYARIT	8.531330	5.833905	7.179828	1.05579	31.97682	22.21201	11.26566	4.826609	2.333905
19	NUEVO LEON	2.148508	3.835243	9.598111	3.30928	42.82526	15.40317	11.53296	6.620746	4.148837
20	OAXACA	24.79368	14.36912	13.87025	0.46532	25.65328	8.881685	5.012030	2.166497	1.032738
21	PUEBLA	12.59005	9.024398	16.89784	2.94563	33.59823	11.45302	6.922151	3.671715	1.710202
22	QUERETARO	7.665211	5.094223	10.91752	3.04504	36.47584	15.81382	10.65350	6.154453	3.284497
23	QUINTANA ROO	7.714320	5.926220	8.068509	2.58594	27.20571	18.65187	14.67920	7.035357	3.190759
24	SAN LUIS POTOSI	10.86961	8.714859	16.52369	4.62178	34.57211	11.32045	6.825880	3.598378	1.789170
25	SINALOA	4.497923	4.570399	6.121757	2.82945	40.13572	18.52338	11.99673	5.984067	2.553165
26	SONORA	1.855309	3.518935	6.507452	2.19600	40.60218	18.71650	13.90699	7.402389	3.257549
27	TABASCO	10.81934	7.358794	18.04953	1.57586	29.10068	13.57050	9.119954	4.222817	2.154872
28	TAMAULIPAS	3.426484	5.714264	13.95851	3.24592	37.64239	18.15265	9.698488	4.608867	2.092469
29	TLAXCALA	8.998062	7.231103	15.40875	1.68863	40.52205	13.52125	6.670091	2.848801	1.383965
30	VERACRUZ	10.27897	7.462076	18.67014	1.87586	35.24819	12.26562	7.577395	3.443200	1.388932
31	YUCATAN	5.982024	13.24014	19.53910	1.84859	34.66294	12.12337	6.703785	3.276648	1.469544
32	ZACATECAS	17.59062	8.823669	12.07268	1.91198	34.02624	11.11975	6.437250	3.101630	1.464045

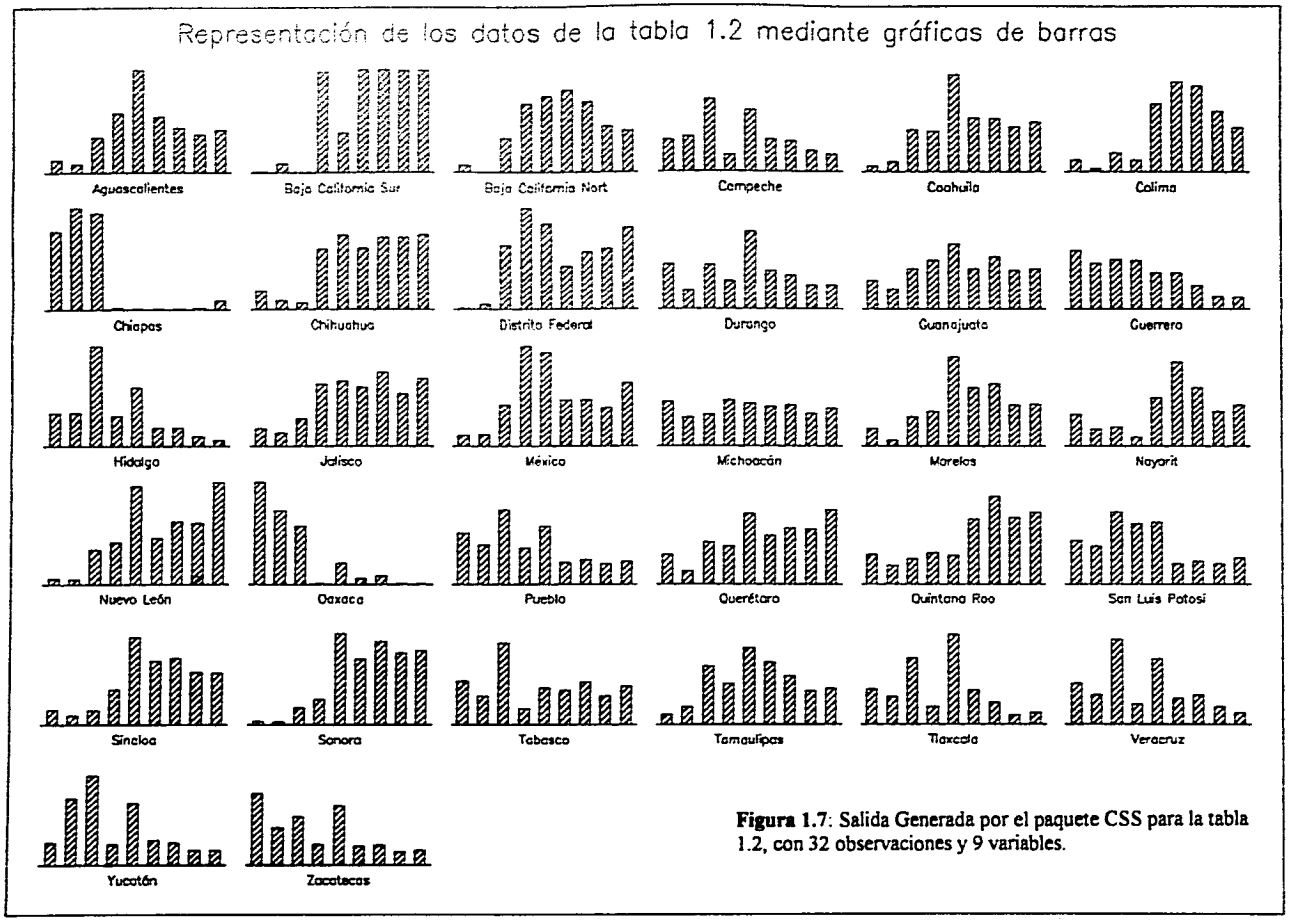


Figura 1.7: Salida Generada por el paquete CSS para la tabla 1.2, con 32 observaciones y 9 variables.

completo. Y hay que tener en cuenta que el salario mínimo en zonas rurales es menor al de zonas urbanas como el D.F..

La mayoría de los estados muestra cierta tendencia a estabilizarse en las variables 4 y 5. En este aspecto es notable la semejanza entre los estados de Tlaxcala y Veracruz. Además, los estados de Baja California Nte., Distrito Federal y Nuevo León presentan un nivel de ingresos bastante elevado respecto a los demás individuos.

En los estados de Quintana Roo, Morelos, Sinaloa y Sonora predominan valores altos en las últimas variables y bajos en las primeras variables. Lo que indica la tendencia de éstos hacia ingresos aceptables.

1.6.3.-GRÁFICAS DE LÍNEAS

Las variables de la tabla 1.3 son:

VARIABLE 1: Porcentaje de la población ocupada que se dedica a la agricultura.
VARIABLE 2: Porcentaje de la población ocupada que se dedica a la minería
VARIABLE 3: Porcentaje de la población ocupada que se dedica a la extracción de recursos naturales.
VARIABLE 4: Porcentaje de la población ocupada que se dedica a la industria de manufacturas.
VARIABLE 5: Porcentaje de la población ocupada que se dedica a la generación electricidad y extracción y tratamiento de agua.

VARIABLE 6: Porcentaje de la población ocupada que se dedica a la industria de la construcción.

VARIABLE 7: Porcentaje de la población ocupada que se dedica al comercio.

VARIABLE 8: Porcentaje de la población ocupada que se dedica a la industria del transporte.

VARIABLE 9: Porcentaje de la población ocupada que se dedica a los servicios financieros.

VARIABLE 10: Porcentaje de la población ocupada que se dedica a los servicios adminsitrativos.

VARIABLE 11: Porcentaje de la población ocupada que se dedica a los servicios comunitarios.

VARIABLE 12: Porcentaje de la población ocupada que se dedica a los servicios profesionales.

VARIABLE 13: Porcentaje de la población ocupada que se dedica a los servicios de hotelería y restaurantes.

VARIABLE 14: Porcentaje de la población ocupada que se dedica a los servicios de mantenimiento.

VARIABLE 15: Porcentaje de la población ocupada que se dedica a actividades no especificadas.

Para este grupo de variables se eligió la representación por gráficas de líneas (figura 1.8) y en ella se muestra de manera inmediata la semejanza entre Chiapas y Oaxaca. En estos lugares más de la mitad de la población se dedica a la agricultura. Nuevamente se presenta contraste con el D.F. donde resaltan actividades tales como: comercio, servicios financieros, administrativos y profesionales.

TABLA 1.3 (CONTINUACIÓN)
VARIABLES SEGUN SECTOR DE ACTIVIDAD
PORCENTAJES RESPECTO A LA POBLACION OCUPADA

	ENTIDAD FEDERATIVA	VARIABLE 9 SERV. FIN.	VARIABLE 10 S. ADMVS.	VARIABLE 11 S. COMUN.	VARIABLE 12 S. PROF.	VARIABLE 13 REST. Y HOTEL.	VARIABLE 14 MANTENIM.	VARIABLE 15 NO ESPECIF.
	ESTADOS UNIDOS MEXICANOS	1.540015	3.956763	8.620900	1.843812	3.277180	9.134718	3.434849
1	AGUASCALIENTES	.255385	5.094059	9.213853	1.577472	3.057942	9.100369	1.916982
2	BAJA CALIFORNIA	.865565	3.817348	8.372666	2.216912	5.816920	11.23788	3.886671
3	BAJA CALIFORNIA SUR	1.581308	3.831972	10.56800	1.762307	7.427770	9.577377	3.412706
4	CAMPECHE	.82098	5.567297	9.019688	1.257475	2.918330	8.178926	4.019122
5	COAHUILA DE ZARAGOZA	.501283	3.208482	10.29147	2.452892	2.776010	9.287828	2.942345
6	COLIMA	.282646	6.621514	9.576396	1.708947	5.867060	9.964487	2.991593
7	CHIAPAS	0.527535	2.867967	6.326808	0.781236	1.463544	5.341160	3.105627
8	CHIHUAHUA	1.363471	3.149139	7.579873	2.339024	3.306299	8.437200	3.809597
9	DISTRITO FEDERAL	3.875618	7.342224	12.88970	3.754219	3.864105	12.85001	4.006229
10	DURANGO	1.001511	4.120653	9.914333	1.230436	2.318911	8.112878	2.719746
11	GUANAJUATO	1.057602	2.225770	6.590723	1.180884	2.601731	8.190086	3.348508
12	GUERRERO	0.696030	3.788117	8.864659	1.022631	6.776895	7.220537	4.118805
13	HIDALGO	0.539006	2.889837	7.880563	0.904290	1.466405	6.823631	3.732706
14	JALISCO	1.752251	2.812061	7.835877	1.871553	4.303368	10.23324	3.415782
15	MEXICO	1.517244	4.882669	7.940402	2.160486	3.075244	10.01186	3.592550
16	MICHOACAN	0.957423	2.368610	7.259890	1.136148	2.521547	7.319315	5.423417
17	MORELOS	0.994956	3.699078	9.997215	1.449662	3.936766	12.01841	2.340128
18	NAYARIT	0.809012	3.813304	9.080686	0.968240	4.053218	7.192703	4.163948
19	NUEVO LEON	2.172478	3.137430	9.484995	2.552041	2.972511	10.77166	3.220039
20	OAXACA	0.393739	3.194596	7.088511	0.773824	1.890614	4.816884	2.364163
21	PUEBLA	0.876866	2.276365	7.211181	1.167556	1.767934	6.715201	3.036845
22	QUERETARO	1.375807	3.243319	7.690471	1.610414	3.122902	9.437220	3.018401
23	QUINTANA ROO	1.530118	5.852687	6.650530	1.555855	14.47699	9.254856	6.061033
24	SAN LUIS POTOSI	0.964621	2.826001	8.542652	1.285594	2.496143	7.825283	3.317669
25	SINALOA	1.404589	3.534849	8.013103	1.364946	3.782994	8.092539	3.716116
26	SONORA	1.882337	4.403025	8.336267	2.047881	3.574057	9.923611	2.801812
27	TABASCO	0.819959	5.235185	8.403950	1.351179	2.304071	8.073272	4.340499
28	TAMAULIPAS	1.442261	4.069972	9.238185	1.760864	3.840040	10.33700	3.193630
29	TLAXCALA	0.538632	3.094975	9.288486	1.090997	1.563000	6.229623	1.875804
30	VERACRUZ	0.723310	2.775684	6.992650	1.142567	2.526563	7.764580	2.641308
31	YUCATAN	1.507842	3.302179	8.812359	1.463898	3.335567	10.69188	1.976250
32	ZACATECAS	0.794340	3.563835	8.774086	0.870752	2.545694	6.569018	3.206229

TABLA 1.3
VARIABLES SEGUN SECTOR DE ACTIVIDAD
PORCENTAJES RESPECTO A LA POBLACION OCUPADA

	ENTIDAD FEDERATIVA	VARIABLE 1 AGRICULTURA	VARIABLE 2 MINERIA	VARIABLE 3 EXTRACCIÓN	VARIABLE 4 MANUFACT.	VARIABLE 5 ELECT. Y AGUA	VARIABLE 6 CONSTRUC.	VARIABLE 7 COMERCIO	VARIABLE 8 TRANSP.
	ESTADOS UNIDOS MEXICANOS	22.64675	0.424010	0.689138	19.19924	0.660027	6.815078	13.28066	4.466835
1	AGUASCALIENTES	14.95820	0.435396	0.094648	24.71970	0.378593	8.584277	14.23728	5.372825
2	BAJA CALIFORNIA	10.36021	0.056413	0.111942	23.19376	0.879797	7.506308	16.42648	4.250085
3	BAJA CALIFORNIA SUR	18.31396	1.248503	0.113854	8.705467	1.009118	7.721650	14.59961	5.126358
4	CAMPECHE	34.29655	0.112579	2.438943	9.223045	0.646073	7.017461	10.97391	3.458391
5	COAHUILA DE ZARAGOZA	12.13600	2.620762	0.163776	25.63919	1.04680	8.188479	13.30734	4.437317
6	COLIMA	23.98294	1.720934	0.314668	9.933769	1.120068	8.298245	12.04129	4.575422
7	CHIAPAS	58.34042	0.091434	0.221855	5.926999	0.415145	4.471181	7.613102	2.505973
8	CHIHUAHUA	17.02367	1.247057	0.127797	26.37925	0.580390	7.580908	13.04423	4.032078
9	DISTRITO FEDERAL	0.663649	0.083437	0.624617	21.27300	0.732388	4.270822	17.15182	6.618155
10	DURANGO	26.56669	1.756533	0.060470	16.96954	0.764811	7.011446	11.68267	3.769347
11	GUANAJUATO	22.97827	0.476819	0.792886	24.97097	0.509532	8.230954	13.35685	3.488390
12	GUERRERO	36.39855	0.374005	0.041519	9.213492	0.562807	6.665903	10.28025	3.975774
13	HIDALGO	37.03191	1.146529	0.781447	15.41489	0.624347	7.271216	9.938072	3.555132
14	JALISCO	15.06668	0.203386	0.091681	23.98258	0.456218	8.016471	15.63157	4.327254
15	MEXICO	8.673263	0.132891	0.364910	28.35637	0.841915	7.137773	15.53550	5.776909
16	MICHOACAN	33.99856	0.143069	0.076692	15.23535	0.446588	7.250808	12.61558	3.246986
17	MORELOS	20.34895	0.283904	0.090424	16.18397	0.695550	10.64138	13.26598	4.053600
18	NAYARIT	38.23218	0.230042	0.039914	9.974248	0.622746	6.766523	10.64377	3.409442
19	NUEVO LEON	6.124799	0.227222	0.545868	29.77473	0.711679	8.932391	14.61641	4.755721
20	OAXACA	52.87622	0.209464	0.970297	10.06502	0.372660	4.795672	7.472706	2.715612
21	PUEBLA	36.92364	0.294194	0.261916	17.76493	0.440646	6.135388	11.58094	3.546383
22	QUERETARO	17.91421	0.559873	0.133566	25.36903	0.525270	10.70091	11.77913	3.519450
23	QUINTANA ROO	19.61701	0.137263	0.026962	6.295116	0.741467	8.331392	13.65218	5.816532
24	SAN LUIS POTOSI	31.12987	1.060648	0.266532	17.31422	0.427019	6.942323	11.63367	3.967743
25	SINALOA	36.72388	0.200785	0.172339	10.62891	0.486302	5.684478	12.14350	4.050657
26	SONORA	22.74238	1.223536	0.190972	16.07916	0.726013	7.191324	14.19167	4.685927
27	TABASCO	35.60775	0.132932	5.230610	8.432418	0.668472	6.042182	10.02328	3.334231
28	TAMAULIPAS	16.27346	0.117157	2.832663	18.96019	0.882331	7.718647	13.87831	5.455262
29	TLAXCALA	28.55922	0.070698	0.092569	25.49222	0.315855	7.934529	9.747264	4.106119
30	VERACRUZ	39.35684	0.325406	2.775741	11.48703	0.910552	5.661521	10.97278	3.943450
31	YUCATAN	27.01866	0.101881	0.131095	15.46287	0.759813	8.068503	13.77311	3.594075
32	ZACATECAS	39.79752	2.394908	0.061808	8.832838	0.382737	9.809208	10.06085	2.336156

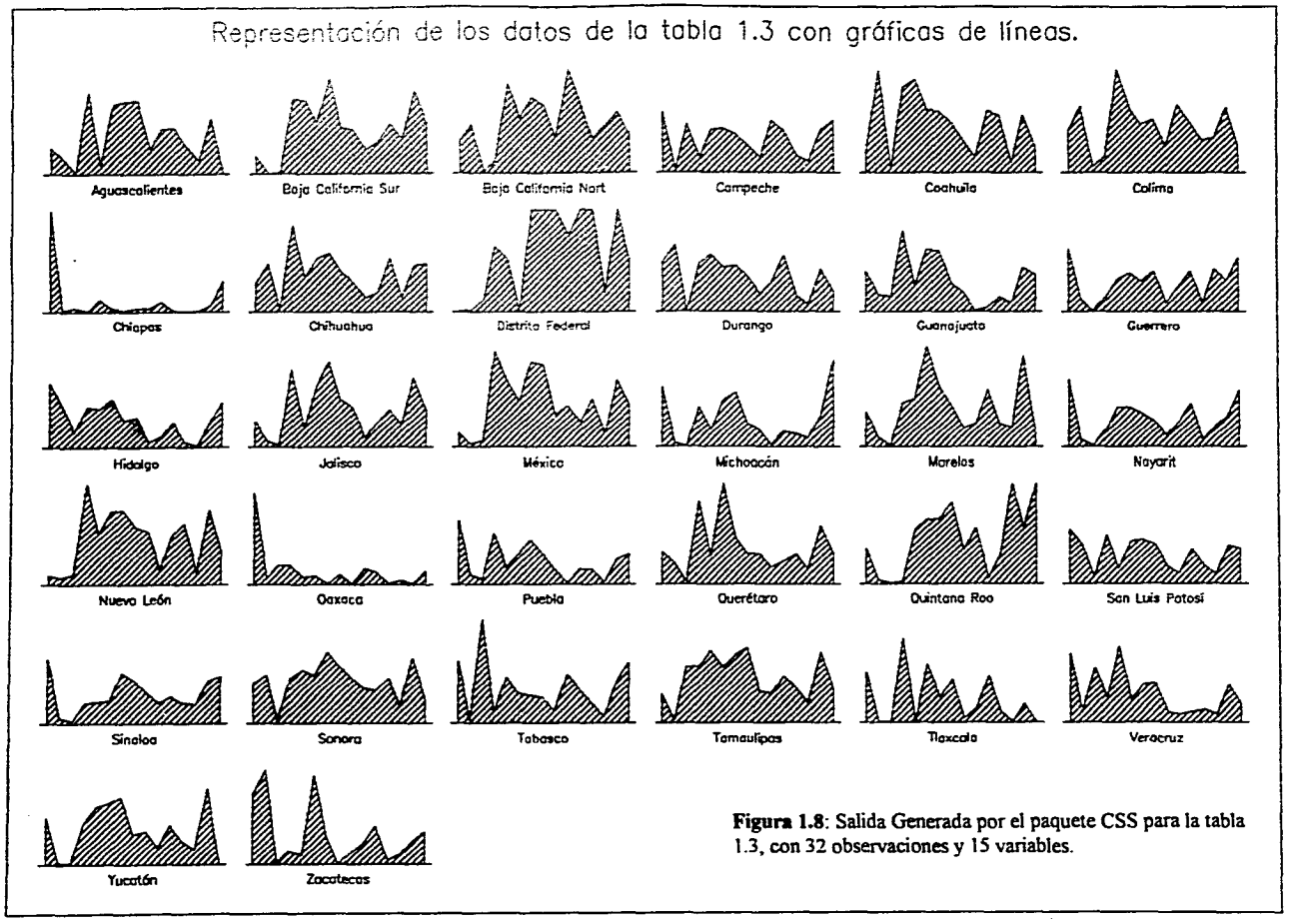


Figura 1.8: Salida Generada por el paquete CSS para la tabla 1.3, con 32 observaciones y 15 variables.

En muchos estados (sobre todo en Campeche, Distrito Federal, Guerrero, Michoacán, Nayarit, Quintana Roo y Tabasco) se presenta alto porcentaje en la variable 15, que corresponde a **sectores no especificados**. En Quintana Roo hay que notar que presenta los índices más altos en cuanto a PEA y población dedicada a sectores no especificados.

1.6.4.-CARITAS DE CHERNOFF

Las variables descritas en la tabla 1.4 son:

VARIABLE 1: Porcentaje de la población ocupada que es obrero.
VARIABLE 2: Porcentaje de la población ocupada que es jornalero o peón.
VARIABLE 3: Porcentaje de la población ocupada que es trabajador por su cuenta.
VARIABLE 4: Porcentaje de la población ocupada que es patrón o empresario.
VARIABLE 5: Porcentaje de la población ocupada que es un trabajador familiar si remuneración.
VARIABLE 6: Porcentaje de la población ocupada que desarrolla una actividad no especificada.

Los datos de la tabla 1.4 correspondientes a variables que indican la actividad desempeñada por la población de cada estado se analizaron mediante las "caritas de Chernoff". Esta técnica, descrita en la sección 1.5, se ejemplificó mediante el uso del paquete CSS, el cual asocia cada variable a un rasgo del rostro de la siguiente manera:

Variable	Asociación
1) Empleado	Ancho de la cara
2) Jornalero	Altura de la cara
3) Trabaja por su cuenta	Excentricidad general de la cara
4) Patrón	Excentricidad mitad superior
5) Trabajo familiar no remunerado	Excentricidad mitad inferior
6) No especificado	Longitud de la nariz

En base a estas asociaciones y si se hace referencia a la figura 1.9 se puede notar lo siguiente:

La similitud entre Chiapas y Oaxaca salta a la vista, ya que ambos presentan una nariz de bastante longitud, lo que indica que , al igual que Michoacán (caso 16), Quintana Roo (caso 23), Zacatecas (caso 32) y Campeche (caso 4) tienen un alto porcentaje de población dedicada a actividades no especificadas.

Por otra parte, la altura o "longitud vertical" de los rostros correspondientes a dichos estados, junto con la de Guerrero son mayores a la de los demás estados, lo que muestra que en ellos hay un alto porcentaje de la población que son jornaleros.

En otros aspectos, el D.F (caso 9), Nuevo León (caso 19), B.C.N. (caso 2), Aguascalientes (caso 1), Coahuila (caso 5), Quintana Roo (caso 23), Colima (caso 6), Sonora (caso 26) y Tamaulipas (caso 28) se encuentran representados

TABLA 1.4
VARIABLES SEGUN EMPLEO DESEMPEÑADO.
PORCENTAJES RESPECTO A LA POBLACION OCUPADA.

	ENTIDAD FEDERATIVA	VARIABLE 1 EMPLEADO	VARIABLE 2 JORNALERO O PEÓN	VARIABLE 3 TRAB. PART.	VARIABLE 4 PATRÓN	VARIABLE 5 TRAB. FAM. SIN REM.	VARIABLE 6 NO ESPECIFICO
	ESTADOS UNIDOS MEXICANOS	57.36748	10.72612	23.35511	2.286025	2.510014	3.755234
1	AGUASCALIENTES	66.98373	10.71551	15.40225	2.683116	1.698019	2.517363
2	BAJA CALIFORNIA	70.54879	7.139181	14.93834	3.975447	0.531238	2.866990
3	BAJA CALIFORNIA SUR	63.22217	11.05164	18.11644	3.574243	1.217364	2.818134
4	CAMPECHE	48.41215	11.34661	30.55812	2.216917	2.536287	4.929892
5	COAHUILA DE ZARAGOZA	69.84910	8.873098	14.88437	2.665290	1.025308	2.702822
6	COLIMA	59.56216	14.01096	18.72949	2.641712	1.595067	3.460599
7	CHIAPAS	25.79683	14.10685	47.31390	1.382178	5.651758	5.748461
8	CHIHUAHUA	65.14163	7.341223	18.95071	2.754494	2.557495	3.254430
9	DISTRITO FEDERAL	77.23802	1.331007	16.07237	2.895756	0.482285	1.980548
10	DURANGO	54.01252	12.72190	22.95673	1.954646	4.470520	3.883665
11	GUANAJUATO	55.86802	13.54071	20.77628	2.232662	3.240758	4.341558
12	GUERRERO	39.91810	11.65221	36.77354	1.555197	4.088074	6.012864
13	HIDALGO	43.15842	22.60624	24.94491	1.326333	2.894702	5.069377
14	JALISCO	62.74245	7.940113	20.89580	2.827320	2.131145	3.463168
15	MEXICO	68.85555	5.689491	19.02868	1.867474	1.341709	3.217083
16	MICHOACAN	40.14248	16.46187	30.00942	2.157818	4.243092	6.985299
17	MORELOS	54.08044	15.88054	22.97700	2.448350	1.663236	2.950421
18	NAYARIT	41.65922	19.21201	27.07210	2.675965	3.714163	5.666523
19	NUEVO LEON	76.32103	3.118809	14.71665	2.498256	0.864019	2.481219
20	OAXACA	27.83515	12.71209	47.08188	1.187583	6.200409	4.982865
21	PUEBLA	42.66726	16.25328	30.63304	1.750781	4.205508	4.490111
22	QUERETARO	58.55900	14.56120	17.82978	2.168557	2.930510	3.950947
23	QUINTANA ROO	60.13971	5.360009	24.17488	2.807770	1.797291	5.720326
24	SAN LUIS POTOSI	48.85088	15.49348	25.53287	1.830001	3.851301	4.441453
25	SINALOA	48.73922	21.83278	20.99968	2.358735	1.702513	4.367042
26	SONORA	60.19904	15.55675	17.45118	3.031547	0.755175	3.006298
27	TABASCO	49.81089	17.32565	23.09764	2.203164	2.883329	4.679310
28	TAMAULIPAS	85.89482	8.856475	18.34519	2.590606	1.412606	2.900299
29	TLAXCALA	50.58364	12.76747	28.78877	1.116429	2.991216	3.772462
30	VERACRUZ	44.14512	18.40500	27.34148	2.135088	4.115883	3.857406
31	YUCATAN	57.13131	12.00799	23.91656	2.235495	1.641638	3.066993
32	ZACATECAS	40.78238	14.02848	30.76703	1.907912	6.803347	5.710831

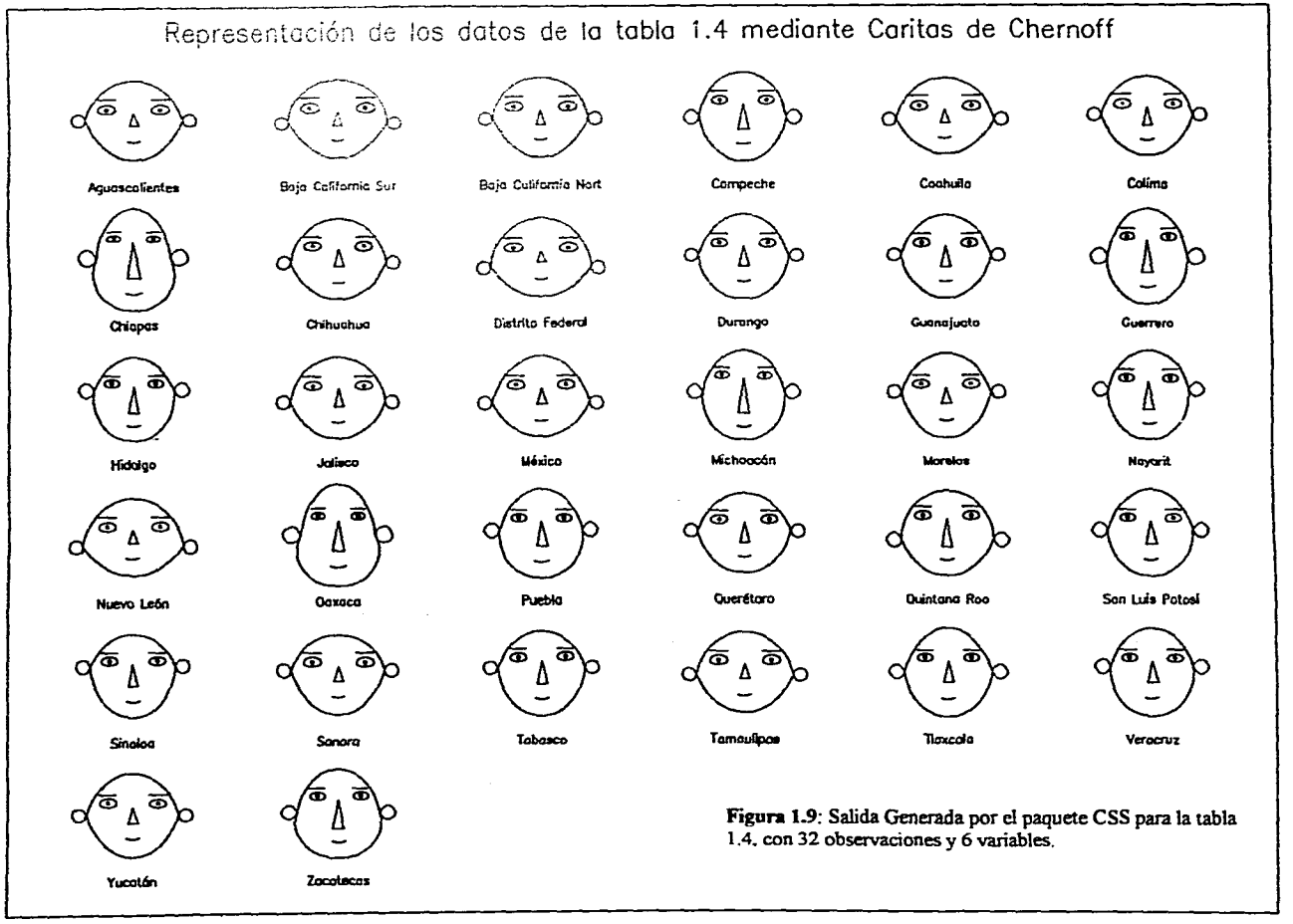


Figura 1.9: Salida Generada por el paquete CSS para la tabla 1.4, con 32 observaciones y 6 variables.

con un ancho de la cara mayor a los demás, lo que revela el hecho de que el porcentaje de población que trabaja como empleado es mayor en dichos estados.

Es notable el hecho de que el D.F. y Nuevo León presentan la menor excentricidad en la parte inferior de la cara, lo que revela que en estos 2 casos hay menor porcentaje de población dedicada a actividades familiares no remuneradas.

La excentricidad en la parte superior de la cara es muy similar en la mayoría de los estados (exceptuando Oaxaca, Chiapas y Campeche donde es menor) indicando que en la mayoría del país el porcentaje de la población que son patrones es muy similar.

1.6.5.-DRAFTSMAN PLOT.

Para ejemplificar estas técnicas se eligieron los datos de la tabla 1.1.

En la figura 1.10 se observa que, en general, las variables tienen un comportamiento aleatorio entre sí, lo cual nos haría pensar que no se presenta una dependencia muy marcada en este conjunto de variables.

Las variables 1 (% de PEA), 2 (% de PEA desocupados) y 3 (Otros inactivos de la PEI) no muestran ninguna relación con respecto a las demás variables. Sin embargo, la variable 4 presenta lo que parece ser una relación lineal inversa con las variables 5 (% que gana de 1 a 3 s.m.), 6 (% que gana más de 3 s.m.), 9 (% de alfabetas) y 8 (% que dispone de electricidad), aunque con esta última la

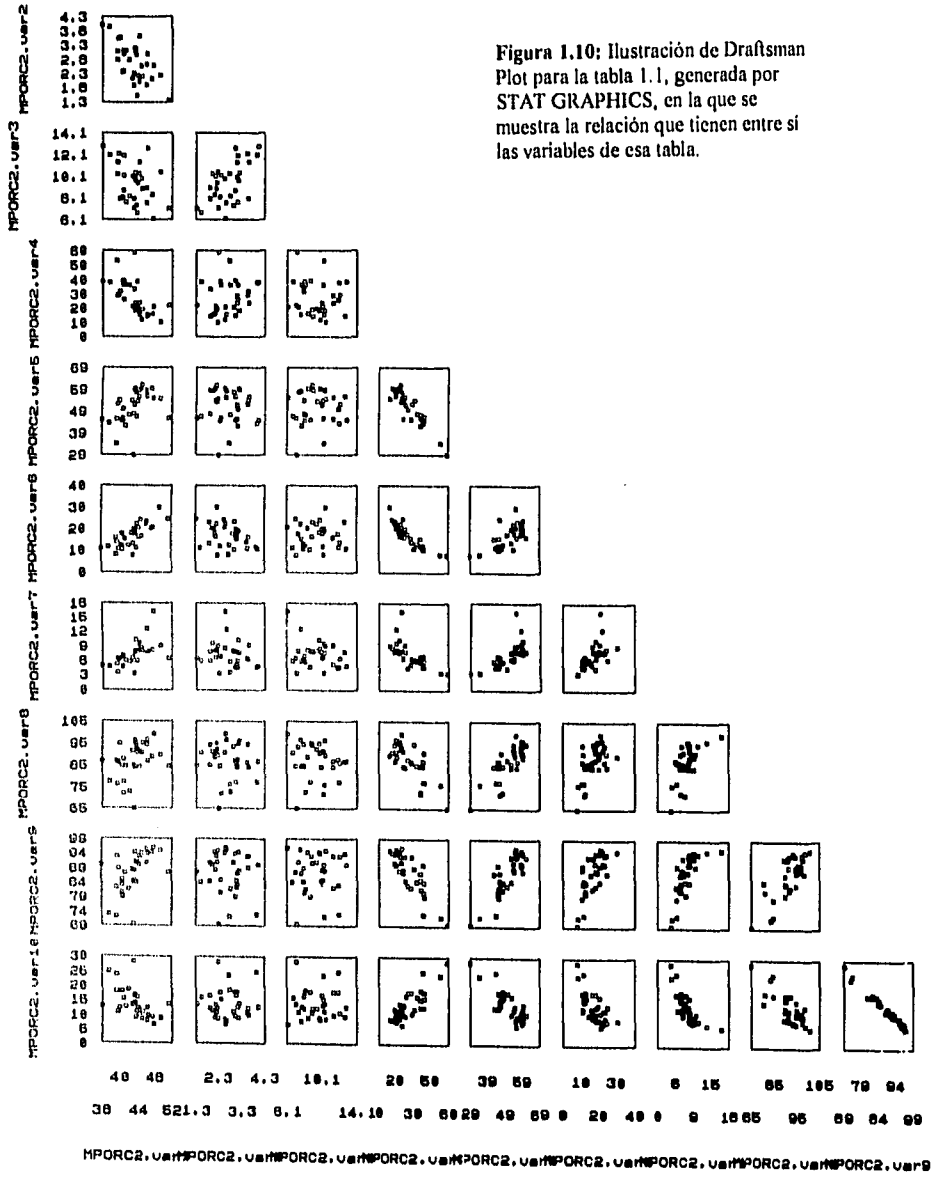


Figura 1.10: Ilustración de Draftsman Plot para la tabla 1.1, generada por STAT GRAPHICS, en la que se muestra la relación que tienen entre sí las variables de esa tabla.

relación no es muy marcada. Con la variable 7 (% de población universitaria) presenta una relación cuadrática y con la 10 (% sin primaria)una relación directa.

Por su parte, la variable 5 muestra relaciones directas con las variables 6, 8 y 9 y una relación inversa con la variable 10, es decir, lo opuesto a la variable 4. Además, respecto a la variable 7 también muestra un comportamiento cuadrático, pero inverso al que tiene la variable 4 con la variable 7.

La variable 6 no muestra ninguna tendencia clara con respecto a las demás variables. Pero la variable 7 parece tener un comportamiento lineal directo con respecto a la variable 8, y una relación cuadrática con respecto a las variables 9 y 10.

Finalmente, se muestra una relación lineal inversa bastante clara entre las variables 9 y 10, que indica que a mayor número de alfabetas menor número de personas que no tienen estudios de primaria. Esta es la relación es lógica y la gráfica lo refleja claramente.

1.6.6. STEAM AND LEAF.

Un ejemplo de este tipo de representaciones se da en la figura 1.11, la cual se aplicó a los datos de la tabla 1.1, específicamente a la columna correspondiente a la variable 1, que se refiere al porcentaje de población económicamente activa respecto a la población mayor de 12 años. Esto se hizo utilizando el paquete CSS en su sección de gráficos (ver apéndice 1 para detallar

su utilización). Aquí se incluyó al renglón correspondiente a "Estados Unidos Mexicanos", pues su inserción no afectaba en mayor medida los resultados.

Como los datos son porcentajes (decimales con 2 enteros y 5 decimales), el paquete descrito redondeó a un decimal. Es importante notar que al hablar de decimales la pérdida de centésimas no es altamente significativa.

css/3: analytic diagrams	Stem and Leaf Diagram of A one leaf=1 case		
steam.leaf	(leaf unit =.100000, e.g. 6.5=6.500000)	class n	percentiles
36.0	. . . . .	1	3.1
37.5	. . . . .	1	6.3
39.124	. . . . .	3	15.6
39.7	. . . . .	1	18.8
40.34	. . . . .	2	25.0
40.55	. . . . .	2	31.3
41.0	. . . . .	1	34.4
41.8	. . . . .	1	37.5
42.689	. . . . .	3	46.9
43.0034	. . . . .	4	59.4
43.68	. . . . .	2	65.6
44.11	. . . . .	2	71.9
44.68	. . . . .	2	78.1
45.99	. . . . .	2	84.4
46.2	. . . . .	1	87.5
47.2	. . . . .	1	90.6
47.6	. . . . .	1	93.8
49.4	. . . . .	1	96.9
51.2	. . . . .	1	100.0
minimum = 36.0007600000			
maximum = 51.2115300000			
Total N		32	

Figura 1.11: Salida generada por el paquete CSS para obtener la representación de "Steam and Leaf". Se realizó dentro de la opción de "CSS-Graphics". La variable de estudio fué el porcentaje de población económicamente activa.

La salida generada por el paquete mencionado presenta 2 partes importantes: en la primera columna se muestra la parte de "steam and leaf". Antes del punto se muestran los enteros correspondientes a los porcentajes observados para cada individuo. Después del punto se muestra la parte decimal (es importante tener en cuenta que se redondeó a un decimal). En la tercera columna se muestra el número de clases para cada bloque presentado en la primera columna, es decir, nos da el número de individuos que tienen alguno de los porcentajes mostrados. En la última columna se da el porcentaje acumulado de observaciones

De dicho cuadro se puede observar lo siguiente: se muestra de manera obvia los valores mínimo y máximo correspondientes a Zacatecas y Quintana Roo respectivamente. Además, hay 5 estados que están alrededor del 43%: Coahuila (43.3), Jalisco (43.8), Estado de México (43.4), Querétaro (43.0) y Yucatán (43.6), además de la República Mexicana con un 43.0. Si se nota que hay 2 clases de 43% es que la primera corresponde a aquellas cifras cuyos decimales están entre 0 y 4, la segunda contiene a los que presentan decimales entre 5 y 9.

De éste cuadro se puede concluir que el porcentaje de la PEA en México es de alrededor del 43% ya que la mayoría de los estados presentan porcentajes muy cercanos a esta medida, exceptuando estados como Zacatecas y Guerrero y a Quintana Roo y Baja California Norte, que respectivamente presentan valores más bajos y más altos de lo común.

1.6.7. BOX PLOT

Para ejemplificar la técnica de box plot se utilizaron los datos de la tabla 1.1 y específicamente los de la variable 7 que corresponde al porcentaje de la población universitaria respecto a la población mayor de 18 años. La gráfica se representa en la figura 1.12.

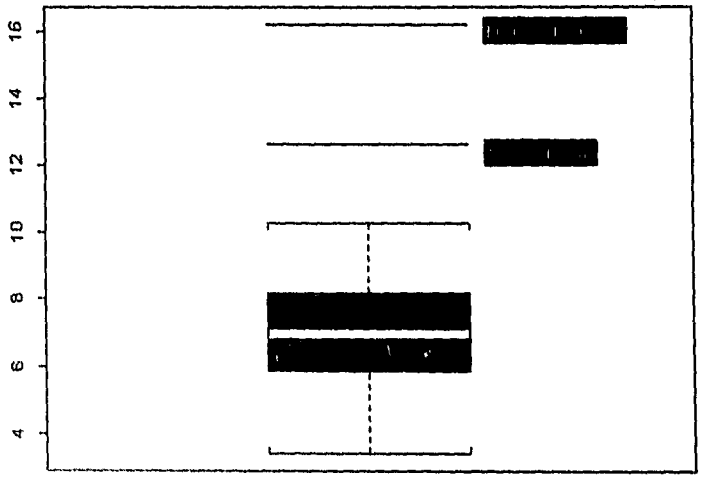


Figura 1.12: Varible 7 de la tabla 1.1. (Porcentaje de población universitaria respecto a la población mayor de 18 años). Salida generada por SPLUS.

La salida mostrada, que fue generada con el paquete SPLUS para Windows muestra claramente al Distrito Federal y a Nuevo León como observaciones discrepantes, en el sentido de que son los que poseen, respectivamente, el más alto porcentaje de población universitaria. Por otro

lado, debido a la escala manejada por el paquete no se puede siquiera apreciar a Chiapas y Oaxaca, los cules se ubican en la parte inferior de la gráfica, ambos con un valor para esta variable menor al 4%.

De esta gráfica, salta a la vista que, nuevamente Chiapas y Oaxaca muestran valores totalmente opuestos a los del D.F. y Nuevo León. Y el hecho de que estos 2 últimos tengan el más alto porcentaje de univeristarios en su población explica algunas de las conclusiones obtenidas a partir de las gráficas anteriores, que muestran a Chiapas y Oaxaca como los estados con menores recursos económicos y al D.F. y Nuevo León como los que tienen mejores perspectivas de desarrollo.

1.7.-CONCLUSIONES

De los datos y gráficas presentados anteriormente se puede concluir lo siguiente:

Con un primer vistazo se puede tomar como patrón para describir a cualquier estado del país como aquellos "que se parecen a Chiapas y Oaxaca" o aquellos "que se parecen al D.F. y Nuevo León". Entendiéndose por esto que presentan características semejantes a los estados tomados como patrón de comparación. Así, podríamos decir que Zacatecas "se parece" al primer grupo y Jalisco al segundo.

Obviamente, no todos los estados tienen que encuadrar en uno u otro grupo, simplemente se señala lo anterior como aspecto descriptivo, y haciendo

alusión a las variables que nos interesen. Así, si decimos que un estado se parece más a Chiapas que a Nuevo León se tiene un indicativo respecto a las variables que se tomaron en las tablas anteriores y que engloban aspectos educativos, de ingresos y laborales.

En Chiapas, Oaxaca y otros estados se presenta el siguiente marco:

1) Altos porcentajes en población analfabeta y sin primaria, lo que repercute en que haya pocos estudiantes a nivel universitario.

2) La mayoría de la población percibe ingresos ínfimos, que dificultan la subsistencia.

3) Alto porcentaje de gente dedicada a la agricultura y que, además, trabaja por su cuenta.

De lo anterior, se observa que lo señalado en el punto 1 se explica por lo presentado en los puntos 2 y 3 y la carencia de servicios energéticos.

El D.F. y Nuevo León presentan aspectos totalmente opuestos.

Finalmente, cabe mencionar que fue más fácil llegar a las conclusiones anteriores a través de las representaciones gráficas que a través de las tablas de datos. Recurriendo a dichas gráficas se consiguió:

CAPITULO 2

COMPONENTES PRINCIPALES

II.- COMPONENTES PRINCIPALES

2.1 OBJETIVOS

El análisis de componentes principales es una técnica multivariada, originalmente desarrollada por Pearson (1901) y posteriormente estudiada por Hotelling (1933); pero debido a la cantidad de cálculos que involucra, en cuanto a manejo de matrices, empezó a tomar auge con el advenimiento de las computadoras.

Este método puede ser visto de 3 maneras:

1) Como un método para transformar variables correlacionadas a otras no correlacionadas.

2) Como una técnica para encontrar combinaciones lineales con variabilidad relativamente grande o relativamente pequeña con respecto a los datos originales.

3) Es una herramienta para reducir la dimensión de los datos.

Jollife[11] sugiere una definición que engloba los tres aspectos anteriores: "Técnica cuyo objetivo es disminuir la dimensión de datos en los cuales hay variables correlacionadas, manteniendo tanto como sea posible la varianza presente en los datos originales".

En muchas ocasiones resulta útil tener un conjunto de datos cuya dimensión es pequeña, respecto a la dimensión de los datos originales, y que logre explicar en un determinado porcentaje lo mismo que explicarían dichos datos. De esta manera, al reducir la dimensión a $\mathbf{R}^2$ o $\mathbf{R}^3$ se puede recurrir a la graficación.

Para ello, se realiza una transformación hacia un conjunto de nuevas variables, obtenidas a partir de las originales, las cuales presentan ciertas características que se describen más adelante. Dichas variables (llamadas **componentes**) se ordenan de tal manera que la primera representa o explica la mayor proporción de la varianza presente en los datos originales.

El cálculo de las componentes principales se reduce, así, a un problema de valores propios y eigenvectores para una matriz semidefinida positiva. Y, aunque el realizar dichos cálculos parece laborioso, es muy rápido si se cuenta con una paquetería adecuada para tal fin.

Para el propósito de analizar los datos del censo, éste método es particularmente útil, ya que al manejar gran cantidad de variables puede presentarse el fenómeno de colinealidad, que consiste en que dos o más variables son lincalmente dependientes y que podamos tener una sola variable en lugar de 2 o más cuando éstas representan lo mismo.

2.2 DESARROLLO

Geométricamente, lo que se pretende con las componentes principales es obtener las proyecciones sobre ciertos ejes, de tal manera que la primera

proyección maximice la varianza, la segunda sea ortogonal a la primera y también maximice la varianza dada la condición de ortogonalidad.

Ya se había mencionado que se pretende transformar el conjunto X_1, $X_2,...,X_p$ a $Y_1, Y_2,...,Y_p$ el cual tiene las siguientes propiedades:

a) Cada Y_i es una combinación lineal de las X_i's :

$$Y_i = \alpha_1^t \cdot X = \alpha_{i1} \cdot X_1 + \alpha_{i2} \cdot X_2 + \cdots + \alpha_{ip} \cdot X_p = \Sigma_i \, \alpha_{ij} \cdot X_j$$

b) Las Y_i's no están correlacionadas entre sí y están ordenadas de acuerdo a su varianza (Y_1 tiene la mayor varianza, le sigue Y_2, etc.)

La idea es que, una vez obtenidas las componentes principales (Y_i's), elegir las primeras de tal manera que expliquen cierto porcentaje de la variabilidad presente en los datos originales. El número de componentes se elige de acuerdo al porcentaje de variación explicada. Así, se obtiene una representación del mismo conjunto de datos que puede ser más conveniente.

De esta manera, el problema se reduce a encontrar las α_{it}. Si llamamos a Σ la matriz de varianzas y covarianzas, la cual se estima con :

$$S = (1/n-1) \, \Sigma(\bar{xi}-x)(\bar{xi}-x)^t$$

La cual puede ser vista como $S = XHX^t$ para alguna matriz H de p x n. Al considerar que $Yi = \alpha_i^t \cdot X$ y que se desea maximizar:

$$Var(Y_i) = Var(\alpha_i^t \cdot \mathbf{X}) = \alpha_i^t \cdot \Sigma \cdot \alpha_i$$

Entonces el problema se reduce a encontrar α_i^t tal que maximice : $\alpha_i^t \cdot \Sigma \cdot \alpha_i = Var(\alpha_i^t \cdot \mathbf{X})$. Dicha igualdad se deduce de la definición de:

$$\alpha_1^t \cdot \mathbf{X} = \alpha_{i1} \cdot X_1 + \alpha_{i2} \cdot X_2 + \cdots + \alpha_{ip} \cdot X_p$$

Si se agrega la restricción de que α_i^t sea normalizado, es decir, $\alpha_i^t \cdot \alpha_i = 1$, se elimina la posibilidad de encontrar dicha maximización para diferentes α_i^t.

De esta manera, el problema será: maximizar $\alpha_i^t \cdot \Sigma \cdot \alpha_i$ sujeto a $\alpha_i^t \cdot \alpha_i = 1$ para i=1,..,p. Mediante un desarrollo con multiplicadores de Lagrange (Ver Jollife[11]) se llega a que la solución al problema así planteado se logra a partir de la solución para cada una de las siguientes ecuaciones:

$$(\Sigma - \lambda_i \mathbf{I})\, \alpha_i \qquad\qquad i=1,..,p$$

Lo anterior equivale a encontrar las raíces del determinante:

$$| \Sigma - \lambda_i \mathbf{I} | = 0$$

Dicha ecuación tiene p raíces, que corresponden a los p valores propios de λ_i. A cada valor propio se asociará el respectivo eigenvector α_i. Y al normalizar el vector: $\alpha=(\alpha_1,...,\alpha_p)$, éste corresponderá a las componentes principales.

Como Σ está semidefinida positiva, no habrá raíces negativas (ver Marriott [11]). Una vez que se ordenen los valores propios de mayor a menor y se obtengan los eigenvectores correspondientes se obtendrán las componentes principales como:

$$Y_1=\alpha_1{}^tX$$
$$Y_2=\alpha_2{}^tX$$
$$\vdots$$
$$Y_p=\alpha_p{}^tX$$

2.3 PROPIEDADES

Una propiedad interesante que presenta este método es que si tenemos la matriz estimada de varianzas y covarianzas (estimada por **S**, descrita anteriormente), al obtener los valores propios λ_i se tiene que $\Sigma(\lambda_i) = \text{traza}(\Sigma)$, es decir, la suma de las varianzas de las componentes principales es la misma que la de las variables originales. En este caso, como se está trabajando con **S** y no con Σ, se obtendrá que:

$$\Sigma(\hat{\lambda}_i)=\text{traza}(S).$$

Donde $\hat{\lambda}$ es un estimador óptimo de λ. Para un desarrollo detallado de la propiedad anterior, y de las siguientes ver Jollife [11].

De lo anterior se sigue que cada componente aporta la proporción de $\lambda_i/\Sigma(\lambda_i)$ al total de la varianza de los datos originales. Pero para que se presenten dichas propiedades es necesario que las raíces o valores propios sean todas diferentes entre sí y distintas de cero.

La existencia de raíces iguales a cero implica que las variables originales son linealmente dependientes, en este caso puede demostrarse (ver Jollife [11]) que la componente principal correspondiente a una raíz igual a cero es cero.

Para dar la interpretación geométrica basta observar que una ecuación de la forma $x^tV^{-1}x = k$ (Siendo V cualquier matriz definida positiva y k un real) representa un elipsoide en p dimensiones, por lo que los cálculos descritos anteriormente para encontrar las componentes principales son los necesarios para encontrar los ejes de la hiperelipsoide representada, en orden de magnitud. Así, la presencia de raíces iguales a cero indicará que la elipsoide es degenerada y puede representarse en menos de p dimensiones.

Con ésta técnica se tiene un criterio para "ignorar" los ejes cuya longitud es relativamente intrascendente. Aunque para eliminar algún eje hay que tener en cuenta el porcentaje de explicación de la variabilidad que presenta cada uno (para ello nos basamos en el valor del valor propio correspondiente).

Una aplicación interesante de Componentes Principales se da en el modelo de Análisis de Regresión lineal. Este modelo consiste en explicar $\mathbf{Y}_{nx1}$ a partir de $\mathbf{X}_{nxp}$. Si de la matriz **X** se toman los primeros r componente ($r<p$) y se hace la regresión sobre ellos en lugar de hacerla sobre **X**, se obtendrán como ventajas:

1) La dimensión se reduce, lo que facilita y agiliza los cálculos numéricos.

2) Las nuevas variables son no correlacionadas y linealmente independientes lo que evita problemas de multicolinealidad.

Obviamente, con las ventajas 1 y 2 es más fácil aplicar el modelo de regresión lineal a un conjunto determinado de datos que tenga una cantidad considerable de elementos, aunque tiene que considerarse que la interpretación del modelo ya no es tan inmediata..

2.4 APLICACIÓN A LOS DATOS DEL CENSO.

Se aplicó esta técnica a los datos de la tabla 1.1 Al hacerlo se observa que los 2 primeros componentes principales se obtiene un 80% de explicación de los datos originales (ver tabla 2.1) y con los 3 primeros se obtiene un 88%, cifra suficientemente alta para hacer inferencia en base a dichos componentes.

Número de Componente	Valor propio respectivo (porcentaje de varianza asociado)	Porcentaje de varianza acumulado.
1	61.0964	61.0964
2	19.2583	80.3547
3	8.3100	88.6647
4	4.2479	92.9127
5	2.7609	95.6737
6	1.7932	97.4669
7	1.6746	99.1415
8	0.7226	99.8642
9	0.0922	99.9565
10	0.0434	100.0000

Tabla 2.1: Datos referentes a los componentes principales de la tabla 1.1.generados por el paquete estadistico STAT GRAPHICS.

Además, los valores propios de la matriz de covarianzas no son muy grandes, lo que parecería indicar que no existe una gran dispersión entre los datos.

Al analizar la tabla 2.2, en la cual se muestran los componentes principales se observa que de la primera componente la mitad tiene signo positivo y la mitad signo negativo. En magnitud (valor absoluto) no se muestra un dominio de alguno de los dos; exceptuando las observaciones correspondientes a los estados de Chiapas y Oaxaca, los cuales, además de tener signo negativo en esta componente, son de mayor magnitud.

De los de signo positivo, poseen mayor magnitud los correspondientes al D.F., Baja California N. y Nuevo León. Así, los estados de Chiapas y Oaxaca por un lado y el D.F., Baja California N. y Nuevo León por otro, representan valores extremos respecto a la primera componente.

Respecto a la segunda componente se tiene que la mayoría son de signo positivo, y en este caso las observaciones correspondientes a Quintana Roo y a Zacatecas son los valores extremos, uno con signo positivo y otro con signo negativo. Es importante destacar, que del capítulo uno se notó que Quintana Roo tenía poco desempleo (con un alto porcentaje de PEA ocupada) y que en Zacatecas se presentaba un alto índice de desempleo. Además, de la tabla 2.3, que muestra los vectores propios, resalta el hecho que el segundo valor propio contraste empleo contra desempleo, ya que los valores de PEA, de desocupados y de PEI son los más altos, el primero con signo positivo y los otros 2 con signo negativo.

No.	1	2	3	4	5	6	7	8	9	10
1	1.9723	0.2558	0.3384	0.9225	0.2272	0.3153	0.3868	0.4245	0.07663	-0.02451
2	3.3485	0.4716	-1.8450	-0.6726	0.0762	-0.4317	-0.3937	-0.0530	1.399E-3	0.0773
3	2.4098	0.8523	-0.2459	0.1485	0.5907	-0.0650	0.6148	0.1681	-0.0168	-0.0258
4	-1.0525	1.3976	0.2410	0.6893	0.3178	0.2449	0.0337	0.0241	0.0614	-0.09731
5	2.1748	-1.0607	0.7908	-0.2117	-0.0960	0.1376	0.0743	0.1204	0.0320	-0.01394
6	2.2896	0.8040	-0.6712	0.8125	-0.2966	0.0773	0.0919	-0.0777	0.0595	0.07257
7	-5.9783	2.7782	-0.2761	-0.8958	-0.1341	-0.1467	0.0673	0.2761	0.2214	-5.969E-4
8	1.9259	-1.3235	-1.6223	-0.6126	0.3548	-0.0806	-0.0737	0.4856	-0.0634	0.0180
9	3.8348	1.2390	2.1433	-1.5044	-0.6511	0.4138	-0.7343	0.1235	-0.0119	-0.0247
10	-0.3129	-2.3381	0.1480	-0.0605	0.9049	0.0437	-0.1451	0.0977	0.1133	0.03593
11	-1.1321	-1.3678	-1.0603	0.6597	-0.5314	-0.1818	-0.1093	-3.798E-3	0.0523	-5.878E-3
12	-4.0951	-1.8010	-0.5689	-0.3839	-1.2944	0.1651	0.2634	0.0350	-0.0462	-0.0551
13	-2.7851	0.1885	0.3488	-0.1215	0.1456	-0.1243	0.5299	-0.0712	-0.2664	0.0321
14	1.5131	0.3046	-0.4272	0.4313	-0.0990	0.0200	-0.3650	-0.3412	0.1575	0.0485
15	1.4255	0.8345	1.0596	0.2980	-0.2588	-0.5564	0.5851	-0.1628	6.338E-4	-0.0108
16	-1.64709	-1.1586	-0.4937	0.4428	-0.4639	-0.0228	-0.4978	-0.2505	-0.0226	-0.2131
17	1.2010	-1.1382	-8.111E-3	0.5167	-0.9343	-0.1474	0.2394	0.2870	0.0201	0.0535
18	0.6691	0.0473	-0.4352	0.9611	0.0771	0.3182	-0.1821	-0.0996	-0.0343	-0.0460
19	3.3154	0.4330	1.0414	-0.6228	-0.3779	0.1131	-0.1755	-0.1586	0.0208	-8.129E-3
20	-5.0082	0.3651	-0.2251	0.8161	-0.1628	0.1612	-0.6194	0.3270	-0.1747	0.1330
21	-1.9871	0.8915	0.7368	0.2644	-0.2415	0.4183	-0.2704	-0.2557	-0.0276	0.0566
22	-0.1134	-0.3020	-0.5196	-0.5911	-0.8058	-0.0216	0.2989	-0.1936	0.1004	0.0628
23	1.2429	3.1573	-1.4586	-0.2853	0.1965	-0.8270	-0.2105	0.1770	-0.0778	-0.1181
24	-1.9395	0.2110	-0.2771	-0.5970	1.0618	0.9631	0.1852	-0.4358	-0.0306	-0.0182
25	1.9475	0.1284	-0.5106	0.3949	-0.1122	0.9691	-0.0237	-0.1197	-0.1217	-0.0310
26	2.6308	-0.3718	-0.5727	0.8278	0.1468	-0.1285	0.1932	-0.4672	-0.0723	0.0880
27	-1.1963	0.1923	0.8952	-0.3353	0.4587	-1.0290	-0.3158	-0.5825	-0.0691	-6.113E-3
28	0.9451	-1.6177	8.4819E-3	-1.1672	0.4489	0.1728	0.3858	0.5137	-0.0393	-0.0489
29	-0.3963	-1.0559	1.9437	0.6314	9.714E-3	-0.7292	0.3694	0.1373	-0.0538	-3.418E-3
30	-2.2820	0.8910	0.0829	-0.9958	0.2305	0.1870	0.9549	-0.2938	0.1120	0.0171
31	-0.9660	2.2474	0.9243	1.2159	0.3427	0.1228	-0.3711	0.3597	-9.682E-3	0.0617
32	-2.02616	-3.1197	0.5067	0.8885	0.8698	-0.2624	-0.7865	0.0145	0.1089	3.548E-4

Tabla 2.2: Componentes principales de la tabla 1.1, generados por STAT GRAPHICS.

En la componente número 3 el D.F. difiere totalmente de las demás observaciones al tener el valor más alto (notar que el D.F. presenta una desproporcionada densidad de población). Tamaulipas y Morelia son las más pequeñas, una con signo positivo y la otra con signo negativo.

De la tabla 2.3, que muestra los 3 primeros vectores propios obtenidos al realizar el análisis de componentes principales, se puede observar que las variables que mayor peso tienen en la primera componente son aquellas que tienen signo negativo y se refieren principalmente a aspectos de productividad (variable1=Porcentaje de PEA), ingresos (Variables 5 y 6), nivel de escolaridad (variables 7 y 9) y disposición de energía eléctrica. En esta primera componente casi no tienen peso las variables referentes a población inactiva y población desocupada, las cuales, junto con la variable 4 (gente que gana menos de 1 salario mínimo) y la 10 (porcentaje de la población mayor de 6 años sin primaria) tienen signo positivo. En este sentido, se distinguen dos grupos de variables: las "favorables" y las "desfavorables" para el país, dándole mayor peso a las primeras.

En la segunda componente se le da mayor peso a las variables de población desocupada y población inactiva, ambas con signo negativo, como se mencionó anteriormente. Finalmente, la variable 3 no muestra un comportamiento definido en cuanto a la asignación de pesos de las variables y obtener conclusiones basándose únicamente en esa componente sería poco confiable.

En la figura 2.1 (Ver gráficas al final del capítulo) que corresponde a componente 1 vs. componente 2 se ve a Chiapas, Oaxaca y Guerrero como observaciones discrepantes, muy cargados al lado derecho y al D.F. muy cargado

hacia el lado izquierdo. Lo que confirma que el comportamiento del D.F. es totalmente contrario al de Chiapas y Oaxaca. Otra observación alejada, pero hacia arriba y por el centro es Quintana Roo.

No	Variable	α_1	α_2	α_3
1	% PEA	-0.59947	0.67650	0.27065
2	% Desocupados	0.22092	-1.01267	-0.22627
3	% PEI	0.09065	-0.97326	0.53243
4	% (0,1] S:M:	0.7878	0.16953	-0.23412
5	% (1,3) S:M:	-0.74762	-0.30392	-0.04666
6	% + 3 S:M:	-0.71588	0.08836	0.44448
7	% Pob. Universitaria	-0.68884	0.07875	-0.31042
8	% Con electricidad	-0.68681	-0.18773	-0.30060
9	% Alfabetas	-0.7723	-0.2461	-0.10644
10	% Sin primaria	0.76799	0.2020	0.16827

Tabla 2.3: Vectores propios de la matriz de correlaciones para la obtención del modelo de componentes principales.

En la figura 2.2 que representa componente 1 vs. 3 nuevamente Chiapas y el D.F. son valores extremos, el primero hacia el lado derecho y el segundo hacia el lado izquierdo.

Lo anterior confirma lo observado en el capítulo anterior al concluir que Chiapas y Oaxaca representan un grupo de estados con características poco favorables de acuerdo a las variables que se estudiaron. Muestra de ello es que sus coordenadas sobre el eje de la primera componente son negativas y de gran magnitud. Asimismo, el D.F. aparece como contraparte de estos casos y su coordenada en el mismo eje es positiva y la de mayor magnitud.

Quintana Roo muestra un comportamiento diferente a los demás estados, ya que posee el valor más alto sobre la segunda componente principal y su coordenada en la primera es positiva, aunque no de gran magnitud. En cambio Zacatecas muestra un comportamiento contrario con respecto a Quintana Roo.

2.5 CONCLUSIONES

Se ha visto cómo la técnica de componentes principales ayuda a describir el comportamiento de las observaciones de una manera más clara y sintetizada. El hecho de poder graficar las componentes que explican mayor variabilidad nos permite visualizar, de manera aproximada, lo que realmente esta ocurriendo en un espacio de 10 dimensiones.

De las figuras 2.1 y 2.2 se observa cómo los puntos correspondientes a Chiapas y Oaxaca son vecinos, en tanto que en el extremo opuesto el D.F. es vecino de Nuevo León.

La figura 2.3 nos muestra cierta diferencia respecto a las anteriores ya que aleja a Oaxaca de Chiapas, esto ocurre quizá debido a que las componentes 2 y 3 explican en conjunto poco más del 27% de la varianza de los datos, razón por la cual dicha figura no podría ser considerada muy relevante.

Finalmente, es de notarse que el caso 14, correspondiente a Jalisco no está tan cerca del D.F. y Nuevo León como se podría esperar.

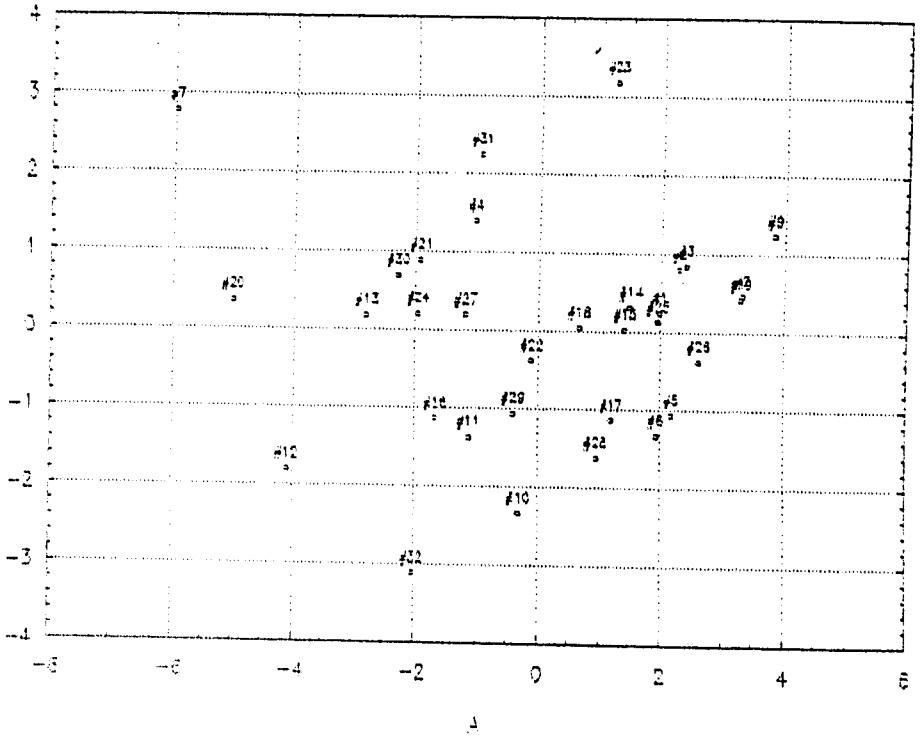


Figura 2.1: Gráfica de las componentes principales 1 (eje X) y 2 (eje Y) de la tabla 1.1, generadas por el paquete CSS.

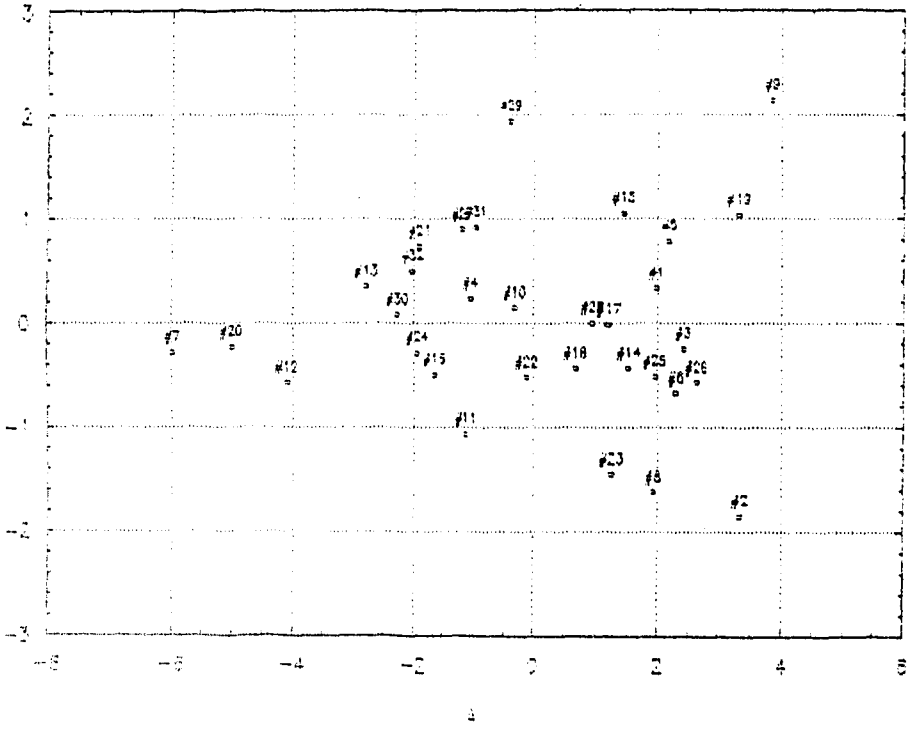


Figura 2.2: Gráfica de las componentes principales 1 (eje X) y 3 (eje Y) de la tabla 1.1, generadas por el paquete CSS.

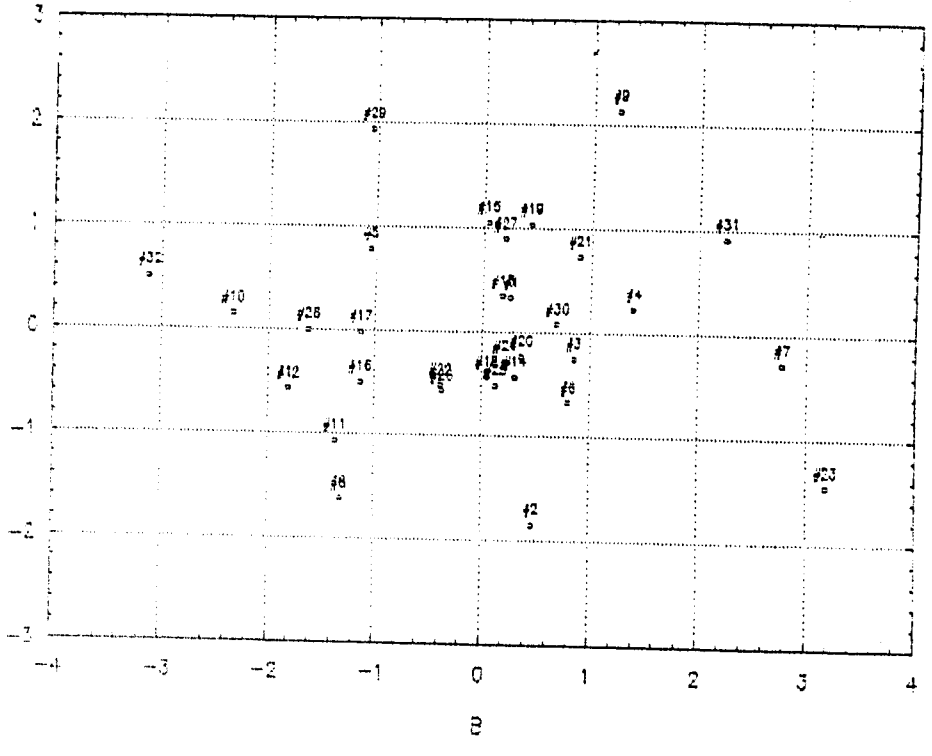


Figura 2.3: Gráfica de las componentes principales 2 (eje X) y 3 (eje Y) de la tabla 1.1, generadas por el paquete CSS.

CAPITULO 3

ESCALAMIENTO MULTIDIMENSIONAL

III.-ESCALAMIENTO MULTIDIMENSIONAL.

3.1 INTRODUCCION

La técnica conocida como escalamiento multidimensional (del inglés Multidimensional Scaling) pretende representar gráficamente (en R^2 o R^3) la situación de un conjunto de n puntos (individuos). En dicha representación se tendrá un "mapa" de las observaciones registradas, y debido al tipo de datos que se manejan (y que se describen adelante) se puede apreciar qué tan "cercanos o lejanos" están los individuos entre sí.

Este método, a diferencia de otros, maneja como datos a una matriz de similitudes o discrepancias. Dicha matriz, de nxn, nos indicará la "distancia" (similitud o diferencia) que hay entre cada par de individuos, y a partir de ella se construirá un modelo gráfico.

La configuración resultante respeta las similitudes o diferencias entre los individuos. A este respecto hay que notar que, dada una matriz de distancias, existe una infinidad de representaciones en el plano, por lo que desde el inicio debe especificarse alguna orientación y localización de los puntos. Aunque si esto no se hace se obtendrá una configuración ya sea reflejada o rotada.

Además de lo anterior, ésta técnica es exploratoria y permite ubicar en el plano (o espacio) a los individuos de una manera tal que se facilite el objetivo de la investigación a realizar. Por ejemplo, si se va a aplicar un análisis de

conglomerados o incluso de componentes principales, siempre es útil tener una idea de cómo se formarán los grupos a estudiar.

La medida de similitud o disimilitud que se menciona anteriormente y sobre la cual se basan todos los cálculos, puede ser obtenida directamente al realizar la medición de alguna cualidad que presenten los individuos en cuestión. Por ejemplo, en el caso de los datos del censo se tienen 32 estados, de los cuales podemos obtener para cada uno su densidad de población; si definimos la "distancia" (o disimilitud) $\mathbf{d}_{ij}$ del estado i al estado j como el valor absoluto de la diferencia entre las densidades de población de los estados i y j, entonces a partir de esas diferencias se puede construir la matriz de disimilitudes.

Sin embargo, para que los cálculos involucrados en una técnica de este tipo sean válidos, las distancias o medidas de disimilitud que se manejen deben cumplir las siguientes características:

1) $d_{ii} = 0$. La distancia de un individuo a sí mismo es cero.

2) $d_{ij} \geq 0$ para todo par (i,j).

3) $d_{ij} \leq d_{ik} + d_{kj}$ para i,j,k.

Esta última propiedad es necesaria para realizar ciertos procedimientos, y se pide que se cumpla porque hay ciertos tipos de medidas, sobre todo las

obtenidas en áreas como la psicología en las cuales no se presenta esta propiedad.

No obstante, no siempre es posible construir la matriz de disimilitudes directamente, como en el caso de la densidad de población,, ya que muchas veces se parte de una matriz de datos con n observaciones a las cuales se han medido p características o variables, como las tablas presentadas en el capítulo 1. Pero aún en esos casos es posible llevar la matriz de datos, denotada por **X** a una matriz de disimilitudes mediante diferentes tipos de métricas. Así, podemos definir

$$d_{ij} = |\Sigma_i (X_{ik} - X_{jk})^r|^{1/r}$$

Donde X_{ik}, X_{jk} son elementos de la matriz de datos observados.

En muchos de los casos se elige r=2, la métrica euclidiana y que cumple con las características necesarias para desarrollar todos los cálculos involucrados.

Para abundar sobre los conceptos de distancia y métrica puede verse Bartlet [1]. En este caso se manejarán ambas como términos comunes.

3.2 DESCRIPCION.

Cuando se tienen 2 puntos en el espacio $\mathbf{R}^p$ y se conocen sus coordenadas, es fácil encontrar la distancia entre ellos de acuerdo a alguna

métrica pre-establecida. Sin embargo, en esta ocasión el problema es el inverso: dadas las distancias entre los puntos hay que obtener sus coordenadas para poder representarlas en el espacio ($\mathbf{R}^2$ o $\mathbf{R}^3$ para permitir la visualización). Así, dada la información sobre las similitudes o disimilitudes (que son lo que se maneja como distancia) entre los puntos hay que encontrar su representación gráfica.

Existen dos tipos de escalamiento: el clásico o métrico y el no métrico

3.2.1 ESCALAMIENTO CLÁSICO

Para llegar a un conjunto de puntos en $\mathbf{R}^k$, a partir de una matriz de distancias o similitudes, que nos permita visualizar la configuración gráfica de los datos se realiza el siguiente proceso:

1) Se forma la matriz A_{nxn} donde $a_{rs} = -(1/2)d_{rs}^2$. Donde d_{rs} es la medida de disimilitud del individuo r al s.

2) A partir de la matriz anterior se forma la matriz **B**, donde

$$b_{rs} = a_{rs} + a_{r.} + a_{.s} + a_{..}$$

donde:

$a_{r.}$ = Promedio del r-ésimo renglón.

$a_{.s}$ = Promedio en la s-ésima columna.

$a_{..}$ = Promedio general.

3) De la matriz anterior se obtienen los vectores propios normalizados y se ordenan de acuerdo a la magnitud de los valores propios. El mapa de los datos estará dado por los k primeros eigenvectores para el espacio $\mathbf{R}^k$. Lo más común es elegir k=2.

Se puede demostrar que el escalamiento clásico cuando se utiliza sobre una matriz de distancias euclidianas es equivalente al análisis de componentes principales de una matriz de covarianzas, en el sentido de que las coordenadas obtenidas con ambos métodos representan lo mismo. (Aunque lo anterior no quiere decir que los valores obtenidos de las coordenadas sean iguales).

3.2.2. ESCALAMIENTO NO MÉTRICO

El objetivo del escalamiento no métrico es el mismo que el del clásico, enfatizando el hecho de que las distancias entre los puntos correspondan a las similitudes observadas. Para ello se define una función objetivo que mide la discrepancia entre dichas similitudes y las distancias ajustadas.

Si δ_{ij} es la discrepancia o disimilitud observada y d_{ij} la distancia ajustada, entonces para disminuir la diferencia entre ambas se utilizará un algoritmo que minimice $SS = \sum_{i,j}(\delta_{ij} - d_{ij})^2$.

Se puede suponer que dij es de la forma $f(\delta_{ij})+\varepsilon_{ij}$, así que se deseará minimizar:

$$SS = \sum_{i,j}(\delta_{ij} - f(\delta_{ij})^2$$

Si representamos a d_{ij} como una transformación monotónica, podemos escribir $d_{ij} = h_{ij} + \varepsilon_{ij}$. Donde ε_{ij} es un error de ajuste y h_{ij} son monotónicas con respecto a δ_{ij}, es decir, si $\delta_{ij} \leq \delta_{ik}$ entonces $h_{ij} \leq h_{ik}$. Para obtener h_{ij} se utiliza un método conocido como **regresión monotónica.**

Al realizar dichos cálculos se tratará de minimizar:

$$SS = \{ (\sum_{i,j}(d_{ij} - h_{ij})^2) / \sum_{i} d_{ij}^2 \}^{1/2}$$

De esta manera, el escalamiento no métrico se compone de 2 partes principales: a) El análisis de regresión monotónica de las distancias sobre las disimilitudes y b) Minimizar SS.

Para ello se han elaborado varios algoritmos (ver Krzanowsky [12]), los cuales operan a grandes rasgos como sigue: comienzan con un conjunto arbitrario de coordenadas, las cuales se eligen de acuerdo a determinados criterios y se calcula SS. A continuación se mueven las coordenadas iniciales de alguna manera determinada por los algoritmos y se vuelve a calcular SS, se observa como va cambiando éste valor, hasta llegar a unas coordenadas que den el mínimo SS. (Este proceso es semejante a cuando se calculan mínimos por el criterio de la primera derivada, es decir, cuando al variar la abscisa cambia el valor de la derivada de negativo a positivo, en ese intervalo hay un mínimo).

Desde luego que al proceder de la forma anterior se corre el riesgo de caer en un mínimo local y no en uno global, lo cual dependerá en mucho de la primera configuración que se tome; aunque algunos métodos proporcionan criterios para seleccionar dicha configuración inicial, esto representa cierta falta de robustez.

En términos generales se considera un buen ajuste cuando $SS \leq 5$.

3.3 APLICACIÓN A LOS DATOS DEL CENSO.

Para propósitos de analizar los datos del censo ésta técnica resulta especialmente atractiva, ya que permite ver gráficamente a los individuos, permitiendo situar a cada estado respecto a los demás.

Cabe aclarar que debido a lo extenso de las tablas que se generaron, éstas se presentan al final del capítulo, junto con las gráficas, para dar mayor facilidad a la lectura.

En este caso se parte de la matriz de datos correspondiente a la tabla 1.1 (raw matrix) y se transformó en una matriz de disimilitudes de 32x32 mediante la distancia euclidiana. La distancia entre dos individuo x,y está dada por:

$$(\Sigma \, |x-y|^2)^{1/2}$$

Donde cada uno de los elementos de la pareja (x,y) son vectores de 1x10. Y a partir de la definición anterior se forma la matriz A_{32x32}.

A esta última matriz (ver tabla 3.1) se le aplicó el escalamiento clásico obteniéndose los valores para las componentes principales finales de la tabla 3.3, la cual genera las coordenadas que se observan en la tabla 3.4, con los cuales se construye la figura 3.1. Haremos enfásis en ésta gráfica más que en el análisis de las componentes principales obtenidas, ya que dicho análisis se explicó en el capítulo 2 para la matriz de datos. En la figura 3.2 se presenta exactamente lo mismo que en la figura 3.1, únicamente que se ajusta la escala para que la visualización sea más clara y se pueda apreciar mejor la ubicación de los individuos.

A partir de ésta gráfica se pueden obtener las siguientes conclusiones:

Nuevamente aparecen alejados del conglomerado principal los estados de Chiapas y Oaxaca (etiquetados con 7 y 20), apareciendo ambos en la parte superior de la gráfica. Un poco más abajo se encuentran Guerrero e Hidalgo (12 y 13) los cuales se acercan más al conglomerado principal.

Dentro del cuadrante inferior se encuentra una distribución que , a simple vista, parece no presentar observaciones discrepantes. Siendo la observación más separada hacia el lado izquierdo la correspondiente a Quintana Roo, que como ya se había notado en los capítulos anteriores, muestra un comportamiento muy particular.

En la parte más baja de la gráfica se sitúan Baja California Norte, Sonora y Nuevo León. En tanto que del lado derecho y hacia abajo se ubica el Distrito Federal, el cual tiene como vecino más cercano a Baja California Sur y a Nuevo León.

En la figura 3.2 se muestra de manera clara que, de acuerdo a las variables seleccionadas en la tabla 1.1, los estados de Chiapas y Oaxaca son los que más alejados están del "comportamiento standard" y, prácticamente no tienen vecinos cercanos, y el que menos dista de ellos es Guerrero, un estado que presentó características semejantes a estos dos estados.

Cabe destacar que anteriormente no había aparecido como vecino del D.F. el estado de B.C.S., lo cual pudiera indicar un comportamiento extraño del estado o una impresición en el paquete.

css/3: cluster analysis	Euclidean distances=SUMA(ABS(x-y)^2)^(1/2)					
linkage Case No.	C:1	C:2	C:3	C:4	C:5	C:6
C:1	.00000	18.95368	8.47784	24.77256	4.51604	8.73436
C:2	18.95368	.00000	12.83525	33.77909	17.25420	12.32179
C:3	8.47784	12.83525	.00000	26.09309	8.19071	8.08606
C:4	24.77256	33.77909	26.09309	.00000	25.55687	27.27682
C:5	4.51604	17.25420	8.19071	25.55687	.00000	8.96323
C:6	8.73436	12.32179	8.08606	27.27682	8.96323	.00000
C:7	66.95155	72.57186	67.13303	42.35718	67.63268	68.52105
C:8	13.13381	9.93124	7.01975	26.10892	11.68993	9.73895
C:9	13.48269	19.14689	14.98922	28.97423	11.08471	14.87149
C:10	17.88923	27.92903	19.13824	13.18804	17.38146	22.08320
C:11	20.26051	27.51900	21.55777	10.99058	20.94958	20.50470
C:12	41.08992	48.24495	42.16713	18.72474	41.94488	42.28524
C:13	35.99730	43.96859	36.68518	12.09696	36.72980	38.31757
C:14	9.97606	14.91459	9.63894	20.79649	9.03861	7.88779
C:15	5.06024	18.99517	9.16324	20.91185	5.91380	9.03405
C:16	24.87779	32.03976	26.27617	8.29031	25.19028	25.71633
C:17	7.26412	19.22086	12.06731	23.37198	8.71227	7.56835
C:18	10.33583	19.91973	12.22636	16.63425	10.81605	11.31042
C:19	9.26425	13.75594	9.26840	29.57599	6.65476	8.51057
C:20	54.30326	61.63637	55.56728	29.83023	54.94412	56.36952
C:21	31.82663	40.70605	33.59283	7.92116	32.42015	34.23839
C:22	18.46319	23.48812	17.71554	12.89476	18.84774	17.70860
C:23	22.34194	19.48241	18.62803	19.08629	22.18914	19.15506
C:24	34.65726	40.52907	33.67085	13.99937	34.77318	36.75400
C:25	7.77083	14.23578	6.73473	24.58628	7.54503	5.26051
C:26	11.61562	9.09920	6.79645	30.39038	9.98591	6.95382
C:27	28.58325	35.34909	29.12656	8.25857	28.32284	30.74206
C:28	13.38285	20.89631	11.47313	17.78458	12.70199	16.29582
C:29	17.45802	32.11959	22.37519	13.51715	18.27332	23.05189
C:30	34.85247	41.30116	34.26842	13.13183	35.51088	36.66477
C:31	27.78926	38.54719	30.70856	8.45429	28.69650	31.24393
C:32	29.51139	39.21692	31.68198	12.80483	29.18926	33.40952

Tabla 3.1: Matriz de disimilitudes obtenida con la distancia euclidiana, generada por CSS.

css/3: cluster analysis	Euclidean distances=SUMA(ABS(x-y)^2)^(1/2)					
linkage Case No.	C:7	C:8	C:9	C:10	C:11	C:12
C:1	66.95155	13.13381	13.48269	17.88923	20.26051	41.08992
C:2	72.57186	9.93124	19.14689	27.92903	27.51900	48.24495
C:3	67.13303	7.01975	14.98922	19.13824	21.55777	42.16713
C:4	42.35718	26.10892	28.97423	13.18804	10.99058	18.72474
C:5	67.63268	11.68993	11.08471	17.38146	20.94958	41.94488
C:6	68.52105	9.73895	14.87149	22.08320	20.50470	42.28524
C:7	.00000	65.89093	69.43713	52.77808	49.69727	29.74740
C:8	65.89093	.00000	17.87984	19.04937	20.07224	40.74155
C:9	69.43713	17.87984	.00000	22.99879	26.57370	46.12591
C:10	52.77808	19.04937	22.99879	.00000	13.63607	28.90705
C:11	49.69727	20.07224	26.57370	13.63607	.00000	22.37399
C:12	29.74740	40.74155	46.12591	28.90705	22.37399	.00000
C:13	31.74590	36.07738	40.37933	22.25232	19.89576	10.55559
C:14	62.00600	9.22077	14.20497	15.63549	14.73905	36.44168
C:15	63.09093	12.61280	13.40934	15.02265	16.46580	37.31108
C:16	44.84257	24.77468	29.41888	14.66604	6.31710	18.32604
C:17	64.95554	13.83304	16.16987	18.72680	16.39346	37.83778
C:18	58.43618	12.52536	17.39742	12.45949	10.48676	32.13110
C:19	71.13592	12.42314	7.86865	22.55009	24.94118	46.10611
C:20	15.45297	54.48235	57.08460	40.09826	36.98066	18.04440
C:21	36.07293	33.29048	35.02361	19.20681	16.24268	13.68328
C:22	51.59065	16.21257	24.06402	14.14400	6.82640	25.28720
C:23	54.15629	16.42581	23.31190	20.73873	16.60062	32.47024
C:24	35.75626	32.47218	38.76447	19.70694	20.64460	16.99845
C:25	65.89497	8.14704	15.45930	18.72506	18.11222	39.36608
C:26	71.28612	7.16050	16.14224	23.04362	24.00952	45.45853
C:27	41.05259	28.36603	30.21941	14.38764	15.68701	21.15227
C:28	57.82563	11.66779	19.16190	9.52531	15.67705	33.44141
C:29	53.56880	24.42965	22.29991	10.90881	15.16450	29.48865
C:30	34.12640	33.44167	39.65363	21.60390	19.73326	13.67506
C:31	41.86908	31.53513	30.25107	17.42323	16.83527	21.50860
C:32	43.40051	31.06682	32.59560	13.11899	18.27680	22.91646

Tabla 3.1 (Continuación)

css/3: cluster analysis	Euclidean distances=SUMA(ABS(x-y)^2)^(1/2)					
linkage Case No.	C:13	C:14	C:15	C:16	C:17	C:18
C:1	35.99730	9.97606	5.06024	24.87779	7.26412	10.33583
C:2	43.96859	14.91459	18.99517	32.03976	19.22086	19.91973
C:3	36.68518	9.63894	9.16324	26.27617	12.06731	12.22636
C:4	12.09696	20.79649	20.91185	8.29031	23.37198	16.63425
C:5	36.72980	9.03861	5.91380	25.19028	8.71227	10.81605
C:6	38.31757	7.88779	9.03405	25.71633	7.56835	11.31042
C:7	31.74590	62.00600	63.09093	44.84257	64.95554	58.43618
C:8	36.07738	9.22077	12.61280	24.77468	13.83304	12.52536
C:9	40.37933	14.20497	13.40934	29.41888	16.16987	17.39742
C:10	22.25232	15.63549	15.02265	14.66604	18.72680	12.45949
C:11	19.89576	14.73905	16.46580	6.31710	16.39346	10.48676
C:12	10.55559	36.44168	37.31108	18.32604	37.83778	32.13110
C:13	.00000	31.98603	32.19582	15.73921	34.35497	27.76074
C:14	31.98603	.00000	6.84968	19.16437	8.29893	5.73579
C:15	32.19582	6.84968	.00000	20.78642	6.45739	6.50232
C:16	15.73921	19.16437	20.78642	.00000	21.52755	15.09259
C:17	34.35497	8.29893	6.45739	21.52755	.00000	7.90508
C:18	27.76074	5.73579	6.50232	15.09259	7.90508	.00000
C:19	40.94794	10.84387	10.30931	29.05880	12.05672	14.82086
C:20	19.81821	49.70053	50.45030	31.79350	52.03081	45.74961
C:21	7.90845	27.62642	27.84937	11.47662	29.79884	23.52806
C:22	21.84662	12.35083	14.62547	11.07142	15.40627	9.52437
C:23	27.89300	14.46663	19.00935	18.69752	20.85695	15.53217
C:24	9.34824	30.31546	31.09008	17.62272	34.03040	26.88743
C:25	35.20218	7.06563	7.31329	23.18533	7.15648	8.90715
C:26	40.89567	10.91685	12.37181	28.96494	12.91500	14.88902
C:27	13.60096	23.31539	24.41399	11.59709	27.63362	20.38481
C:28	27.24447	11.66507	11.33470	18.95897	15.34894	10.36592
C:29	23.49717	17.77227	14.59523	15.68786	17.52946	13.60959
C:30	6.91023	30.66447	31.20881	16.96296	33.62222	26.92626
C:31	15.09426	24.91787	24.22611	13.61556	26.65648	21.02208
C:32	16.41320	26.22728	26.16667	14.65369	28.98399	22.57221

Tabla 3.1 (Continuación)

css/3: cluster analysis	Euclidean distances=SUMA(ABS(x-y)^2)^(1/2)					
linkage Case No.	C:19	C:20	C:21	C:22	C:23	C:24
C:1	9.26425	54.30326	31.82663	18.46319	22.34194	34.65726
C:2	13.75594	61.63637	40.70605	23.48812	19.48241	40.52907
C:3	9.26840	55.56728	33.59283	17.71554	18.62803	33.67085
C:4	29.57599	29.83023	7.92116	12.89476	19.08629	13.99937
C:5	6.65476	54.94412	32.42015	18.84774	22.18914	34.77318
C:6	8.51057	56.36952	34.23839	17.70860	19.15506	36.75400
C:7	71.13592	15.45297	36.07293	51.59065	54.15629	35.75626
C:8	12.42314	54.48235	33.29048	16.21257	16.42581	32.47218
C:9	7.86865	57.08460	35.02361	24.06402	23.31190	38.76447
C:10	22.55009	40.09826	19.20681	14.14400	20.73873	19.70694
C:11	24.94118	36.98066	16.24268	6.82640	16.60062	20.64460
C:12	46.10611	18.04440	13.68328	25.28720	32.47024	16.99845
C:13	40.94794	19.81821	7.90845	21.84662	27.89300	9.34824
C:14	10.84387	49.70053	27.62642	12.35083	14.46663	30.31546
C:15	10.30931	50.45030	27.84937	14.62547	19.00935	31.09008
C:16	29.05880	31.79350	11.47662	11.07142	18.69752	17.62272
C:17	12.05672	52.03081	29.79884	15.40627	20.85695	34.03040
C:18	14.82086	45.74961	23.52806	9.52437	15.53217	26.88743
C:19	.00000	58.83000	36.35790	22.01983	22.29615	38.82291
C:20	58.83000	.00000	22.76722	39.96435	44.03027	26.01142
C:21	36.35790	22.76722	.00000	19.00904	25.08899	13.88364
C:22	22.01983	39.96435	19.00904	.00000	12.93304	20.59462
C:23	22.29615	44.03027	25.08899	12.93304	.00000	25.81970
C:24	38.82291	26.01142	13.88364	20.59462	25.81970	.00000
C:25	9.67726	53.74564	31.57198	14.90109	18.62213	33.14096
C:26	8.73597	59.48310	37.52180	20.50323	21.11911	37.88932
C:27	31.74944	28.63464	9.72286	17.42578	20.35068	14.02807
C:28	16.99347	46.18797	24.89492	12.30092	17.30711	23.42986
C:29	23.32429	39.83714	18.25121	17.52464	24.43713	24.92247
C:30	39.52264	24.44282	12.35772	19.84294	25.72998	5.65133
C:31	32.47167	28.45693	8.61126	19.72794	24.16378	19.60679
C:32	33.92628	29.81238	13.10117	21.67825	27.56220	16.98348

Tabla 3.1 (Continuación)

css/3: cluster analysis	Euclidean distances=SUMA(ABS(x-y)^2)^(1/2)					
linkage Case No.	C:25	C:26	C:27	C:28	C:29	C:30
C:1	7.77083	11.61562	28.58325	13.38285	17.45802	34.85247
C:2	14.23578	9.09920	35.34909	20.89631	32.11959	41.30116
C:3	6.73473	6.79645	29.12656	11.47313	22.37519	34.26842
C:4	24.58628	30.39038	8.25857	17.78458	13.51715	13.13183
C:5	7.54503	9.98591	28.32284	12.70199	18.27332	35.51088
C:6	5.26051	6.95382	30.74206	16.29582	23.05189	36.66477
C:7	65.89497	71.28612	41.05259	57.82563	53.56880	34.12640
C:8	8.14704	7.16050	28.36603	11.66779	24.42965	33.44167
C:9	15.45930	16.14224	30.21941	19.16190	22.29991	39.65363
C:10	18.72506	23.04362	14.38764	9.52531	10.90881	21.60390
C:11	18.11222	24.00952	15.68701	15.67705	15.16450	19.73326
C:12	39.36608	45.45853	21.15227	33.44141	29.48865	13.67506
C:13	35.20218	40.89567	13.60096	27.24447	23.49717	6.91023
C:14	7.06563	10.91685	23.31539	11.66507	17.77227	30.66447
C:15	7.31329	12.37181	24.41399	11.33470	14.59523	31.20881
C:16	23.18533	28.96494	11.59709	18.95897	15.68786	16.96296
C:17	7.15648	12.91500	27.63362	15.34894	17.52946	33.62222
C:18	8.90715	14.88902	20.38481	10.36592	13.60959	26.92626
C:19	9.67726	8.73597	31.74944	16.99347	23.32429	39.52264
C:20	53.74564	59.48310	28.63464	46.18797	39.83714	24.44282
C:21	31.57198	37.52180	9.72286	24.89492	18.25121	12.35772
C:22	14.90109	20.50323	17.42578	12.30092	17.52464	19.84294
C:23	16.62213	21.11911	20.35068	17.30711	24.43713	25.72998
C:24	33.14096	37.88932	14.02807	23.42986	24.92247	5.65133
C:25	.00000	7.52439	28.29616	12.31215	20.78125	33.34103
C:26	7.52439	.00000	32.89312	16.11740	26.32616	38.58074
C:27	28.29616	32.89312	.00000	20.44865	16.52278	15.32496
C:28	12.31215	16.11740	20.44865	.00000	16.44963	24.81978
C:29	20.78125	26.32616	16.52278	16.44963	.00000	25.08346
C:30	33.34103	38.58074	15.32496	24.81978	25.08346	.00000
C:31	29.26670	35.05645	10.40542	23.35700	13.22098	18.71820
C:32	30.51528	35.27980	9.94447	22.03683	15.04407	19.13437

Tabla 3.1 (Continuación)

css/3: cluster analysis	Euclidean distances SUM(ABS(x-y)^2)^(1/2)	
linkage Case No.	C:31	C:32
C:1	27.78926	29.51139
C:2	38.54719	39.21692
C:3	30.70856	31.68198
C:4	8.45429	12.80483
C:5	28.69650	29.18926
C:6	31.24393	33.40952
C:7	41.86908	43.40051
C:8	31.53513	31.06682
C:9	30.25107	32.59560
C:10	17.42323	13.11899
C:11	16.83527	18.27680
C:12	21.50860	22.91646
C:13	15.09426	16.41320
C:14	24.91787	26.22728
C:15	24.22611	26.16667
C:16	13.61556	14.65369
C:17	26.65648	28.98399
C:18	21.02208	22.57221
C:19	32.47167	33.92628
C:20	28.45693	29.81238
C:21	8.61126	13.10117
C:22	19.72794	21.67825
C:23	24.16378	27.56220
C:24	19.60679	16.98348
C:25	29.26670	30.51528
C:26	35.05645	35.27980
C:27	10.40542	9.94447
C:28	23.35700	22.03683
C:29	13.22098	15.04407
C:30	18.71820	19.13437
C:31	.00000	12.99474
C:32	12.99474	.00000

Tabla 3.1 (Continuación)

css/3: multidim scaling	Starting Configuration (Guttman-Lingoes)	
	DIM. 1	DIM. 2
C_1	-.070251	-.079647
C_2	-.079400	-.109571
C_3	-.071394	-.083434
C_4	-.040800	.006152
C_5	-.071067	-.081351
C_6	-.072592	-.086324
C_7	1.318537	-.292891
C_8	-.070291	-.081561
C_9	-.074835	-.093543
C_10	-.054692	-.036340
C_11	-.051560	-.028093
C_12	.022979	.165111
C_13	-.002093	.088928
C_14	-.066298	-.068888
C_15	-.066734	-.069701
C_16	-.043745	-.002362
C_17	-.068306	-.074542
C_18	-.062615	-.058323
C_19	-.074888	-.093239
C_20	.241804	1.203528
C_21	-.021673	.052803
C_22	-.055122	-.039559
C_23	-.059963	-.055107
C_24	-.018200	.058379
C_25	-.069522	-.077890
C_26	-.075362	-.094473
C_27	-.036993	.016979
C_28	-.062265	-.057993
C_29	-.054764	-.035315
C_30	-.013087	.066427
C_31	-.036853	.021105
C_32	-.037958	.020738

```
32 cases from file  C:\USR\ADRIAN\MPORC2.M in 2 dim.
start config.: (Guttman-Lingoes)
last iteration computed:   53;  best iteration:  20
D-star: raw stress = 2.539398;  alienation = .0497830
D-hat:  raw stress = 2.035797;  stress     = .0445879
```

Tabla 3.2: Configuración inicial dada por CSS utilizando un algoritmo metrico.

css/3: multidim scaling	Distances in Final Configuration D-star: raw stress = 2.539398; alienation = .0497830 D-hat: raw stress = 2.035797; stress = .0445879					
	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0.000000	.787631	.356716	1.268204	.046443	.279905
C_2	.787631	0.000000	.478741	1.765927	.742514	.508173
C_3	.356716	.478741	0.000000	1.339715	.320314	.134356
C_4	1.268204	1.765927	1.339715	0.000000	1.298402	1.404886
C_5	.046443	.742514	.320314	1.298402	0.000000	.235589
C_6	.279905	.508173	.134356	1.404886	.235589	0.000000
C_7	3.593790	3.998886	3.630044	2.325751	3.623210	3.712551
C_8	.511020	.440960	.167380	1.325701	.479138	.298004
C_9	.377613	.990453	.676495	1.546740	.386966	.554327
C_10	.806022	1.434870	.965746	.507928	.842973	.995991
C_11	.957077	1.383599	.972706	.386295	.979625	1.051178
C_12	2.188438	2.595066	2.216735	.927014	2.216001	2.299247
C_13	1.879263	2.317647	1.922780	.613726	1.907982	1.999843
C_14	.316641	.760790	.291018	1.058736	.317452	.347191
C_15	.197433	.857525	.381168	1.076128	.222697	.373420
C_16	1.210975	1.646141	1.240546	.169636	1.236467	1.317901
C_17	.118604	.866558	.405544	1.152516	.157772	.362509
C_18	.433512	.983377	.515217	.851426	.456310	.558491
C_19	.280389	.613285	.351608	1.533955	.241155	.217281
C_20	2.897264	3.358011	2.962053	1.632702	2.929149	3.035036
C_21	1.645249	2.148685	1.726308	.386624	1.677359	1.790738
C_22	.851954	1.198765	.810598	.582000	.866823	.902670
C_23	1.026621	1.018961	.809714	.987380	1.019710	.939416
C_24	1.808022	2.132795	1.785392	.640326	1.829365	1.880612
C_25	.256475	.612863	.134585	1.234605	.231577	.172367
C_26	.436885	.354371	.204895	1.541546	.390997	.163582
C_27	1.404218	1.934364	1.499730	.176980	1.437225	1.558135
C_28	.557350	.932799	.509317	.836654	.566744	.594335
C_29	.887072	1.593893	1.115701	.635535	.930152	1.119981
C_30	1.836188	2.187411	1.828673	.633398	1.859305	1.920024
C_31	1.436242	2.034128	1.579552	.360242	1.473752	1.621585
C_32	1.469999	2.111157	1.645925	.513371	1.510269	1.675813

Tabla 3.3: Matriz de distancias finales a partir de la tabla 1.1, después de aplicar un algoritmo no métrico, generada por CSS.

css/3: multidim scaling	Distances in Final Configuration D-star: raw stress = 2.539398; alienation = .0497830 D-hat: raw stress = 2.035797; stress = .0445879					
	C_7	C_8	C_9	C_10	C_11	C_12
C_1	3.593790	.511020	.377613	.806022	.957077	2.188438
C_2	3.998886	.440960	.990453	1.434870	1.383599	2.595066
C_3	3.630044	.167380	.676495	.965746	.972706	2.216735
C_4	2.325751	1.325701	1.546740	.507928	.386295	.927014
C_5	3.623210	.479138	.386966	.842973	.979625	2.216001
C_6	3.712551	.298004	.554327	.995991	1.051178	2.299247
C_7	0.000000	3.579326	3.860275	2.813091	2.661373	1.413546
C_8	3.579326	0.000000	.843556	1.005608	.944991	2.169782
C_9	3.860275	.843556	0.000000	1.047585	1.281865	2.473599
C_10	2.813091	1.005608	1.047585	0.000000	.393688	1.429869
C_11	2.661373	.944991	1.281865	.393688	0.000000	1.248129
C_12	1.413546	2.169782	2.473599	1.429869	1.248129	0.000000
C_13	1.715665	1.885757	2.159429	1.114844	.950261	.315168
C_14	3.366040	.348126	.693602	.675743	.704940	1.953016
C_15	3.400591	.490153	.550213	.631263	.759649	1.993361
C_16	2.395049	1.209539	1.517835	.523677	.267843	.982618
C_17	3.478266	.538827	.448708	.687419	.850149	2.074671
C_18	3.171507	.558944	.773883	.451744	.523705	1.761803
C_19	3.856431	.512716	.378818	1.084091	1.203133	2.446583
C_20	.732081	2.927029	3.147614	2.104056	1.989477	.774530
C_21	1.956432	1.709412	1.904089	.856676	.765191	.593728
C_22	2.820724	.764579	1.204039	.493984	.197685	1.408332
C_23	3.019115	.678372	1.404118	.956909	.643949	1.645673
C_24	1.867622	1.720042	2.128737	1.130267	.851341	.496722
C_25	3.540217	.259963	.615000	.840023	.878845	2.126889
C_26	3.834938	.303913	.653851	1.150644	1.177056	2.421618
C_27	2.200573	1.493526	1.659930	.612520	.560848	.828209
C_28	3.121789	.491879	.924453	.544676	.463564	1.708270
C_29	2.845749	1.182861	1.046430	.245929	.629226	1.506312
C_30	1.811518	1.769544	2.148405	1.134269	.879697	.422864
C_31	2.230479	1.594833	1.650077	.630935	.705104	.923991
C_32	2.278049	1.676204	1.649807	.680906	.828515	1.030254

Tabla 3.3 (Continuación)

css/3: multidim scaling	Distances in Final Configuration D-star: raw stress = 2.539398; alienation = .0497830 D-hat: raw stress = 2.035797; stress = .0445879					
	C_13	C_14	C_15	C_16	C_17	C_18
C_1	1.879263	.316641	.197433	1.210975	.118604	.433512
C_2	2.317647	.760790	.857525	1.646141	.866558	.983377
C_3	1.922780	.291018	.381168	1.240546	.405544	.515217
C_4	.613726	1.058736	1.076128	.169636	1.152516	.851426
C_5	1.907982	.317452	.222697	1.236467	.157772	.456310
C_6	1.999843	.347191	.373420	1.317901	.362509	.558491
C_7	1.715665	3.366040	3.400591	2.395049	3.478266	3.171507
C_8	1.885757	.348126	.490153	1.209539	.538827	.558944
C_9	2.159429	.693602	.550213	1.517835	.448708	.773883
C_10	1.114844	.675743	.631263	.523677	.687419	.451744
C_11	.950261	.704940	.759649	.267843	.850149	.523705
C_12	.315168	1.953016	1.993361	.982618	2.074671	1.761803
C_13	0.000000	1.652759	1.685300	.682513	1.764548	1.455856
C_14	1.652759	0.000000	.178028	.971084	.267109	.224374
C_15	1.685300	.178028	0.000000	1.014409	.102012	.236107
C_16	.682513	.971084	1.014409	0.000000	1.099897	.780701
C_17	1.764548	.267109	.102012	1.099897	0.000000	.330005
C_18	1.455856	.224374	.236107	.780701	.330005	0.000000
C_19	2.140766	.508079	.458954	1.464647	.398099	.684988
C_20	1.041296	2.687949	2.707587	1.721734	2.780138	2.484075
C_21	.288758	1.445015	1.456158	.508851	1.528126	1.235455
C_22	1.121196	.566038	.658621	.448198	.757450	.430524
C_23	1.406449	.713045	.877773	.824039	.976060	.726152
C_24	.358515	1.541654	1.610608	.614737	1.701442	1.374510
C_25	1.827772	.175880	.246909	1.145684	.280317	.393816
C_26	2.127266	.485384	.535030	1.444891	.525633	.705866
C_27	.514622	1.214540	1.216976	.338816	1.286653	1.000507
C_28	1.413469	.257612	.374776	.731352	.476771	.193631
C_29	1.194649	.837605	.746813	.704470	.773506	.626243
C_30	.276088	1.578080	1.638859	.630571	1.727349	1.403201
C_31	.630221	1.288952	1.261526	.529763	1.317679	1.066280
C_32	.755579	1.355263	1.307197	.680850	1.352583	1.130889

Tabla 3.3 (Continuación)

css/3: multidim scaling	Distances in Final Configuration D-star: raw stress = 2.539398; alienation = .0497830 D-hat: raw stress = 2.035797; stress = .0445879					
	C_19	C_20	C_21	C_22	C_23	C_24
C_1	.280389	2.897264	1.645249	.851954	1.026621	1.808022
C_2	.613285	3.358011	2.148685	1.198765	1.018961	2.132795
C_3	.351608	2.962053	1.726308	.810598	.809714	1.785392
C_4	1.533955	1.632702	.386624	.582000	.987380	.640326
C_5	.241155	2.929149	1.677359	.866823	1.019710	1.829365
C_6	.217281	3.035036	1.790738	.902670	.939416	1.880612
C_7	3.656431	.732081	1.956432	2.820724	3.019115	1.867622
C_8	.512716	2.927029	1.709412	.764579	.678372	1.720042
C_9	.378818	3.147614	1.904089	1.204039	1.404118	2.128737
C_10	1.084091	2.104056	.856676	.493984	.956909	1.130267
C_11	1.203133	1.989477	.765191	.197685	.643949	.851341
C_12	2.446583	.774530	.593728	1.408332	1.645673	.496722
C_13	2.140766	1.041296	.288758	1.121196	1.406449	.358515
C_14	.508079	2.687949	1.445015	.566038	.713045	1.541654
C_15	.458954	2.707587	1.456158	.658621	.877773	1.610608
C_16	1.464647	1.721734	.508851	.448198	.824039	.614737
C_17	.398099	2.780138	1.528126	.757450	.976060	1.701442
C_18	.684988	2.484075	1.235455	.430524	.726152	1.374510
C_19	0.000000	3.166119	1.914986	1.073946	1.152504	2.047502
C_20	3.166119	0.000000	1.252026	2.162486	2.420118	1.268260
C_21	1.914986	1.252026	0.000000	.955080	1.314483	.519708
C_22	1.073946	2.162486	.955080	0.000000	.470891	.977948
C_23	1.152504	2.420118	1.314483	.470891	0.000000	1.157844
C_24	2.047502	1.268260	.519708	.977948	1.157844	0.000000
C_25	.355488	2.863614	1.620845	.731734	.806736	1.709577
C_26	.276893	3.166108	1.928154	1.015310	.982015	1.989124
C_27	1.675913	1.494115	.244166	.757627	1.162036	.639901
C_28	.765627	2.453040	1.222218	.309369	.532690	1.286873
C_29	1.165769	2.122053	.913501	.739887	1.200837	1.275824
C_30	2.081410	1.197063	.462857	1.018523	1.224090	.084915
C_31	1.714901	1.506230	.343345	.901084	1.336279	.829886
C_32	1.750374	1.547331	.479874	1.019469	1.469239	.985000

Tabla 3.3 (Continuación)

css/3: multidim scaling	Distances in Final Configuration D-star: raw stress = 2.539398; alienation = .0497830 D-hat: raw stress = 2.035797; stress = .0445879					
	C_25	C_26	C_27	C_28	C_29	C_30
C_1	.256475	.436885	1.404218	.557350	.887072	1.836188
C_2	.612863	.354371	1.934364	.932799	1.593893	2.187411
C_3	.134585	.204895	1.499730	.509317	1.115701	1.828673
C_4	1.234605	1.541546	.176980	.836654	.635535	.633398
C_5	.231577	.390997	1.437225	.566744	.930152	1.859305
C_6	.172367	.163582	1.558135	.594335	1.119981	1.920024
C_7	3.540217	3.834938	2.200573	3.121789	2.845749	1.811518
C_8	.259963	.303913	1.493526	.491879	1.182861	1.769544
C_9	.615000	.653851	1.659930	.924453	1.046430	2.148405
C_10	.840023	1.150644	.612520	.544676	.245929	1.134269
C_11	.878845	1.177056	.560848	.463564	.629226	.879697
C_12	2.126889	2.421618	.828209	1.708270	1.506312	.422864
C_13	1.827772	2.127266	.514622	1.413469	1.194649	.276088
C_14	.175880	.485384	1.214540	.257612	.837605	1.578080
C_15	.246909	.535030	1.216976	.374776	.746813	1.638859
C_16	1.145684	1.444891	.338816	.731352	.704470	.630571
C_17	.280317	.525633	1.286653	.476771	.773506	1.727349
C_18	.393816	.705866	1.000507	.193631	.626243	1.403201
C_19	.355488	.276893	1.675913	.765627	1.165769	2.081410
C_20	2.863614	3.166108	1.494115	2.453040	2.122053	1.197063
C_21	1.620845	1.928154	.244166	1.222218	.913501	.462857
C_22	.731734	1.015310	.757627	.309369	.739887	1.018523
C_23	.806736	.982015	1.162036	.532690	1.200837	1.224090
C_24	1.709577	1.989124	.639901	1.286873	1.275824	.084915
C_25	0.000000	.312060	1.389850	.422774	.982859	1.748248
C_26	.312060	0.000000	1.699484	.713930	1.281822	2.033089
C_27	1.389850	1.699484	0.000000	1.002398	.680768	.609786
C_28	.422774	.713930	1.002398	0.000000	.758949	1.325692
C_29	.982859	1.281822	.680768	.758949	0.000000	1.265447
C_30	1.748248	2.033089	.609786	1.325692	1.265447	0.000000
C_31	1.460017	1.772071	.217707	1.105208	.615843	.787212
C_32	1.520920	1.831433	.381900	1.193982	.603444	.937360

Tabla 3.3 (Continuación)

css/3: multidim scaling	Distances in Final Configuration D-star: raw stress = 2.539398; alienation = .0497830 D-hat: raw stress = 2.035797; stress = .0445879	
	C_31	C_32
C_1	1.436242	1.469999
C_2	2.034128	2.111157
C_3	1.579552	1.645925
C_4	.360242	.513371
C_5	1.473752	1.510269
C_6	1.621585	1.675813
C_7	2.230479	2.278049
C_8	1.594833	1.676204
C_9	1.650077	1.649807
C_10	.630935	.680906
C_11	.705104	.828515
C_12	.923991	1.030254
C_13	.630221	.755579
C_14	1.288952	1.355263
C_15	1.261526	1.307197
C_16	.529763	.680850
C_17	1.317679	1.352583
C_18	1.066280	1.130889
C_19	1.714901	1.750374
C_20	1.506230	1.547331
C_21	.343345	.479874
C_22	.901084	1.019469
C_23	1.336279	1.469239
C_24	.829886	.985000
C_25	1.460017	1.520920
C_26	1.772071	1.831433
C_27	.217707	.381900
C_28	1.105208	1.193982
C_29	.615843	.603444
C_30	.787212	.937360
C_31	0.000000	.164436
C_32	.164436	0.000000

Tabla 3.3 (Continuación)

css/3: multidim scaling	Final Configuration D-star: raw stress = 2.539398; alienation = .0497830 D-hat: raw stress = 2.035797; stress = .0445879	
	DIM. 1	DIM. 2
C_1	-.81979	.184866
C_2	-1.23123	-.486760
C_3	-.87435	-.167653
C_4	.44351	.073388
C_5	-.85252	.151916
C_6	-.95501	-.060209
C_7	2.75568	-.177513
C_8	-.82056	-.326154
C_9	-1.04707	.486421
C_10	-.01963	.281936
C_11	.09560	-.094508
C_12	1.34234	-.153434
C_13	1.04480	-.049512
C_14	-.60796	-.050484
C_15	-.63134	.126003
C_16	.36279	-.075816
C_17	-.70209	.199489
C_18	-.40761	.050539
C_19	-1.09001	.110046
C_20	2.07607	.094641
C_21	.82478	.137494
C_22	-.06485	-.209974
C_23	-.22584	-.652493
C_24	.89874	-.376924
C_25	-.78287	-.068938
C_26	-1.07924	-.166638
C_27	.58442	.180461
C_28	-.36586	-.138537
C_29	-.00143	.527190
C_30	.94890	-.308409
C_31	.60061	.397565
C_32	.60101	.562001

Tabla 3.4: Configuración final que sirve para obtener la gráfica en dos dimensiones para la tabla 1.1. (Ver figuras 3.1 y 3.2).

css/3: multidim scaling	Final Configuration D-star: raw stress = 2.539398; alienation = .0497830 D-hat: raw stress = 2.035797; stress = .0445879

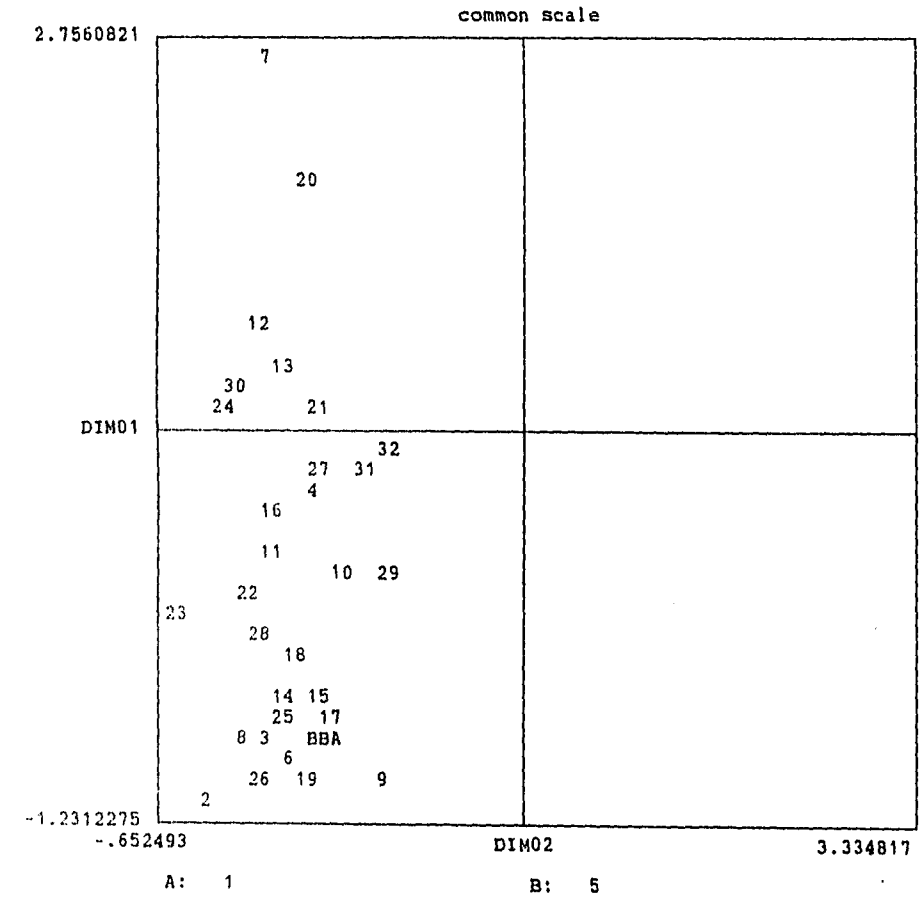


Figura 3.1: Configuración final sin realizar ajuste de escala, generada por el paquete CSS, a partir de la tabla 3.4.

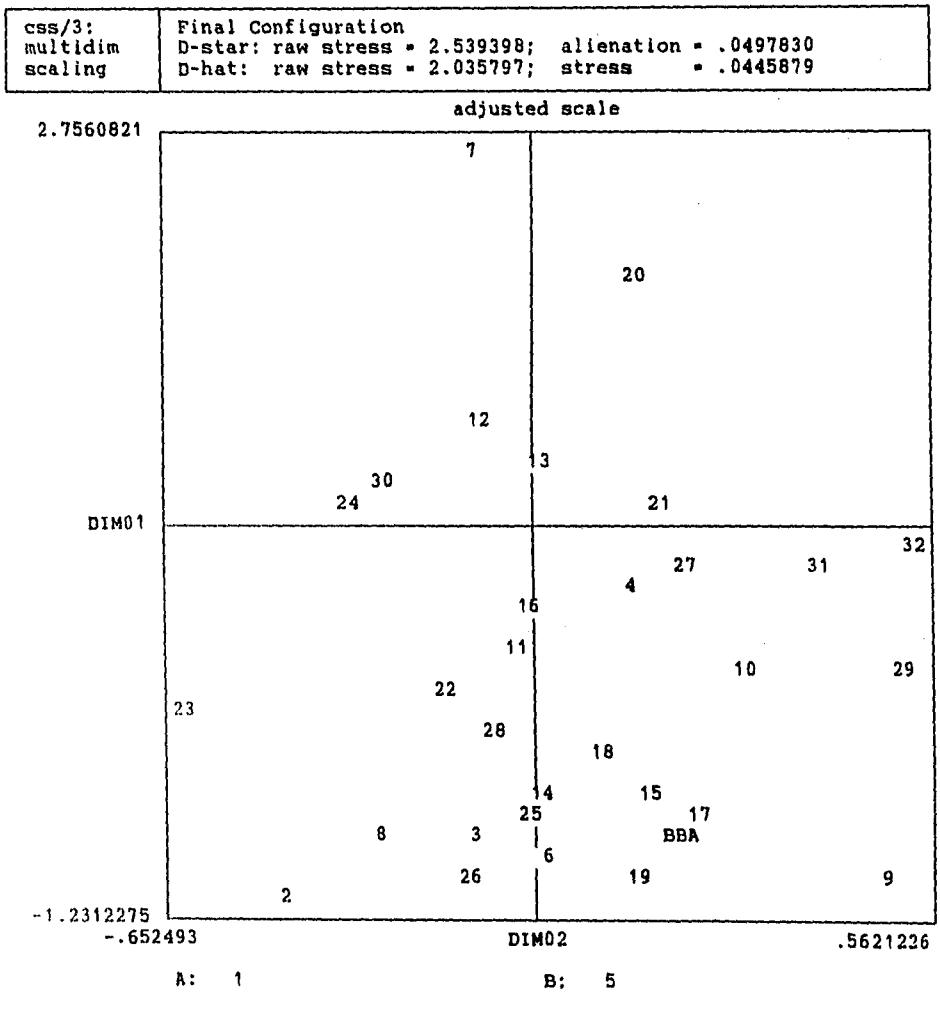


Figura 3.2: Gráfica final obtenida ajustando la escala original para mejorar la visualización de los datos en un espacio bidimensional, generada por CSS a partir de la tabla 3.4.

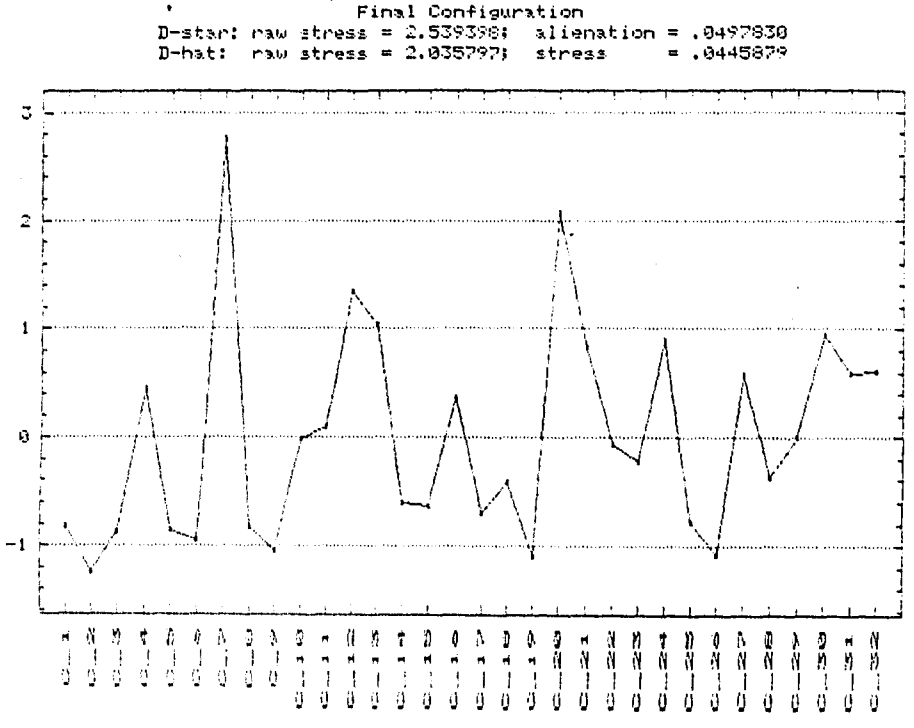


Figura 3.3: Representación directa, donde el eje de las ordenadas representa el valor de DIM1 (de la tabla 3.4) y el eje de las abscisas el número del estado. Aquí se observa nuevamente cómo Chiapas (7) y Oaxaca (20) difieren del común de los demás estados.

CAPITULO 4

ANÁLISIS DE CONGLOMERADOS

IV.-ANÁLISIS DE CONGLOMERADOS.

4.1 OBJETIVO

El objetivo del análisis de conglomerados es agrupar un conjunto de datos en un esquema que permita hacer inferencia acerca de su comportamiento. Es decir, pretende encontrar conglomerados o grupos homogéneos que permitan dividir el conjunto de datos para realizar algún estudio determinado. Así, este método puede ser utilizado como una de las técnicas "exploratorias" para datos multivariados, que permite hacer predicciones basadas en los grupos o conglomerados obtenidos. Aunque también es utilizado como un método para establecer particiones sobre un conjunto de datos, es decir, puede caer también dentro de las técnicas confirmatorias.

Aunque lo más común es agrupar a los individuos u observaciones, existen técnicas para realizar agrupamientos sobre las variables, pero el interés principal será el agrupamiento de individuos.

Una gran parte de ejemplos de aplicación de ésta técnica se encuentran en el área de la biología, ya que el agrupamiento de especies, tanto vegetales como animales es muy frecuente en estudios de los seres vivos. También se presentan en el área de psicología donde la clasificación de los disturbios mentales puede ayudar a aplicar métodos de terapia. En la investigación de mercado se puede agrupar a los encuestados de acuerdo a su opinión acerca de algún producto.

De esta manera, cada grupo tendrá elementos que compartan determinadas características.

Dentro del análisis de conglomerados se pueden manejar tres tipos principales de datos:

1) Un conjunto con n individuos y p variables. Como el caso estudiado en el primer capítulo.

2) Una matriz de disimilitudes (de nxn) donde la entrada a_{ij} es la medida de similitud o disimilitud entre los individuos i y j.

3) Datos ordenados en base a determinados criterios.

De cualquier manera, los tres tipos de datos pueden transformarse a una matriz de similitudes (como lo realizado en el capítulo 3) y aplicar a ella los algoritmos de ésta técnica; se puede hacer esto, debido a que el análisis de conglomerados es un agrupamiento de individuos con base en determinadas características que presenten en común. Sin embargo, algunos autores prefieren aplicar métodos diferentes según sea el tipo de datos que se esté manejando.

Es importante recalcar que el agrupamiento que se obtenga dependerá de los criterios utilizados en el algoritmo, ya que puede presentarse el caso de que se obtenga una configuración no muy evidente al ojo humano.

Debido al gran desarrollo que ha tenido esta técnica en diferentes campos tales como la psicología, botánica, antropología, etc. han aparecido artículos refiriéndose al análisis de conglomerados en cada una de esas áreas, lo que ha ocasionado que haya una gran producción de distintos métodos, muchos de los cuales no tienen una justificación teórica suficiente.

Sin embargo, se pueden distinguir 2 principales tipos de agrupamientos: los jerárquicos y los no jerárquicos. A continuación se describen brevemente cada uno de ellos.

4.2 MÉTODOS JERÁRQUICOS:

A grandes rasgos, estos métodos sugieren que partir de los datos iniciales se forman determinados conglomerados (grupos) y, una vez que se tienen, se van conjuntando en "conglomerados de conglomerados" hasta llegar a uno solo. La idea es ir fusionando un conglomerado con otro conglomerado o individuo e ir formando una especie de árbol o diagrama. La secuencia sería de la siguiente manera:

a) En una etapa inicial, cada individuo forma un conglomerado. (conglomerados de un solo elemento).

b) En una segunda etapa aparecerán algunos conglomerados en los cuales sus elementos son las parejas que muestran mayor afinidad entre sí. Desde luego que en este segundo nivel pueden seguir apareciendo conglomerados de un solo elemento.

Al ir avanzando las etapas, a algunos conglomerados se agregará ya sea un individuo o algún otro conglomerado.

c) En la etapa final se tendrá un único conglomerado que agrupa a todos los elementos del conjunto de datos. Dicha agrupación tendrá una estructura de árbol.

4.3 MÉTODOS NO JERÁRQUICOS:

En estos métodos, el investigador puede fijar el número de conglomerados que desee. Aquí la jerarquía no se da una manera general como en el caso anterior, sino que se presenta únicamente en los primeros niveles (de acuerdo a como los define el investigador). Ocurre que al obtener una segunda solución o aumentar de nivel, los nuevos conglomerados no tendrán necesariamente subconjuntos de los conglomerados del nivel anterior. La mayoría de los métodos de este tipo operan sobre una matriz de datos, aunque, como se describió al principio de este capítulo, es equivalente a una matriz de disimilitudes.

De acuerdo a la configuración que generan, los algoritmos pueden ser vistos también como:

De partición: Los individuos se agrupan en conglomerados ajenos. Aquí no se llega al diagrama del método jerárquico, sino más bien, a un conjunto de conglomerados que no se intersectan.

Conglomerados no ajenos: En este caso los conglomerados si se pueden intersectar.

Igualmente, existen diversos algoritmos para cada tipo de agrupamiento.

Para nuestro propósito, se analizarán los datos de la tabla 1.1 utilizando un método jerárquico.

El método utilizado generará una salida conocida como **dendograma**, para ello utiliza un algoritmo conocido como "el vecino más cercano" (o de liga sencilla).

Dicho método se basa en una matriz de disimilitudes. Para convertir una matriz de correlación o covarianzas en una matriz de disimilitudes basta definir una medida adecuada de similitud que cumpla con las características descritas en el capítulo anterior. Por ejemplo. se puede definir d(i,j) como:

$$d(i,j)= |var(i) - var(j)|$$

Donde i y j se refieren al número de observación o individuo. Para una mayor descripción de medidas de similitud ver Chatfield [4].

A partir de lo anterior, se irán fusionando grupos de una matriz de nxn, definiendo

$d_{c1c2}=\min\{d_{rs} : r \_ c1, s \_ c2\}$
donde c1 y c2 son dos grupos.

De esta manera, se comienza con conglomerados de dos elementos, los cuales corresponderán a las menores medidas de disimilitud de la matriz original. Se mide la distancia de los demás elementos a ese conglomerado inicial, el cual será (i,j), por lo que se procede a calcular:

$$d_{k,(i,j)} = \min \{d_{ki}, d_{kj}\} \text{ para k distinto de i,j.}$$

Las distancias anteriores se toman como nuevas d_{rs}. De la matriz original desaparece la columna j y el renglón i, agregando la columna C_1 cuyas entradas d_{kc_1} serán las calculadas anteriormente. Se procede de la misma manera hasta agotar la matriz. Al final se tendrá algo semejante a la figura 4.1.

En esta gráfica la escala vertical dará una idea del tamaño del conglomerado.

Existe un algoritmo semejante en el cual las distancias no se toman como los mínimos descritos anteriormente sino como los máximos. En este caso también se generará un dendograma, sólo que su interpretación será diferente. A este método se le conoce como **liga completa**. Y, además, existe otra técnica que toma como distancia al promedio de la distancia mínima y la distancia máxima, a éste se le cono ce como **liga del promedio** o algoritmo del promedio

La aplicación de métodos jerárquicos puede ser cuestionada, ya que si no se aplica correctamente puede producir resultados poco congruentes. A este respecto, algunas personas aducen que no se puede aplicar un algoritmo jerárquico a un conjunto de datos que no sabemos si tiene o no estructura

jerárquica. Por ello prefieren los métodos no jerárquicos que involucran cierta medida para conocer qué tan bueno es un ajuste realizado. Dicha medida funciona muy parecido al SS del capítulo 3 (escalamiento multidimensional), descrito en la sección 3.2.2.

4.4 APLICACIÓN A LOS DATOS DEL CENSO.

Esta técnica se aplicó a los datos de la tabla 1.1 en el paquete CSS utilizando la medida de similitud descrita anteriormente (del capítulo de escalamiento multidimensional) y obteniéndose la misma matriz de disimilitudes, con la cual se corrió dicho paquete. Es decir, a partir de la tabla 3.1, en lugar de obtener coordenadas en $\mathbf{R}^2$ para los puntos, se construye un dendograma, que es directamente una representación gráfica.

Al utilizar las 10 variables se obtuvo la salida de la figura 4.1, la cual revela las siguientes características:

Se vuelve a corroborar cómo los estados de Chiapas y Oaxaca muestran características muy parecidos, razón por la cual son "vecinos muy cercanos" en la configuración obtenida. A su vez, se observa que se alejan del comportamiento usual y difieren con respecto a los demás estados.

En el dendograma obtenido se observa que Quintana Roo se encuentra muy cercano al grupo formado por Chiapas y Oaxaca; y se pueden distinguir, a grandes rasgos 3 grupos:

1) Aguascalientes, Coahuila, Estado de México, Morelos, Jalisco, Nayarit, Nuevo León, Baja California Sur, Colima, Sinaloa, Sonora, Chihuahua y el D.F , los cuales son estados del norte y centro del país

2) Baja California Norte, Campeche, Hidalgo, San Luis Potosí, Veracruz, Puebla, Tabasco, Guanajuato, Michoacán, Querétaro, Yucatán, Zacatecas, Durango, Tamaulipas, Guerrero y Tlaxcala, la mayoría de los cuales son del centro y sur del país

3) Quintana Roo, Chiapas y Oaxaca. Que son estados del sur del país.

No se esperaba que Quintana Roo apareciera como vecino de Chiapas; ni tampoco que el D.F. y Nuevo León estuvieran tan alejados. Las aparentes inconsistencias pueden ser explicadas un poco si se observa la gráfica obtenida en el escalamiento multidimensional, en el cual Quintana Roo, aunque muy separado del conglomerado principal se encuentra "relativamente cerca" de Chiapas y Oaxaca.

Sin embargo, se puede ver que se sigue cierta tendencia a mantener grupos de niveles económicos y educativos altos y otros grupos con niveles bajos.

Lo que no muestra ningún cambio es la similitud entre los estados de Chiapas y Oaxaca y su distanciamiento del D.F.

CSS/3: CLUSTER ANALYSIS 11-10-94 20:33:17 PAGE 6
DENDOGRAMA

data file: C:\USR\ADRIAN\MPORC2.CSS [32 cases with 10 variables]

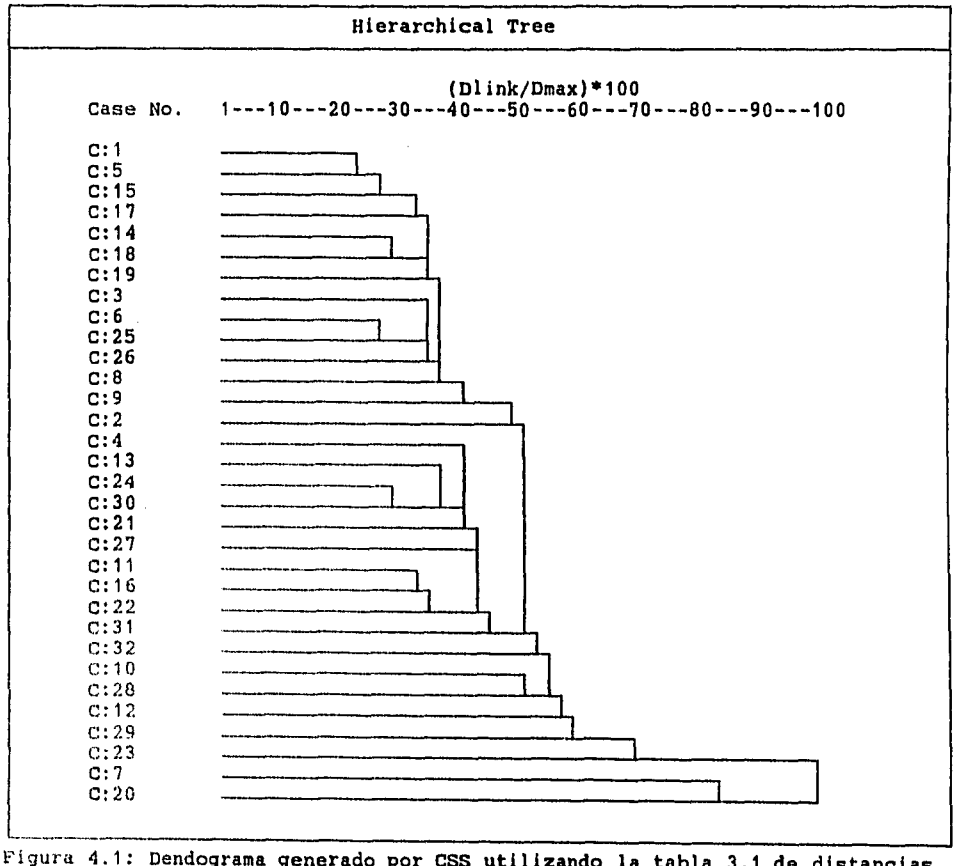


Figura 4.1: Dendograma generado por CSS utilizando la tabla 3.1 de distancias euclidianas empleando un método de liga simple.

CAPÍTULO 5

ANÁLISIS DE CORRESPONDENCIA

V.- ANÁLISIS DE CORRESPONDENCIA

5.1 INTRODUCCIÓN

El análisis de correspondencia es una herramienta estadística que, a semejanza del escalamiento multidimensional, permite una representación gráfica de nuestros datos (que en general, corresponderán a una matriz de incidencias) en una dimensión que permite tener una mejor perspectiva respecto a ellos.

Esta técnica comenzó a desarrollarse hace más de 50 años por Richardson y Kuder en 1933 y por Horst en 1935, pero es hasta hace aproximadamente 20 años que empezó a ser utilizada. En los 50's Fisher y Guntman desarrollaron separadamente el análisis de correspondencia, uno en el contexto psiquiátrico y otro en el contexto biométrico, que son los campos en los que cada uno desarrolló sus aplicaciones.

La disposición de herramientas como las computadoras, que permiten hacer en segundos cálculos complejos, permite que éste análisis se aplique con mayor frecuencia en la actualidad.

Esta técnica utiliza como datos de entrada a una matriz de incidencia, cuyas entradas son no negativas, conocida también como tabla de contingencias. El objetivo de este método es obtener una representación gráfica simplificada de dicha matriz en una dimensión menor (dicha dimensión será 2 o 3) que permita una comprensión más fácil y rápida del comportamiento de los datos; teniendo en cuenta que la relación entre los números es más interesante que los números en sí mismos.

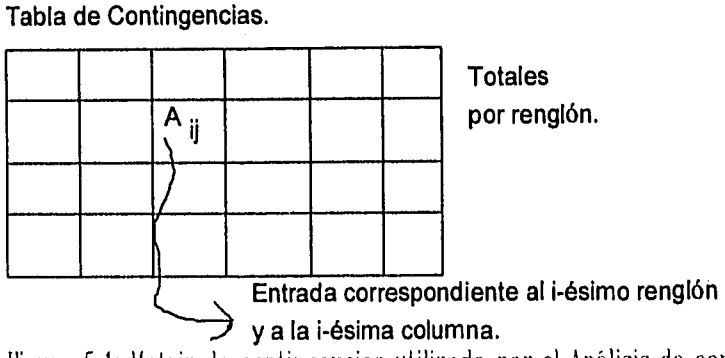


Figura 5.1: Matriz de contingencias utilizada por el Análisis de correspondencia.

En general se hace referencia al conjunto de los renglones como **I** y al de las columnas como **J**.

Si la matriz de contingencias es de mxn, se puede identificar a los renglones como puntos en $\mathbf{R}^n$; entonces, haciendo semejanza con componentes principales, se obtendrá una representación (o proyección) de esos puntos sobre el plano euclidiano para que puedan ser visualizados gráficamente con sus coordenados respecto a los dos primeros ejes (que se denominan ejes principales).

Para las columnas se hace algo análogo, pudiendo obtenerse una visualización de puntos m-dimensionales y n-dimensionales en $\mathbf{R}^2$ o $\mathbf{R}^3$ .De esta manera, se pueden obtener conclusiones útiles del comportamiento de las columnas y renglones, los cuales pueden ser concebidos como "puntos", razón por la cual en muchas ocasiones se hace referencia a ellos como puntos renglón o puntos columna. Para un mayor detalle ver Greenacre [9].

5.2 DESARROLLO

Una vez que se tiene conformada la matriz de contingencias se procede de la siguiente manera:

1) Obtener totales por renglón y columna, así como un total general. Esta matriz **B** tendrá $\mathbf{B}_i$ como total del renglón i y $\mathbf{B}_j$ como total de la columna j. El análisis subsecuente se desarrollará para renglones únicamente, aunque puede realizarse para columnas de manera muy semejante.

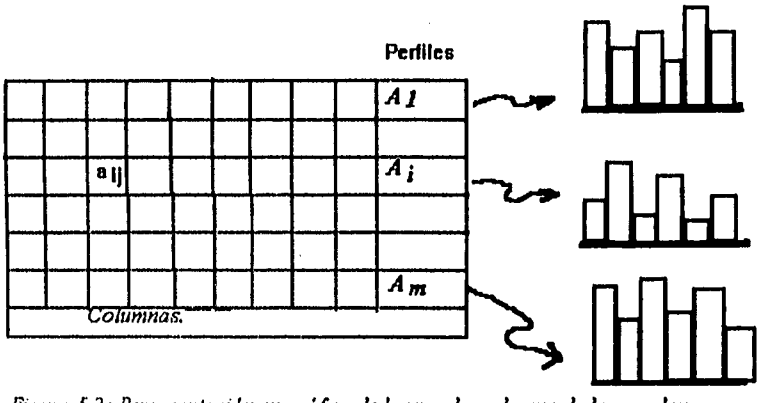


Figura 5.2: Representación en gráfica de barras de cada uno de los renglones de frecuencias relativas.

2) Obtener las frecuencias relativas para cada entrada. Donde $\mathbf{a_{ij}} = \mathbf{b_{ij}}/\mathbf{B_{i.}}$,es la proporción con la que se presenta la variable i en la variable j. Con lo que se conforma la matriz **A** (de frecuencias relativas). En esta etapa, generalmente se

construyen gráficas de barras para cada uno de los renglones de frecuencias relativas.

3) Se obtienen los llamados perfiles con $A_i=B_i/\sum B_i$. Que son porcentajes de cada renglón respecto a la suma total, algunos autores llaman a esto el peso de cada perfil.

Los perfiles son vectores de frecuencias relativas (respecto al total general), miden la importancia relativa de cada uno y describen su relación con respecto a todas las columnas. A continuación se procede a realizar los cálculos con la matriz de frecuencias relativas.

$\mathbf{D_r}$= Matriz diagonal cuyos elementos son los pesos de cada perfil.

$\mathbf{D_c}$= Matriz diagonal cuyos elementos son el centroide., denotado por **c** (total general).

$$\mathbf{c=Dc1}$$

donde $\mathbf{1}^t = [1,1,...1]$

Las coordenadas de los perfiles en una representación óptima en 2 dimensiones son las siguientes, aunque también puede considerarse una o tres dimensiones que son las más comunes:

$$\mathbf{F_{(2)}=N_{(2)}\, D_{\mu(2)}}$$

Donde $N_{(2)}$ y $D_{\mu_{(2)}}$ son las dos primeras columnas de las submatrices obtenidas de la descomposición generalizada en valor singular de la matriz:

$$A - 1c^t$$

Que es la matriz A omitiendo la dimensión trivial. Y donde:

$$N^t D_r N = M^t Dc\text{-}1M = I$$

Como comentario cabe decir que la descomposición generalizada en valor singular muestra que una matriz A puede ser descompuesta o factorizada como:

$$A = N D_\mu M$$

donde $N^t \Omega N = M^t \Phi M = I$.
Siendo Ω y Φ matrices definidas positivas.

Así, al obtener la representación en dimensión 2 se procede a graficar los puntos obtenidos para visualizarla.

5.3 EL PROBLEMA DUAL.

Lo que se hizo en la sección anterior para renglones puede hacerse para las columnas, ya que columnas y renglones guardan una relación geométrica directa. En este caso, al hacer la división de cada entrada de la matriz original por el total de la columna respectiva se pueden obtener perfiles por columnas.

Si a la matriz de frecuencias le llamamos **C** y $\mathbf{D_r}$= Matriz diagonal cuyos elementos son los pesos de cada columna y $\mathbf{r=D_r1}$.

Entonces al obtener la descomposición generalizada en valor singular de la matriz

$$\mathbf{C - 1 \cdot r^t = PD\mu Q^t}$$

donde $\mathbf{S^t D_C S = Q^t D_r - 1 Q = I}$

Se calcula $G_{(2)}$ como: $G_{(2)} = S_{(2)} D_{\mu(2)}$

Donde $G_{(2)}$ y $D_{\mu(2)}$ son las dos primeras columnas de las submatrices obtenidas de la descomposición generalizada en valor singular de la matriz:

$$\mathbf{C - 1 \cdot r^t}$$

Con lo que se obtiene la configuración en R^2 de las columnas de la matriz de I x J.

Si las matrices son no singulares, $\mathbf{D_\mu}$ es la misma tanto en renglones como en columnas y las matrices se definen positivas, de las relaciones $\mathbf{F_{(2)}=N_{(2)}D_{\mu(2)}}$. y de $G_{(2)} = S_{(2)} D_{\mu(2)}$ se puede llegar a la conclusión de que $\mathbf{G = SN^{-1}F}$ y $\mathbf{F = NS^{-1}G}$. De esto se desprende que los ejes asociados a las columnas y a los renglones, en estos casos, tengan la misma inercia y puedan representarse ambos en una sola gráfica, pero no se puede hablar de distancia entre puntos renglón y puntos columna, aunque estén en la misma gráfica.

Una vez que se han obtenido los vectores $\mathbf{F}_{(2)}$ o $\mathbf{G}_{(2)}$, dependiendo del caso, se procede a la construcción de la gráfica lo que permitirá una interpretación mucho más fácil de asimilar.

En este punto, puede mencionarse que existe un resultado (ver Benzécri [2]) que indica que si dos renglones de frecuencias (perfiles) son distribucionalmente idénticos (es decir, al dividir la entrada correspondiente entre el total por renglón se obtienen los mismos resultados) entonces se puede eliminar cualquiera de ellos sin que afecte la configuración geométrica ya que las coordenadas resultantes serán las mismas para ambos. Un principio idéntico se aplica a columnas con frecuencias iguales.

5.4.- INERCIA.

Dentro del análisis de correspondencia se maneja un concepto conocido como **inercia**, el cual, haciendo comparación con componentes principales (capítulo 2) viene siendo el equivalente al porcentaje de explicación de la varianza que presentan los datos.

Así, es de esperarse un valor alto en los resultados de la inercia, ya que esto indicaría que el comportamiento de los datos está siendo bien explicado por la representación obtenida.

La fórmula para obtenerla es:

$$in(I) = \Sigma w_i d_i^2.$$

Donde $w_i = n_i/n$, es la entrada del i-ésimo perfil, $d_i^2 = (P_i-c)^t\mathbf{D}(P_i-c)$. En esta última igualdad, P_i es la entrada i-ésima del vector de perfiles, c= centroide y **D** es la matriz de distancias, en este caso euclidianas.

De esta manera, cada eje principal, tendrá asociado cierto porcentaje de la inercia total. Lo deseable es que los dos primeros ejes principales acumulen un alto porcentaje de la inercia total. De otra manera, indicaría que existen puntos (los cuales representan a los perfiles) que no están siendo bien representados con respecto a los dos primeros ejes principales.

Como, al final de cuentas, lo que se obtiene es una proyección sobre el plano $\mathbf{R}^2$, no se puede saber qué tan alejados estén realmente los puntos de dicho plano. Para analizar ésta cuestión se hace uso del ángulo entre el eje principal en cuestión y el punto que representa al vector perfil (ver figura 5.1).

Mediante un análisis detallado (ver Benzécri [2]) se puede obtener que, si θ es el ángulo entre el punto y el eje principal, entonces $\cos^2\theta$ representará la contribución de los ejes a la inercia del punto, de tal manera que si éste valor es alto quiere decir que el eje en cuestión explica muy bien la inercia del punto.

Figura 5.1. Dos puntos con la misma proyección sobre el eje X tiene distancias muy diferentes a él.

Paralelo al concepto de inercia, se maneja el concepto de masa, que representa la importancia relativa que tienen cada uno de los puntos para contribuir a la configuración final. El valor de $\cos^2\theta$ también se conoce como contribución relativa porque es independiente de la masa del punto.

5.5 APLICACIÓN A LOS DATOS DEL CENSO.

Para ejemplificar el análisis de correspondencia, se eligió una tabla en la cual, tanto las columnas como los renglones representaran variables, y a partir de ellas se pudiera formar una tabla de contingencias. De esta manera, a partir de los datos que presenta el códice 90 se construyó la tabla 5.1, la cual muestra en las columnas variables que corresponden al tipo de actividad desarrollada por parte de la **PEA** y que a continuación se describen :

Tipo de actividad desarrollada:

C1 = Empleado u obrero.
C2 = Jornalero o Peón.
C3 = Trabajador por su cuenta.
C4 = Patrón o empresario.
C5 = Trabajador familiar no remunerado.
C6 = No especificado.

Así, cada una de las variables anteriores se contrastó con el número de horas trabajadas a la semana por cada uno de los tipos de empleo representados en las variables anteriores. Los renglones de la tabla 5.1 corresponden a los siguientes valores:

Número de horas trabajadas
durante la semana de referencia.

R1 = 0

R2 = Menos de 8

R3 = De 9 a 16

R4 = De 17 a 24

R5 = De 25 a 32

R6 = De 33 a 40

R7 = De 41 a 48

R8 = De 49 a 56

R9 = Mas de 56

R10 = No especificó

En la tabla 5.1, que corresponde a una matriz de incidencias puede observarse la relación que se establece entre el número de horas trabajadas por una persona y el trabajo al que se dedica.

Actividad/ horas trabajadas	Empleado u obrero	Jornalero o Peón	Por su cuenta	Empres. o Patrón	Familiar sin remun.	No especific..
0	185025	4877	108627	5969	12370	116937
< 8	237081	59910	160518	11310	13299	21076
9 a 16	273079	66320	213940	13905	19421	19204
17 a 24	420025	99209	307398	16156	33631	25662
25 a 32	878418	137042	450444	28000	50725	40939
33 a 40	3500126	341206	942673	101647	91946	90471
41 a 48	4485871	1074169	1666931	140724	196408	165511
49 a 56	1484274	282828	582313	80882	50180	48056
Más 56	1760609	347219	855680	123209	70480	70894
N. E.	210442	53599	177370	13206	48969	280103

Tabla 5.1 Tabla de relación entre actividad y horas trabajadas a la semana.

De lo anterior se tiene que, si la matriz formada por los datos anteriores se denota por A, la entrada A(i,j) corresponderá al número de personas que trabajan i horas a la semana y que se dedican al tipo de trabajo j.

Es importante resaltar que en esta tabla no fue necesario obtener porcentajes por estado debido a que los datos son totales para el país, y por lo tanto no se tienen resultados parciales. Además, dicha tabla refleja la situación en la semana de referencia, es decir, hace referencia a los días en que se llevó a cabo el censo.

En primera instancia se realizó el análisis para los datos tal como se muestran en la tabla 5.1, de éste análisis se obtuvo que el renglón y columna que se refieren al rubro de **"No especificados"** tienen una importancia y un peso demasiado significativos para el análisis y configuración final. Así, éstos datos son los que determinan en gran medida los resultados que se obtienen.

Sin embargo, para obtener conclusiones que reflejen información útil referente a los datos, no sirve de mucho decir, por ejemplo, que "Quienes se dedican a un empleo no especificado trabajan un número de horas no especificado". Esta conclusión realmente no sirve de nada, puesto que se desea obtener alguna asociación entre empleo y número de horas trabajadas.

Tomando en cuenta lo anterior, se decidió que un análisis más objetivo y útil se obtendría dándole un peso menor a los datos que estuvieran dentro de los **"No especificados"**.

Para realizar lo anterior, se utilizó el paquete estadístico SAS, el cual tiene la opción de que cuando se realice un análisis de correspondencia se pueda dejar una columna como suplementaria y a un renglón darle un peso menor a los demás.

Con esto, la tabla quedó como sigue: columnas 1 a 5 como principales, columna 6 como suplementaria; renglones 1 a 9 con un peso de 1000, renglón 10 con un peso de -1000. De esta manera, se garantiza que la importancia que tengan los puntos referentes a **No especificados** no repercutirá de manera significativa en el análisis y configuración final.

De la tabla 5.1 se obtiene la tabla 5.2, que corresponde a la matriz de perfiles, la cual, como ya se había mencionado presenta los porcentajes respecto a los totales por renglones.

Actividad/ horas trabajadas	Empleado u obrero	Jornalero o Peón	Por su cuenta	Empresario	Familiar sin remuneración	No especificado. (Suplemen -- tario.)	Peso para cada uno de los renglones.
0	38.732	10.210	22.739	1.249	2.589	24.478	100
< 8	47.115	11.905	31.899	2.247	2.642	4.188	100
9 a 16	45.072	10.946	35.311	2.295	3.205	3.169	100
17 a 24	46.561	10.997	34.073	1.791	3.728	2.844	100
25 a 32	55.400	8.643	28.409	1.765	3.199	2.582	100
33 a 40	69.062	6.732	18.600	2.005	1.814	1.785	100
41 a 48	58.034	13.896	21.565	1.820	2.541	2.141	100
49 a 56	58.701	11.185	23.029	3.198	1.984	1.900	100
Más 56	54.540	10.756	26.507	3.816	2.183	2.196	100
N. E.	26.003	6.918	22.895	1.704	6.321	36.156	-100

Tabla 5.2 Matriz de Perfiles de la tabla 5.1 (Porcentajes con respecto al total por renglones).

De la tabla 5.2 se obtiene la figura 5.2, en la cual se presentan los perfiles en una representación de gráficas de barras. De dicha figura se observa cierta

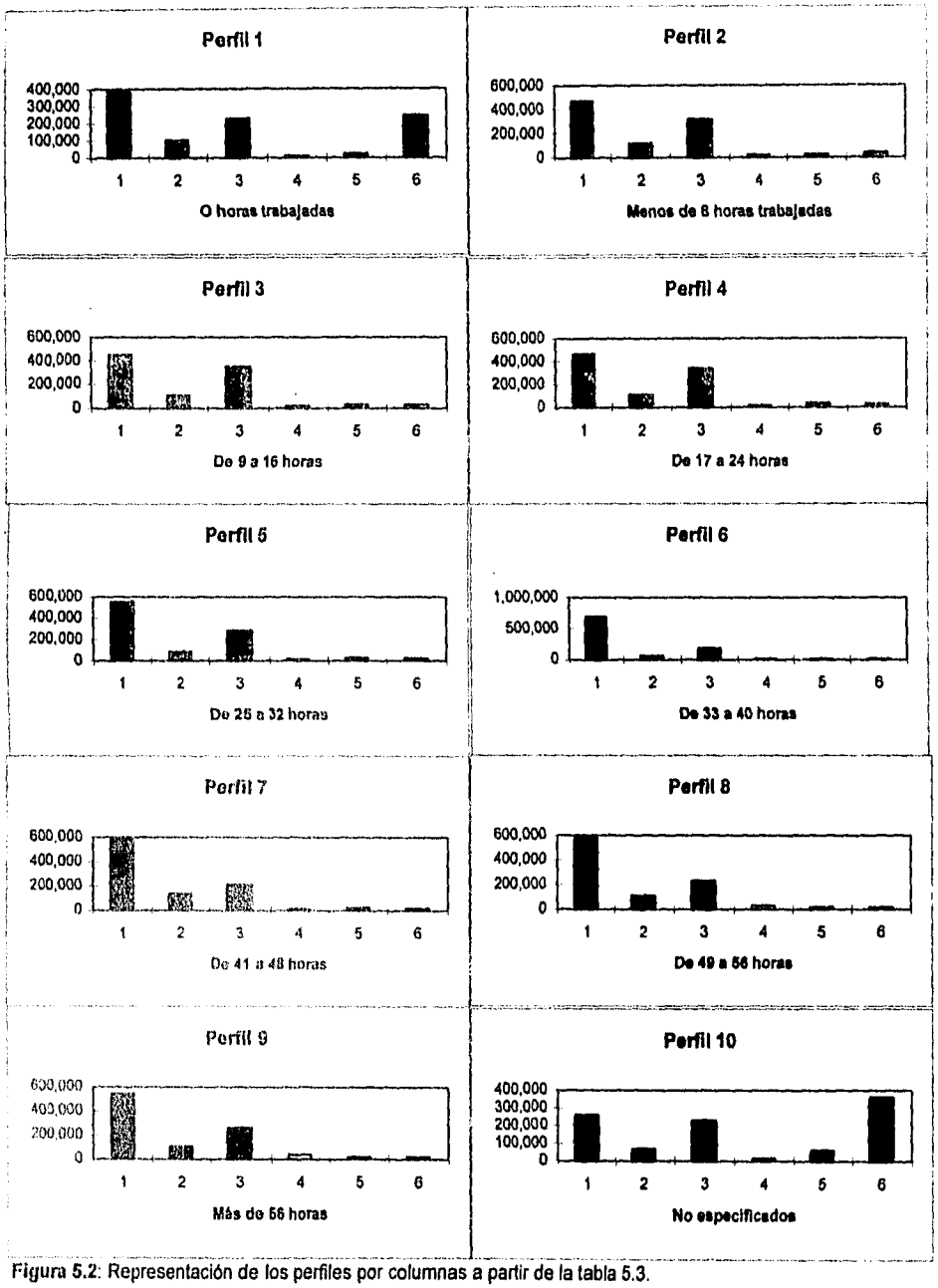


Figura 5.2: Representación de los perfiles por columnas a partir de la tabla 5.3.

discrepancia del perfil 10 con respecto a los demás, debido a que presenta valores más altos en la columna 6. Lo mismo se observa en el perfil uno, por lo que podía esperarse una relación entre puntos renglón 1 y 10 con punto columna 6, es decir, se puede concluir que "Quienes trabajan en una actividad no especificada trabajan un número indeterminado de horas y quienes trabajan 0 horas se dedican a un empleo indeterminado", esto es lo que se mencionaba anteriormente y que fue la razón por la cual columna 6 y renglón 10 que corresponden a no especificados fueron relegados en cierta manera. Los demás perfiles muestran valores altos para las columnas 1 y 3 y bajos para las demás columnas, no pudiéndose obtener alguna otra asociación a primera vista.

Una vez que se obtuvieron las gráficas de la matriz de perfiles se procedió a aplicar la técnica de Análisis de correspondencia mediante el paquete estadístico SAS, el cual tomó como datos de entrada la tabla 5.1 y, además de proporcionar la tabla de perfiles (5.2), generó las tablas que a continuación se describen, tanto para puntos renglón como para puntos columna. Entendiendo por punto renglón a las coordenadas que representan al vector de perfiles cuando se toman por renglón; los puntos columna serán, así, el equivalente a vectores columna de la matriz de perfiles.

En primer lugar, se obtiene la tabla 5.3 que corresponde a la descomposición de la inercia. En ella se muestra la inercia principal y el porcentaje de varianza total explicado por cada uno de los ejes principales ordenados en forma ascendiente.

En esta tabla se muestra la inercia principal y el porcentaje de varianza total explicado por cada uno de los ejes principales ordenados de manera ascendente.

Se observa que el primer eje explica en un 66.34% la varianza total de los datos originales y, tomándolo conjuntamente con el segundo eje principal, explican el 90.2%. Este porcentaje es bastante alto sobre todo si se toma en cuenta que se está trabajando con datos reales de nuestro país.

Por otra parte, se obtiene también una tabla donde se presenta la masa e inercia para cada uno de los puntos, así como la calidad del ajuste con que se representa cada punto. La tabla 5.4A se refiere a puntos columna (que se refiere a actividad o empleo). Esta tabla muestra que la mayor importancia relativa la tienen los "Empleados u obreros" seguidos de "Trabajadores por su cuenta" y, en menor medida los "Jornaleros o Peones". Lo anterior se debe a que muestran los valores más altos en cuanto a **masa**.

Valor Singular	**Inercia Principal**	**$\chi 2$**	**Proporción de varianza explicada.**
0.13324	0.01775	3.911E8	66.34%
0.07990	0.00638	1.406E8	23.86%
0.05061	0.00256	5.643E7	9.57%
0.00774	0.00006	1320487	0.22%
	0.02676	5.895E8	(Grados de libertad = 32)

Tabla 5.3 Salida generada por el paquete SAS donde se muestra la descomposición de la inercia para los ejes principales en orden ascendiente.

Por otro lado, son los "Trabajadores por su cuenta", los "Jornaleros o Peones" y los "Empleados u obreros" (en ese orden) los que contribuyen más a la inercia asociada al primer eje principal. Además, a excepción de los "Patrones o empresarios" los demás puntos tienen un buen ajuste dentro de la representación final.

En esta tabla no se muestran los valores para masa e inercia de los "No especificados" debido a que no tienen significancia ni repercusión en el esquema obtenido,

Para puntos renglón se puede hacer el mismo análisis através de la tabla 5.4B, donde se muestra que los puntos que tienen mayor importancia relativa dentro del análisis son los relativos a "41 a 48 hrs." y "33 a 40 hrs.", seguidos por "Más de 56 hrs." y por "49 a 56 hrs.".

Tipo Empleo	Calidad de ajuste	Masa	Inercia
Empleado u obrero	0.997647	0.600296	0.260886
Jornalero o Peón	0.992674	0.111515	0.263677
Por su cuenta	0.994272	0.240060	0.341088
Patrón o Empresario	0.141884	0.023686	0.091601
Fam. sin remuneración	0.652233	0.024442	0.042748
No especificó	.074783		

Tabla 5.4A Masa e inercia obtenida por SAS para las variables correspondientes a tipo de empleo (puntos columna).

Horas trabajadas	Calidad de ajuste	Masa	Inercia
0	0.898753	0.016376	0.024116
< 8	0.970761	0.021885	0.046969
9 a 16	0.979578	0.026630	0.097584
17 a 24	0.921193	0.039783	0.125734
25 a 32	0.743272	0.070115	0.056562
33 a 40	0.991469	0.225947	0.395123
41 a 48	0.981602	0.343355	0.144043
49 a 56	0.004318	0.112596	0.017950
Más 56	0.523060	0.143314	0.091917
N. E.	0.520775		

Tabla 5.4B Masa e inercia obtenida por SAS para las variables correspondientes a horas trabajadas (puntos renglón).

Los que contribuyen más a la inercia asociada al primer eje son "33 a 40 hrs." seguido por "41 a 48 hrs." y por "17 a 24 hrs.". Además, a excepción de los puntos "49 a 56" y "Más de 56 hrs." se mantiene una buena representación para los puntos.

De los datos anteriores, quien dijera que en México se trabajan 40 horas a la semana estaría dando un dato que cae dentro del rango posible para esta variable.

En las tablas 5.5A y 5.5B se muestra los valores correspondientes a cosenos cuadrados y a ángulo formando con los dos primeros ejes principales para puntos columna y renglón, respectivamente.

Tipo Empleo	Dimensión 1	Dimensión 2	Ángulo dim 1	Ángulo dim 2
Empleado u obrero	0.997091	0.000556	3.097	88.6488
Jornalero o Peón	0.335974	0.656700	54.5752	35.8678
Por su cuenta	0.831681	0.162591	24.2216	66.2200
Patrón o Empresario	0.035075	0.106809	79.2057	70.9243
Fam. sin remuneración	0.651324	0.000910	36.1916	88.2713
No especificó	0.074777	0.000006	74.1301	89.8596

Tabla 5.5A Cosenos cuadrados y ángulos correspondientes a los puntos columna.

De la tabla 5.5A se observa que los puntos que tienen menor ángulo respecto al primer eje son: "Empleados u obreros" y "Trabajadores por su cuenta", que además son de los puntos con mayor importancia relativa; por esta razón influyen de manera determinante en la configuración final y en la orientación del primer eje principal.

Lo análogo para renglones se presenta en la tabla 5.5B, donde el punto de menor ángulo con respecto al primer eje principal es el de "Menos de 8 hrs.", seguido por "33 a 40", "9 a 16" y "17 a 24" (en ese orden), lo cual indica que estos puntos influyen en la orientación del primer eje. Si se observa que de la tabla 5.4B el punto "33 a 40 hrs." es uno de los de mayor importancia relativa, se tiene que este punto es sumamente importante para la configuración final.

Horas trabajadas	Dimensión 1	Dimensión 2	Ángulo para dim 1	Ángulo para dim 2
0	0.894231	0.004522	18.9789	86.1441
< 8	0.918344	0.052417	16.6039	76.7648
9 a 16	0.847642	0.131936	22.9710	68.7013
17 a 24	0.819532	0.101661	25.1389	71.4070
25 a 32	0.281126	0.462146	57.9801	41.1709
33 a 40	0.916544	0.074925	16.7913	74.1139
41 a 48	0.014258	0.967344	83.1421	10.4111
49 a 56	0.004279	0.000040	86.2493	89.6376
Más 56	0.356944	0.166116	53.3126	65.9475
N. E.	0.481887	0.038888	46.0380	78.6267

Tabla 5.5B Cosenos cuadrados y ángulos correspondientes a los puntos renglón.

Como comentario respecto a las tablas 5.5A y 5.5B, cabe decir que los ángulos respecto al segundo eje son complementarios para el primero. Y, además, los puntos renglón y los puntos columna generan dos gráficas independientes que pueden ser analizados en una sola. Así, los primeros determinan la orientación y configuración para la gráfica referente a renglones y los segundos para la referente a columnas.

Una salida un poco más detallada de la contribución parcial a la inercia se muestra en la tablas 5.6A y 5.6B. Donde nuevamente el punto renglón de "33 a 40 hrs." se destaca por su contribución al primer eje; y, de los puntos columna aparecen

los "Trabajadores por su cuenta", "Empleados u obreros" y "Jornaleros o Peones" como los más importantes. Los resultados para el segundo eje son complementarios, de tal manera que aquellos que tienen mayor importancia en el primer eje no la tienen en el segundo e inversamente, lo que es muy relevante en el segundo eje no lo es en el primero.

Empleado u obrero	0.392084	0.000608
Jornalero o Peón	0.133528	0.725773
Por su cuenta	0.427579	0.232448
Patrón o Empresario	0.004843	0.041008
Fam. sin remuneración	0.041967	0.000163
No especificó		

Tabla 5.6A Contribución parcial a la inercia por parte de los puntos columna.

Horas trabajadas		
0	0.032505	0.000457
< 8	0.065015	0.010319
9 a 16	0.124677	0.053964
17 a 24	0.155315	0.053576
25 a 32	0.023967	0.109564
33 a 40	0.545857	0.124086
41 a 48	0.003096	0.584032
49 a 56	0.000116	0.000003
Más 56	0.049453	0.063999
N. E.		

Tabla 5.6B Contribución parcial a la inercia para los puntos renglón.

De las dos tablas anteriores es de esperarse que no aparezcan los valores referentes a "No especificados" puesto que el modelo se planteó de tal manera que éstos no influyeran en los resultados y esquema final.

En las tablas 5.7A y 5.7B se muestran las coordenadas de la configuración final, tanto para puntos renglón como para puntos columna.

Ubicación en la figura 5.2	[redacted]	[redacted]	[redacted]
A	**Empleado u obrero**	-.107681	-.002542
B	**Jornalero o Peón**	0.145798	0.203836
C	**Por su cuenta**	0.177820	-.078623
D	**Patrón o Empresario**	0.060246	-.105132
E	**Fam. sin remuneración**	0.174588	0.006524
F	**No especificó**	0.388254	-0.003401

Tabla 5.7A Coordenadas finales generadas por el paquete SAS al aplicar análisis de correspondencia sobre columnas.

De éstas últimas dos tablas se construye la figura 5.3 que representa a los puntos renglón y columna para visualizar posibles asociaciones entre ellos.

Ubicación en la gráfica 5.2	[redacted]	[redacted]	[redacted]
1	**0**	0.187716	0.013349
2	**< 8**	0.229652	-.054866
3	**9 a 16**	0.288295	-.113740
4	**17 a 24**	0.263263	-.092722
5	**25 a 32**	0.077900	-.099880
6	**33 a 40**	-.207095	-.059211
7	**41 a 48**	0.012651	0.104206
8	**49 a 56**	-.004272	-.000411
9	**Más 56**	0.078268	-.053394
10	**N. E.**	0.409845	-0.116427°

Tabla 5.7B Configuración final para renglones obtenida mediante el paquete SAS.

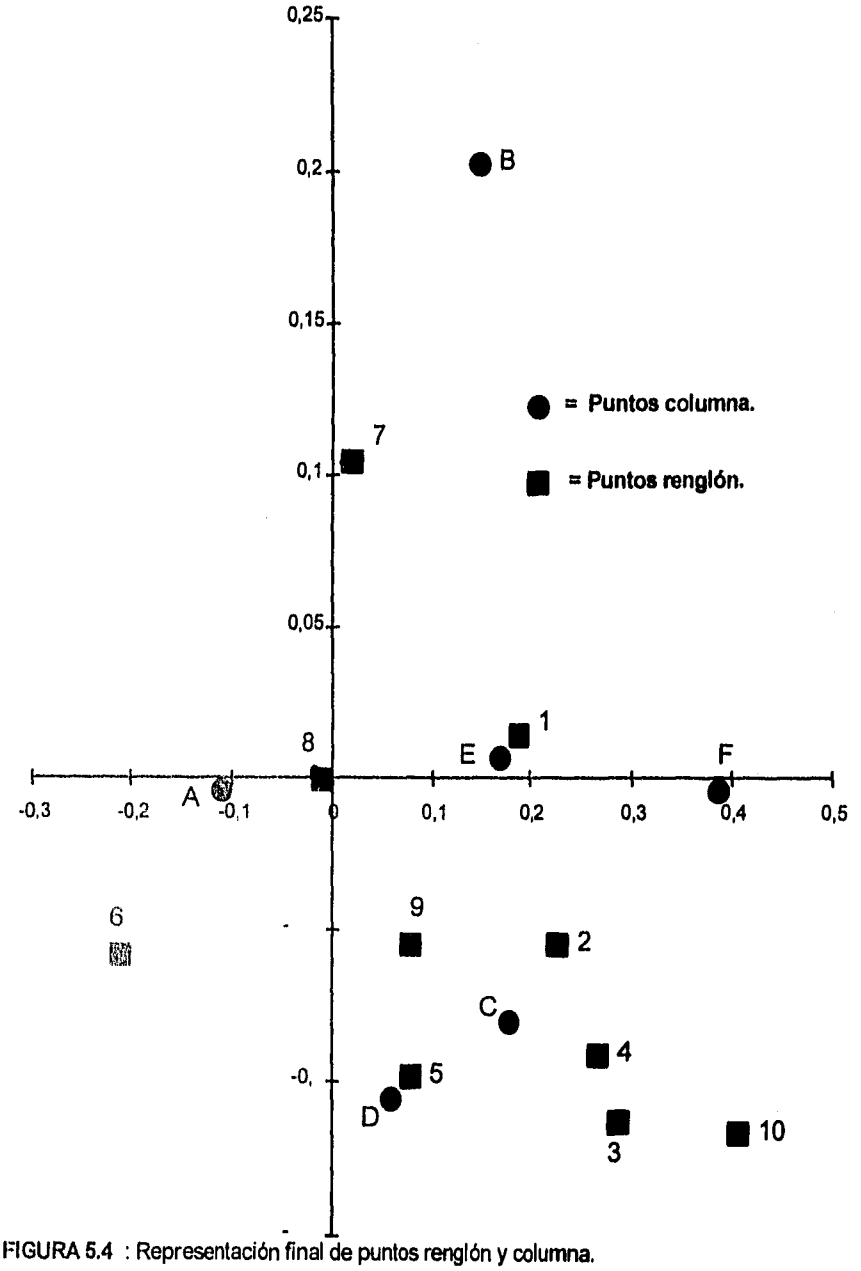


FIGURA 5.4 : Representación final de puntos renglón y columna.

De la figura 5.3 se observa lo siguiente: En la parte superior del primer cuadrante se encuentra como un grupo separado los puntos "41 a 48 hrs." y "Jornaleros o Peones". Es decir que, según los datos obtenidos es factible una asociación entre "Jornaleros y Peones" y personas que trabajen de 41 a 48 horas.

Sobre el eje de las abcisas del lado izquierdo se nota una asociación entre "Empleados u obreros" y personas que trabajan de "49 a 56 hrs.", pero también se puede asociar con "33 a 40 hrs.". Del lado derecho se observa que el punto correspondiente a "Cero horas" se encuentra formando un grupo junto con los que se dedican a trabajos familiares sin remuneración.

En la parte inferior de la gráfica se observa una asociación clara entre "Empresarios o Patrones" y quienes trabajan de "25 a 32 hrs.". Lo que indicaría, hasta aquí, que "Empleados u obreros" y "Jornaleros o peones" trabajan más que los "Empresarios o patrones".

Finalmente, los "Trabajadores por su cuenta" no muestran una relación clara con los puntos renglón, pues gráficamente se ubica dentro de un grupo formado por los puntos columna "Menos de 8 hrs.", "9 a 16 hrs.", "17 a 24 hrs." y "Más de 56 horas". Es decir, podría darse una asociación entre "Trabajadores por su cuenta" y quienes trabajan de 0 a 24 horas o más de 56 horas a la semana. Es decir, la relación se da con quienes trabajan muy poco o con quienes trabajan mucho.

Finalmente, tanto para columnas como para renglones, los que representan a los "No especificados" no muestran ninguna asociación aparente. Además, por la manera en que se realizó el análisis y por lo mencionado anteriormente, de nada serviría (al menos para este análisis) saber la cantidad de horas trabajadas por una

persona si no se sabe a que se dedica; o análogamente, que se supiera el trabajo pero no las horas que se trabajan.

Por lo tanto, puede decirse que el análisis realizado refleja únicamente la información que se considera útil y elimina aquellos datos que podrían cambiar el esquema proporcionando información sin interés.

Como se mencionó al principio de ésta aplicación, lo que primero se hizo fue hacer el análisis dándole el mismo peso a todos los datos, de ahí se generó la figura 5.4 y comparándola con la figura 5.3 se observan las siguientes diferencias:

En la última gráfica (5.4) los puntos no especificados no se asocian claramente a ninguna actividad. Es decir el asociar una actividad no especificada a un determinado número de horas trabajadas o un número sin especificar de horas a una actividad determinada es información que no revela nada y, por lo tanto se elimina del segundo análisis.

La asociación entre empleados u obreros y personas que trabajan de "33 a 40" hrs se conserva, pero se agrega la asociación más clara con "49 a 56" horas. Los puntos renglón referentes a "menos de 8", "9 a 16" y "17 a 24" que tenían ordenadas positivas cambian a ordenadas negativas y sufren una especie de reflexión sobre el primer eje, aunque gráficamente se siguen ubicando en un grupo muy bien especificado.

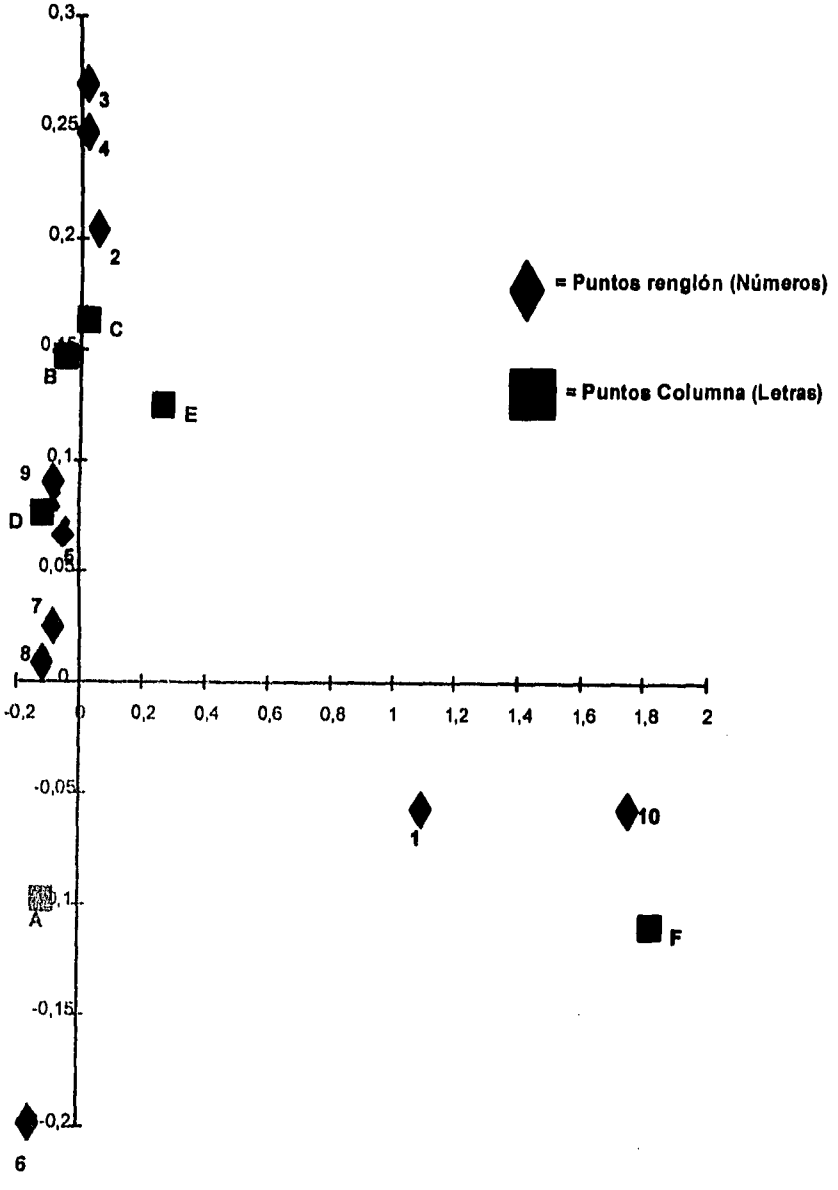
La asociación para Trabajadores por su cuenta aumenta, pues en el primer análisis (figura 5.4) se relacionaba con "menos de 8 horas" y ahora, además lo hace con "más de 56", de "9 a 16" y "17 a 24". Es decir, de los primeros resultados se

hubiera concluido que los Trabajadores por su cuenta se asocian con personas que trabajan muy poco. Ahora, se puede decir , que los Trabajadores por su cuenta se asocian con personas que trabajan poco o con personas que trabajan mucho.

La asociación de Patrones o Empresarios con personas que trabajan más de 56 horas a la semana se pierde en la figura 5.3 o se vuelve menos evidente y se acentúa con aquellas que trabajan de 25 a 32 horas. Así, si de la figura 5.4 se había concluido que este punto columna se asociaba con un número más alto de horas trabajadas que los Empleados u Obreros ahora éstos junto con Jornaleros o Peones se asocian a un número mayor de horas trabajadas.

Finalmente, si en el primer análisis (figura 5.4) no se visualizaba una asociación clara para los trabajadores familiares sin remuneración ahora (figura 5.3) se muestra una con quienes trabajan cero horas. Este dato, aunque parece muy extraño se refleja de la misma información del Censo, es decir que si esta asociación parece confusa habría que analizar la forma en que se realizaron los cuestionarios y ver cómo se tomó la cantidad de horas trabajadas a la semana.

Figura 6.5: Configuración de puntos renglón y columna al dar el mismo peso a todas las observaciones.



CONCLUSIONES.

Al realizar un análisis con técnicas multivariadas sobre un conjunto de datos, se pueden obtener resultados de gran utilidad, ya sea en etapas exploratorias o confirmatorias. Existen ciertas técnicas que se aplican dependiendo del objetivo del análisis, ya sea exploratorio, confirmatorio o ambos.

A partir de las técnicas exploratorias (como métodos gráficos, escalamiento multidimensional, etc.) se obtiene una panorámica general acerca de la forma en que se está comportando el conjunto de datos en estudio con respecto a las variables seleccionadas. Siempre será útil tener una perspectiva inicial para realizar los ajustes que sean necesarios al llevar a cabo un análisis sobre ciertos datos.

Las técnicas exploratorias que se describieron fueron: diversas representaciones gráficas, escalamiento multidimensional, análisis de correspondencia y, en cierta forma componentes principales y conglomerados, los cuales también puede servir como métodos confirmatorios. Aplicando estas técnicas a los datos censales se pudo hacer una distinción entre cada uno de los estados de acuerdo a las variables que se tomaron en la tabla 1.1.

Los análisis efectuados indican que hay estados (Chiapas y Oaxaca por un lado y Nuevo león y el D.F. por otro) que pueden servir como indicadores de la situación general del país, de acuerdo a criterios de tipo económico, laboral y escolar.

En cada uno de los estudios efectuados sobresalen Chiapas y Oaxaca como los estados donde las condiciones de vida son más precarias respecto a los demás estados. En ambos estados se encuentran las siguientes características:

a) Bajos ingresos. Un gran porcentaje de la población no percibe siquiera un salario mínimo completo.

b) Como consecuencia de lo anterior se desprenden varias características: Bajo nivel escolar y escasa población escolar, las actividades se centran en una mal remunerada agricultura.

Con lo anterior, se observa que Chiapas y Oaxaca se alejan del comportamiento medio de los estados de la República.

El D.F. y Nuevo León se alejan un poco del comportamiento medio, pero a un nivel diferente ya que según los resultados obtenidos, en dichos estados se presentan mejores situaciones económicas y escolares.

Las técnicas aplicadas permiten también identificar individuos que se comportan como observaciones discrepantes. Es el caso de Quintana Roo, quien muestra una situación diferente ya que no se acerca a las características de Chiapas, ni a las del D.F. ni a las del común de los estados. Aunque presenta un alto porcentaje de PEA y un nivel de ingresos relativamente alto tiene un bajo porcentaje en cuanto a población escolar.

Los resultados que se obtuvieron pueden ser utilizados si se quiere realizar estudios más profundos o realizar análisis más complejos, pudiendo emplear lo anteriormente descrito como punto de partida.

Como se ha venido mencionando, los métodos que se utilizaron fueron aplicados en computadoras PC compatibles en diversos paquetes estadísticos, lo cual agilizó la aplicación de cada una de las técnicas.

El empleo y desarrollo de software estadístico ha venido tomando un auge en los últimos 5 años, lo que ha permitido que la aplicación de la mayoría de las técnicas de análisis multivariado a diferentes campos sea más frecuente.

Finalmente, se puede decir, que el análisis multivariado es una de las herramientas estadísticas más completas para el análisis de datos, ya que puede ser utilizado para extraer conclusiones que ayuden de manera determinante en la toma de decisiones que involucren a los individuos en cuestión.

APÉNDICE.

A lo largo del presente trabajo se hizo enfásis acerca del uso de algunos paquetes estadísticos. A continuación se da una breve descripción de la manera en que se utilizaron dichos paquetes.

A.1 CSS. (Complete Statistical System)

El CSS es un paquete muy potente y versátil, que tiene una gran variedad de opciones para aplicar a diferentes áreas de la estadística. Se encuentran aplicaciones de estadísitica elemental, análisis multivariado (componentes principales, clusters, escalamiento multidimensional, factores, etc.), manejo de matrices, entre otras.

En este paquete se corrieron las aplicaciones referentes a clusters, escalamiento multidimensional y lo referente a métodos gráficos.

Para utilizar este paquete se teclea desde el prompt el comando: c:> CSS. Al hacerlo se entrara a un menú con diferentes opciones. Cuando se posiciona el cursor en alguna opción aparece un recuadro en la parte inferior izquierda de la pantalla donde se da una explicación de lo que realiza dicha opción. Al digitar la tecla "Enter" se activa la opción elegida.

Generalmente, cada una de las opciones operan del siguiente modo: Se pide una serie de datos, los cuales pueden ser tecleados directamente o importados de algún archivo. (El cual puede tener formato Lotus o ASCII). Una vez hecho lo anterior se procede a ingresar los valores que se solicitan y una vez hecho esto se

ejecutal la aplicación generando la salida correspondiente. Para importar datos se utiliza la opción "**Data managment**".

Este paquete posee, entre otras, las opciones: "**Clusters**" y "**Multidimensional Scaling**". para análisis de conglomerados y para escalamiento multidimensional. En la opción para conglomerados se puede transformar una matriz de datos (raw matrix) a una matriz de distancias o disimilitudes.

Para utilizar métodos gráficos se elige la opción "**Graphics-CSS**".

A.2 STAT-GRAPHICS.

El Stat-graphics es un paquete con menor potencia en cuanto a manejo de datos comparado con CSS. Fue utilizado para la parte de métodos gráficos y componentes principales para utilizar datos en este paquete se pueden teclear directamente o importarlos de archivos ASCII o Lotus. Para correr el paquete se ejecuta el comando: **statgraf.**

Para importar un archivo se procede de la siguiente manera:

1) Copiar el archivo al directorio \statg\data.

2) Entrar a la opción **Data managment.**

3) Dentro de Data managment elegir la opción **importar archivos.**(File import)

4) Una vez que se define el archivo a importar y el archivo en el que se quiere salvar los datos con formato de stat-graphics se oprime la tecla F6.

5) Regresar al menú principal y desarrollar la opción que se desee.

Es importante mencionar que el stat-graphics presenta ciertas limitaciones en cuanto a la cantidad de datos que puede manejar. Por ejemplo, para obtener eigenvectores y eigenvalores de matrices solo puede hacerlo con las que tengan dimensión menor o igual a 9.

A.3 SYSTAT5.

Al igual que en los paquetes anteriores, se permite importar archivos con formato Lotus. Al ejecutar el comando **systat** aparece una pantalla en la que se muestra en la parte superior un menú el cual se puede accesar con las teclas de movimiento del cursor.

Al elegir **Archivo** aparecen opciones que permiten importar o exportar archivos. Se elige la opción de importación de archivos y se digitan los parámetros que se soliciten.

Una vez hecho lo anterior se regresa al menú inicial y se eligen las aplicaciones que se quieran desarrollar con el archivo en uso. Para ello, dentro de esta misma opción se pueden visualizar los archivos disponibles para su utilización.

En los paquetes descritos anteriormente se pueden también introducir los datos directamente desde el teclado.

Por último, en lo que respecta a la impresión de los resultados que se obtengan, los paquetes anteriores tienen una opción para mandar a imprimir lo que se tenga en pantalla o para salvar determinada información en otros archivos.

A.4 SAS (Statistical Analysis System)

El paquete estadistico SAS es uno de los más completos que hay. Es muy versátil y flexible, además de que tiene una eficiencia bastante alta.

Una vez que se ha iniciado el sistema, aparecen en la pantalla 3 ventanas: Una para editar texto, otra para desplegar salidas de datos y una más para desplegar errores.

Para aplicar algún método multivariado a datos que se encuentren en archivos o sean proporcionados desde el teclado se realiza un programa en el cual se hace uso de todas las funciones que tiene habilitadas SAS para ello. La longitud del programa dependerá de la complejidad del análisis y de los datos que se deseen.

La estructura de dicho programa consiste en construir un "dataset", posteriormente aplicarle las funciones y procedimientos que se encuentran disponiblen en el paquete para su análisis. La sintaxis de lo anterior puede verse en el menú de ayuda que se obtiene al oprimir la tecla F1.

Una vez que se tiene el programa se corre con la tecla F3. Si hubo errores de sintaxis, referencias a archivos no existentes o cualquier otro tipo de error éste se desplagará en la ventana de errores. Si el programa fue correctamente escrito y genera una salida de datos, ésta se desplegará en la ventana de salida (output) en caso de que no se direccione a otro lado (a un archivo o a la impresora).

Oprimiendo la tecla F2 aparece una ventana que indica la función que tienen cada una de las teclas, además es útil para moverse de ventana en ventana.

Si al estar situado en cualquiera de las tres ventanas se desea enviar el contenido de alguna de ellas a un archivo o a la impresora se digita el comando en la primera linea señalada con "Command ==>".

Cuando se corre un programa de la ventana de edición, al regresar a ella se encuentra vacía y para visualizar el programa que se tenía se oprime la tecla F9, con lo que el programa (que se guarda en la memoria) aparece de nuevo en la ventana para ser modificado o almacenado.

Actualmente existen dos versiones de SAS, una par DOS y otra para Windows. Las diferencias entre uno y otro son mínimas, pero hay que tener cuidado con los cambios de sintaxis de uno a otro. Aunque cuando se corra un programa de una versión DOS en Windows (o viceversa) la ventana de errores indicará que es lo que falta o está de más para que el programa se ejecute exitosamente.

A.5 PRUEBAS REALIZADAS AL PAQUETE ESTADÍSTICO CSS.

El paquete estadístico que más importancia tuvo en el desarrollo de la tesis fue el CSS (descrito en el apéndice), ya que se utilizó más que los otros paquetes descritos debido a su facilidad de manejo y a su gran variedad de opciones estadísticas, además de que se acoplaba perfectamente a los objetivos que se pretendía alcanzar.

Se realizaron varias pruebas en los módulos que se utilizaron para poder detectar posibles fallas que pusieran en duda la total confiabilidad del sistema, las conclusiones que se obtuvieron fueron:

- Los datos erróneos que manda el sistema ocurren únicamente cuando se proporcionan cantidades que, por sus características, caen fuera de la capacidad de procesamiento de la computadora, es decir, cuando se supera la cantidad de decimales que puede manejar la computadora.

En este caso, los errores cometidos son inherentes a la capacidad de redondeo que tenga la máquina. Dichos errores se producen en cualquier ámbito computacional y no dependen de la calidad del paquete utilizado sino, más bien, de la potencia y capacidad de la computadora.

Cuando CSS detecta datos que ya no puede manejar puede ocurrir que mande un error y no pueda completar la corrida o que proporcione datos disparatados (cómo pueden ser puros ceros o unos). A este tipo de problemas se les conoce como

underflow, cuando se majena un número tan pequeño que no puede ser porcesado y *overflow* cunado se proporciona un número demasiado grande.

- Los módulos que se utilizaron como son Componentes Principales, Análisis de Conglomerados y Escalamiento Multidimensional utilizan entre sus principales procesos inversión y escalamiento de matrices, empleando para ello algoritmos muy generales e implementados en la mayoría de los paquetes que manejen matrices (hablando de paquetería estadística, matemática o de ingeniería), razón por la cual los problemas que se encuentren en dichos módulos por manejo de matrices serán los que se encuentren en la mayoría de los paquetes computacionales.

A este respecto cabe hacer notar que el paquete (y, en general cualquier paquete que maneje matrices) emitirá datos erróneos cuando, por ejemplo, al tratar de invertir alguna matriz se proporcionen renglones o columnas "casi" linealmente dependientes. El "casi" se refiere a datos, que por su longitud de decimales y su magnitud hagan que al ser redondeados o truncados provoquen como resultado renglones o columnas linealmente dependientes y al ser manejados por la computadora provoque los errores antes mencionados.

De esta manera, los errores que se detectaron se asocian con la capacidad del sistema de punto flotante que maneje la computadora y no con fallas en la programación del mismo.

Se intentó introducir ejemplos que fueran un poco "capciosos" y que provocaran algún error en el paquete, pero éste respondió con los mensajes ya descritos.

Por otro lado, para el capítulo de Análisis de Correspondencia se utilizó el paquete SAS, obteniéndose los mismos resultados. A este respecto, cabe señalar que este paquete cuenta con un gran prestigio dentro del ámbito estadístico y computacional, existiendo en Estados Unidos un Instituto surgido a raíz de la creación de este Instituto. Lo anterior, aunque no garantiza el 100% de efectividad del paquete, indica que existe una mayor especialización de las personas que lo programan, haciendo con ello más difícil la detección de errores.

BIBLIOGRAFÍA

[1] Bartlet, Robert G.
"Introducción al análisis matemático"
Limusa.
México 1989.

[2] Benzécri, Jean Paul.
"Correspondence Analysis Handbook"
Decker.
EUA 1989

[3] **Códice 90.**
Editado por el Instituto Nacional de Estadística Geografía e Informática (INEGI).
México D.F. 1993

[4] Chernoff.
"About Chernoff faces".
Revista **JASA** vol. 68, año 73, página 361.
E.U.A.

[5] Chatfield S y Collins.
"Introduction to Multivariate Analysis"
Limusa.
EUA 1984.

[6] Diloria, Frank C.
"SAS Applications **Programming (A gentle introduction)"**
PWS-KENT.
EUA 1991.

[7] Drapper-Smith.
"Applied Regresion Analysis"
John Wiley & Sons.
EUA 1966.

[8] Everitt Brian S. -Graham Dum.
"Applied Multivariate Data Analysis"
De. Edward Arnold.
EUA 1984.

[9] Greenacre Michael J.
"Theory and applications of correspondence analysis"
ISBN.
EUA 1984.

[10] Hoaglin David C. /Tukey John W. /Mosteller Frederick.
"Understanding Robust and Exploratory Data Analysis"
John Wiley & Sons.
EUA 1983

[11] Jollife Y. T.
"Principal Component Analysis"
Springer Series in Statistics.
EUA 1986.

[12] Krzanowsky W. J.
"Principles of Multivariate Analysis"
Oxford Science Publications.
Inglaterra 1990.

[13] Marriot F.H.C.
"The interpretation of multiple observations"
Academmic Press Inc.
Londres. 1974