

01168



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

**DIVISIÓN DE ESTUDIOS DE POSGRADO
FACULTAD DE INGENIERÍA**

**APLICACIÓN DEL ANÁLISIS MULTIVARIADO A
LOS DATOS DEL CENSO DE 2000 EN MÉXICO**

T E S I S

PRESENTADA POR:

ADRIÁN AVALOS PADILLA

PARA OBTENER EL GRADO DE:

MAESTRO EN INGENIERÍA

(INVESTIGACIÓN DE OPERACIONES)

DIRIGIDA POR:

M.I. RUBÉN TÉLLEZ SÁNCHEZ



CIUDAD UNIVERSITARIA

ENERO 2005

m 340279



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

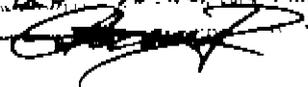
Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Autorizo a la Dirección General de Bibliotecas de la UNAM y al personal que se encuentre a cargo el contenido de la obra que se encuentra en custodia.

NOMBRE: Adrián Avalos Padilla

FECHA: 20 - Enero - 2005

FIRMA: 

ÍNDICE

INTRODUCCIÓN	5
ANTECEDENTES	5
PROBLEMÁTICA	6
OBJETIVOS	6
HIPÓTESIS	7
ESTRUCTURA DEL PRESENTE TRABAJO	8
CAPÍTULO 1	9
LOS CENSOS EN MÉXICO	9
UN POCO DE HISTORIA	9
CUADRO 1. CENSOS Y CONTEOS REALIZADOS EN MÉXICO	10
CENSO DE 1990	12
ALGUNOS COMPARATIVOS	12
CENSOS EN AMÉRICA LATINA	13
CUADRO 2. FECHA DE LOS CENSOS DE AMÉRICA LATINA, 1950-2000	14
TRADICIÓN CENSAL EN MÉXICO	15
POBLACIÓN EXISTENTE	15
CAPÍTULO 2.	18
DESARROLLO CENSAL EN MÉXICO	18
METODOLOGÍA DEL CENSO	18
UNIVERSO DE DATOS DISPONIBLES	19
CUADROS BASE DE INFORMACIÓN	20
MÉTODOS MULTIVARIADOS A EMPLEAR	22
¿QUÉ ES EL ANÁLISIS MULTIVARIADO?	23
CAPÍTULO 3.	25
ANÁLISIS DE CONGLOMERADOS	25
OBJETIVO	25
ELEMENTOS NECESARIOS PARA APLICAR EL ANÁLISIS DE CONGLOMERADOS	25
MÉTODOS JERÁRQUICOS	26
MÉTODOS NO JERÁRQUICOS	27
APLICACIÓN A LOS DATOS DEL CENSO	27
RESULTADOS OBTENIDOS EN EL PAQUETE SPSS	30

ANÁLISIS DE RESULTADOS	34
CONSIDERACIONES ADICIONALES	35
EJEMPLO ADICIONAL	41
CAPÍTULO 4.	47
ANÁLISIS FACTORIAL	47
OBJETIVO	47
ANTECEDENTES	47
MODELO BÁSICO DE FACTORES	48
TIPOS DE MODELOS FACTORIALES	48
ANÁLISIS DE COMPONENTES PRINCIPALES	48
PASOS PARA LA ELABORACIÓN DEL ANÁLISIS FACTORIAL	49
OBTENCIÓN DE LAS COMPONENTES PRINCIPALES	50
APLICACIÓN A LOS DATOS DEL CENSO	51
RESULTADOS OBTENIDOS CON EL PAQUETE ESTADÍSTICO SPSS	54
EJEMPLO ADICIONAL	62
CAPÍTULO 5.	77
ESCALAMIENTO MULTIDIMENSIONAL	77
OBJETIVO	77
DEBARROLLO	77
ESCALAMIENTO CLÁSICO	77
ESCALAMIENTO MÉTRICO Y NO MÉTRICO	79
APLICACIÓN A LOS DATOS DEL CENSO	80
RESULTADOS OBTENIDOS AL REALIZAR EL TRATAMIENTO	83
CAPÍTULO 6	88
RESULTADOS DE LOS MÉTODOS ANALIZADOS	88
ANÁLISIS DE CONGLOMERADOS	88
ANÁLISIS FACTORIAL Y CONGLOMERADOS	88
CASO DISTRITO FEDERAL	89
GRUPOS	90
¿CÓMO CONTINUAR?	91
CAPÍTULO 7	93
COMPARATIVO CON LO OBTENIDO EN 1990	93
CONCLUSIONES Y RECOMENDACIONES	94
APÉNDICE	97

OTRAS APLICACIONES	97
TASA DE PIRATERÍA DE SOFTWARE A NIVEL MUNDIAL	97
GASTO EN TECNOLOGÍAS DE INFORMACIÓN	101
BIBLIOGRAFÍA	106

Introducción

Antecedentes

En México, cada 10 años se realiza un censo de la población con el fin de obtener información estadística relevante que muestre la situación del país en un momento determinado del tiempo.

Aunque el censo no es la única fuente de información sobre la población, es la más importante y se complementa con conteos generales, información a nivel estatal o municipal y encuestas que se aplican de manera general de acuerdo a la metodología del Instituto Nacional de Estadística, Geografía e Informática (INEGI), en el ámbito nacional o regional.

Estos censos se han llevado a cabo desde 1895 y la manera de aplicarlos y analizar la información generada ha evolucionado a lo largo del tiempo. Aunque antes de ese año se habían realizado recuentos poblacionales, el primero del que se tiene conocimiento fue en el año 1116, en la época prehispánica.

El antecedente inmediato al censo de 1895 fue un conteo que se llevó a cabo en 1831. Sin embargo, con la creación de la Dirección General de Estadística en 1882 se sentaron las bases para que el censo de 1895 fuera llevado a cabo; y no se limitó a contar a la población sino que intentó consolidar información más allá del simple conteo y fue el inicio de una tradición censal en México, la cual no existía anteriormente.

De ese entonces a la fecha se han llevado a cabo 12 censos (incluido el del 2000), un conteo de población en 1995 y en 1997 una encuesta nacional de dinámica demográfica.

En 1990, el INEGI generó un CD con todos los resultados del censo de ese año. A ese CD se le nombró *código-90*, al momento de escribir el presente trabajo todavía no se había generado un CD con las mismas características para el censo del año 2000, sin embargo, los resultados tabulares por estado ya se han publicado en la página de Internet del INEGI y es de ahí de donde se provee el presente documento de la información que se analizará. La dirección de dicha página es <http://www.inegi.gob.mx>. El uso de la difusión a través de este medio es una característica novedosa que en ningún censo o conteo anterior se había llevado a cabo de manera inmediata.

El censo de 1990, junto con el conteo general de población y vivienda realizado en 1995 son los antecedentes inmediatos del censo del año 2000, la velocidad de procesamiento se va acelerando en cada período y se aprovechan las

experiencias del pasado para generar los resultados definitivos en menor tiempo y con mayor calidad para su explotación.

Problemática

El hecho de contar con información relativa al Censo de Población y Vivienda nos enfrenta, prácticamente, a un mundo de información que muchas veces es difícil de explotar para fines específicos. Dada la extensión y alcances de los datos obtenidos, existe la posibilidad de perderse en números que por sí solos no representan ninguna información valiosa.

En otros aspectos, la utilización de esos datos para la obtención de resultados que expliquen lo que el país está viviendo en un momento determinado del tiempo, es tarea laboriosa que requiere enfocar adecuadamente la información vertida en los resultados tabulares del Censo.

Asimismo, identificar en base a esos datos cuál o cuáles son las características de las diferentes regiones y estados que conforman el país es de suma utilidad y es posible realizarlo a partir de la información censal, siempre y cuando se utilicen los mecanismos adecuados de explotación.

Por último, la población en general únicamente conoce los resultados a un nivel de estadística descriptiva, y en ocasiones la información relativa al comportamiento que guardan cada uno de los estados con respecto a los demás es poco utilizada, ya que generalmente se emplean los datos que se obtienen a nivel local.

Objetivos

El objetivo del presente trabajo es la aplicación del ANÁLISIS MULTIVARIADO a los datos del censo de México del año 2000, y a partir de ello, obtener información que sea útil en la toma de decisiones o que revele perfiles importantes de la situación del país en cada una de sus entidades federativas.

El énfasis que se dio a lo largo de esta tesis es la relevancia que tiene la información y su tratamiento, así como las conclusiones que se pueden obtener a partir de ella. Sin embargo, también se explican las técnicas multivariadas, detallando sus conceptos esenciales y su contexto estadístico y matemático sin llegar a la profundidad de análisis minucioso, ya que para ello se necesitaría dedicar un trabajo completo a cada uno de los tópicos que aquí se tratan.

En este escrito, más bien, lo que se desea es que el lector conozca las técnicas de análisis multivariado, su utilidad en casos reales para una adecuada explotación de la información y que sirva de apoyo para la toma de decisiones a nivel país.

Dentro de este estudio, no se pretende realizar una estadística descriptiva de los datos del censo, el cual entenderemos a los largo del trabajo como "censo

realizado en México en el año 2000. La idea es, en todo caso, realizar un estudio comparativo entre las diferentes entidades federativas que conforman la República Mexicana. Sin embargo, considero que la estadística descriptiva siempre es un buen parámetro de comparación o un punto de partida, por ello se hará un esbozo de la estadística descriptiva que se genera a partir de los datos censales.

Dentro de los objetivos que también se pretenden alcanzar con esta tesis, se encuentra la conformación de grupos o conjuntos, cuyos elementos serán los estados del país, donde cada conjunto represente una región socio-económico-demográfica, en incluso geográfica si es posible, del país con elementos comunes a cada uno de los estados pertenecientes a ese grupo.

Con lo anterior, se estima posible la asignación de recursos del ámbito federal a aquellos estados o regiones del país que, de acuerdo, a los resultados obtenidos se encuentren en condiciones inferiores o de desventaja respecto a las demás regiones o estados del país. Esto también podría servir para la propuesta de programas sectoriales de desarrollo o la adecuación de los existentes y permitir, así, una mejor canalización de los recursos financieros y de infraestructura que se asignan a cada región o estado del país.

Hipótesis

Atendiendo a uno de los objetivos primordiales de este trabajo, que es la conformación de regiones o grupos de estado de acuerdo a los niveles generales de vida que privan en cada uno de ellos, se plantea el supuesto que se obtendrán al menos tres conjuntos.

En el primer grupo, se encontrarán los estados con niveles generales de vida de mayor calidad: mayor alfabetismo, condiciones de vivienda, servicios, etc. Supongo que estados, como el D.F. (si lo consideramos como tal), Nuevo León, Jalisco y Quintana Roo se encontrarán en este subconjunto.

En el segundo grupo se encontrarán estados con los niveles generales de vida más precarios. A este grupo podrían pertenecer Chiapas, Oaxaca, Guerrero y Zacatecas.

Existirá un tercer grupo con estados que presentan carencias en algunos aspectos, pero que en otros pueden ser más sólidos. A este conjunto quizá ingresen Aguascalientes, Tabasco y Veracruz.

Es muy difícil suponer la conformación total de las regiones o grupos de estados debido a que las variables a analizar son numerosas y el comportamiento se espera con una varianza muy alta, lo que dificulta la conformación de los conjuntos pretendidos.

También se espera que existan variables que resultarán muy importantes y significativas y que, por su propia naturaleza, sean imprescindibles para el análisis pretendido, dichas variables serían las de educación, económicas y de salud, por lo que se complementarán con variables tal vez menos significativas pero que ayuden a redondear dicho análisis.

En resumen, se espera encontrar grupos o conjuntos, en los que los elementos (estados) cuenten con características homogéneas en cuanto las variables de decisión, presentando cierta independencia de la ubicación geográfica.

Estructura del presente trabajo

Esta tesis pretende presentar en su primera parte el marco referencial y el teórico de lo relativo al análisis de los datos censales referidos. Posteriormente, se aplicarán las técnicas del análisis multivariado, y cada capítulo se conformará de la siguiente manera:

1. Una breve descripción de la técnica y el algoritmo utilizado para la ejemplificación de los datos censales.
2. Aplicación a datos del censo previamente tratados para utilizar algún paquete específico.
3. Análisis de los resultados y conclusiones o inferencias acerca del comportamiento de los datos.

Al final se presentarán algunas conclusiones y recomendaciones derivadas de la aplicación de los métodos multivariados al censo del 2000.

Capítulo 1

Los Censos en México

Como se mencionó en la introducción, el antecedente inmediato del censo del año 2000 es el conteo de población y vivienda de 1995, y el censo inmediato anterior es el realizado en 1990. De esos años a la fecha, los procedimientos han evolucionado significativamente resultando más rápidos y confiables. Desde luego el apoyo de herramientas computacionales ha influido enormemente en agilizar el proceso de preparación, levantamiento, procesamiento y generación de resultados definitivos.

Un poco de historia

En el cuadro 1 se muestra la información relativa a los censos o conteos poblacionales que se han llevado a cabo en nuestro país a lo largo de nuestra historia.

Del referido cuadro se observa una regularidad en la realización de estos trabajos, lo que bien puede llevar a decir que se ha creado una cultura en lo referente a Censos en nuestro país. De lo cual, el INEGI ha sido pilar en el desarrollo y permanencia de los trabajos destinados a obtener información que indiquen cuántos y cómo somos los mexicanos.

Si bien es cierto que hasta 1921 se llevaba a cabo el levantamiento mediante autoempadronamiento, lo cual generaba bastantes errores de origen que perturbaban la veracidad de la información, las técnicas y procedimientos para levantar la información se han ido refinando con el paso del tiempo, y las experiencias adquiridas a lo largo de los años, permiten que cada censo sea llevado en menos tiempo y con mayor calidad de información que los anteriores.

De igual manera, han evolucionado las boletas y los cuestionarios. Se aprecia en el mismo cuadro que hasta 1980 el levantamiento se realizaba en un solo día. En ese año se emplearon más de un millón de empadronadores. En 1990, se observa que el período de levantamiento aumenta a una semana, lo que se hizo pretendiendo reducir errores en el procesamiento de información e iniciando con el análisis previo en cada estado para, posteriormente, consolidarlo, en ese 1990 se emplearon cerca de 500 mil encuestadores. Para el año 2000 el período se amplió a dos semanas, utilizando cerca de 260 mil empadronadores. En esta ocasión, por primera vez se les remuneró económicamente, de igual forma, recibieron una capacitación mayor lo que permitió agilizar de manera importante el proceso de análisis y tratamiento de los datos. De esta manera, aunque se tarda una semana más en levantar las encuestas, los resultados definitivos se obtienen con muchos meses de anterioridad si se compara con años anteriores.

Cuadro 1. Censos y conteos realizados en México

Año	Nombre oficial	Fecha de levantamiento	Fecha censal	Tipo de boleta	Forma de empadronamiento
1895 ¹	Censo General de la República Mexicana	20 octubre	20 octubre	Familiar	Autoempadronamiento
1900 ²	Censo General de la República Mexicana	28 octubre	28 octubre	Familiar	Autoempadronamiento
1910 ³	Tercer Censo de Población de los Estados Unidos Mexicanos	27 octubre	27 octubre	Familiar	Autoempadronamiento
1921 ⁴	Censo General de Habitantes	30 noviembre	30 noviembre	Familiar	Autoempadronamiento
1930	Quinto Censo de Población	15 mayo	15 mayo	Colectiva (100 hab.)	Entrevista
1940	Sexto Censo de Población 1940	6 marzo	6 marzo	Colectiva (80 hab.)	Entrevista
1950	Séptimo Censo General de Población	6 junio	6 junio	Colectiva (45 hab.)	Entrevista
1960	VIII Censo General de Población, 1960	8 junio	8 junio	Colectiva (50 hab.)	Entrevista

1970	IX Censo General de Población, 1970	28 enero	28 enero	Por vivienda (14 hab.)	Entrevista
1980	X Censo General de Población y Vivienda, 1980	4 junio	4 junio	Cuadernillo por vivienda	Entrevista
1990	XI Censo General de Población y Vivienda, 1990	12 al 16 de marzo	12 marzo	Cuadernillo por vivienda	Entrevista
1995 ^a	Conteo de Población y Vivienda 1995	23 de octubre al 2 de diciembre	5 noviembre	Cuestionario por vivienda (básico y ampliado)	Entrevista
2000 ^b	XII Censo General de Población y Vivienda, 2000	7 al 18 de febrero	14 febrero	Cuestionario por vivienda (básico y ampliado)	Entrevista

¹ Se dividió la población en tres categorías: Población presente, de paso y ausente, usando una boleta para cada tipo.

² Se dividió a la población en presente, de paso y ausente, usando tres boletas.

³ Se usó una sola boleta y se realizó el censo de edificios como labor previa y no simultánea como en 1895 y 1900; se clasificó por primera vez a la población urbana y rural.

⁴ Se introdujo el municipio como división administrativa de las entidades federativas.

⁵ Este es un recuento intercensal en el que se combinan las características de un censo y una encuesta. La enumeración se realizó del 23 de octubre al 18 de noviembre y la encuesta, del 6 de noviembre al 2 de diciembre.

⁶ Se realizó una encuesta a la par en cuyas viviendas seleccionadas se aplicó el cuestionario ampliado.

Censo de 1990

Como se ha mencionado, el censo inmediato anterior se llevó a cabo en 1990, y una de las ventajas del avance tecnológico es la conservación de los datos históricos y su comparación con información actualizada.

A continuación, se presentan algunos de los resultados más relevantes que se obtuvieron con el censo de 1990.

En ese año, se contabilizó la población del país en 81.2 millones de habitantes.

Existían cuatro estados, que representaban los polos opuestos de la configuración de los estados de la República Mexicana: en el extremo inferior se encontraban Chiapas y Oaxaca; en el extremo superior estaban el D.F. y Nuevo León.

Ambos grupos se establecían como indicadores y puntos comparativos para determinar el grado de nivel de vida de cada entidad federativa. Así, si se comentaba que determinado estado se "parecía" más a Oaxaca y otro a Nuevo León, estaba dándose por sentado, una identificación de cada uno.

Las características generales del grupo inferior eran: Bajos ingresos dentro de la población, lo que generaba problemas migratorios y se relacionaba con escasa población escolar, esto explica muchas de las consecuencias de tipo social que se dieron en la década de los 90's.

Algunos estados, Quintana Roo entre ellos, se comportaban como observaciones discrepantes, pues presentaba un nivel de ingresos alto, pero baja población escolar. Además que, en aquel entonces, era uno de los estados menos poblados del país. En aquel año ya existía una alta actividad turística en ese estado y esta era la principal fuente de ingresos de la población.

Algunos comparativos

Uno de los censos, tal vez más significativos en cuanto a la información que se obtuvo de ellos es el de 1970. Una de las razones principales es que en ese año se identificó un crecimiento relativo de la población negativo. Es decir, por primera vez desde 1910 se empezaba a obtener un éxito relativo en el control de la natalidad, aunque este fuera pequeño.

Lo anterior no quiere decir que la población absoluta disminuyera, sino que el crecimiento comparativo con años anteriores fue menor.

También es importante el comportamiento que ha tenido la población a lo largo del tiempo. De 1990 hacia atrás, la distribución de la población por edades se comportaba como una pirámide. Sin embargo, a lo largo del tiempo se ha ido estrechando en la base y ensanchando en las edades mayores a 15 años, lo que indica que se ha iniciado un proceso de envejecimiento de la población.

La migración ha ido creciendo de 1970 al 2000. Es decir, en términos generales, la población que no permanece en sus lugares de nacimiento ha aumentado.

El porcentaje de viviendas que cuenta con los servicios básicos de agua entubada, drenaje y energía eléctrica ha ido incrementándose, lo que parece indicar que cada vez se alcanza a cubrir a mayor cantidad de personas con este tipo de servicios.

En algunos aspectos y de manera general, se ha mejorado la prestación de los servicios a la población, el alfabetismo ha aumentado en términos porcentuales, el promedio de años de estudio en el país pasó de 3.4 en 1970 a 7.6 en el 2000, lo que muestra a una población con mayores niveles de estudio.

Sin embargo, no habría que limitarse a los números y contrastar los aspectos sociales y la percepción de la sociedad respecto a la realidad. Tal vez, los números indiquen que se ha mejorado en el aspecto social, pero si la percepción es que en el terreno económico se ha deteriorado el poder adquisitivo de la población, esto trae como consecuencia un nivel de vida menor. Aunque este es uno solo de los aspectos debido a que el análisis demográfico del país involucra considerar muchos aspectos que tienen que ver con economía, psicología social, seguridad social, sociología, finanzas nacionales, etc.

Otro de los aspectos que también es importante considerar es el comparativo con respecto a otros países. Cuando se presenta este planteamiento, generalmente surge el interés de compararse con los Estados Unidos. En este caso, convendría compararse con países que tengan factores comunes a los nuestros, y es en los países de América Latina donde mejores afinidades se pueden encontrar y de donde es posible extraer elementos que permitan tomar decisiones para mejorar los niveles de vida del país.

Censos en América Latina

A continuación se presenta el cuadro 2, que contiene las fechas en las que se han realizado los censos en algunos países de América Latina. Dicho cuadro fue obtenido de la revista Demos del año 2000.

Cuadro 2. Fecha de los censos de América Latina, 1950-2000

<i>País</i>	<i>50's</i>	<i>60's</i>	<i>70's</i>	<i>80's</i>	<i>90's</i>	<i>2000 (estimado)</i>
Argentina	10-may-1947	30-sep-1960	30-sep-1970	22-oct-1980	15-may-1991	2001
Bolivia	5-sep-1950	-	29-sep-1976	-	3-jun-1992	2001
Brasil	1-jul-1950	1-sep-1960	1-sep-1970	1-sep-1980	1-sep-1991	1-ago-2000
Colombia	9-may-1951	15-jul-1964	24-oct-1973	15-oct-1985	24-oct-1993	2002
Costa Rica	22-may-1950	31-mar-1963	14-may-1973	11-jun-1984	-	28-jun-2000
Cuba	28-ene-1953	-	6-sep-1970	11-sep-1981	-	2002
Chile	24-abr-1952	29-nov-1960	22-abr-1970	21-abr-1982	22-abr-1992	2002
Ecuador	29-nov-1950	25-nov-1962	8-jun-1974	28-nov-1982	25-nov-1990	2001
El Salvador	13-jun-1950	2-may-1961	28-jun-1971	-	27-sep-1992	2002
Guatemala	18-abr-1950	18-abr-1964	26-mar-1973	23-mar-1981	17-abr-1994	2004
Haití	7-ago-1950	-	31-ago-1971	30-ago-1982	-	2001
Honduras	18-jun-1950	17-abr-1961	6-mar-1974	29-may-1988	-	2001
México	6-jun-1950	8-jun-1960	28-jun-1970	4-jun-1980	12-mar-1990	14-feb-2000
Nicaragua	31-may-1950	31-may-1963	20-abr-1971	-	25-abr-1995	2005
Panamá	10-dic-1950	11-dic-1960	10-may-1970	11-may-1980	13-may-1990	14-may-2000
Paraguay	28-oct-1950	14-oct-1962	9-jul-1972	11-jul-1982	26-ago-1992	2002
Perú	23-abr-1905	2-feb-1961	4-jun-1972	12-jul-1981	11-jul-1993	2001
Rep. Dom.	6-ago-1950	7-ago-1960	9-ene-1970	12-dic-1981	24-sep-1993	2001
Uruguay	22-mar-1905	16-oct-1963	21-may-1975	23-oct-1985	22-may-1996	2004
Venezuela	26-nov-1950	26-feb-1961	2-nov-1971	20-oct-1981	21-oct-1990	2001

De este cuadro se observa que México y Panamá son los únicos países que han sido constantes en la elaboración de censos cada 10 años exactamente.

Se observa un interés muy remarcado en la elaboración de censos en América Latina, lo que indica que los países están interesados en conocer la forma en la que han ido evolucionando.

Tradición censal en México

A través del tiempo se han definido estrategias propias del INEGI y se aprovechan las experiencias que se tienen al ejecutar cada censo para mejorar el siguiente. De esta manera, la información generada tiene mayor rapidez de procesamiento y sirve para tomar decisiones gubernamentales que impactan a toda la sociedad en su conjunto

Por lo mismo, la manera de elaborar los cuestionarios, de capacitar a los encuestadores y la logística de operación es mejorada durante cada censo, debido a que se aprovecha la experiencia adquirida. Es importante resaltar que la cantidad de tiempo para levantar los datos, procesar la información y generar los resultados ha disminuido sensiblemente. En el 2000 se levantó la información durante febrero, y para julio ya se tenían los primeros resultados preliminares, lo que habla de la reducción de tiempos.

Por lo anterior, y a manera de preámbulo, se enlistan los principales resultados de la estadística descriptiva que el INEGI publicó en su página de internet y que dio a conocer a los medios de comunicación. Dichos resultados, conocidos como *preliminares* se convirtieron en *definitivos* y dan idea de la situación general del país. El objetivo, es particularizar esos resultados generales y obtener información específica sobre cada estado del país.

PRINCIPALES RESULTADOS DEL CENSO DEL 2000

Población existente	97.5 millones de habitantes
Lugar entre los países con más población	11
Estado del país con mayor población	Estado de México con 13.1 millones de habitantes
Estado menos poblado	Baja California Sur con 0.4 millones de habitantes
Porcentaje de hombres	48.8% (47.58 millones)
Porcentaje de mujeres	51.2% (49.92 millones)
Población mayor a 5 años que habla una lengua Indígena	6 millones
Población mayor a 5 años que sólo habla una lengua Indígena	1 millón
Estados con mayor población de hablantes de alguna lengua	Oaxaca y Yucatán

<i>indígena</i>	
<i>Estado con mayor porcentaje de No católicos</i>	Chiapas
<i>Estado con mayor índice alfabetismo</i>	Distrito Federal con 97.1%
<i>Estado con menor índice de alfabetismo</i>	Chiapas con 77%
<i>Promedio de años de estudio de la población</i>	7.6 años
<i>Actividades laborales con predominio de hombres</i>	Operadores de transporte, actividades agropecuarias y vigilancia
<i>Actividades laborales con predominio de mujeres</i>	Trabajadoras domésticas, oficinistas y trabajadoras de la educación
<i>Personas con alguna discapacidad</i>	1.8 millones
<i>Porcentaje de hogares que cuentan con aparato de TV</i>	86.6%
<i>Porcentaje de hogares que cuentan con computadora</i>	9.5%
<i>Porcentaje de hogares que cuenta con energía eléctrica</i>	95.4%
<i>Porcentaje de hogares cuyo jefe de familia es hombre</i>	78.94%
<i>Porcentaje de la población que no es derechohabiente de alguna Institución de seguridad social</i>	57%

Los datos anteriores nos describen como país, dan información que refleja la situación actual. Esta información nos permite identificar algunos estados que sobresalen por tener alguna característica muy particular. Sin embargo, esta información es una parte muy pequeña de todo el universo de variables que obtuvo el INEGI. Se podrían construir muchísimas tablas semejantes a la anterior, dependiendo de las variables en la que se tenga interés.

El contar con elementos que permitan comparar las diferencias o semejanzas que existen entre los 32 estados del país nos puede generar una visión de "grupos",

cada uno de ellos formado por estados que cuentan con características semejantes y que no necesariamente tienen por qué pertenecer a la misma región geográfica.

En este momento es importante mencionar que a lo largo de este trabajo se entenderá por "estado" a cada una de las entidades federativas del país.

En la presentación que hizo el INEGI para mostrar los resultados definitivos del censo realizó algunas comparaciones con los censos de 1970 a 1990. Dichas comparaciones nos dan idea de la forma en que hemos cambiado como país.

Aunque dichas comparaciones son interesantes por sí mismas, este trabajo no pretende hacer un comparativo con datos históricos, más bien pretende hacer un comparativo actual y determinar la situación que reflejan los números obtenidos y tabulados.

El INEGI menciona que uno de los usos que puede tener el censo es la asignación de recursos para programas presupuestales. Quizá si se determinara que el índice de analfabetismo es alto en un estado en particular, tal vez se podría decidir asignar más recursos al ramo de la educación en ese estado (¿Chiapas?). Además, esos recursos se podrían canalizar de mejor manera y ser más efectivos si se parte del hecho que el analfabetismo es más alto en la región sureste, probablemente de esa manera una idea general se particulariza y genera resultados directos, en menor tiempo y con menor inversión. Así, aunque es sumamente útil tener la visión global de cómo estamos como país, también es muy importante particularizar nuestras fortalezas y debilidades para destinar los recursos donde sea más necesario y donde se puedan obtener mejores resultados en el corto plazo.

El hecho de mostrar la metodología que se utilizó para la realización del censo, quizá sea demasiado extenso y requiera un espacio mayor que el dedicado al análisis de la información, pues este trabajo parte del hecho de que ya se cuenta con la información y lo que interesa es explotarla, por ello se comenta someramente en el siguiente capítulo.

Capítulo 2

Desarrollo censal en México

Hace pocos años no existía literatura en español acerca de métodos multivariados, el conocimiento que se tenía acerca de estas técnicas era meramente teórico. Sin embargo, con el advenimiento de las computadoras y el software adecuado, ha sido posible la aplicación de esta gama de técnicas para el tratamiento de información.

El enfoque que tiene este trabajo es explicar de manera breve el fundamento de cada técnica del análisis multivariado, indicar para qué sirve, cuándo es recomendable su uso y, principalmente, ejemplificar a través de los datos del censo del 200 para nuestro país su utilización y la interpretación de los resultados que se obtengan.

Sin duda alguna, cuando se tengan múltiples variables, grandes cantidades de datos o una cantidad importante de individuos, es casi seguro que la aplicación de alguna, o algunas, de las técnicas del análisis multivariado serán de gran ayuda y obtendremos resultados que ayudarán al proceso de toma de decisiones.

Antes de iniciar la descripción de los datos que generó el censo, describiremos someramente la metodología que utilizó el INEGI para su elaboración.

Metodología del censo

- El censo fue de derecho o *jure*, es decir, la población se censó en su lugar de residencia.
- Las unidades de análisis fueron las viviendas, sus residentes y los hogares.
- La información se captó a través de una entrevista directa al jefe del hogar, su cónyuge o una persona mayor de 15 años residente en la vivienda.
- La fecha censal fue establecida por el INEGI el 14 de febrero del 2000.

Los temas y variables fueron los siguientes:

Características de la vivienda: Tipo y clase, disponibilidad de agua entubada, drenaje y electricidad, material de construcción predominante, disponibilidad de cocina y servicio sanitario, combustible usado para cocinar, número de cuartos, tenencia y bienes en la vivienda.

Características de las personas: Parentesco, sexo y edad; lugar de nacimiento, derechohabencia a servicio médico, discapacidad, migración, lengua indígena; características económicas y educativas, religión, estado conyugal, fecundidad y mortalidad.

Antes de iniciar el levantamiento de información se actualizó la cartografía con apoyo de imágenes satelitales y fotografía aérea para un adecuado tránsito de los encuestadores.

Asimismo, se realizó una amplia campaña de publicidad con el fin de sensibilizar a la población y lograr una cooperación que permitiera captar la información de una manera rápida, eficiente y confiable.

Universo de datos disponibles

La materia prima para esta tesis son los datos del censo, los cuales son especificados a continuación.

Los elementos disponibles en la página de Internet del INEGI relativos a los tabuladores por entidad federativa para el censo, y que se utilizarán posteriormente son relacionados a los siguientes puntos:

Población, fecundidad, mortalidad, migración, lengua indígena, religión, educación, servicios de salud, discapacidad, estado conyugal, empleo, hogares y vivienda.

Analizar todas y cada una de las variables que se recopilaron en el censo sería, además de tardado y fatigoso, un poco complicado, ya que se podrían presentar situaciones como la multicolinealidad o la redundancia. Es decir, tendríamos tanta información que en lugar de ayudarnos a obtener parámetros útiles nos complicarían el análisis.

Ciertamente hay algunas variables, por ejemplo las relacionadas con la educación, que por sí solas servirían para obtener resultados significativos acerca de cada uno de los estados del país.

El hecho de comparar el porcentaje de analfabetismo nos presenta un indicador acerca del grado de desarrollo que puede presentar cada estado. Sin embargo, si se agregan las variables adecuadas se pueden obtener mejores resultados.

La idea es no caer en una gran cantidad de variables que compliquen el análisis, pero tampoco realizar análisis con poca información. Se pretende realizar un análisis integral que proporcione la panorámica general de los niveles de vida que existen en los estados del país. A este respecto, es oportuno aclarar que el D.F. se tratará como un estado más.

En términos generales se describirá la situación específica en un momento determinado del tiempo para cada estado de acuerdo a las variables que se analicen y se realizarán comparaciones que permitan identificar grupos con "semejanzas". Para ello, las variables que se seleccionen al aplicar cada una de las técnicas descritas nos arrojarán la descripción comparativa para la formación de estos grupos.

Cuadros base de información

Se utilizará una tabla para cada técnica que se describa. La forma general será la siguiente:

Entidad Federativa	Variable 1	Variable 2	Variable n
1	$X_{1,1}$	$X_{2,1}$	$X_{n,1}$
2	$X_{1,2}$	$X_{2,2}$	$X_{n,2}$
.....
31	$X_{1,31}$	$X_{2,31}$	$X_{n,31}$
32	$X_{1,32}$	$X_{2,32}$	$X_{n,32}$

Es importante enfatizar que cada una de las variables X_{ij} presentadas en la tabla anterior se encuentran en totales brutos. Sin embargo, muchas veces no es ilustrativo realizar comparaciones directas. Por ejemplo, si decimos que en el D.F. hay x millones de personas económicamente activas (PEA), tal vez no diga mucho. Sin embargo, si indicamos que el D.F. tiene y porcentaje del total de PEA del país, entonces estamos reflejando la importancia relativa que tiene cada estado respecto a la variable que se está utilizando.

Por lo anterior, es más fácil trabajar con los porcentajes que con los números en bruto. Ya que ello resulta más claro y las conclusiones pueden ser más directas.

Para llegar a la tabla que se ejemplificó anteriormente, deben seleccionarse las variables a analizar, obtener los porcentajes correspondientes y construirla. Con la información que se encontró en la página de internet del INEGI, referida anteriormente, resulta más sencillo, pues los datos tienen formato PDF o XLS. En este caso, el formato que nos conviene manejar es el formato XLS, de excel.

La elaboración de dicha tabla involucra cierta cantidad de tiempo, ya que los resultados vienen desglosados muchas veces por grupos quinquenales de edad o por agrupamientos específicos, propios de las variables que se trate; por lo que hay que integrar los datos para obtener los totales por estado y posteriormente los

porcentajes correspondientes. Esta tarea, si bien puede reducirse al uso de ciertas fórmulas o macros en excel, es importante y minuciosa debido a la consolidación que debe darse por estado.

Como ejemplo, se presenta la tabla 1, que contiene una sola variable, la cual corresponde a la población total de cada estado. Agregamos la columna del porcentaje respecto al total de la población, que en este caso fue de **97.5 millones de habitantes**.

Como se observa, el estado con mayor población es el Estado de México. Esto puede determinarse directamente al analizar la tabla. Asimismo, se puede determinar el estado con menor población. Si consideramos que se trata de 32 individuos, tal vez podamos conocer el promedio de población por estado.

Este último dato, quizá no sirva de mucho si consideramos como factor importante la extensión territorial de cada estado. Para ello, deberíamos de referirnos a la densidad de población.

Lo que resulta interesante al analizar la tabla 1 es que no se requiere más que un análisis visual de ella para empezar a obtener una idea del comportamiento de cada estado de acuerdo a esta variable. Al incluir n variables empezamos a tratar con el análisis multivariado.

Las proporciones o porcentajes quizá fueran más útiles si no conociéramos el total de la población. Al aplicar técnicas matemáticas y/o estadísticas es más cómodo y más útil emplear los porcentajes, por las razones que ya se comentaron.

Si de la tabla referida observamos que Baja California Sur tiene el 0.43% de la población, inmediatamente deducimos que es de los estados menos poblados.

Sin embargo, si mencionamos que Baja California Sur tiene casi medio millón de habitantes, quizá pareciera que está "medianamente" poblada.

Ya se mencionó que Quintana Roo era de los estados menos poblados en 1990, observamos que este hecho se sigue presentando 10 años después. Si seguimos revisando la tabla, encontramos por ejemplo, que Tlaxcala tiene un poco más habitantes que Quintana Roo, sin embargo, las extensiones territoriales son muy diferentes.

Cons.	Estado	Población	Porcentaje
		Total	
1	AGUASCALIENTES	944,286	0.97%
2	BAJA CALIFORNIA	2,497,387	2.55%
3	BAJA CALIFORNIA SUR	424,041	0.43%
4	CAMPECHE	890,889	0.71%
5	COAHUILA DE ZARAGOZA	2,298,070	2.36%
6	COLIMA	542,527	0.56%
7	CHIAPAS	3,920,892	4.02%
8	CHIHUAHUA	3,062,907	3.13%
9	DISTRITO FEDERAL	8,905,239	8.83%
10	DURANGO	1,448,561	1.49%
11	GUANAJUATO	4,963,032	4.78%
12	GUERRERO	3,079,849	3.16%
13	HIDALGO	2,235,591	2.29%
14	JALISCO	6,322,002	6.48%
15	MÉXICO	13,098,888	13.43%
16	MICHOCÁN DE OCAMPO	3,966,857	4.09%
17	MORELOS	1,666,299	1.80%
18	NAYARIT	920,185	0.94%
19	NUEVO LEÓN	3,834,141	3.93%
20	OAXACA	3,438,755	3.53%
21	PUEBLA	5,078,988	5.21%
22	QUERÉTARO DE ARTEAGA	1,404,309	1.44%
23	QUINTANA ROO	874,083	0.90%
24	SAN LUIS POTOSÍ	2,299,360	2.38%
25	SINALOA	2,536,844	2.60%
26	SONORA	2,218,989	2.27%
27	TABASCO	1,891,529	1.94%
28	TAMAULIPAS	2,763,222	2.82%
29	TLAXCALA	962,648	0.99%
30	VERACRUZ-LLAVE	6,908,975	7.09%
31	YUCATÁN	1,658,210	1.70%
32	ZACATECAS	1,363,610	1.39%
	Total	97,483,412	100.00%

Tabla 2.1

Métodos multivariados a emplear

Se utilizará el paquete estadístico SPSS, versión 10.0. En este paquete se encuentran disponibles los análisis que se pretenden aplicar:

1. Análisis de conglomerados
2. Análisis factorial o de factores
3. Escalamiento multidimensional.

Cada punto se analizará en los siguientes capítulos, y dependiendo de la amplitud del campo de aplicación se determinará el número de variables y ejemplos a describir.

Antes de iniciar el análisis multivariado, propiamente dicho, es ilustrativo considerar un ejemplo de un breve análisis descriptivo de la información, considerando una sola variable. En este caso, se tomó la variable población, la de la tabla 1. Al aplicar la opción de estadística descriptiva se obtuvo, considerando 32 observaciones, el siguiente resultado.

Descriptive Statistics	N	Minimum	Maximum	Mean	Std. Deviation
Población	32	424041.00	13000000.00	3046356.6250	2664751.0545
Valid N (listwise)	32				

Tabla 2.2

Con esta simple tabla se pueden obtener conclusiones interesantes. En principio, la varianza es muy alta. Asimismo, la diferencia entre la observación máxima y la mínima es muy pronunciada, por lo que se puede determinar que la distribución de la población en cada uno de los estados es muy poco uniforme. Desde luego, aquí no se está considerando el factor territorial.

Los datos de origen para la tabla 2.2, se tomaron de la tabla 2.1. En este caso se tomaron los valores tal cual y no en porcentajes, y se hizo así únicamente para fines ilustrativos.

El paquete SPSS acepta datos "copiados" y "pegados", también acepta abrir los datos desde un archivo de excel, pero esto debe tener la estructura correcta para que pueda ser leído directamente por el paquete.

En la medida de lo posible, se especificará el uso del paquete, indicando los pasos necesarios para su operación, en la bibliografía se recomienda un libro que explica la utilización de este paquete.

En el siguiente capítulo iniciaremos la aplicación de las técnicas correspondientes, pero cabe hacerse una pregunta antes de continuar:

¿Qué es el análisis multivariado?

Sin duda alguna, y debido a que el título del presente trabajo incluye la frase "análisis multivariado", y con el objeto de evitar perdemos en lo que realmente se está aplicando, es importante una breve definición de la técnica estadística que se emplea en este documento.

De manera general, el análisis multivariado es la indagación dentro de la estructura de las relaciones que guardan entre sí múltiples medidas (datos, variables, etc.).

Sin embargo, lo anterior es muy genérico y puede aplicarse a una amplia gama de ramas de la estadística, y se requiere una definición un poco más específica, por lo que puede decirse que es **el estudio mediante modelos lineales de las relaciones entre variables múltiples.**

En otras palabras, El propósito del análisis multivariado es medir, explicar y predecir el grado de relación que existe entre la variación (combinación lineal ponderada de las variables).

El carácter multivariado del análisis descansa no sólo en el número de variables sino en las múltiples combinaciones existente entre las variables

Las conclusiones que se obtienen a partir del análisis multivariado son muy importantes en ramas de la física, biología, psicología y en las ciencias sociales, particularmente la economía. Aunque para que las conclusiones que se obtengan a partir de dicho análisis tengan validez, debe cuidarse que la matriz de datos (mencionada en el "cuadro base de información") tenga una estructura que permita evitar los "vicios" más comunes cuando hablamos de variables estadísticas, es decir fenómenos como la colinealidad, la dependencia (lineal o no), varianzas muy grandes, etc.

Para finalizar este punto, es importante mencionar que la aplicación real del análisis multivariado no sería posible sin el uso de las computadoras, ya que la cantidad de cálculos y manipulación de matrices es una limitante para cualquier cálculo manual.

Asimismo, la elección de un paquete específicamente creado para este fin es importante. Y como ya se mencionó, se utilizará el SPSS versión 10.0, que cuenta con varios años de estar en el mercado y que constantemente se encuentra en actualización.

Capítulo 3.

Análisis de Conglomerados

Objetivo

El análisis de conglomerados, conocido en inglés como "*cluster analysis*" o también como análisis de grupos pretende la agrupación de individuos en conglomerados. Dentro de cada conglomerado los individuos son "semejantes" entre sí, y cada conglomerado es diferente o "dismil" de los demás conglomerados.

Si se tienen n individuos, a cada uno de los cuales se les asocian p variables, se obtiene una matriz de $n \times p$. Entonces, será materia del análisis de conglomerados obtener grupos entre los cuales la distancia entre los renglones respectivos sea mínima y, al mismo tiempo, que la distancia entre cada grupo sea "grande".

Además de la matriz descrita anteriormente, este método también puede utilizar una matriz de similitudes o disimilitudes entre individuos.

Los usos de este método van desde la psicología, botánica, antropología, sociología, etc. Lo que ha ocasionado que surjan algoritmos especializados en alguna de estas ramas de la ciencia.

Elementos necesarios para aplicar el análisis de conglomerados

1. Elección de las unidades de datos, es decir, Individuos.
2. Elección de las variables.
3. Decidir si se realizará el agrupamiento por Individuos o por variables, lo cual también es útil en algunos campos de la ciencia.
4. Homogeneizar las variables, es decir, que vayan enfocadas en el mismo sentido.
5. Medidas similares. Cada variable, para ser comparada con otra debe contar con medidas equiparables.
6. Criterio de agrupamiento. Este punto va relacionado con el algoritmo que se utilice.
7. Elección del algoritmo y herramientas computacionales a utilizar.
8. Número de conglomerados deseables.
9. Interpretación de los resultados.

Se han desarrollado varios algoritmos para la aplicación de este método, a grandes rasgos se pueden agrupar en *métodos jerárquicos y no jerárquicos*.

Métodos Jerárquicos

Estos métodos se basan en agrupaciones sucesivas y ocurren en varias etapas. En la primera fase se tienen tantos conglomerados como individuos. En la segunda etapa aparecerán conglomerados de uno o dos individuos, y así sucesivamente hasta llegar a un conglomerado único.

Los criterios para determinar la distancia entre cada conglomerado pueden ser variados. Por ejemplo, si consideramos que cada individuo es un vector en un espacio p -dimensional, en la primera etapa la distancia euclídea podría indicarnos los individuos más cercanos.

Para la siguiente etapa se podría tomar el promedio de distancia entre cada par de individuos pertenecientes a diferentes conglomerados. Aunque este es sólo un ejemplo ya que la medida que se tome para la distancia entre un individuo y/o entre conglomerados es un elemento importante, ya que marca la pauta para la formación de los conglomerados.

Un algoritmo jerárquico es el conocido como *liga sencilla* o "vecino más cercano", el cual se basa en una *matriz de disimilitudes*. Al tener n individuos y p variables, esta matriz será de $n \times n$. En ella, la entrada (i,j) será la disimilitud o distancia que hay entre el individuo i , y el individuo j . En caso que $i=j$, $d(i,j)=0$.

Por ejemplo, se puede definir $d(i,j)$ como $|\text{var}(i) - \text{var}(j)|$, o también puede ser la distancia euclídea. En términos generales, la medida de distancia entre los pares de individuos puede tomarse de diversas maneras, sin embargo, la distancia euclídea o ésta al cuadrado son las mayormente utilizadas, aunque debido a la naturaleza de los estudios que se realicen se pueden emplear otras medidas de disimilitud siempre y cuando sea justificable.

A partir de esta matriz, se definirán grupos para los cuales la distancia entre ellos será el mínimo de las distancias de cada par de individuos pertenecientes a los grupos que se tengan conformados. Es decir, si se tienen dos grupos C_1 y C_2 , entonces $d(C_1, C_2) = \min \{ d(r,s) : r \in C_1, s \in C_2 \}$

Al principio, se parte de conglomerados de dos individuos a lo más. Para cada conglomerado de dos individuos se mide la distancias de los demás elementos a ese grupo de la siguiente manera: $d(k, (i,j)) = \min \{ d(k,i), d(k,j) \}$ para $k \neq i, j$.

Estas distancias se toman como nuevas $d(r,s)$, por lo que de la matriz original desaparecerán el renglón i y la columna j , y se agregará la columna C_1 , donde las entradas $d(k, C_1)$ serán las calculadas anteriormente. (En este primer caso $C_1=(i,j)$).

Se procede de manera similar y subsiguiente hasta agotar la matriz. Con estos resultados se genera un dendograma, el cual se ejemplificará posteriormente.

Cuando este método, en lugar de tomar los mínimos de las distancias toma los máximos se conoce como *liga completa*. Cuando toma el promedio de la distancia mínima y la máxima se conoce como *liga promedio*.

Métodos no jerárquicos

En estos métodos el número de conglomerados se fija a priori por parte del investigador. Tomando como referencia el caso anterior, en estos métodos los conglomerados del siguiente nivel no necesariamente tendrán subconjuntos del nivel anterior.

De acuerdo a los resultados que generan, se pueden identificar como:

De partición: En este caso no se llega a un dendograma, sino a una serie de subconjuntos, cada uno de los cuales será un conglomerado, y no habrá intersección entre ellos.

Conglomerados no ajenos: En este caso, se generan subconjuntos no ajenos.

Un método no jerárquico es el conocido como de las *k medias*. Este método consta de los siguientes pasos:

1. Elegir el número de conglomerados k .
2. Determinar el centro de cada uno de los k conglomerados, cuando se desconozcan, estimarlos.
3. Con base en el centro más próximo, agrupar a los individuos en conglomerados.
4. A partir de este último paso, se vuelven a calcular los centros de cada agrupación.
5. Se repiten los pasos 3 y 4 hasta que se cumpla el criterio de parada, que debe establecerse previamente.

Aplicación a los datos del censo

Para ejemplificar éste procedimiento se utilizarán las variables relacionadas con la vivienda. En este caso, no se utilizarán porcentajes porque las unidades serán el número de viviendas, el cual es común a cada uno de los individuos. Además, al partimos de que según estos datos hay aproximadamente 100 millones de habitantes en México y que se cuenta con poco más de 21 millones de viviendas, no resultaría significativo el porcentaje porque el número de habitantes que hay en cada vivienda no es proporcional en todos los Estados.

Así, se pretende establecer una agrupación en base al número de viviendas y su tipo que existen en cada uno de los estados del país. Cada tipo de vivienda

representa una variable, en este caso las variables son: *casa independiente, departamento en edificio, vivienda en vecindad, cuarto de azotea, local no construido para habitación, vivienda móvil, refugio, no especificado y vivienda colectiva.*

1	AGUASCALIENTES	200,873
2	BAJA CALIFORNIA	810,087
3	BAJA CALIFORNIA SUR	105,228
4	CAMPECHE	167,172
5	COAHUILA DE ZARAGOZA	644,880
6	COLIMA	132,330
7	CHIAPAS	808,551
8	CHIHUAHUA	758,959
9	DISTRITO FEDERAL	2,132,413
10	DURANGO	325,309
11	GUANAJUATO	625,254
12	GUERRERO	657,968
13	HIDALGO	494,317
14	JALISCO	1,394,025
15	MÉXICO	2,893,387
16	MICHOACÁN DE OCAMPO	855,512
17	MORELOS	387,399
18	NAYARIT	220,115
19	NUEVO LEÓN	888,552
20	OAXACA	741,005
21	PUEBLA	1,068,882
22	QUERÉTANO DE ARTAGA	298,372
23	QUINTANA ROO	213,568
24	SAN LUIS POTOSÍ	482,014
25	SINALOA	575,292
26	SONORA	530,435
27	TABASCO	412,634
28	TAMAULIPAS	683,058
29	TLAXCALA	194,549
30	VERACRUZ-LLAVE	1,606,184
31	YUCATÁN	373,432
32	ZACATECAS	299,483

TOTAL DE VIVIENDAS: 21,954,713

Tabla 3.1. Total de viviendas habitadas por estado.

Tabla 3.2. Total de viviendas por tipo, utilizada para el análisis de conglomerados.

ENTIDAD FEDERATIVA	CASA		DEPARTAMENTO	VEJEDAD	CUARTO DE AZOTEA	LOCAL NO CONSTRUIDO PARA HABITACIÓN	VIVIENDA MÓVIL	REFUGIO	NO ESPECIFICADO	VIVIENDA COLECTIVA
	INDEPENDIENTE	EN EDIFICIO								
1 AGUASCALIENTES	178,373	11,641	2,265	125	238	26	28	7,833	144	
2 BAJA CALIFORNIA	460,159	37,644	31,238	604	604	3,167	143	76,325	390	
3 BAJA CALIFORNIA SUR	92,355	4,224	4,031	135	185	389	41	3,704	165	
4 CAMPECHE	150,817	15,817	1,335	26	427	103	46	3,601	118	
5 COAHUILA DE ZARAGOZA	506,833	4,048	7,696	346	379	103	82	24,835	338	
6 COLIMA	116,605	2,558	2,524	141	255	38	43	10,037	129	
7 CHIAPAS	732,572	5,785	15,941	406	1,246	140	166	49,865	420	
8 CHIHUAHUA	668,237	8,840	23,149	391	593	146	236	53,787	580	
9 DISTRITO FEDERAL	1,243,468	572,862	210,463	12,843	3,028	130	332	88,284	1,003	
10 DURANGO	307,569	1,485	1,657	60	165	43	34	14,111	185	
11 GUANAJUATO	847,849	29,728	9,795	318	807	73	127	37,168	419	
12 GUERRERO	595,682	21,823	15,506	909	705	54	245	22,673	412	
13 HIDALGO	456,611	10,227	12,365	271	364	34	68	15,243	134	
14 JALISCO	1,206,615	100,121	28,550	1,770	2,135	220	293	53,071	1,251	
15 MÉXICO	2,265,800	187,883	196,137	3,704	2,133	213	222	236,723	542	
16 MICHOACÁN DE OCAMPO	793,508	14,237	9,473	390	884	165	180	35,951	724	
17 MORELOS	316,135	14,924	13,853	582	493	45	53	21,145	169	
18 NAYARIT	207,781	3,321	2,489	89	591	66	55	5,447	289	
19 NUEVO LEÓN	811,207	22,817	14,366	916	885	105	64	37,727	465	
20 OAXACA	694,556	5,863	17,018	351	912	64	125	21,662	454	
21 PUEBLA	877,459	73,765	55,004	853	1,086	72	88	57,043	503	
22 QUERÉTARO DE ARTEAGA	269,847	7,353	6,925	118	239	37	41	13,615	197	
23 QUINTANA ROO	164,286	16,978	22,447	458	690	58	17	8,393	239	
24 SAN LUIS POTOSÍ	455,040	17,143	4,900	145	482	36	135	14,769	264	
25 SINALOA	533,834	9,818	14,806	191	596	157	17	15,458	363	
26 SONORA	496,692	6,160	3,190	129	635	549	95	20,487	496	
27 TABASCO	374,123	11,340	13,494	868	893	35	48	11,851	182	
28 TAMAULIPAS	596,318	23,149	31,473	688	760	211	114	29,938	417	
29 TLAXCALA	176,558	6,916	4,876	70	134	16	14	6,068	77	
30 VERACRUZ-Llave	1,468,839	31,058	60,416	1,675	2,447	167	266	40,629	697	
31 YUCATÁN	361,819	1,088	517	90	612	38	59	8,951	188	
32 ZACATECAS	280,042	4,905	1,936	50	225	50	45	11,896	234	

Resultados obtenidos en el paquete SPSS

Cluster

Case Processing Summary^a

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
32	97.0	1	3.0	33	100.0

a. Average Linkage (Between Groups)

Vista parcial de la matriz de disimilitudes:

Case	Proximity Matrix																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1		E+10	C+08	E+08	E+11	E+09	E+11	E+11	E+12	E+10	E+11	C+11	E+10	E+12	E+12	E+11	E+10	E+08	E+11	E+11	E+11	E+11
2	E+10		E+11	C+11	E+09	E+11	E+10	E+10	E+11	E+10	E+11	E+10	C+09	E+11	E+12	E+11	E+10	E+10	E+11	E+10	E+11	E+11
3	E+09	E+11		E+09	E+11	E+08	E+11	E+11	E+12	E+10	E+11	E+11	C+11	E+12	E+12	E+11	E+10	C+10	E+11	E+11	E+11	E+11
4	E+08	E+11	E+09		E+11	E+09	E+11	E+11	E+12	E+10	E+11	E+11	C+10	E+12	E+12	E+11	E+10	E+09	E+11	E+11	E+11	E+11
5	E+11	E+09	E+11	C+11		E+11	E+10	E+10	E+11	E+10	E+11	E+09	C+09	E+11	E+12	E+10	E+10	E+10	E+10	E+10	E+11	E+11
6	E+09	E+11	E+08	C+09	E+11		E+11	E+11	E+12	E+10	E+11	E+11	E+11	E+12	E+12	E+11	E+10	C+09	E+11	E+11	E+11	E+11
7	E+11	E+10	E+11	E+11	E+10	E+11		E+08	C+11	E+11	E+10	E+10	E+10	E+11	E+12	E+09	E+11	E+08	E+09	E+10	E+10	E+10
8	E+11	E+10	E+11	C+11	E+10	E+11	E+09		E+11	E+11	E+10	E+09	C+10	E+11	E+12	E+10	E+11	C+11	E+11	E+10	E+09	E+10
9	E+12	E+11	E+12	C+12	E+11	E+12	E+11	E+11		E+12	E+11	E+11	E+11	E+11	E+12	E+11	E+12	E+12	E+11	E+11	E+11	E+11
10	E+10	E+10	E+10	E+10	E+10	E+10	E+11	E+11	E+12		E+11	E+10	E+10	C+11	E+12	E+11	E+08	E+10	E+11	E+11	E+11	E+11
11	C+11	E+11	E+11	E+11	E+11	E+11	E+10	E+10	E+11	C+11		E+10	E+11	E+11	C+12	E+09	E+11	E+11	E+09	E+10	E+09	E+10
12	E+11	E+10	E+11	E+11	E+08	C+11	E+10	E+09	E+11	E+10	E+10		E+10	E+11	E+12	E+10	E+10	E+11	E+10	E+10	E+10	E+10
13	E+10	E+08	E+11	E+10	E+09	E+11	C+10	E+10	E+11	E+10	E+11	E+10		E+11	E+12	E+11	E+10	E+10	E+11	E+11	E+11	E+11
14	E+12	E+11	E+12	E+12	E+11	E+12	E+11	E+11	E+11	E+11	E+11	C+11	E+11		E+12	E+11	E+11	E+12	E+11	E+11	E+11	E+11
15	E+12	E+12	E+12	E+12	E+12	E+12	E+12	E+12	E+12	E+12	E+12	E+12	E+12	E+12		E+12						
16	E+11	E+11	C+11	E+11	E+10	E+11	C+08	E+10	E+11	E+11	E+08	E+10	E+11	E+11	E+12		E+11	E+11	E+08	E+10	E+10	C+11
17	E+10	E+10	C+10	E+10	E+10	E+10	C+11	E+11	E+12	E+08	E+11	E+10	E+10	E+11	E+12	E+11		E+10	E+11	E+11	E+11	C+11
18	E+08	E+10	C+10	E+09	E+10	E+09	C+11	E+11	E+12	E+10	E+11	E+11	E+10	E+12	E+12	E+11	E+10		E+11	E+11	E+11	C+11
19	E+11	E+11	E+11	E+11	E+10	E+11	E+08	C+10	E+11	E+11	E+09	E+10	E+11	E+11	E+12	E+08	E+11	E+11		E+10	E+08	E+10
20	E+11	E+10	C+11	E+11	E+10	E+11	E+08	E+09	E+11	E+11	E+10	E+10	E+10	E+11	E+12	E+10	E+11	E+11	E+10		E+10	C+11
21	E+11	E+11	E+11	E+11	E+11	E+10	E+10	E+10	E+11	E+11	E+08	E+10	E+11	E+11	E+12	E+10	E+11	E+11	E+09	E+10		E+10
22	E+08	E+10	E+10	E+10	E+10	C+11	E+11	E+12	E+09	E+11	E+11	E+10	E+11	E+11	E+12	E+11	E+08	E+09	E+11	E+11	E+11	E+11
23	E+08	E+10	E+08	E+08	E+11	E+09	E+11	C+11	E+12	E+10	E+11	E+11	E+10	E+12	E+12	E+11	E+10	E+09	E+11	E+11	E+11	E+11
24	E+10	E+09	E+11	C+10	E+09	E+11	E+10	E+10	E+11	E+10	E+11	E+10	E+08	E+11	E+12	E+11	E+10	E+10	E+11	E+10	E+11	E+11
25	E+11	E+10	E+11	E+11	C+08	E+11	E+10	E+10	E+11	C+10	E+10	E+09	E+09	E+11	E+12	E+10	E+10	E+11	E+10	E+10	E+11	E+11
26	C+11	E+00	E+11	E+08	E+11	E+10	E+10	E+11	E+10	E+11	E+09	E+09	E+11	E+12	E+11	E+10	E+10	E+10	E+10	E+10	E+11	E+11
27	E+10	E+10	E+10	E+10	E+10	E+11	E+10	E+12	C+09	E+11	E+10	E+09	E+09	E+11	E+12	E+11	E+10	E+10	E+10	E+10	E+10	E+11
28	E+11	E+10	E+11	E+11	C+08	E+11	E+10	E+09	E+11	E+10	E+10	E+08	E+10	E+11	E+12	E+10	E+10	E+11	E+10	E+10	E+10	E+10
29	E+07	E+10	E+09	E+08	C+11	E+09	E+11	E+11	E+12	E+10	E+11	E+11	E+10	E+11	E+12	E+11	E+10	E+08	E+11	E+11	C+11	E+11
30	E+12	E+12	E+12	E+12	E+10	C+12	E+11	E+11	E+12	E+11	E+11	E+12	E+11	E+12	E+11	E+11	E+12	E+12	E+11	E+11	E+11	E+11
31	E+10	E+10	E+10	E+10	E+10	C+10	E+11	E+10	E+12	E+09	E+11	E+10	E+09	E+11	E+12	E+11	E+09	E+10	E+11	E+11	E+11	E+11
32	E+10	E+10	E+10	E+10	E+10	C+10	E+11	E+11	E+12	E+08	E+11	E+11	E+10	E+11	E+12	E+11	E+09	E+09	E+11	C+11	E+11	E+11

This is a dissimilarity matrix

Liga promedio (Entre grupos)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	29	34496424	0	0	9
2	13	24	104159182	0	0	18
3	5	26	110282872	0	0	12
4	22	32	137183284	0	0	13
5	12	28	310007616	0	0	23
6	16	19	414327424	0	0	16
7	27	31	431736960	0	0	20
8	10	17	452683328	0	0	13
9	1	23	604167104	1	0	11
10	3	6	633345408	0	0	19
11	1	4	842137408	9	0	15
12	5	25	1.166E+09	3	0	18
13	10	22	1.549E+09	8	4	20
14	8	20	1.771E+09	0	0	17
15	1	18	1.919E+09	11	0	16
16	11	16	2.303E+09	0	6	22
17	7	8	3.229E+09	0	14	23
18	5	13	3.712E+09	12	2	21
19	1	3	5.701E+09	15	10	24
20	10	27	6.118E+09	13	7	24
21	2	5	8.581E+09	0	18	25
22	11	21	9.125E+09	16	0	27
23	7	12	1.200E+10	17	5	25
24	1	10	2.959E+10	19	20	28
25	2	7	3.469E+10	21	23	27
26	14	30	7.470E+10	0	0	29
27	2	11	8.479E+10	25	22	28
28	1	2	1.947E+11	24	27	30
29	9	14	3.143E+11	0	26	30
30	1	9	9.308E+11	28	29	31
31	1	15	3.238E+12	30	0	0

Vista parcial de las aglomeraciones sucesivas

Cosa	Horizontal locale																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
15	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
30	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
14	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
8	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
21	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
10	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
16	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
11	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
20	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
12	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
10	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
6	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
7	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
24	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
13	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
21	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
16	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
6	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
11	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
27	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
10	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
21	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
17	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
10	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
6	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
15	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
21	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
20	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

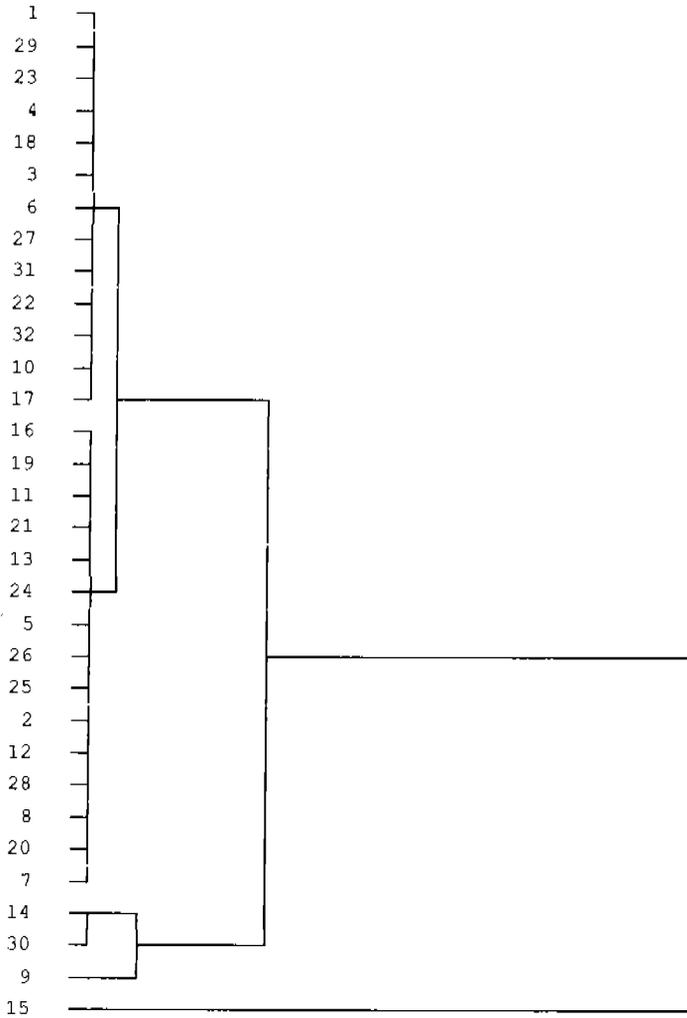
Dendrograma

* * * * * ANALISIS JERÁRQUICO DE CLUSTERS * * * * *

Dendrograma usando ligas promedio (Entre Grupos)

Distancia de cluster con escala ajustada

C A S O 0 5 10 15 20 25
Etiqueta Num +-----+-----+-----+-----+-----+



Análisis de Resultados

Uno de los elementos más importantes y relevantes es el dendograma, ya que a partir de él se puede visualizar el comportamiento de cada uno de los estados de acuerdo a los demás.

Cada número ubicado a la izquierda representa a cada uno de los estados de acuerdo a la tabla 3.2.

Inmediatamente se nota al Estado de México (16) como una observación discrepante. Asimismo el Distrito Federal (9) no muestra muchas similitud con los demás estados. Jalisco (14) y Veracruz(30), aunque también son discrepantes muestran cierto "parecido" entre ellos.

Por otro lado, se observan claramente dos grupos, los cuales se conforman de la siguiente manera:

Grupo I	Grupo II
Michoacán (18)	Aguascalientes (1)
Nuevo León (19)	Tlaxcala (29)
Guanajuato (11)	Quintana Roo (23)
Puebla (21)	Campeche (4)
Hidalgo (13)	Nayarit (18)
San Luis Potosí (24)	Baja California Sur (3)
Coahuila (6)	Colima (8)
Sonora (26)	Tabasco (27)
Sinaloa (25)	Yucatán (31)
Baja California(2)	Querétaro (22)
Guerrero (12)	Zacatecas (32)
Tamaulipas (28)	Durango (10)
Chihuahua (8)	Morelos (17)
Oaxaca (20)	
Chiapas (7)	

Si se intentara realizar un comparativo geográfico, el grupo I englobaría la región norte, centro y sur, con algunas excepciones. El segundo grupo correspondería al centro-sureste.

Si olvidamos la situación geográfica, el grupo II se presenta más homogéneo de acuerdo a los estados que lo conforman. Es decir, en cuanto a desarrollo económico no parece presentarse ninguna disparidad.

Sin embargo, el grupo I es muy heterogéneo, ya que engloba estados como Nuevo León y Oaxaca o Chiapas, quienes claramente presentan niveles de desarrollo bastante diferenciados.

Recordemos que los grupos indican a aquellos estados que tienen similar composición de ocupantes de vivienda, por lo que no hay que perder de vista las variables que se utilizaron en la tabla 3.2.

Un hecho que destaca es que los estados del grupo I cuentan con su población viviendo en casas de tipo no especificado en una proporción significativamente mayor que en los demás tipos de vivienda.

Considerando los primeros cuatro estados como un grupo aparte, se puede decir que, de acuerdo al tipo de vivienda, se pueden conformar 3 grupos con los 32 estados del país.

Consideraciones adicionales

Como corolario del análisis anterior, y tomando en cuenta que a primera vista parece que las variables *vivienda* y *población* estén estrechamente relacionadas, podemos suponer que existe una relación directa entre ellas, es decir, que el estado que cuenta con mayor número de viviendas es el que cuenta con mayor número de habitantes.

Para corroborarlo, nos auxiliaremos de las tablas 2.1 y 3.1 y realizaremos un simple comparativo entre ellas, ordenándolas de mayor a menor.

Lo anterior se muestra en la página siguiente en la cual claramente se observa que se presenta lo mencionado y que efectivamente la relación población-viviendas se presenta de una manera directa. Si bien al realizar el ordenamiento no se observa que el comportamiento sea exacto en ambas tablas nos podemos percatar que la variación de lugares no es significativa en su mayoría.

Por ejemplo, Michoacán se ubica en la posición número 7 de acuerdo al total de habitantes, y en la tabla de total de viviendas se ubica en el 8. Casos similares ocurren con otros estados en los cuales su variación de tabla a tabla es de uno o dos lugares, encontrando que Tlaxcala y San Luis Potosí se ubican entre los de mayor cambio con 3 posiciones entre tabla y tabla.

El ejercicio previo nos permite corroborar que existe una colinealidad entre las variables utilizadas (número de habitantes y número de viviendas) y que la realización de un dendograma con cualquiera de las dos variables nos arrojaría resultados semejantes y que sería prácticamente ocioso realizar un análisis de conglomerados con ambas variables, ya que con solo una de ellas sería suficiente.

Sin embargo, otra cosa sería si se incluyera el desglose por tipo de vivienda contra la población de cada estado. Lo cual requeriría un análisis más específico para determinar la inclusión de las variables.

Disciplinas como la mercadotecnia se sirven mucho de este tipo de análisis.

ENTIDAD FEDERATIVA		POBLACION TOTAL	ENTIDAD FEDERATIVA	POBLACION TOTAL
1	AGUASCALIENTES	200,673	15	MEXICO
2	BAJA CALIFORNIA	610,057	16	DISTRITO FEDERAL
3	BAJA CALIFORNIA SUR	105,129	17	VERACRUZ-Llave
4	CAMPECHE	157,172	18	JALISCO
5	COAHUILA DE ZARAGOZA	541,660	19	PUEBLA
6	COLIMA	132,330	20	QUERETARO DE ARTEAGA
7	CHIAPAS	806,551	21	SINALOA
8	CHIHUAHUA	755,959	22	SONORA
9	DISTRITO FEDERAL	3,052,307	23	HIDALGO
10	DURANGO	1,448,961	24	SAN LUIS POTOSI
11	GUANAJUATO	3,078,849	25	BAJA CALIFORNIA
12	GUERRERO	526,284	26	SAN LUIS POTOSI
13	HIDALGO	494,317	27	COAHUILA DE ZARAGOZA
14	JALISCO	2,893,367	28	HIDALGO
15	MEXICO	885,512	29	SONORA
16	MICHOACÁN DE OCAIMPO	367,359	30	NAYARIT
17	MORELOS	220,118	31	AGUASCALIENTES
18	NAYARIT	888,532	32	QUINTANA ROO
19	BAJA CALIFORNIA SUR	741,005	33	VERACRUZ-Llave
20	OAXACA	1,065,882	34	YUCATAN
21	PUEBLA	296,372	35	ZACATECAS
22	QUERETARO DE ARTEAGA	213,566	36	MORELOS
23	QUINTANA ROO	492,914	37	DURANGO
24	SAN LUIS POTOSI	575,292	38	QUERETARO DE ARTEAGA
25	SINALOA	530,435	39	TTLAXCALA
26	SONORA	412,634	40	QUINTANA ROO
27	TABASCO	583,068	41	AGUASCALIENTES
28	TAMULIPIAS	194,549	42	QUINTANA ROO
29	TTLAXCALA	1,606,194	43	CAMPECHE
30	VERACRUZ-Llave	299,432	44	COAHUILA
31	YUCATAN	758,483	45	BAJA CALIFORNIA SUR
32	ZACATECAS	275,573	46	TOTAL DE VIVIENDAS
				27,557,737

ENTIDAD FEDERATIVA		POBLACION TOTAL	ENTIDAD FEDERATIVA	POBLACION TOTAL
15	MEXICO	2,893,367	15	MEXICO
16	DISTRITO FEDERAL	3,052,307	16	DISTRITO FEDERAL
17	VERACRUZ-Llave	1,065,882	17	VERACRUZ-Llave
18	JALISCO	2,206,002	18	JALISCO
19	PUEBLA	5,078,696	19	PUEBLA
20	QUERETARO DE ARTEAGA	4,663,032	20	QUERETARO DE ARTEAGA
21	SINALOA	3,985,667	21	SINALOA
22	SONORA	3,834,141	22	SONORA
23	HIDALGO	3,438,765	23	HIDALGO
24	SAN LUIS POTOSI	3,079,649	24	SAN LUIS POTOSI
25	BAJA CALIFORNIA	3,052,907	25	BAJA CALIFORNIA
26	SAN LUIS POTOSI	2,753,222	26	SAN LUIS POTOSI
27	COAHUILA DE ZARAGOZA	2,638,844	27	COAHUILA DE ZARAGOZA
28	HIDALGO	2,487,367	28	HIDALGO
29	SONORA	2,289,360	29	SONORA
30	NAYARIT	2,286,070	30	NAYARIT
31	AGUASCALIENTES	2,236,591	31	AGUASCALIENTES
32	QUINTANA ROO	2,216,959	32	QUINTANA ROO
33	VERACRUZ-Llave	1,891,629	33	VERACRUZ-Llave
34	YUCATAN	1,698,210	34	YUCATAN
35	ZACATECAS	1,595,296	35	ZACATECAS
36	MORELOS	1,448,661	36	MORELOS
37	DURANGO	1,404,306	37	DURANGO
38	QUERETARO DE ARTEAGA	1,353,610	38	QUERETARO DE ARTEAGA
39	TTLAXCALA	982,646	39	TTLAXCALA
40	QUINTANA ROO	944,296	40	QUINTANA ROO
41	AGUASCALIENTES	920,165	41	AGUASCALIENTES
42	QUINTANA ROO	874,983	42	QUINTANA ROO
43	CAMPECHE	890,889	43	CAMPECHE
44	COAHUILA	542,827	44	COAHUILA
45	BAJA CALIFORNIA SUR	444,041	45	BAJA CALIFORNIA SUR
				27,557,737

Previo al ordenamiento

Tablas comparativas entre el número de viviendas de cada estado y su número de habitantes.

Ordenamiento de mayor a menor

A continuación tomaremos un grupo de variables, únicamente para continuar con la idea de formar grupos en los estados, y para verificar el comportamiento cuando consideramos elementos diferentes a la población y tipos de vivienda. Esta tabla es la 3.3 y en ella no se contempló la utilización de porcentajes porque la población que se contempla es la mayor a 5 años. Es probable que resulte algún agrupamiento interesante al tomar la cuestión religiosa y en este caso, tal vez sea posible que los grupos se formen de acuerdo a la situación geográfica.

El dendograma que resulta de aplicar la técnica de cluster jerárquico con la distancia euclídea se muestra en la página siguiente y de él se pueden obtener las siguientes impresiones a primera vista.

El Distrito Federal y el Estado de México parecen observaciones atípicas, sin embargo entre ellos se muestra similitud. Tal vez por la proximidad geográfica. Pero otro grupo claro lo forman Aguascalientes, Tlaxcala, Nayarit, Campeche, Colima, Baja California Sur, Quintana Roo, Querétaro, Zacatecas, Durango, Morelos y Yucatán. Es importantísimo revisar los grupos formados en la sección anterior, ya que todos ellos se ubicaron en el grupo dos.

Otro grupo está formado por Guanajuato, Puebla, Michoacán y Nuevo León, que en nuestro análisis anterior se ubicaron en el grupo uno. A este se le une posteriormente el formado por Jalisco y Veracruz.

También se forma un grupo por Tabasco, Sinaloa, Tamaulipas, Coahuila, Sonora, Hidalgo, San Luis Potosí, Baja California, Chihuahua, Guerrero, Oaxaca y Chiapas.

Algo importante que podría ser útil a posteriori es que los estados de Chiapas, Oaxaca y Guerrero en ambos dendogramas han aparecido en el mismo grupo y definitivamente en grupos distintos a los que pertenecen al estado de México y el D.F. Lo cual, de acuerdo a los conglomerados que se forman, indica que los tres primeros estados se parecen mucho entre sí (al igual que los dos últimos entre ellos), pero que ambos grupos son diferentes entre ellos.

Ahora bien, se observa también que en ambos dendogramas Aguascalientes y Tlaxcala no solamente aparecen en el mismo grupo, sino que lo hacen uno al lado del otro. Aquí cabe resaltar que ambos se encuentran entre los estados con menor población y menos número de viviendas, lo cual se explica por su menor extensión geográfica, aunque haya otros estados que se muestren en igual condición, llama la atención que estos dos estados, al que se puede agregar Nayarit, Campeche Colima y Baja California Sur formen un grupo que geográficamente se muestra disperso.

Tal vez aquí cabría aventurar la hipótesis que el número y tipo de viviendas (que va de la mano con el número de habitantes) tenga alguna relación con la conformación de las creencias religiosas entre los diferentes estados, incluso independientemente de su ubicación geográfica.

Tabla 3.4. Población de 5 años y más, según sus creencias religiosas.

1	AGUASCALIENTES	765,614	15,857	6,154	58	940	6,723	8,056
2	Baja CALIFORNIA	1,637,065	159,974	54,459	369	4,852	123,743	31,464
3	Baja CALIFORNIA SUR	333,156	75,063	7,271	51	682	13,632	4,380
4	CAJON DE OTEPEQUE	432,457	79,994	26,407	47	1,050	56,973	4,785
5	COMUNIDAD DE ZARAGOZA	1,743,978	137,368	35,340	133	2,481	76,574	22,158
6	COLIMA	425,954	13,274	6,546	79	485	8,036	3,484
7	CHAMPAS	2,088,246	457,736	261,734	178	1,446	429,803	38,824
8	CHIHUAHUA	2,216,778	185,955	51,839	133	3,017	133,482	28,202
9	DISTRITO FEDERAL	8,888,482	277,400	103,885	18,380	58,281	223,086	58,893
10	DURANGO	1,142,324	48,794	23,346	17	966	36,730	11,834
11	GUANAJUATO	3,804,423	53,330	26,447	317	5,838	27,697	29,837
12	GUERRERO	2,359,763	117,511	62,028	875	10,832	81,386	23,659
13	HIDALGO	1,791,931	102,748	25,875	391	8,309	31,147	13,467
14	JALISCO	5,265,970	116,413	49,131	683	4,962	48,825	40,196
15	MÉJICO	10,122,231	423,068	172,313	14,084	82,137	187,663	65,990
16	MICHOACÁN DE OCAMPO	3,287,058	63,726	37,728	228	5,264	46,358	28,986
17	MORELOS	1,118,040	97,880	41,258	1,788	8,559	57,461	13,526
18	MAYAGUIT	748,578	24,313	10,535	72	1,722	23,807	8,435
19	NEWYO LEÓN	2,982,592	211,482	87,118	685	3,780	84,309	32,181
20	ONAYACA	2,387,807	234,150	89,855	1,759	8,345	120,150	28,063
21	PUEBLA	3,673,395	180,586	82,770	2,251	15,693	60,065	34,861
22	QUERÉTARO DE ARTEAGA	1,166,221	23,461	16,861	85	2,178	11,349	9,501
23	QUINTANA ROO	532,745	84,319	34,819	587	1,772	72,588	8,812
24	SAN LUIS POTOSÍ	1,860,806	83,257	28,015	57	3,149	30,705	14,548
25	SINALOA	1,946,228	85,346	45,383	75	2,038	159,828	22,822
26	SONORA	1,718,888	84,487	35,457	58	1,801	85,207	28,740
27	TABASCO	1,172,488	226,683	83,111	114	1,287	168,963	13,708
28	TAMAULIPAS	2,012,177	210,021	57,703	182	4,530	118,866	23,740
29	TLAXCALA	781,264	24,200	12,103	88	3,782	8,174	7,225
30	VERACRUZ-LIAME	5,070,065	422,973	201,193	1,334	10,508	303,951	47,853
31	YUCATÁN	1,241,108	123,162	44,057	377	1,880	50,841	11,218
32	ZACATECAS	1,130,872	23,088	11,434	37	748	12,987	9,548
		74,812,373	4,408,158	1,751,810	45,288	281,183	2,982,828	792,830

***** ANALISIS HIERARQUICO DE CLUSTERS *****

Dendograma usando liqas promedio (Creencias Religiosas)

Distancia de cluster con escala ajustada

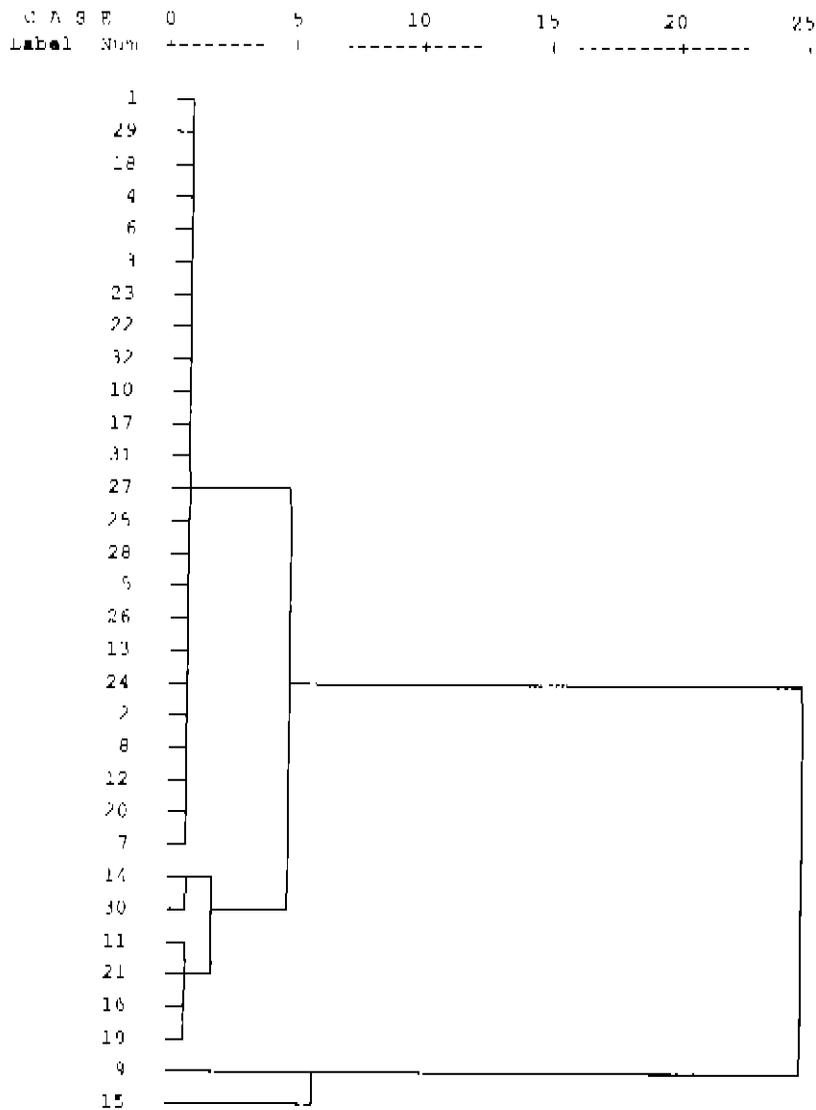


Tabla 3.4. Distribución de la población mayor de 12 años según condición conyugal.

POBLACIÓN MAYOR DE 12 AÑOS SEGÚN ESTADO CONYUGAL	EN						NO ESPECIFICADO
	SOLTERA	CASADA CIVIL Y/O RELIGIOSAMENTE	UNIÓN LIBRE	SEPARADA	DIVORCIADA	VUEDA	
1 AGUASCALIENTES	256,825	32,634	26,172	10,174	6,871	23,887	1,461
2 BAJA CALIFORNIA	578,383	667,214	258,546	52,238	29,579	59,061	10,030
3 BAJA CALIFORNIA SUR	110,090	136,726	40,073	8,057	1,668	9,501	1,528
4 CAMPECHE	172,505	235,878	44,467	14,117	4,743	18,285	983
5 COAHUILA DE ZARAGOZA	567,004	851,334	106,004	31,049	20,954	30,252	4,734
6 COLIMA	142,733	163,671	39,615	10,446	4,913	15,648	865
7 CHIHUAHUA	938,627	976,239	486,130	65,467	11,314	66,344	3,307
8 CHIHUAHUA	728,867	955,449	279,268	52,158	39,263	90,199	6,493
9 DISTRITO FEDERAL	2,558,746	2,718,423	607,692	236,834	177,541	301,261	13,261
10 DURANGO	368,684	468,819	95,519	23,009	11,033	46,587	2,841
11 GUANAJUATO	1,272,871	1,567,323	407,451	48,743	18,878	65,844	3,577
12 GUERRERO	754,237	938,354	196,083	61,028	13,424	107,156	5,457
13 HIDALGO	578,113	645,725	201,240	30,351	10,664	40,004	2,300
14 JALISCO	1,812,095	2,160,332	236,766	84,587	41,066	192,646	10,325
15 MEXICO	5,371,286	6,064,713	608,533	211,800	71,000	300,000	15,000
16 MICHOACAN DE OCAMPO	1,064,141	1,384,719	139,831	50,606	15,775	124,466	8,046
17 MORELOS	588,481	652,116	147,305	30,365	12,487	52,000	2,800
18 NAYARIT	239,055	263,842	104,820	19,838	4,865	30,550	1,172
20 OAXACA	895,120	1,033,061	250,826	53,802	10,087	133,896	6,441
22 QUERÉTARO DE ARTEAGA	395,302	465,379	59,250	15,659	8,540	35,726	3,022
24 SAN LUIS POTOSÍ	614,148	762,266	117,005	33,564	9,133	69,194	3,335
26 SONORA	586,702	680,035	211,446	44,591	24,080	65,540	4,723
28 TAMAULIPAS	715,356	924,338	213,412	53,996	24,720	85,285	5,267
30 VERACRUZ-Llave	1,790,087	1,920,044	814,129	170,414	35,695	244,300	9,893
32 ZACATECAS	358,564	492,962	42,646	14,930	6,051	39,167	2,010

Ejemplo adicional

Como ejemplo adicional y para mostrar la diversidad de opciones con las que se puede contar al emplear estos métodos, utilizaremos la tabla 3.4 para agregar un ejemplo de esta técnica, empleando ahora variables de otra índole, en este caso, el estado conyugal.

Aunque la tabla, por sí misma nos proporciona ciertos datos interesantes, como es el hecho que el Distrito Federal y el Estado de México tienen el mayor número de solteros, revisemos los datos que proporciona la "corrida" en el paquete estadístico de los datos anteriores.

La generación de clusters se hará considerando la distancia euclídea, para simplificar los procesos y obtener algo ya conocido de acuerdo a lo antes mencionado.

A continuación el encabezado de la salida del paquete.

Cluster

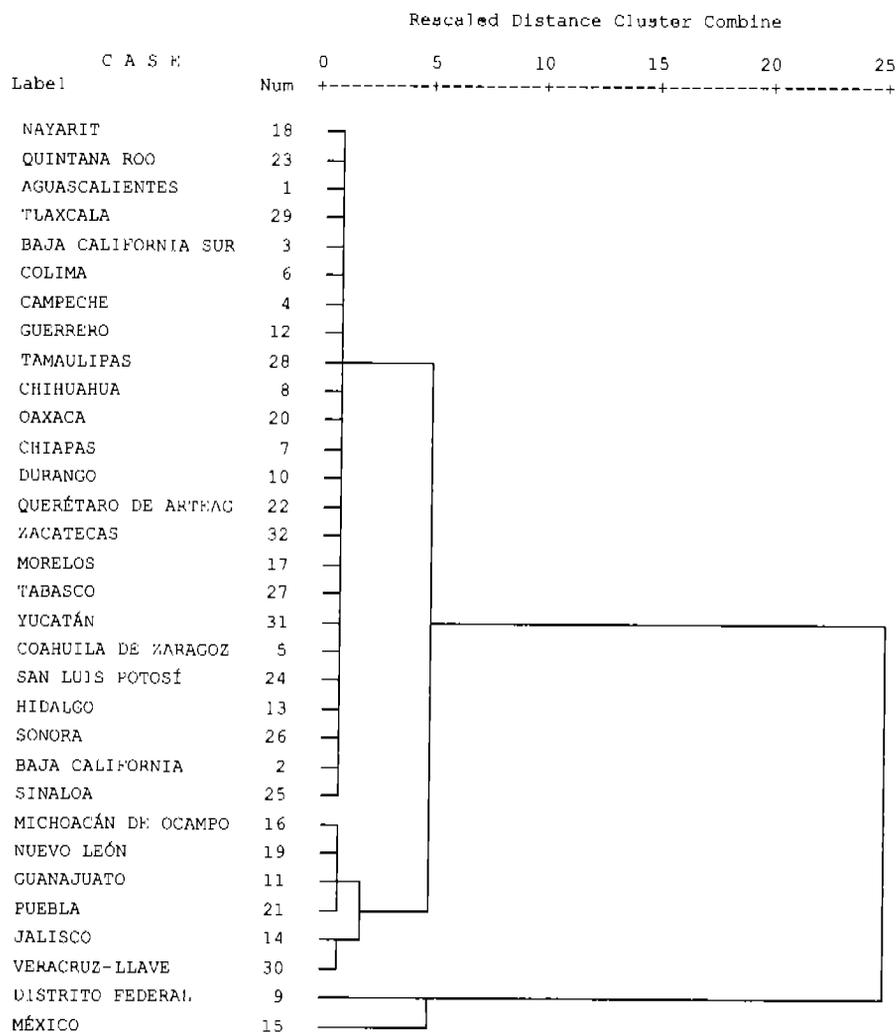
Notes

Output Created	22-SEP-2004 22:02:43	
Comments		
Input	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	32
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
Syntax	CLUSTER SOLTERA CASADA_CIVIL YO_RELIGIOSAME NTE EN_UNIN_LIBRE SEPARADA DIVORCIADA VIUDA NO_ESPECIFICADO /METHOD BAVERAGE /MEASURE= SEUCLID /ID=V1 /PRINT SCHEDULE /PLOT DENDROGRAM VICICLE.	
Resources	Elapsed Time	0:00:00.05

Y finalmente, el dendograma obtenido se muestra en la siguiente página.

Aplicación del análisis multivariado a los datos del censo del 2000 en México

***** H I E R A R C H I C A L C L U S T E R A N A L Y S I S *****
 Dendrogram using Average Linkage (Between Groups)



De nuestra observación a simple vista, se corrobora que el Estado de México y el Distrito Federal forman un grupo semejante y separado de los demás, aunque más acentuada la diferencia del Estado de México, el cual en otro tipo de análisis y dependiendo del tipo de técnica que se analice, pudiera incluso ser considerado una observación discordante.

En un primer grupo, no necesariamente homogéneo, se ubican Nayarit, Quintana Roo, Aguascalientes, Tlaxcala, Baja California Sur, Colima, Campeche y Guerrero.

Un segundo grupo lo forman Tamaulipas, Chihuahua, Oaxaca, Chiapas, Durango, Querétaro, Zacatecas, Morelos, Tabasco, Yucatán, Coahuila, San Luis Potosí, Hidalgo, Sonora, Baja California y Sinaloa.

Y, finalmente, un tercer grupo se vería conformado por Michoacán, Nuevo León, Guanajuato, Puebla, Jalisco y Veracruz. Estos tres grupos, adicionales al integrado por el Estado de México y el Distrito Federal.

Sin embargo, hay algo que no debe perderse de vista, y es que los resultados anteriores fueron obtenidos en base a números absolutos de cada estado. Desde luego, como el Estado de México es quien tiene mayor número de población, seguido por el Distrito Federal, es lógico que su comportamiento en números absolutos sea similar. De igual manera, los demás grupos si se realiza un estudio muy detallado, podría arrojar que en el fondo coinciden en algunos aspectos con lo obtenido al analizar la tabla de número de habitantes por estado.

Para solventar lo anterior, se recurre a realizar el mismo análisis, pero esta vez utilizando la tabla en términos porcentuales, es decir, en lugar de tomar el dato "número de solteras en el estado X", se utilizará el "porcentaje de solteras en el estado X, respecto a su población total". No hay que perder de vista que el número de solteros y población total para este cuadro, corresponde únicamente a la población mayor de 12 años.

Los porcentajes se obtendrán con base en la siguiente tabla, que muestra el total de habitantes mayores de 12 años de cada estado.

AQUASCALIENTES	689,883
BAJA CALIFORNIA	1,855,051
BAJA CALIFORNIA SUR	810,577
CAMPECHE	490,978
COAHUILA DE ZARAGOZA	1,684,815
COLIMA	377,891
CHIAPAS	2,571,828
CHIHUAHUA	2,151,697
DISTRITO FEDERAL	6,674,874
DURANGO	1,016,482
GUANAJUATO	3,243,850
GUERRERO	2,076,739
HIDALGO	1,588,892
JALISCO	4,537,817
MÉXICO	9,085,039
MICHOACÁN DE OCAMPO	2,787,584
MORELOS	1,084,887
NAYARIT	664,142
NUEVO LEÓN	2,889,264
OAXACA	2,383,233
PUEBLA	3,470,870
QUERÉTARO DE ARTEAGA	982,878
QUINTANA ROO	812,331
SAN LUIS POTOSÍ	1,808,645
SINALOA	1,634,749
SONORA	1,617,117
TABASCO	1,344,814
TAMAULIPAS	2,022,374
TLAXCALA	688,475
VERACRUZ-Llave	4,984,552
YUCATÁN	1,212,854
ZACATECAS	956,340

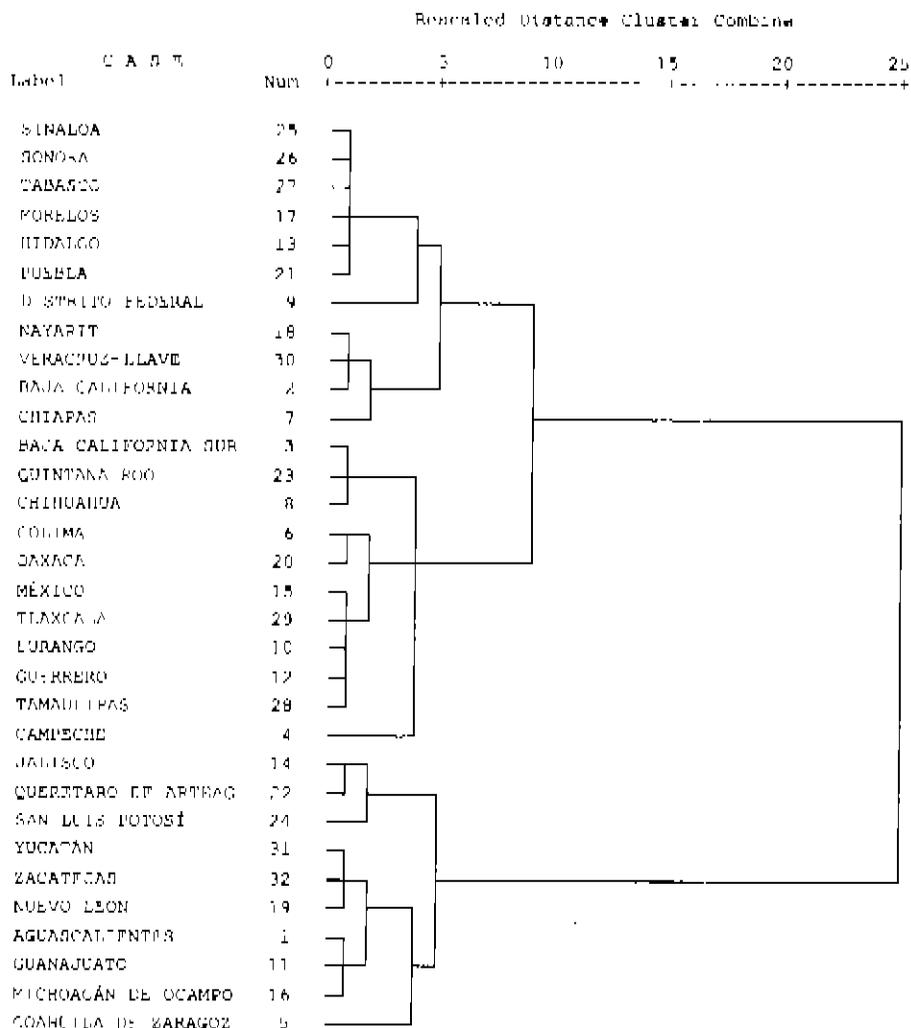
Tabla 3.5. Distribución porcentual de la población mayor de 12 años según condición conyugal

	SOLTERA	CASADA CIVIL Y/O EN UNIÓN RELIGIOSAMENTE LIBRE	SEPARADA	DI/CRUJADA	VUJDA	NO ESPECIFICADO
1 AGUASCALIENTES	38.62%	50.37%	4.27%	1.65%	0.07%	0.22%
2 BAJA CALIFORNIA	34.95%	40.31%	15.62%	3.16%	1.79%	3.57%
3 BAJA CALIFORNIA SUR	35.85%	43.63%	12.69%	2.89%	1.51%	3.20%
4 CAMPECHE	35.13%	48.04%	9.06%	2.88%	0.97%	0.20%
5 COAHUILA DE ZARAGOZA	54.10%	51.57%	6.45%	2.29%	1.25%	4.11%
6 COLIMA	37.77%	43.31%	10.48%	2.76%	1.30%	4.14%
7 CHIHUAHUA	36.84%	37.78%	18.89%	2.56%	0.67%	1.87%
8 CHIHUAHUA	33.87%	44.40%	12.98%	2.42%	1.82%	4.19%
9 DISTRITO FEDERAL	35.35%	40.72%	16.22%	3.58%	1.94%	4.55%
10 DURANGO	36.27%	46.12%	9.40%	2.26%	1.09%	4.58%
11 GUANAJUATO	39.22%	51.40%	3.31%	1.37%	0.55%	3.10%
12 GUERRERO	36.34%	45.21%	9.45%	2.94%	0.65%	5.16%
13 HIDALGO	36.39%	40.65%	14.56%	2.57%	0.67%	4.57%
14 JALISCO	39.93%	47.61%	5.22%	1.86%	0.90%	4.25%
15 MEXICO	38.85%	44.73%	10.79%	2.47%	0.55%	3.73%
16 MICHOACAN DE OCAMPO	38.17%	49.67%	5.02%	1.82%	0.57%	4.47%
17 MORELOS	38.94%	41.31%	12.45%	3.25%	1.17%	4.13%
18 NAYARIT	35.99%	39.73%	15.78%	2.99%	0.73%	4.60%
19 NAYARIT	37.55%	43.15%	11.15%	2.55%	0.75%	4.15%
20 OAXACA	37.56%	43.35%	10.52%	2.26%	0.42%	5.62%
21 OAXACA	37.56%	43.35%	10.52%	2.26%	0.42%	5.62%
22 QUERÉTARO DE ARTEAGA	40.22%	47.35%	6.03%	1.59%	0.87%	3.63%
23 QUERÉTARO DE ARTEAGA	40.22%	47.35%	6.03%	1.59%	0.87%	3.63%
24 SAN LUIS POTOSI	38.18%	47.39%	7.27%	2.09%	0.57%	4.30%
25 SAN LUIS POTOSI	38.18%	47.39%	7.27%	2.09%	0.57%	4.30%
26 SONORA	36.28%	42.05%	13.08%	2.76%	1.49%	4.05%
27 SONORA	36.28%	42.05%	13.08%	2.76%	1.49%	4.05%
28 TAMAULIPAS	35.37%	45.71%	10.55%	2.67%	1.22%	4.22%
29 TAMAULIPAS	35.37%	45.71%	10.55%	2.67%	1.22%	4.22%
30 VERACRUZ-Llave	35.91%	38.52%	16.33%	3.42%	0.72%	4.90%
31 VERACRUZ-Llave	35.91%	38.52%	16.33%	3.42%	0.72%	4.90%
32 ZACATECAS	37.49%	51.55%	4.46%	1.56%	0.63%	4.10%
33 ZACATECAS	37.49%	51.55%	4.46%	1.56%	0.63%	4.10%

Directamente podemos ver el dendrograma que se genera al aplicar la técnica sobre la tabla anterior.

***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Average Linkage (Between Groups)



Inmediatamente se observa que la conformación de grupos cambia drásticamente. En primera instancia ya no vemos al Distrito Federal en el mismo grupo que el Estado de México. Asimismo, Oaxaca y Chiapas, que en principio son semejantes en muchas condiciones, aquí parecen variar su comportamiento.

En principio se observan hasta nueve conglomerados, sin embargo, para materializar un agrupamiento que vaya de acuerdo con nuestros objetivos, nos concretaremos en la formación de 3 grupos.

Los grupos quedarían de la siguiente manera: En el grupo uno se incluye a Sinaloa, Sonora, Tabasco, Morelos, Hidalgo y Puebla. Aquí se agregaría al Distrito Federal, y con un comportamiento no tan homogéneo, se agregan Nayarit, Veracruz, Baja California y Chiapas.

En un segundo grupo se encuentran Baja California Sur, Quintana Roo y Chihuahua, además de Colima, Oaxaca, Estado de México, Tlaxcala, Durango, Guerrero, Tamaulipas y Campeche.

Finalmente, el tercer grupo lo forman Jalisco, Querétaro, San Luis Potosí, Yucatán, Zacatecas, Nuevo León, Aguascalientes, Guanajuato, Michoacán y Coahuila.

De estos grupos, se observa que en algunos casos la situación geográfica marca una semejanza en comportamiento, sin embargo, hay casos en los que no pareciera lógica la asociación o pertenencia a ciertos grupos. Por ejemplo, el grupo bien identificado que forman Yucatán, Zacatecas y Nuevo León, nos habla de estados aparentemente heterogéneos, sin embargo, según este estudio, se muestra que el comportamiento respecto a las variables en estudio es muy similar en estos 3 casos. Adicionalmente, hay que observar que el Distrito Federal no se encuentra "cerca" de Nuevo León.

También llama la atención que Chiapas y Oaxaca pertenezcan a diferentes grupos, pues uno esperaría que en primera instancia pertenecieran al mismo, sin embargo, habría que observar de manera muy importante (aunque este dato no se ha mencionado de manera explícita) que Chiapas es el estado con menor número de católicos en el país, lo cual incide de manera directa en el régimen conyugal bajo el cual viven sus habitantes.

Capítulo 4.

Análisis Factorial

Objetivo

Esta técnica multivariada parte del hecho que se cuenta con variables que presentan cierta correlación entre ellas. El objetivo es reducir la dimensión del espacio de variables, de tal manera que se tenga un menor número de variables, las cuales surgen o se crean a partir de las originales. Este nuevo conjunto de variables explica la mayor parte de la variabilidad del total de los datos.

Con ello, se tiene una matriz de datos de menor dimensión, lo que la hace más manejable para cuestiones de manipulación numérica.

Adicionalmente, puede ser visto como un método para ortogonalizar un conjunto p -dimensional de datos.

Antecedentes

Muchos autores mencionan que el campo donde se originó esta técnica es el de la psicología, pues a principios del siglo XX se consideraba que existían aspectos desconocidos vinculados con la personalidad y la Inteligencia, a los que se podía acceder por medio de evaluaciones indirectas, cuyo análisis profundo debería detectar la presencia de ciertos esquemas básicos responsables del comportamiento. De esta manera, el análisis factorial pretende superar las múltiples y complicadas apariencias externas en busca de un conjunto reducido de factores internos explicativos.

Uno de los primeros en aplicar esta técnica fue el psicólogo Charles Spearman, quien intentó resolver el problema acerca de si la inteligencia respondía a un solo factor general o respondía a un conjunto de habilidades específicas.

Así, aunque fueron los matemáticos (Pearson en 1901 y Hotelling en 1933) quienes llevaron a cabo los planteamientos iniciales, fueron los psicólogos los primeros en aplicarlos en su práctica profesional.

Posterior a las aplicaciones en el área de la psicología, esta técnica se aplicó en la mercadotecnia, aunque su principal desarrollo en esta área data de mediados de los años 80 (del siglo XX). Adicionalmente, se ha aplicado en diversas áreas, tales como la sociología, demografía, etc.

Modelo básico de factores

Si consideramos que estamos partiendo de una matriz que contiene n individuos y p variables, estamos refiriéndonos, al igual que en el capítulo anterior, a una matriz de $n \times p$. En este caso, como se planteó al principio de este capítulo, el objetivo será disminuir la dimensión de nuestra matriz.

Dicha disminución puede darse en los individuos o en las variables. El caso más común y que aquí mencionaremos es el de la disminución del espacio p -dimensional, es decir, el de las variables.

Para llevar a cabo esta tarea, se parte de un supuesto muy importante: **existe correlación entre las variables**. De otra manera, no sería posible la reducción de los datos sin pérdida significativa de información.

Tipos de modelos factoriales

Existen básicamente dos tipos de modelos factoriales, el **exploratorio** y el **confirmatorio**.

La diferencia básica que existe entre ellos es la rigidez de los supuestos que se utilizan. En el análisis exploratorio se deben satisfacer ciertos criterios matemáticos, que no son muy "fuertes". En cambio, en el análisis confirmatorio, satisfacer de manera sustantiva los criterios matemáticos es fundamental.

Además, en el primer caso se trabaja con la matriz de varianza total y en el segundo con la de la varianza común. En términos muy generales, se utilizará el análisis exploratorio para la reducción de dimensiones y el segundo para determinar grados de correlación entre variables u ortogonalización de las mismas, aunque esto no se debe tomar como una regla.

El tipo de modelo que se analizará en el presente capítulo, corresponde al relativo al análisis exploratorio, en particular, la técnica de **componentes principales**.

Análisis de Componentes Principales

Este método consiste en, partiendo de una matriz de datos como la descrita en el capítulo anterior, obtener un conjunto de p variables, cada una de las cuales será una combinación lineal de las otras.

Estas nuevas variables serán los *componentes*, y serán ordenadas de tal manera que la primera tenga el mayor grado de explicación de la variabilidad en los datos originales y la última sea la que menos la explica.

De esta manera, serán seleccionados los primeros k componentes, los cuales explicarán el mayor porcentaje de la variabilidad de la matriz de datos inicial.

Desde el punto de vista de obtener las combinaciones lineales anteriormente descritas, el problema se reduce en encontrar los eigenvalores y eigenvectores para una matriz definida semipositiva.

Pasos para la elaboración del análisis factorial

Aunque no es una metodología general, es importante considerar los siguientes pasos al elaborar un análisis factorial.

1. Elaboración de la matriz de correlaciones

Lo primero es obtener una matriz donde se obtengan las correlaciones entre todas las variables consideradas. Algunos criterios de determinar si existe dicha correlación son:

- **Test de esfericidad de Bartlett:** Se utiliza para probar la hipótesis nula que afirma que las variables no están correlacionadas en la población. Se pueden tomar como válidos aquellos resultados que indiquen un valor elevado del test y cuya confiabilidad sea menor a 0.05. En este caso se rechaza la hipótesis nula y se continúa con el método.
- **Índice de Kaiser-Meyer-Olkin:** Mide la adecuación de la muestra. Indica qué tan apropiado es aplicar el análisis factorial. Los valores entre 0.5 y 1 indican que es adecuado.

La "comunalidad" es la cantidad de varianza que una variable comparte con las demás. Los eigenvalores pueden interpretarse como la cantidad de varianza explicada por cada factor. La "carga factorial" es la correlación entre las variables y los factores.

2. Extracción de los factores iniciales

Es la aplicación del método de componentes principales, que se describe en la siguiente sección. Dicho método identifica el factor que explique la mayor cantidad de la varianza en la matriz de correlación. Este recibe el nombre de "factor principal". Esta varianza explicada se resta de la matriz original, produciendo una matriz residual, en la que se aplica el mismo procedimiento, hasta que la varianza no explicada sea mínima (cero si es posible). Los factores así obtenidos son ortogonales.

3. Rotación de los factores iniciales

Los factores obtenidos se rotan para una más fácil interpretación. Existen varios métodos para llevar a cabo esta tarea, entre los de mayor uso se encuentran **quartimax** y **varimax**, este último utilizado por el paquete SPSS.

Obtención de las componentes principales

Se ha mencionado, que se pretende transformar un conjunto de variables :

$$X_1, X_2, X_3, \dots, X_n \text{ a un nuevo conjunto } Y_1, Y_2, Y_3, \dots, Y_p$$

Las Y_i 's son una combinación lineal de las X_i 's, por lo que se pueden expresar de la siguiente manera:

$$Y_j = \alpha_j X = \alpha_{j1} X_1 + \alpha_{j2} X_2 + \dots + \alpha_{jp} X_p = \sum_{i=1}^p \alpha_{ji} X_i$$

Las Y 's no están correlacionadas, y se ordenan de acuerdo a la varianza que representan, es decir, Y_1 representará la mayor varianza y Y_p la menor.

El objeto de disminuir la dimensión de los datos se consigue al seleccionar las primeras k variables ($k < p$) para obtener un nuevo conjunto, el cual representa la mayor variabilidad de los datos originales, pero que es más fácil de manejar. Lo ideal es conseguir $k=2$ ó 3 , ya que de esta manera se puede obtener una representación gráfica, lo que ayuda a la identificación de comportamientos específicos dentro del conjunto original de datos.

Con lo anterior, el problema consiste en encontrar las α_i^1 . Para ello, será importante contar con la matriz de varianzas y covarianzas, la cual generalmente se denota por Σ y se estima de la siguiente manera:

$$S = \left(\frac{1}{n} - 1 \right) \sum_{j=1}^n (x_j - \bar{X})(x_j - \bar{X})^T$$

Esta matriz puede ser vista como $S = XHX^T$, donde H es una matriz de $p \times n$.

Si consideramos que $Y = \alpha^1 X$ y que se desea maximizar la siguiente expresión:

$$\text{Var}(Y) = \text{Var}(\alpha^1 X) = \alpha^1 \Sigma \alpha$$

Nuestro problema se reduce a encontrar los α_i^1 de tal manera que se maximice dicha expresión, la cual se deduce del planteamiento inicial para las Y_i 's.

Si agregamos la restricción que α_i^t sea normalizado, es decir, que se cumpla la igualdad $\alpha_i^t \cdot \alpha_i = 1$, se elimina el hecho de obtener la maximización deseada para diferentes α_i^t .

Con lo anterior, el problema se plantea como:

$$\text{Maximizar } \alpha_i^t \Sigma \alpha_i$$

$$\text{Sujeto a: } \alpha_i^t \cdot \alpha_i = 1$$

Mediante el uso de multiplicadores de Lagrange, la solución al problema se da con la solución del siguiente sistema de ecuaciones:

$$(\Sigma - \lambda_i I) \cdot \alpha_i = 0 \quad \text{lo cual equivale a encontrar las raíces del determinante } |(\Sigma - \lambda_i I)| = 0.$$

Dicha matriz tiene p raíces, que corresponde a los valores propios de λ_i ; a cada valor propio se asocia el respectivo eigenvector α_i y al normalizar el vector α_i , éste corresponderá a las componentes principales.

Como Σ está definida semipositiva no habrá raíces negativas. Una vez que se ordenan los valores propios de mayor a menor y se obtengan los eigenvectores correspondientes, las componentes principales estarán dadas por:

$$Y_1 = \alpha_1^t X; Y_2 = \alpha_2^t X; \dots; Y_p = \alpha_p^t X$$

Finalmente, cabe mencionar la siguiente propiedad de los valores propios λ_i .

$$\sum \lambda_i = \text{traza}(\Sigma)$$

$$\sum \bar{\lambda}_i = \text{traza}(S), \text{ donde } \bar{\lambda}_i \text{ es un estimador óptimo de } \lambda_i.$$

Aplicación a los datos del censo

Para aplicar esta técnica a los datos del censo, se eligieron variables relacionadas con la fecundidad por grupos quinquenales de edad, mortalidad y distribución porcentual de hombre y mujeres en cada uno de los estados.

Las variables utilizadas fueron las siguientes:

Mortalidad, que es igual al total de hijos fallecidos entre el total de nacidos vivos.

Promedio de hijos por cada mujer mayor de 12 años, que es el total de hijos nacidos vivos entre el total de la población femenina mayor de 12 años.

Tasa de fecundidad, distribuida por grupos quinquenales de edad: 12-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49 años.

La tasa global de fecundidad en cada estado.

Porcentaje de hombres, respecto a la población total de cada estado.

Porcentaje de mujeres, respecto al total de la población por estado.

Las 13 variables, todas relacionadas con una misma característica de la población de cada estado, pueden ser muy útiles para describir el comportamiento poblacional, ya que de la mano con la fecundidad y la mortalidad, van aspectos como la salud, educación y empleo.

En la tabla 4.1 se describe, en forma tabular, los valores de las variables descritas, por lo que el objetivo del análisis será reducir la dimensión de esta matriz, en un número menor de variables.

Sin embargo, el objetivo de reducir el número de variables es para algún fin específico. No se reduce la dimensión de los datos sin aplicar después algún método o darle algún tratamiento a la información.

En este caso, ya observamos en el capítulo anterior cómo funciona el análisis de clusters o conglomerados. Por ello, y para aprovechar las ventajas que el paquete estadístico SPSS nos ofrece, una vez reducida la dimensión de los datos, en un nuevo conjunto de variables, aplicaremos nuevamente la técnica de análisis de conglomerados, con ello, obtendremos elementos para realizar una confrontación de los grupos obtenidos en el capítulo anterior.

Este nuevo conjunto, como se ha venido mencionado, contendrá un cierto porcentaje de la varianza de los datos originales. Aunque se espera que sea la mayor posible y nos proporcione una descripción bastante completa.

Tabla 4.1: Variables relacionadas con la fecundidad, mortalidad y distribución por sexo.

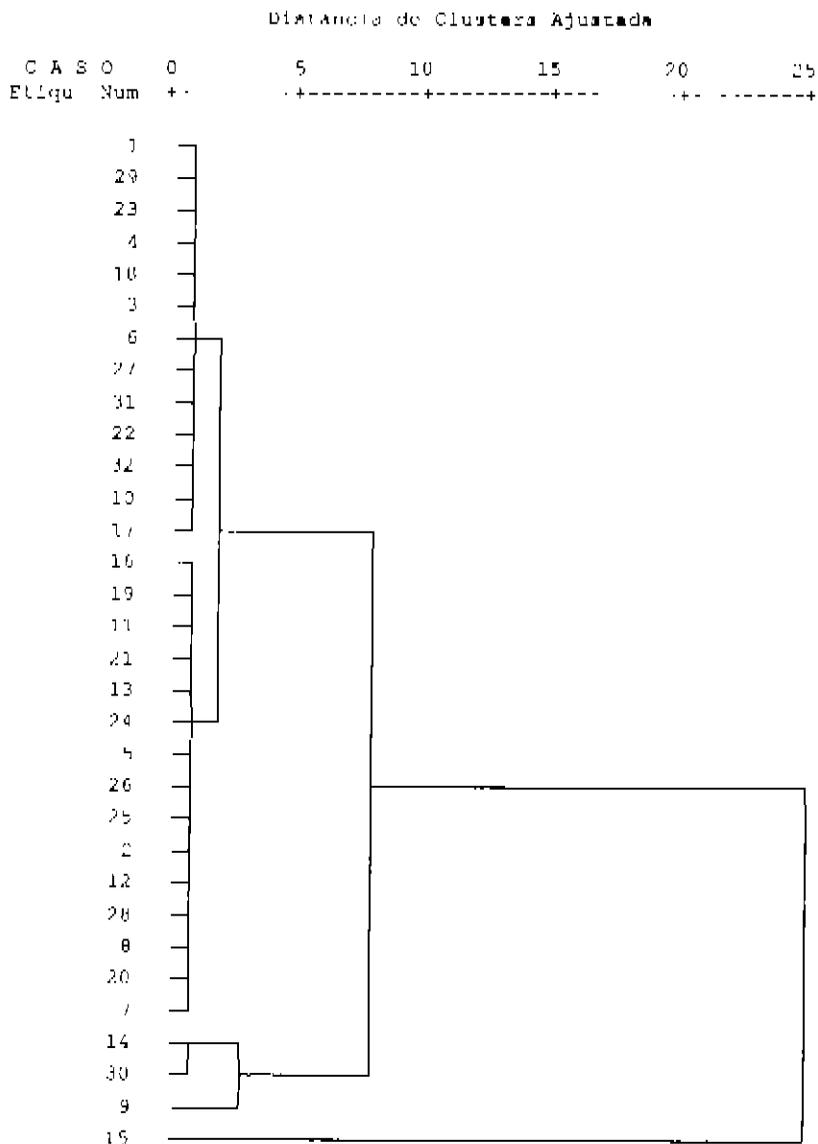
ENTIDAD FEDERATIVA	PORCENTAJE (HIJOS FALLECIDOS / NACIDOS VIVOS)	PROMEDIO de hijos nacidos vivos/pop. Femenina >12	12 - 14 AÑOS	15 - 19 AÑOS	20 - 24 AÑOS	25 - 29 AÑOS	30 - 34 AÑOS	35 - 39 AÑOS	40 - 44 AÑOS	45 - 49 AÑOS	Tasa Global de fecundidad	%Hombres	%Mujeres
02 BAJA CALIFORNIA	7.94	2.39	0.0018	0.0839	0.165	0.1517	0.1072	0.057	0.0197	0.005	2.9518	0.5035771	0.496422924
04 CAMPECHE	9.47	2.59	0.0018	0.0819	0.167	0.151	0.0988	0.0519	0.0191	0.005	2.8812	0.498537	0.501463032
06 COLIMA	10.53	2.63	0.0011	0.0623	0.15	0.1416	0.1003	0.0505	0.0152	0.004	2.6232	0.4942474	0.50575257
08 CHIHUAHUA	10.14	2.55	0.0015	0.0814	0.165	0.1514	0.1053	0.0549	0.0169	0.004	2.8996	0.4978769	0.502123058
10 DURANGO	11.62	2.91	0.0008	0.0719	0.186	0.1643	0.1176	0.064	0.0223	0.005	3.0704	0.4897771	0.510222889
12 GUERRERO	13.82	3	0.0016	0.0783	0.191	0.186	0.1407	0.0892	0.0394	0.012	3.6861	0.4842393	0.515760725
14 JALISCO	10.55	2.65	0.0008	0.0528	0.149	0.1611	0.1264	0.0725	0.026	0.005	2.9685	0.4856438	0.514356212
16 MICHOACÁN DE OCAMPO	11.89	2.97	0.0011	0.0608	0.16	0.1672	0.1301	0.0813	0.0352	0.009	3.2203	0.4794876	0.520512376
18 NAYARIT	11.75	2.92	0.0012	0.0757	0.168	0.1539	0.1113	0.0577	0.0202	0.004	2.9565	0.4956866	0.504333368
20 OAXACA	14.27	2.91	0.0011	0.0678	0.168	0.1656	0.123	0.0785	0.0371	0.01	3.2538	0.4819771	0.518022895
22 QUERÉTARO DE ARTEAGA	10.56	2.59	0.0006	0.0501	0.148	0.1629	0.1244	0.0732	0.0307	0.007	2.9816	0.4849128	0.515087168
24 SAN LUIS POTOSÍ	11.62	2.91	0.0009	0.0613	0.162	0.1646	0.1354	0.0833	0.0338	0.007	3.2388	0.4874561	0.512543925
26 SONORA	8.15	2.55	0.0014	0.0745	0.16	0.1562	0.1103	0.0522	0.0167	0.003	2.8675	0.5009497	0.49905028
28 TAMAULIPAS	8.6	2.42	0.0014	0.065	0.146	0.1457	0.1027	0.0489	0.015	0.003	2.637	0.4939209	0.506079059
30 VERACRUZ-Llave	11.65	2.63	0.0016	0.0694	0.156	0.1383	0.0954	0.0531	0.0197	0.006	2.688	0.485624	0.514376011
32 ZACATECAS	12.68	3.11	0.0007	0.0556	0.161	0.1677	0.1323	0.0824	0.0314	0.007	3.1901	0.4828444	0.517155606

Resultados obtenidos con el paquete estadístico SPSS

Dendograma del Análisis de Factores

* * * * * ANÁLISIS DE CLUSTER JERARQUICOS * * * * *

Dendograma usando liga promedio (entre Grupos)



Correlación Matrix

	PORCENTAJE DE HIJOS FALLECIDOS / NACIDOS VIVOS)	PROMEDIO de hijos nacidos vivos/pop. Femenina (>12)	12 - 14 AÑOS	15 - 19 AÑOS	20 - 24 AÑOS	25 - 29 AÑOS	30 - 34 AÑOS	35 - 39 AÑOS	40 - 44 AÑOS	45 - 49 AÑOS	Tasa Global de fecundidad	Hombres	Mujeres
Correlix PORCENTAJE DE HIJOS FALLECIDOS / NACIDOS VIVOS)	1.000	.822	-.275	-.055	.493	.601	.608	.687	.754	.725	.571	-.639	.939
PROMEDIO de hijos nacidos vivos/pop. Femenina (>12)	.822	1.000	-.208	.077	.593	.774	.713	.691	.691	.602	.769	-.416	.416
12 - 14 AÑOS	-.275	-.208	1.000	.859	.503	-.087	-.344	-.251	-.178	.137	.109	.655	-.655
15 - 19 AÑOS	-.055	.077	.859	1.000	.740	.153	-.204	-.184	-.123	.175	.300	.659	-.659
20 - 24 AÑOS	.493	.593	.503	.740	1.000	.696	.382	.414	.468	.629	.805	.234	-.234
25 - 29 AÑOS	.601	.774	-.087	.153	.696	1.000	.884	.813	.774	.677	.939	-.191	.191
30 - 34 AÑOS	.608	.713	-.344	-.204	.382	.884	1.000	.951	.876	.869	.844	-.448	.448
35 - 39 AÑOS	.687	.691	-.251	-.184	.414	.813	.951	1.000	.967	.813	.859	-.517	.517
40 - 44 AÑOS	.754	.691	-.178	-.123	.468	.774	.876	.967	1.000	.869	.858	-.556	.556
45 - 49 AÑOS	.725	.602	.137	.175	.629	.677	.669	.813	.969	1.000	.839	-.375	.375
Tasa Global de fecundidad	.571	.769	.109	.300	.605	.939	.844	.859	.858	.839	1.000	-.176	.176
%Hombres	-.639	-.416	.655	.659	.234	-.191	-.448	-.517	-.556	-.375	-.176	1.000	-1.000
%Mujeres	.939	.416	-.655	-.659	-.234	.191	.448	.517	.556	.375	.176	-1.000	1.000

Lo primero que arroja el paquete es la matriz de correlaciones. Un dato importante que se observa de esta matriz, es la alta correlación que existe entre la tasa global de fecundidad y la tasa de fecundidad en mujeres de 25-29 años. Este simple dato, nos pudiera dar pauta para suponer que las mujeres en México están decidiendo tener hijos a una edad mayor a los 25 años.

Component Matrix^a

	Component		
	1	2	3
40 - 44 AÑOS	.956	-2.5E-02	-1.7E-02
35 - 39 AÑOS	.945	-5.6E-02	-.201
30 - 34 AÑOS	.908	-6.4E-02	-.394
Tasa Global de fecundidad	.902	.408	-.124
25 - 29 AÑOS	.866	.286	-.323
45 - 49 AÑOS	.853	.235	.227
PORCENTAJE (HIJOS FALLECIDOS / NACIDOS VIVOS)	.852	-9.6E-02	.398
PROMEDIO de hijos nacidos vivos/pob. Femenina >12)	.845	8.91E-02	8.63E-02
15 - 19 AÑOS	-7.0E-02	.948	.232
12 - 14 AÑOS	-.237	.853	.281
20 - 24 AÑOS	.555	.784	.133
%Mujeres	.566	-.744	.312
%Hombres	-.566	.744	-.312

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Posteriormente se realiza una rotación, que en este caso se eligió fuera por el método varimax. Para un mayor detalle respecto a dicho método referirse a Bernstein [6].

De dicha rotación, se obtiene la siguiente tabla.

Rotated Component Matrix^a

	Component		
	1	2	3
25 - 29 AÑOS	.982	8.30E-02	5.34E-02
30 - 34 AÑOS	.945	-.270	.115
Tasa Global de fecundidad	.941	.267	.200
35 - 39 AÑOS	.901	-.195	.292
40 - 44 AÑOS	.842	-9.9E-02	.442
PROMEDIO de hijos nacidos vivos/pob. Femenina >12)	.729	5.25E-02	.442
45 - 49 AÑOS	.712	.239	.520
15 - 19 AÑOS	6.41E-02	.970	-.118
12 - 14 AÑOS	-.118	.904	-.144
20 - 24 AÑOS	.817	.735	.141
PORCENTAJE (HIJOS FALLECIDOS / NACIDOS VIVOS)	.583	-5.2E-03	.759
%Mujeres	.198	-.817	.743
%Hombres	-.198	.817	-.743

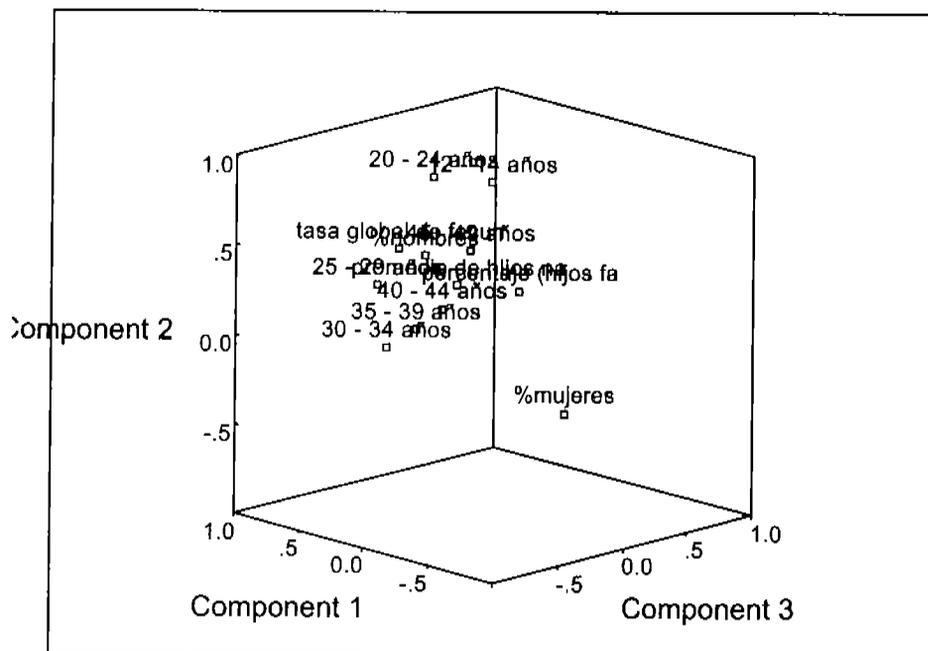
Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

En nuestro espacio tridimensional, rotado, el conjunto de variables se observa de la siguiente manera.

Component Plot in Rotated Space



Como se mencionó anteriormente, ya obtuvimos 3 componentes, que explican más del 90% de la variabilidad de los datos originales, pero para darle una verdadera utilidad al conjunto de datos así obtenido, se aplicará la técnica de conglomerados. Con ello, observaremos cómo se agrupan los estados en base a estas variables, que son derivadas de las observadas.

En primer lugar, la tabla 4.2 nos muestra los valores de las variables obtenidas por el método de componentes principales.

Tabla 4.2. Variables obtenidas mediante Componentes Principales

Estado	C1	C2	C3
01 AGUASCALIENTES	1.15198	-1.37601	-0.76018
02 BAJA CALIFORNIA	-0.1133	1.13928	-1.19787
03 BAJA CALIFORNIA	-0.83183	1.17516	-1.67823
04 CAMPECHE	-0.67477	1.35768	-0.00886
05 COAHUILA DE ZAR	-0.69372	0.29233	-0.58182
06 COLIMA	-0.83978	-0.09181	0.10806
07 CHIAPAS	1.01017	2.2332	0.88743
08 CHIHUAHUA	-0.41335	0.93398	-0.26079
09 DISTRITO FEDERA	-2.91408	-1.86822	1.06724
10 DURANGO	0.43346	-0.00354	0.11313
11 GUANAJUATO	1.66348	-1.06857	0.26378
12 GUERRERO	1.97374	1.05605	1.05536
13 HIDALGO	-0.69829	0.00007	1.27426
14 JALISCO	0.5536	-1.27118	-0.43272
15 MÉXICO	-0.71272	-0.45088	0.15082
16 MICHOACÁN DE OC	0.94901	-0.475	0.94828
17 MORELOS	-0.95298	-0.32315	0.90702
18 NAYARIT	-0.05117	0.83458	0.16558
19 NUEVO LEÓN	-0.81971	-1.16555	-1.80441
20 OAXACA	0.71143	0.25701	1.74089
21 PUEBLA	0.49897	-0.04621	1.24308
22 QUERÉTARO DE AR	0.85974	-1.4439	-0.30445
23 QUINTANA ROO	0.14987	1.43769	-1.86128
24 SAN LUIS POTOSÍ	1.29854	-0.68388	-0.21805
25 SINALOA	0.05438	0.47148	-0.83404
26 SONORA	-0.04859	0.34084	-1.41541
27 TABASCO	-0.87205	0.8857	0.57739
28 TAMAULIPAS	-0.89586	-0.11498	-0.51008
29 TLAXCALA	0.21248	-0.54922	0.18888
30 VERACRUZ-LLAVE	-1.40383	0.89172	1.75295
31 YUCATÁN	0.15282	-1.15909	-1.02882
32 ZACATECAS	1.16238	-0.81537	0.45907

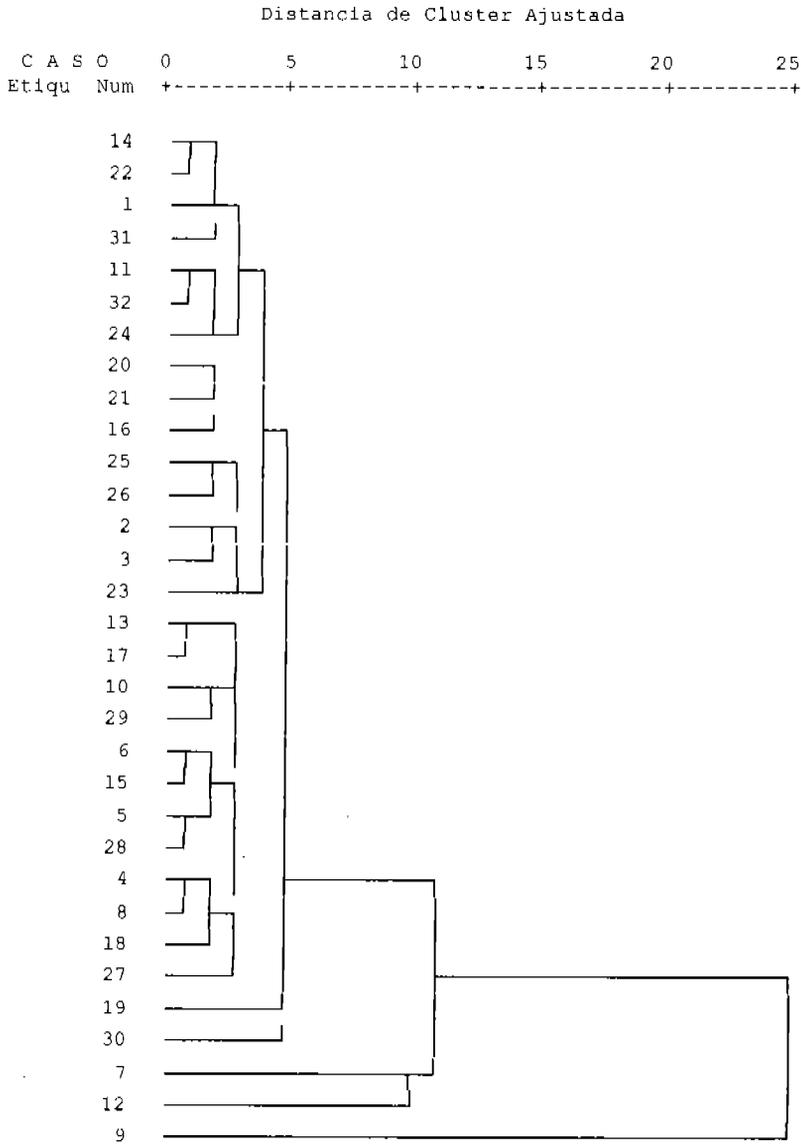
Este conjunto de variables es el que utilizaremos para aplicar la técnica de conglomerados.

Para ello, utilizamos el método de vecino más cercano, con una liga simple.

De los resultados que proporciona el paquete estadístico mencionado, nos reduciremos a estudiar el dendograma, que es la representación gráfica que se obtiene al realizar el agrupamiento de los individuos en las diferentes etapas.

***** ANÁLISIS DE CLUSTER JERÁRQUICO *****

Dendograma usando liga simple



Si observamos la tercera etapa, se distinguen 3 grupos. El más diferenciado de todos, y en el que incluso sus miembros no se unen sino hasta las etapas finales está formado por los estados 30,7, 12 y 9.

El segundo, estaría formado por los estados 13,17,10,29,6,5,15,28,4,8,18,27 y 19.

El otro grupo lo forman los estados 14,22,1,31,11,32,24,20,21,16,25,26,2,3,23.

Si detallamos los estados por grupos quedarían de la siguiente manera:

Grupo A	Grupo B	Grupo C
D.F	Hidalgo	Jalisco
Guerrero	Morelos	Querétaro
Chiapas	Durango	Aguascalientes
Veracruz	Tlaxcala	Yucatán
	Colima	Guanajuato
	Coahuila	Zacatecas
	Edo. México	San Luis Potosí
	Tamaulipas	Oaxaca
	Campeche	Puebla
	Chihuahua	Michoacán
	Nayarit	Sinaloa
	Tabasco	Sonora
	Nuevo León	Baja California
		Baja California Sur
		Quintana Roo

Los integrantes del grupo C, son explicables en el sentido de la cercanía geográfica de estados como Baja California, Baja California Sur, Sonora y Sinaloa. Algo que llama la atención respecto a este grupo, que de hecho es el primero que se forma al estar realizando el análisis, es que a excepción de Aguascalientes, San Luis Potosí, Puebla y Sonora, los demás estados pertenecen al grupo de gran variedad turística, y en gran medida el turismo forma parte esencial de los ingresos que se obtienen en esas entidades federativas.

A este respecto, cabría especular la influencia que tiene en el comportamiento poblacional, en lo relativo a la fecundidad, el turismo, tanto nacional como extranjero que durante todo el año visita dichos estados. Sin embargo, también cabe aclarar que los puntos turísticos se ubican en lugares específicos de dichas entidades, por lo que para afirmar esta supuesta influencia se debe llevar a cabo un estudio más detallado y específico.

El grupo B es bastante heterogéneo, destacando el Estado de México, ya que es el de mayor población del país.

En lo relativo al grupo A, no se debe interpretar como que "el comportamiento de Chiapas es parecido al del D.F.". Más bien habría que hacer mención del comportamiento singular de cada uno de los estados que pertenecen a este grupo.

La interpretación, más bien iría en el sentido que el comportamiento de estas entidades federativas, de acuerdo a las variables originales seleccionadas, no se parece a los demás estados.

Comparativamente con el cuadro obtenido en el capítulo anterior, Guerrero y Chiapas siguen apareciendo en el mismo grupo.

Si se desea ser más exhaustivo en el análisis de los grupos obtenidos, se puede identificar regiones geográficas dentro de los grupos C y B.

Ejemplo adicional

El análisis factorial o de factores, como podemos ver, puede tener múltiples aplicaciones cuando se trata de reducir el espacio dimensional de la matriz de datos.

Para corroborar y reafirmar lo ilustrado anteriormente, se anexa otro ejemplo, para ir corroborando las hipótesis o refutándolas, respecto al comportamiento que tienen los grupos que se han ido identificando a lo largo de este trabajo.

Ahora utilizaremos variables de tipo económico, y que se refieren al nivel de ingresos de la población ocupada en cada estado. El ingreso está determinado en niveles de salario mínimo y cada variable representará el total de la población (o el porcentaje correspondiente) que tiene un determinado nivel de ingresos de acuerdo al número de salarios mínimos que recibe.

En la tabla 4.3 se tienen los datos directos y resumidos, y se incluye en la primera columna el valor "población ocupada". Las siguientes columnas contienen la distribución de esa población ocupada según el ingreso que reciben, en términos de salario mínimo.

Sin embargo, como ya vimos en el último ejemplo del capítulo anterior, esta distribución no siempre es la más adecuada para representar cantidades en ocasiones muy disparadas, por tal motivo, para el análisis y aplicación de la técnica descrita en este capítulo se utilizará la tabla 4.4, que contiene los valores en términos porcentuales, respecto a la población ocupada de cada estado, lo cual es más significativo.

Tabla 4.3. Población ocupada por estado distribuida según ingreso por trabajo en salario mínimo

AGUILA/ALBERTO	201,883	14,827	6,824	15,582	7	102,723	71,593	58,698	31,238	85,412	14,342
Baja California	508,369	11,374	7,428	18,171	4	103,411	214,861	297,153	152,867	87,285	82,726
Baja California Sur	98,914	4,267	2,467	4,842	2	46,219	34,778	7,189	18,728	7,189	8,176
Campeche	243,323	31,588	16,880	37,776	17	89,578	25,822	27,187	17,361	4,363	4,868
Coahuila de Zaragoza	822,008	28,625	11,380	28,227	82	223,983	261,848	100,880	82,582	49,779	45,886
Colima	158,182	10,248	4,467	15,486	3	82,394	35,981	34,178	17,343	7,428	8,972
Chiapas	1,295,821	128,075	128,075	274,648	8	245,178	79,424	165,523	58,282	19,653	19,653
Chihuahua	1,117,747	51,611	12,812	27,857	14	327,487	471,571	208,658	117,199	60,298	68,278
Ciudad de México	3,882,781	77,418	64,082	237,983	547	1,144,807	678,931	528,072	418,147	238,171	228,312
Coahuila	443,811	43,689	14,765	24,320	28	144,778	63,774	68,848	34,444	14,880	21,314
Guerrero	1,488,194	121,598	38,723	82,780	34	438,882	287,675	228,842	102,065	51,765	68,308
Guatemala	688,028	177,233	52,474	84,190	31	288,888	122,883	81,888	39,373	13,843	43,224
Hidalgo	728,728	82,888	38,363	13,817	32	248,511	88,794	77,363	37,475	14,712	33,883
Jalisco	1,332,771	133,271	68,851	130,879	32	641,170	84,243	324,648	288,622	83,160	133,822
México	4,462,367	285,347	100,130	361,580	21	1,591,738	1,083,184	144,818	338,031	185,761	280,276
Micrópolis de Oaxaca	1,228,008	136,228	52,828	112,751	38	387,082	183,846	162,883	73,484	32,360	71,278
Morales	550,831	30,584	18,784	47,514	8	182,385	85,914	72,224	36,878	18,115	28,765
Nuevo Léon	318,887	14,317	14,317	27,423	5	148,820	57,123	43,107	22,887	7,382	9,013
Oaxaca	1,477,887	55,185	15,184	37,228	11	328,483	426,327	228,427	144,881	88,621	75,828
Puebla	1,088,658	386,977	78,577	151,146	12	246,624	85,478	84,888	41,842	13,282	48,078
Querétaro de Arzobispo	1,659,821	228,843	88,428	225,887	27	544,872	280,880	184,884	87,284	48,752	78,083
Quintana Roo	478,988	9,284	9,284	22,128	3	148,888	82,137	81,613	48,283	29,078	38,884
San Luis Potosí	348,778	28,382	7,781	18,822	8	85,001	88,211	72,304	37,826	14,227	13,688
Sinaloa	682,285	38,284	28,773	88,103	20	215,178	85,284	88,228	48,885	21,178	48,788
Sonora	810,824	18,282	14,281	40,371	8	328,482	108,880	143,788	68,208	38,828	43,419
Tlaxcala	883,178	18,282	32,087	123,380	7	258,822	128,822	128,848	77,423	34,038	42,281
Veracruz	1,013,228	41,886	21,087	62,921	6	148,432	88,287	84,038	47,417	28,038	24,888
Yucatán	328,585	38,778	12,928	38,178	6	118,278	62,148	62,185	38,887	14,682	16,388
Zacatecas	2,288,117	288,787	163,185	482,844	183	788,344	258,206	221,881	141,881	58,821	100,881
Zahuatepec	818,448	57,188	48,328	88,414	128	214,728	88,023	87,281	37,851	18,088	23,182
Zapotlán	353,828	63,728	12,354	28,024	5	114,214	51,528	41,324	23,719	8,228	18,628

Tabla 4.4. Porcentaje de población ocupada por estado distribuida según ingreso por trabajo en salario mínimo

AGUASCALIENTES	4.470%	2.003%	4.786%	8.027%	31.826%	17.778%	9.427%	4.658%	4.331%
BATAH CALIFORNIA	1.265%	0.842%	2.115%	8.080%	31.874%	25.067%	14.853%	4.244%	7.031%
BATAH CALIFORNIA SER	2.441%	1.458%	4.730%	8.012%	28.754%	28.677%	11.738%	4.217%	5.425%
CAJEME	13.54%	6.852%	15.525%	8.075%	30.122%	24.547%	11.905%	3.437%	5.552%
COAHUILA DE ZARAGOZA	2.505%	1.377%	3.653%	8.075%	27.274%	19.538%	10.857%	5.055%	6.481%
COCHILA	5.859%	3.253%	7.748%	8.019%	19.538%	17.144%	8.849%	3.752%	3.873%
CHIHUAHUA	22.461%	10.422%	22.679%	8.027%	30.314%	6.852%	7.857%	1.381%	4.280%
CHIHUAHUA	4.571%	1.452%	2.655%	8.085%	29.312%	21.812%	14.671%	5.391%	8.184%
CIRCUITO FEDERAL	1.788%	1.788%	8.016%	8.015%	31.833%	30.459%	11.865%	4.424%	6.318%
CHILANAGO	9.700%	2.282%	5.477%	8.083%	32.538%	15.075%	7.264%	3.354%	4.285%
CHIHUAHUA	7.423%	2.818%	8.352%	8.027%	30.816%	39.714%	18.683%	3.641%	8.811%
CHILANAGO	19.358%	8.087%	14.045%	8.085%	30.257%	12.707%	14.475%	1.487%	4.267%
CHILANAGO	12.747%	5.233%	15.016%	8.044%	31.537%	23.179%	14.015%	2.018%	4.832%
CHILANAGO	5.841%	2.572%	5.547%	8.085%	27.425%	24.734%	8.241%	3.940%	6.030%
CHILANAGO	4.871%	2.373%	6.753%	8.019%	36.546%	19.782%	14.454%	3.477%	6.276%
CHILANAGO	12.247%	4.253%	8.192%	8.085%	31.555%	15.804%	6.817%	2.537%	5.810%
CHILANAGO	7.183%	3.587%	8.989%	8.019%	34.827%	17.281%	13.114%	3.476%	5.284%
CHILANAGO	11.277%	1.822%	2.828%	8.087%	31.875%	17.911%	7.144%	2.386%	2.838%
CHILANAGO	2.472%	7.442%	12.285%	8.087%	22.873%	29.277%	11.432%	4.880%	5.157%
CHILANAGO	28.788%	7.442%	12.285%	8.011%	33.807%	8.846%	1.241%	1.241%	4.570%
CHILANAGO	5.887%	5.887%	13.625%	8.011%	30.816%	12.527%	11.965%	5.247%	4.746%
CHILANAGO	5.887%	4.888%	4.888%	8.085%	29.716%	14.828%	9.825%	4.158%	4.746%
CHILANAGO	4.284%	4.284%	5.338%	8.085%	24.324%	19.714%	18.732%	5.228%	3.864%
CHILANAGO	4.405%	2.833%	4.405%	8.085%	29.716%	13.370%	10.510%	3.007%	5.864%
CHILANAGO	2.283%	1.782%	4.878%	8.085%	36.857%	14.880%	7.817%	3.832%	4.332%
CHILANAGO	1.633%	5.345%	20.549%	8.085%	24.547%	16.541%	8.538%	4.188%	5.213%
CHILANAGO	4.058%	2.107%	8.178%	8.085%	34.759%	13.589%	7.887%	3.574%	4.425%
CHILANAGO	11.794%	3.943%	11.818%	8.085%	36.852%	15.812%	8.444%	4.142%	5.146%
CHILANAGO	12.711%	6.514%	17.533%	8.044%	32.270%	14.816%	11.847%	4.746%	4.284%
CHILANAGO	8.243%	7.989%	15.508%	8.085%	34.751%	14.816%	9.243%	2.102%	4.382%
CHILANAGO	15.185%	3.483%	7.824%	8.014%	32.282%	14.585%	8.704%	2.838%	5.824%

A continuación, simplemente para ejemplificar se muestra el resultado obtenido por el paquete estadístico. En este caso se definieron 3 componentes principales, pero se reitera que el número se determina en base a las variables y tipo de estudio que se esté llevando a cabo.

Factor Analysis

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
NO RECIBE INGRESOS	.091387	.0633717	32
HASTA EL 50% DE UN S. M.	.038798	.0233198	32
MÁS DEL 50% HASTA MENOS DE UN S.M.	.088490	.0532945	32
UN S.M.	.000030	.0000585	32
MÁS DE 1 HASTA 2 S.M.	.298811	.0451884	32
MÁS DE 2 HASTA MENOS DE 3 S.M.	.172771	.0540568	32
DE 3 HASTA 5 S.M.	.143884	.0404740	32
MÁS DE 5 HASTA 10 S.M.	.079230	.0256455	32
MÁS DE 10 S.M.	.036884	.0162499	32
NO ESPECIFICADO	.050708	.0098539	32

Correlation Matrix^a

	NO RECIBE INGRESOS	HASTA EL 50% DE UN S.M.	MÁS DEL 50% HASTA MENOS DE UN S.M.	UN S.M.	MÁS DE 1 HASTA 2 S.M.	MÁS DE 2 HASTA MENOS DE 3 S.M.	DE 3 HASTA 5 S.M.	MÁS DE 5 HASTA 10 S.M.	MÁS DE 10 S.M.	NO ESPECIFICADO
Correlation	1.000	.433	.701	.017	-.149	-.830	-.768	-.701	-.797	-.374
	.433	1.000	.911	.197	-.081	-.804	-.688	-.720	-.758	-.812
	.701	.911	1.000	.213	-.047	-.880	-.748	-.713	-.698	-.474
	.017	.197	.213	1.000	.304	-.229	-.203	-.201	-.148	-.178
	-.149	-.081	-.047	.384	1.000	-.088	-.282	-.383	.327	-.113
	-.830	-.804	-.888	-.229	-.088	1.000	.848	.788	.774	.417
	-.768	-.688	-.848	-.203	-.282	.848	1.000	.987	.829	.403
	-.701	-.720	-.713	-.201	-.383	.788	.807	1.000	.936	.384
	-.797	-.758	-.698	-.148	-.327	.774	.829	.936	1.000	.486
	-.374	-.812	-.474	-.178	-.113	.417	.403	.384	.486	1.000

a. Determinant = .000

b. The results are not positive definite

Communalities

	Initial	Extraction
NO RECIBE INGRESOS	1.000	.878
HASTA EL 50% DE UN S. M.	1.000	.847
MÁS DEL 50% HASTA MENOS DE UN S.M.	1.000	.867
UN S.M.	1.000	.817
MÁS DE 1 HASTA 2 S.M.	1.000	.842
MÁS DE 2 HASTA MENOS DE 3 S.M.	1.000	.880
DE 3 HASTA 5 S.M.	1.000	.909
MÁS DE 5 HASTA 10 S.M.	1.000	.948
MÁS DE 10 S.M.	1.000	.904
NO ESPECIFICADO	1.000	.493

Extraction Method: Principal Component Analysis.

Component Matrix^a

	Component		
	1	2	3
NO RECIBE INGRESOS	-.870	-.328	-.129
HASTA EL 50% DE UN S. M.	-.938	-.209	.155
MÁS DEL 50% HASTA MENOS DE UN S.M.	-.894	-.164	.203
UN S.M.	-.268	.662	.560
MÁS DE 1 HASTA 2 S.M.	-.165	.899	-.328
MÁS DE 2 HASTA MENOS DE 3 S.M.	.933	.088	-.046
DE 3 HASTA 5 S.M.	.941	-.127	.088
MÁS DE 5 HASTA 10 S.M.	.923	-.138	.277
MÁS DE 10 S.M.	.903	-.118	.274
NO ESPECIFICADO	.542	-.072	-.440

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Rotated Component Matrix^a

	Component		
	1	2	3
NO RECIBE INGRESOS	-.925	-.148	-.029
HASTA EL 50% DE UN S. M.	-.814	-.178	.283
MÁS DEL 50% HASTA MENOS DE UN S.M.	-.864	-.165	.333
UN S.M.	-.012	.332	.841
MÁS DE 1 HASTA 2 S.M.	-.043	.956	.102
MÁS DE 2 HASTA MENOS DE 3 S.M.	.908	.019	-.242
DE 3 HASTA 5 S.M.	.898	-.233	-.225
MÁS DE 5 HASTA 10 S.M.	.013	-.328	.062
MÁS DE 10 S.M.	.898	-.309	-.051
NO ESPECIFICADO	.424	.102	-.550

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Component Transformation Matrix

Component	1	2	3
1	.962	-.085	-.260
2	.197	.874	.444
3	.189	-.478	.857

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.

Component Score Coefficient Matrix

	Component		
	1	2	3
NO RECIBE INGRESOS HASTA EL 50% DE UN S. M.	- .206	- .108	- .190
MÁS DEL 50% HASTA MENOS DE UN S.M.	- .138	- .197	.131
UN S.M.	- .115	- .198	.191
MÁS DE 1 HASTA 2 S.M.	.172	.081	.769
MÁS DE 2 HASTA MENOS DE 3 S.M.	.022	.714	- .051
DE 3 HASTA 5 S.M.	.145	.064	- .058
MÁS DE 5 HASTA 10 S.M.	.147	- .137	.011
MÁS DE 10 S.M.	.186	- .247	.199
NO ESPECIFICADO	.184	- .234	.202
	- .023	.196	- .485

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.
 Component Scores.

Component Score Covariance Matrix

Component	1	2	3
1	1.000	.000	.000
2	.000	1.000	.000
3	.000	.000	1.000

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.

Component Scores.

Al multiplicar la tabla 4.4 por la matriz de componentes se obtendrán los valores para reducir nuestra matriz original a una de dimensión 32x3.

Los datos así obtenidos, generan los siguientes componentes.

AGUASCALIENTES	0.78891	0.04011	0.21152
BAJA CALIFORNIA	1.82118	-2.37136	-0.59253
BAJA CALIFORNIA SUR	1.01658	-0.60482	-0.28655
CAMPECHE	-0.72239	-0.90689	1.43973
COAHUILA DE ZARAGOZA	1.31249	-0.38148	0.52094
COLIMA	0.33589	0.04784	0.31657
CHIAPAS	-2.3495	-2.08579	-0.23121
CHIHUAHUA	0.96578	0.01859	-0.77174
DISTRITO FEDERAL	1.32061	0.35201	1.47448
DURANGO	0.24675	0.71051	0.12575
GUANAJUATO	0.11806	0.79911	-1.30602
GUERRERO	-1.29081	0.34341	-0.63666
HIDALGO	-0.98318	0.32879	0.09786
JALISCO	0.48416	-0.16133	-0.71039
MÉXICO	0.20236	1.47686	-1.03619
MICHOACÁN DE OCAMPO	-0.5457	0.67481	-0.8202
MORELOS	-0.15114	1.078	-0.59963
NAYARIT	-0.34032	0.04253	0.52412
NUEVO LEÓN	1.4817	-1.31406	-0.0148
OAXACA	-1.98418	-1.04451	-0.80339
PUEBLA	-0.98981	0.1727	-0.37586
QUERÉTARO DE ARTEAGA	0.71005	-0.08792	-0.82342
QUINTANA ROO	0.79914	-1.55601	0.6808
SAN LUIS POTOSÍ	-0.58373	0.0595	-0.48023
SINALOA	0.35851	1.38274	-0.5324
SONORA	0.75155	0.55738	-0.42691
TABASCO	-0.84583	-1.55214	0.57756
TAMAULIPAS	0.47911	0.85757	-0.5121
TLAXCALA	-0.18242	1.71295	2.48586
VERACRUZ-LLAVE	-1.1426	0.16313	0.41838
YUCATÁN	-0.4443	0.50915	3.1242
ZACATECAS	-0.61692	0.7187	-1.01732

Tabla derivada de Componentes Principales

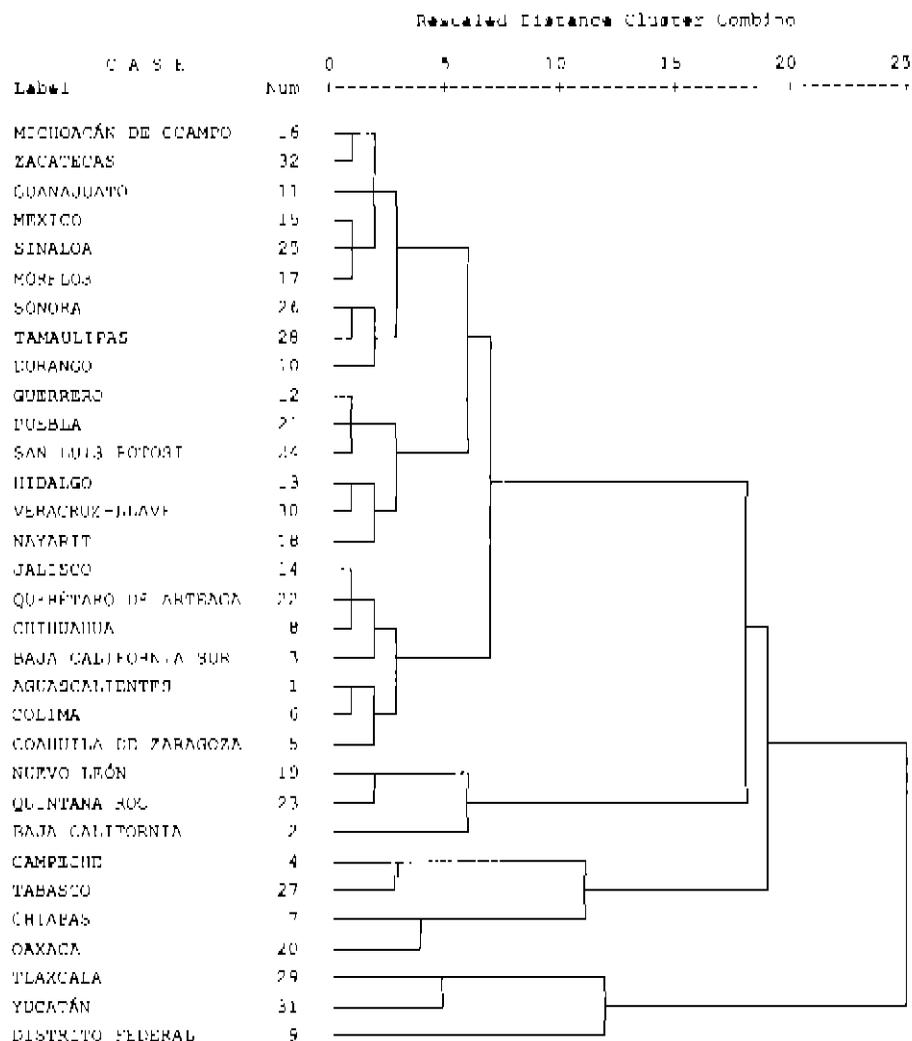
Nuevamente, y como en el caso anterior, para no quedarnos a este nivel con la aplicación de la técnica, utilizaremos esta tabla para realizar un agrupamiento.

Directamente, a continuación se muestra el dendograma que se obtiene al aplicar la técnica del capítulo 3, con lo cual continuaremos nuestro objetivo de conformar grupos de acuerdo a comportamientos comunes.

Dendógrama

***** HIERARCHICAL CLUSTERS ANALYSIS *****

Dendrogram using Average Linkage (Between Groups)



Aunque el agrupamiento debiera ser muy detallado, en términos generales podemos analizar en cada nivel los grupos que se forman, y aunque puede ser algo tedioso, conviene revisarlo como a continuación se hace.

En un primer nivel se forman grupos pequeños.

Grupo	Integrantes
1	Michoacán y Zacatecas
2	Guanajuato
3	Estado de México, Sinaloa y Morelos
4	Sonora y Tamaulipas
5	Durango
6	Guerrero, Puebla y San Luis Potosí
7	Hidalgo y Veracruz
8	Nayarit
9	Jalisco, Querétaro y Chihuahua
10	Baja California Sur
11	Aguascalientes y Colima
12	Coahuila
13	Nuevo León
14	Quintana Roo
15	Baja California
16	Campeche
17	Tabasco
18	Chiapas
19	Oaxaca
20	Tlaxcala
21	Yucatán
22	Distrito Federal

En un siguiente nivel algunos grupos se unen, formando nuevos conglomerados

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos
2	Sonora, Tamaulipas, Durango
3	Guerrero, Puebla, San Luis Potosí
4	Hidalgo, Veracruz, Nayarit
5	Jalisco, Querétaro, Chihuahua, Baja California Sur
6	Aguascalientes, Colima, Coahuila
7	Nuevo León, Quintana Roo
8	Baja California
9	Campeche
10	Tabasco
11	Chiapas
12	Oaxaca
13	Tlaxcala
14	Yucatán
15	Distrito Federal

Siguiente nivel.

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos, Sonora, Tamaulipas, Durango
2	Guerrero, Puebla, San Luis Potosí, Hidalgo, Veracruz, Nayarit
3	Jalisco, Querétaro, Chihuahua, Baja California Sur, Aguascalientes, Colima, Coahuila
4	Nuevo León, Quintana Roo
5	Baja California
6	Campeche, Tabasco
7	Chiapas
8	Oaxaca
9	Tlaxcala
10	Yucatán
11	Distrito Federal

Cuarto Nivel

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos, Sonora, Tamaulipas, Durango
2	Guerrero, Puebla, San Luis Potosí, Hidalgo, Veracruz, Nayarit
3	Jalisco, Querétaro, Chihuahua, Baja California Sur, Aguascalientes, Colima, Coahuila
4	Nuevo León, Quintana Roo
5	Baja California
6	Campeche, Tabasco
7	Chiapas, Oaxaca
8	Tlaxcala
9	Yucatán
10	Distrito Federal

Quinto Nivel

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos, Sonora, Tamaulipas, Durango
2	Guerrero, Puebla, San Luis Potosí, Hidalgo, Veracruz, Nayarit
3	Jalisco, Querétaro, Chihuahua, Baja California Sur, Aguascalientes, Colima, Coahuila
4	Nuevo León, Quintana Roo
5	Baja California
6	Campeche, Tabasco
7	Chiapas, Oaxaca
8	Tlaxcala, Yucatán
9	Distrito Federal

Sexto Nivel

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos, Sonora, Tamaulipas, Durango, Guerrero, Puebla, San Luis Potosí, Hidalgo, Veracruz, Nayarit
2	Jalisco, Querétaro, Chihuahua, Baja California Sur, Aguascalientes, Colima, Coahuila
3	Nuevo León, Quintana Roo, Baja California
4	Campeche, Tabasco
5	Chiapas, Oaxaca
6	Tlaxcala, Yucatán
7	Distrito Federal

Séptimo Nivel

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos, Sonora, Tamaulipas, Durango, Guerrero, Puebla, San Luis Potosí, Hidalgo, Veracruz, Nayarit, Jalisco, Querétaro, Chihuahua, Baja California Sur, Aguascalientes, Colima, Coahuila
2	Nuevo León, Quintana Roo, Baja California
3	Campeche, Tabasco
4	Chiapas, Oaxaca
5	Tlaxcala, Yucatán
6	Distrito Federal

Octavo Nivel

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos, Sonora, Tamaulipas, Durango, Guerrero, Puebla, San Luis Potosí, Hidalgo, Veracruz, Nayarit, Jalisco, Querétaro, Chihuahua, Baja California Sur, Aguascalientes, Colima, Coahuila
2	Nuevo León, Quintana Roo, Baja California
3	Campeche, Tabasco, Chiapas, Oaxaca
4	Tlaxcala, Yucatán
5	Distrito Federal

Noveno nivel

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos, Sonora, Tamaulipas, Durango, Guerrero, Puebla, San Luis Potosí, Hidalgo, Veracruz, Nayarit, Jalisco, Querétaro, Chihuahua, Baja California Sur, Aguascalientes, Colima, Coahuila
2	Nuevo León, Quintana Roo, Baja California
3	Campeche, Tabasco, Chiapas, Oaxaca
4	Tlaxcala, Yucatán, Distrito Federal

Décimo nivel

Grupo	Integrantes
1	Michoacán, Zacatecas, Guanajuato, Estado de México, Sinaloa y Morelos, Sonora, Tamaulipas, Durango, Guerrero, Puebla, San Luis Potosí, Hidalgo, Veracruz, Nayarit, Jalisco, Querétaro, Chihuahua, Baja California Sur, Aguascalientes, Colima, Coahuila, Nuevo León, Quintana Roo, Baja California
2	Campeche, Tabasco, Chiapas, Oaxaca
3	Tlaxcala, Yucatán, Distrito Federal

Conviene dejar la agrupación a este nivel y empezar a detectar puntos importantes.

En este caso, se observa cómo Tlaxcala, Yucatán y el Distrito Federal son los "más diferentes" de los estados de acuerdo a las variables consideradas. Lo cual al ubicarse el Distrito Federal en este grupo, permite aventurar la hipótesis que son las entidades federativas con mejores ingresos por trabajo. En este caso, Tlaxcala puede justificarse por la cercanía geográfica con el D.F., pero no es tan trivial la explicación para Yucatán.

En un segundo grupo, se ubican Campeche, Tabasco, Chiapas y Oaxaca, que desde luego, permite suponer que son los estados con menores ingresos. En este grupo extraña la presencia de Tabasco, pues al ser un estado con buena actividad petrolera no supondríamos que se encontrara en este conglomerado.

Finalmente, se encuentran el resto de los estados, los cuales en primera instancia mostrarían un comportamiento estándar; sin embargo, al observar como se fueron formando los grupos en los primeros niveles de conglomeración, es factible obtener mejores parámetros de comportamientos similares (o subgrupos) aún dentro del mismo grupo 1.

Por ejemplo, Michoacán y Zacatecas se unen desde el primer nivel, tal vez influya que ambos estados presentan un alto índice de migración de parte de sus habitantes hacia Estados Unidos.

Asimismo, llama la atención que Chiapas y Oaxaca se unan hasta el cuarto nivel de aglomeración, pues se esperaría que se unieran en el primer o segundo nivel, aunque esto no impide observar claramente que son muy similares y prácticamente pertenecen a un grupo perfectamente definido y muy diferente del resto.

Capítulo 5.

Escalamiento Multidimensional

Objetivo

Esta técnica multivariada pretende obtener una representación (básicamente en \mathbb{R}^2 , aunque también es posible en \mathbb{R}^3) de un conjunto de individuos, en base a un número n de variables asociadas a cada uno de ellos.

Para llegar a esta representación, se basa en una matriz de discrepancias o en una de similitudes.

En el caso que nos ocupa, que es el análisis de los datos censales, obtendremos, en primera instancia, una matriz de similitudes (o discrepancias) de 32×32 . De esta matriz, se partirá para obtener una representación gráfica que nos permita ubicar a cada uno de los estados en el plano cartesiano.

Desarrollo

Una de las medidas más comúnmente usadas para la construcción de la matriz de similitudes es la distancia euclidiana. Esta medida se obtiene para dos individuos (i, j) de la siguiente manera:

$$d_{(i,j)} = \left[\sum_{k=1}^d (X_{ik} - X_{jk})^2 \right]^{1/2}$$

Esta medida sirve para el caso de d dimensiones, donde X_{i1}, \dots, X_{id} y X_{j1}, \dots, X_{jd} son elementos de los vectores X_i y X_j respectivamente.

Escalamiento clásico

Dadas las coordenadas de n puntos en el espacio p -dimensional, se puede calcular la distancia euclídea para cada par de puntos a partir de la matriz X usando la expresión $B = XX^t$, de donde se obtiene

$$b_{ij} = \sum_{k=1}^p x_{ik} x_{jk}$$

por lo que

$$d^2_{ij} = b_{ii} + b_{jj} - 2b_{ij}$$

Si a partir de la matriz de disimilitudes se trata de encontrar las coordenadas, hay que tener en cuenta que no existe una representación gráfica única, ya que la ubicación u "ordenamiento" de los individuos puede variar, por lo que se obtendrán representaciones rotadas o reflejadas. Lo que significa que cualquier configuración que se obtenga será sujeta de una transformación ortogonal arbitraria, lo que facilitará la interpretación de la solución.

Para encontrar las coordenadas, se presentan dos pasos principales. El primero se relaciona en encontrar la matriz **B** referida anteriormente. De la expresión anterior, obtener las b_{ij} en términos de las d_{ij} involucra invertir la ecuación y se obtienen soluciones que no son únicas, por lo que se debe incluir la restricción para establecer un punto "fijo de rotación", es decir, $\bar{x} = 0$, por lo que se debe cumplir lo siguiente:

$$\sum_{j=1}^n x_{ij} = 0$$

Utilizando estas dos últimas expresiones se tiene que la suma de términos en cualquier renglón de **B** es cero. Por ello, si sumamos la anterior ecuación sobre i y sobre j y finalmente ambas sobre i y j se obtienen las siguientes tres ecuaciones:

$$\sum_{i=1}^n d^2_{ij} = D + nb_{jj}$$

$$\sum_{j=1}^n d^2_{ij} = D + nb_{ii}$$

$$\sum_{i=1}^n \sum_{j=1}^n d^2_{ij} = 2nD$$

donde $D = \sum_{i=1}^n b_{ii}$ es la traza de la matriz **B**.

Resolviendo las expresiones anteriores, se obtiene la siguiente ecuación:

$$b_{ij} = \frac{1}{2} [d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2]$$

donde

$$d_{i.}^2 = \frac{1}{n} \sum_{j=1}^n d_{ij}^2; \quad d_{.j}^2 = \frac{1}{n} \sum_{i=1}^n d_{ij}^2; \quad d_{..}^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2$$

La expresión de b_{ij} nos da los elementos de la matriz B en términos de las distancias euclídeas.

Para factorizar B en la forma XX' se obtienen los eigenvectores de B y se escalan de tal manera que la suma de sus cuadrados sea igual a los correspondientes eigenvalores. Entonces, la matriz X está dada por

$$X = [c_1, c_2, \dots, c_n]$$

Que son los eigenvectores adecuadamente escalados de B . En caso de buscar la representación en una dimensión k , $k \leq n$, simplemente se seleccionan los k primeros eigenvectores.

Escalamiento métrico y no métrico

Uno de los puntos centrales del escalamiento multidimensional es que las distancias entre los puntos representen las similitudes o proximidades entre los individuos observados. Con esto en mente, varios autores han enfocado el problema definiendo una función objetivo, la cual mida las discrepancias entre las proximidades observadas y las distancias ajustadas.

Si δ_{ij} es la discrepancia o disimilitud para cada par de individuos y d_{ij} es la distancia calculada, una función que podría servir al propósito es la siguiente:

$$SS = \sum_{i=1}^{n-1} \sum_{j=i+1}^n (\delta_{ij} - d_{ij})^2$$

Si las distancias calculadas fueran todas iguales a las correspondientes disimilitudes reales, la función valdría cero. Entre mayores sean esas discrepancias, el valor de SS se incrementa.

Considerando que d_{ij} puede ser visto como un valor funcional de δ_{ij} , mas un error, es decir:

$$d_{ij} = f(\delta_{ij}) + \varepsilon_{ij}$$

Esta situación nos lleva a lo que los expertos llaman una relación monótonica entre las distancias calculadas y las disimilitudes reales, lo que permite expresar la distancia como:

$$d_{ij} = \hat{d}_{ij} + \varepsilon_{ij}$$

donde los \hat{d}_{ij} son un conjunto de números que son monótonicos con los δ_{ij} , es decir, si ordenamos los δ_{ij} y tenemos:

$$\delta_{1|11} < \delta_{1|22} < \dots < \delta_{N|NN}$$

Entonces ocurrirá que

$$\hat{d}_{1|11} < \hat{d}_{1|22} < \dots < \hat{d}_{N|NN}$$

La función de estrés es definida, entonces por:

$$Stress = \sqrt{\frac{\left[\sum_{i=1}^{n-1} \sum_{j=i+1}^n (d_{ij} - \hat{d}_{ij})^2 \right]}{\left[\sum_{i=1}^{n-1} \sum_{j=i+1}^n d_{ij}^2 \right]}}$$

Se ha elaborado la siguiente tabla para calificar la calidad del ajuste de acuerdo a este valor:

Stress (%)	Calidad del ajuste
20	Pobre
10	Aceptable
5	Bueno
2.5	Excelente
0	Perfecto

Aplicación a los datos del censo

Para la aplicación de esta técnica se eligieron 13 variables relacionadas con la educación, las cuales son las siguientes:

- X1 = Población de cada estado entre el rango de edades de 6 a 14 años
- X2 = Población de 6 a 14 años que sabe leer y escribir
- X3 = Población de 6 a 14 años que no sabe leer y escribir
- X4 = Población de 6 a 14 años que no especificó si sabe leer y escribir
- X5 = Población mayor de 5 años en cada estado
- X6 = Población mayor de 5 años que asiste a la escuela}
- X7 = Población mayor de 5 años que no asiste a la escuela
- X8 = Población mayor de 5 años que no especificó si asiste a la escuela o no
- X9 = Población mayor de 15 años
- X10 = Población mayor de 15 años alfabeta
- X11 = Población mayor de 15 años analfabeta
- X12 = Población mayor de 15 años que no especificó su estado de alfabetismo
- X13 = Población mayor de 5 años sin ningún grado de Instrucción

En la tabla 5.1 se identifican las variables que servirán como entrada al modelo de escalamiento al utilizar el paquete estadístico SPSS.

Tabla 5.1. Variables utilizadas para el escalamiento multidimensional relacionadas con la educación (porcentajes)

ENTIDAD FEDERATIVA	(X1)POBLACIÓN DE 6 A 14 AÑOS	(X2)SABE LEER Y ESCRIBIR	(X3)SABE LEER Y ESCRIBIR	(X4)NO ESPECIFICADO SI SABE LEER Y ESCRIBIR	(X5)POBLADO N MAYOR DE 5 AÑOS	(X6)ASISTE A LA ESCUELA	(X7)NO ASISTE A LA ESCUELA	(X8)NO ASISTE A LA ESCUELA	(X9)POBLADO N MAYOR DE 15 AÑOS	(X10)POB. > 15 ALFABETA
02 BAJA CALIFORNIA	0.07943381	0.02225482	0.01993616	0.03928887	0.023774929	0.027742785	0.024985867	0.068441206	0.02494785	0.02881133
04 CAMPECHE	0.0074537	0.007185403	0.006154646	0.037462572	0.037154337	0.00754934	0.009383525	0.006543082	0.00705513	0.036874059
06 CHIHA	0.005202106	0.005033024	0.0046234703	0.009106613	0.0093966687	0.002478578	0.00509464519	0.004828151	0.005467701	0.009599653
08 CHIHUAHUA	0.029407647	0.030512638	0.02136635	0.003924059	0.003810772	0.029283622	0.031664732	0.32473384	0.001387241	0.030336511
10 DURANGO	0.019935847	0.018789339	0.014773566	0.07434073	0.014836765	0.01909483	0.014847785	0.013209159	0.014535569	0.01509193
12 GUERRERO	0.039710389	0.032676137	0.003796979	0.038076375	0.0312265428	0.033841704	0.0206969026	0.030813767	0.029056725	0.028365703
14 JALISCO	0.03520479	0.048991625	0.060224517	0.066983777	0.0653051916	0.065454043	0.065472863	0.062434556	0.054139599	0.067608486
16 H. CHILCAN DE OCAJUPA	0.040397532	0.044203751	0.050965873	0.045165795	0.041029837	0.043285037	0.041354191	0.044484865	0.038663911	0.037654938
18 NAYARIT	0.009912785	0.006365509	0.006959212	0.009496553	0.008614579	0.008883031	0.006515584	0.008130519	0.006540167	0.006594156
20 OAXACA	0.041254144	0.038429776	0.06124682	0.040511812	0.039904958	0.037110037	0.041877449	0.046191715	0.036889886	0.029204662
22 OJURENARO DE ARTEAGA	0.016309122	0.016340054	0.015527985	0.015994615	0.014435944	0.015027798	0.0141446749	0.015652108	0.014090183	0.014000966
24 SAN LUIS POTOSI	0.009093985	0.0029836	0.002685132	0.004018949	0.023710737	0.024784802	0.023215793	0.027663219	0.022956266	0.022490964
26 SONORA	0.02155482	0.021317089	0.021089732	0.024847829	0.023074823	0.022704697	0.023195395	0.021362576	0.023683797	0.024893337
28 TAMAULIPAS	0.025554479	0.026175025	0.021090074	0.027990076	0.028925789	0.028645586	0.029552863	0.001699676	0.029636693	0.031006062
30 VERACRUZ-Llave	0.073878045	0.068164472	0.105133605	0.064436217	0.072182218	0.070676802	0.072798073	0.067087071	0.071736422	0.087483612
32 ZACATECAS	0.016337731	0.015564532	0.014018436	0.013666848	0.014118889	0.013866541	0.014287498	0.013366115	0.013575433	0.013801881

Resultados Obtenidos al realizar el tratamiento

Las características del modelo utilizado las despliega el SPSS de la siguiente manera:

Miscel Procedure Options

Data Options-

Number of Rows (Observations/Matrix) . . . 32
Number of Columns (Variables) . . . 32
Number of Matrices 1
Measurement Level Ordinal
Data Matrix Shape Symmetric
Type Dissimilarity
Approach to Ties Untie
Conditionality Matrix
Data Cutoff at000000

Model Options-

Model Euclid
Maximum Dimensionality 2
Minimum Dimensionality 2
Negative Weights Not Permitted

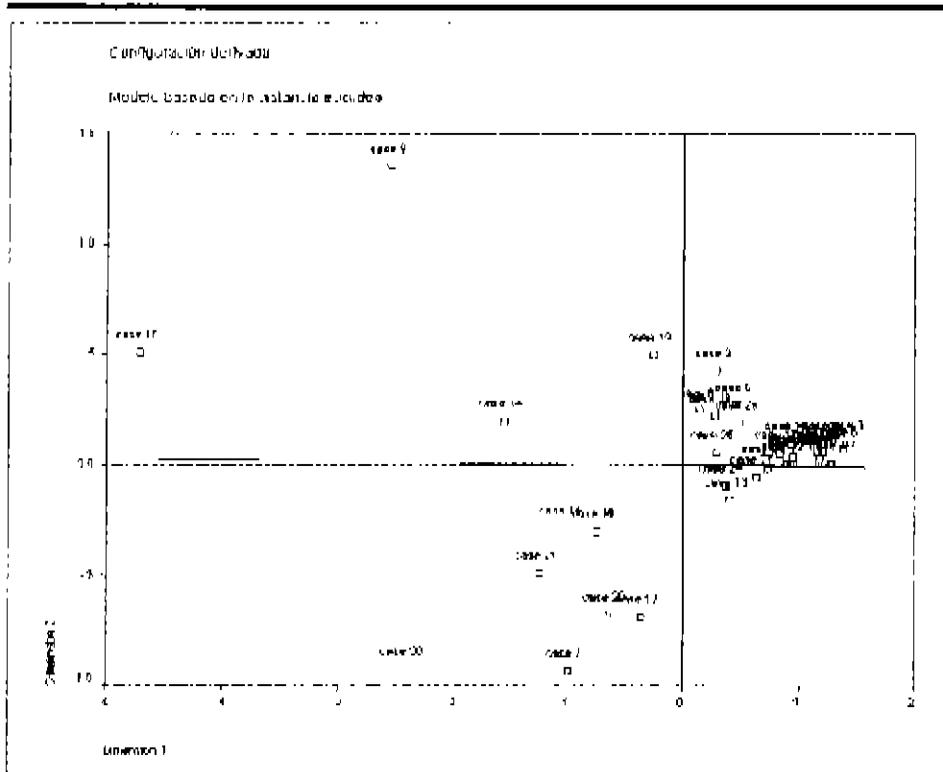
Output Options-

Job Option Header Printed
Data Matrices Printed
Configurations and Transformations . . . Plotted
Output Dataset Not Created
Initial Stimulus Coordinates Computed

Algorithmic Options-

Maximum Iterations 30
Convergence Criterion00100
Minimum S-stress00500
Missing Data Estimated by Unbounded
Tolerance 496

Aquí observamos las definiciones que le dimos al paquete para que le diera tratamiento a los datos originales. El paquete también muestra la matriz de disimilitudes, sin embargo, ya definimos como se obtiene y, si la desplegamos en este documento tal vez no tendría mucha significancia. Sin embargo, haremos énfasis en el aspecto gráfico del modelo, el cual se muestra en la siguiente gráfica que genera el modelo.



Distrito Federal (9)
 Estado de México (15)
 Nuevo León (19)
 Jalisco (14)

Grupo 1

Baja California (2)
 Chihuahua(8)
 Tamaulipas(28)

Coahuila(5)
 Sonora(26)
 Sinaloa(25)

Durango(10)
 Morelos(17)
 Querétaro(22)
 Zacatecas(32)

Aguascalientes(1)
 Campeche(4)
 Nayarit (18)
 Quintana Roo(23)
 Tlaxcala(29)

Baja California Sur(3)
 Colima(6)

Grupo 2

Grupo 3: Nivel Bajo

Grupo 4: Nivel medio-bajo.

Más del 50% de los elementos se encuentra en el grupo 2. Este punto es relevante, ya que nos indica que en la mayoría de los Estados el comportamiento de acuerdo a las variables elegidas, relacionadas con la educación, es hasta cierto punto homogéneo en esas 17 entidades federativas.

Sin embargo, es importante resaltar que debido a que es un grupo algo numeroso, se puede analizar un siguiente nivel de agrupamiento en ese mismo sector. Llama la atención el hecho que geográficamente, el único agrupamiento se da en aquel donde se ubican Coahuila, Sonora y Sinaloa.

Como era de esperarse, el D.F. se ubica en un sector muy diferente al de los demás. En este caso, y de acuerdo a los análisis realizados en capítulos anteriores, el hecho de ubicar al D.F. dentro de la gráfica, nos da un punto de apoyo si consideramos que es el Estado con mejores niveles educativos y establecer así una escala visualmente comparativa.

En otras palabras, aquellos elementos que se encuentren muy cercanos al D.F. o en su mismo cuadrante son los que estarán en "mejores" condiciones educacionales. Este punto de comparación no siempre es posible establecerlo, sin embargo, en este caso es posible de acuerdo a las características bien conocidas de la capital de nuestro país.

Nuevo León y Jalisco era de esperarse que estuvieran "cercaños" al D.F., aunque a ello hay que agregarle al Estado de México, quien por su situación geográfica que lo ubica como vecino inmediato del D.F. podríamos realizar un análisis más detallado para determinar si son los municipios del área metropolitana quienes "elevan" el nivel educativo de ese Estado.

Es importante observar que a pesar de estar en el mismo cuadrante, aún entre esos cuatro estados hay distancias significativas, lo que sin duda alguna, al compararlo con los demás grupos, demuestra que las diferencias educacionales entre los estados son bastante pronunciadas.

Esto nos lleva a considerar que existe una varianza muy alta en los datos presentados en la tabla analizada. Lo que traducido significa una desigualdad entre los niveles educacionales de cada uno de los Estados.

Del grupo 3 hay varios aspectos que son importantes de resaltar. El primero de ellos, es que como se esperaba, Guerrero, Oaxaca y Chiapas son de un comportamiento semejante.

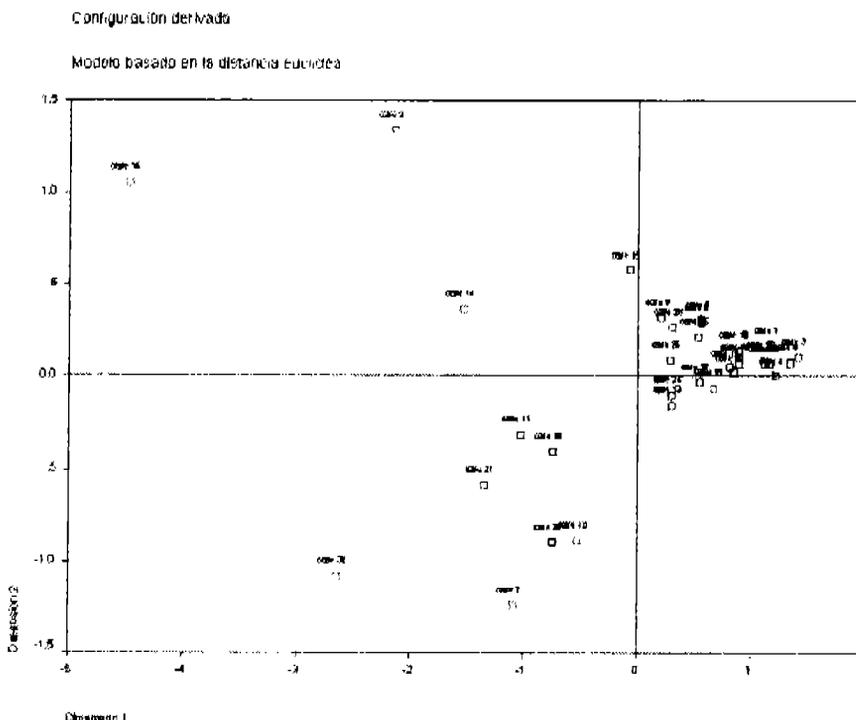
En segundo lugar, Es de extrañar que Guanajuato sea "vecino" de Oaxaca. En general, se esperaría que dada la condición de la vida académica que ha existido

en dicho estado, se ubicará en "mejor" posición de acuerdo a las variables seleccionadas. Una situación similar podría considerarse para Veracruz y Puebla.

En tercer lugar, Michoacán, geográficamente contiguo a Guanajuato, muestra un comportamiento muy semejante a éste, por lo que podría establecerse un segundo agrupamiento geográfico en la zona occidente del país, ya que Guerrero es vecino geográfico de Michoacán. Lo cual es pertinente de analizar, toda vez que, a pesar del hecho que Jalisco esté en el primer cuadrante, gráficamente no se encuentra "muy lejano" de esta región mencionada.

El último grupo muestra dos estados del sureste y dos estados del centro del país, que si bien se encuentran en un cuadrante diferente al grupo 2, es posible considerar un análisis más detallado para determinar si existen variables que al ser eliminadas o agregadas cambian la configuración de la gráfica, haciendo los grupos más definidos.

Si consideramos únicamente las variables X2, X3, X6, X7, X10 y X11, evltando variables que podrían ocasionar "ruido" se obtiene la siguiente gráfica, que en términos generales es muy semejante a la anterior y de la cual podemos obtener prácticamente las mismas conclusiones.



Capítulo 6

Resultados de los métodos analizados

Análisis de conglomerados

Grupo I	Grupo II
Michoacán (16)	Aguascalientes (1)
Nuevo León (19)	Tlaxcala (29)
Guanajuato (11)	Quintana Roo (23)
Puebla (21)	Campeche (4)
Hidalgo (13)	Nayarit (18)
San Luis Potosí (24)	Baja California Sur (3)
Coahuila (5)	Colima (6)
Sonora (28)	Tabasco (27)
Sinaloa (25)	Yucatán (31)
Baja California(2)	Querétaro (22)
Guerrero (12)	Zacatecas (32)
Tamaulipas (28)	Durango (10)
Chihuahua (8)	Morelos (17)
Oaxaca (20)	
Chiapas (7)	

Grupo 3: D.F. Estado de México, Jalisco y Veracruz.

Análisis factorial y conglomerados

Grupo A	Grupo B	Grupo C
D.F.	Hidalgo	Jalisco
Guerrero	Morelos	Querétaro
Chiapas	Durango	Aguascalientes
Veracruz	Tlaxcala	Yucatán
	Colima	Guanajuato
	Coahuila	Zacatecas
	Edo. México	San Luis Potosí
	Tamaulipas	Oaxaca
	Campeche	Puebla
	Chihuahua	Michoacán
	Nayarit	Sinaloa
	Tabasco	Sonora

	Nuevo León	Baja California Baja California Sur Quintana Roo
--	------------	--

Escalamiento multidimensional

Distrito Federal (9) Estado de México (15) Nuevo León (19) Jalisco (14)	Baja California (2) Chihuahua(8) Tamaulipas(28) Coahuila(5) Sonora(26) Sinaloa(25) Durango(10) Morelos(17) Querétaro(22) Zacatecas(32) Aguascalientes(1) Campeche(4) Nayarit (18) Quintana Roo(23) Tlaxcala(29) Baja California Sur(3) Colima(6)
Guanajuato (11) Michoacán(16) Puebla (21) Oaxaca(20) Guerrero(12) Veracruz (30) Chiapas (7)	Tabasco (27) Yucatán (31) Hidalgo (13) San Luis Potosí (24)

Dentro de las conclusiones que se pueden obtener a partir de los resultados arrojados por la aplicación de las diferentes técnicas multivariadas, nos podemos enfocar a los grupos que se conformaron en cada caso.

Caso Distrito Federal

En primer lugar, el D.F. se observa como un estado diferente al de los demás, por lo que la primera conclusión es que de acuerdo a las variables consideradas (vivienda, demografía y educación) aún existe un desarrollo diferente para la capital y cada uno de los estados.

Si bien, en un principio se pensaba que el estado que probablemente sería más parecido al D.F. era Nuevo León, los resultados nos mostraron que es el Estado de México quién mostró mayor parecido al D.F.

En este aspecto, lo más relevante es lo referido a vivienda y a educación, ya que tanto el Estado de México como el D.F. comparten una amplia zona metropolitana que incide notablemente en el desarrollo del primero. Asimismo, para asegurar categóricamente que el nivel educativo en el Estado de México es parecido al del D.F. tendrían que realizarse estudios particulares sobre aquél para identificar si el desarrollo educativo en ese estado es uniforme o existe una disparidad notoria en los municipios colindantes al D.F. (Ecatepec, Coacalco, Nezahualcóyotl y Naucalpan).

Llama la atención que estados como Tlaxcala, Puebla y Morelos, que también colindan con el D.F. no aparecen tan cercanos al D.F. como se esperaría. Aquí cabría determinar si la situación geográfica de dichos estados es propicia para el intercambio de elementos que permitan el desarrollo de tales estados. Principalmente para aprovechar la cercanía y la red de comunicaciones con que se cuenta, la cual, aparentemente no ha sufrido la suficiente explotación.

A este punto, es necesario recalcar que en lo relativo al análisis factorial, el D.F. se muestra en el mismo grupo que Chiapas, Guerrero y Veracruz, lo cual no quiere decir que es un estado que "se parezca" a alguno de ellos tres; más bien se integró a dicho grupo porque discrepa del común, pero "hacia arriba", es decir, hacia condiciones más favorables que los demás no tienen. Recordemos que lo analizado en esta técnica fue lo concerniente a variables asociadas con la demografía y, ciertamente, el D.F. presenta una composición demográfica muy especial, pues se encuentra habitado, en su mayoría, por individuos originarios de otros estados del país.

Grupos

En general, son los siguientes estados los que permanecen con mejores condiciones de acuerdo a las variables seleccionadas: D.F., Estado de México, Jalisco y Nuevo León.

En el caso opuesto, se encuentran Oaxaca, Guerrero y Chiapas. Sin embargo, con los análisis efectuados no es posible identificar estados "con bajos niveles de desarrollo" de acuerdo a las variables elegidas.

Estos tres últimos estados se encuentran muy cercanos en cada uno de los análisis que se efectuaron, denotando que, si bien geográficamente son "vecinos" o elementos de una misma región, de acuerdo a este análisis, comparten muchas cosas en común.

En caso de asignación de recursos, enfocados a programas de vivienda, educación y salud, es importante considerar a esta región como prioritaria, y puede considerarse como homogénea en los tres estados, ya que presentan condiciones muy semejantes.

Un punto que llama la atención, es que de acuerdo a las variables de educación y el análisis desarrollado para tal efecto, se encuentra a Chlapas como uno de los estados de menor nivel, no obstante que en 2001 se publicó un informe en el cual Chlapas presentaba el menor índice de reprobación en educación primaria.

Este punto puede llamara a la reflexión respecto a si se está alcanzando la calidad deseable en la educación que están recibiendo los niños de esta región, sobre todo en lo relativo a la educación primaria.

Esta región geográfica, que representa zonas con gran riqueza de recursos naturales, manifiesta también una desventaja frente a grupos como los representados por el D.F., ya que el desarrollo en ambos casos es muy desigual.

Mención especial merece el estado de Veracruz, quién comparte con estos tres últimos estados la característica de tener una gran riqueza de recursos naturales, sin embargo, se muestra con un comportamiento muy cercano a este grupo.

En un principio no se pensaría en Veracruz para integrar el grupo de los niveles inferiores. Antes bien, se suponía que es un estado con grandes posibilidades de ingresar a los grupos con niveles superiores.

Sin embargo, una realidad (paralela a este estudio) es el hecho que la migración ha sufrido un incremento considerable, y este puede ser considerado como un parámetro para mostrar que tal vez "las cosas no andan muy bien por Veracruz".

Es importante hacer mención que tratándose de asignaciones de recursos a nivel estatal es recomendable tomar en cuenta los resultados de este estudio, pero se puede tomar también como punto de partida para el análisis a nivel municipal, ya que se puede reproducir el mismo estudio hacia el interior de cada entidad federativa, arrojando resultados válidos para municipios de cada estado.

¿Cómo continuar?

Uno de los objetivos que se plantearon al incluir este trabajo fue constituir un material que sirviera como base para una posible aplicación inmediata, o bien para continuar con un análisis más profundo, dependiendo de los alcances que se pretendieran.

Este documento puede continuarse de dos maneras. La primera consiste en un enfoque general del país, el cual se puede conseguir incluyendo más grupos de variables y, tal vez, algunas técnicas de análisis multivariado más. Esto

complementarla el trabajo, y se podría canalizar hacia algún concepto en particular.

Tal vez se podría realizar el análisis de un segmento en particular, por ejemplo, las relativas a educación, y al mismo grupo de variables, aplicarles diferentes técnicas y valorar los resultados.

Otro posible enfoque es el que contempla el análisis más particular de ciertas regiones del país. Por ejemplo, ya hablamos considerado a Guerrero, Oaxaca y Chiapas como una región con un comportamiento similar, sin embargo, el análisis que se realizó con cada uno de los estados, es factible de llevarse a cabo con cada uno de los municipios de esta región.

Con ello, se obtendrán resultados mucho más específicos para esta región, y pueden utilizarse para llevar a cabo programas de manera más dirigida. Esto es considerable porque, si se diera el caso de asignar recursos a esta región, cabría la pregunta ¿a qué calidades de esta región debe darse prioridad?, ya que el hecho de que sean únicamente tres estados no facilita la distribución en la práctica.

Finalmente, se pretende que el enfoque que aquí se utilizó se aplique en otras cuestiones de índole similar. Se puede realizar un análisis parecido para estudiar el comportamiento de empresas, entidades gubernamentales, instituciones en general, que puedan considerar a cada uno de sus componentes como individuos a los cuales se les puedan medir ciertas variables comunes.

Los alcances de la aplicación de estas técnicas dependen de la posibilidad de adaptar un problema del mundo real a un modelo multivariado, ya que una vez superada esta etapa es relativamente fácil y rápido obtener resultados apoyados por el software computacional, siempre y cuando se ubique conceptualmente qué es lo que se está haciendo, ya que eso nos permitirá interpretar adecuadamente los resultados.

Capítulo 7

Comparativo con lo obtenido en 1990

En 1990 los resultados mostraban una configuración relativamente diferente a la que muestra actualmente. Llama la atención que en el 2000 el estado de Veracruz también se comporte como una observación que tiene cierta "discrepancia" y que en algunos casos se "asimila" al Distrito Federal.

Una de las posibles razones por las cuales se puede dar esta probable semejanza entre el D.F. y Veracruz en lo relativo a vivienda y aspectos demográficos es que ambos estados se encuentran dentro de los que tienen menores tasas de crecimiento poblacional, con 8.6 y 6.9 millones de habitantes respectivamente.

Esta conformación no se daba en 1990, en donde la "cercanía" de Chiapas y Oaxaca era muy evidente, en esta ocasión, si bien dichos estados se mantienen muy cerca uno de otro, se incorporan otros elementos al grupo o incluso se empiezan a identificar ciertas particularidades entre ellos.

Si bien es cierto que el desarrollo económico de estos dos estados, de alguna manera presenta circunstancias muy similares, básicamente por la situación geográfica, en este año a pesar de que se encuentran en los mismos grupos (en general) se observan con ciertas diferencias. En este caso, una de las discrepancias básicas como individuos se observó en el análisis relativo a las variables de fecundidad y mortalidad.

Es significativo el hecho que el comportamiento que se puede derivar al analizar las variables demográficas incide de manera directa en aspectos tales como Inmigración, educación y nivel de vida. A este respecto, podríamos proyectar una posible "separación" entre estos dos estados. Aunque, desde luego, se requiere un estudio mucho más profundo en esta región como para afirmar lo anterior.

Dentro de este mismo grupo, el estado de Guerrero se muestra también consistente con lo que presentaba en 1990. Tal vez es importante mencionar que hace 10 años Guerrero se "parecía" más a Oaxaca y ahora se "parece" más a Chiapas.

En lo relativo al D.F. es notorio que sigue comportándose como un estado "líder", que sigue representando los mejores niveles de vida, en promedio, con respecto a los demás estados del país.

Sin embargo, hace 10 años era muy claro que pocos estados tenían similitudes con el D.F. Ahora, hay estados que se "asemejan" a él en algunos aspectos. Sin duda alguna, Nuevo León presentó un comportamiento relativo al D.F. muy similar a lo ocurrido en el censo anterior.

Se observa en este año, que los grupos son más dispersos y que de acuerdo a las variables seleccionadas se dan "mezclas" entre estados que aparentemente no tienen muchos factores comunes entre sí.

El hecho que en algunos análisis se muestre a Veracruz muy "cercano" al D.F. no se presentaba en 1990. En ese año, se identificó a Quintana Roo como una observación discrepante, con elementos que en ocasiones se acercaban al D.F. y en ocasiones a Oaxaca.

Ahora, una de las observaciones discrepantes fue precisamente Veracruz. Básicamente en lo referente a variables demográficas, se muestra muy cercano al D.F.

Una posible explicación a lo anterior reside en el hecho que ambos estados presentan las menores tasas de crecimiento poblacional. Es decir, que la población crece menos en estos dos estados que en los demás.

La explicación para tal acontecimiento muy seguramente es diferente para ambos casos, ya que en 1990 se captó al D.F. como uno de los estados con mayor inmigración, y su población se integraba por un mosaico diverso proveniente de prácticamente todas las regiones del país. Ante el crecimiento poblacional en años anteriores, la política de población en la capital del país tuvo que enfocarse a la reducción del crecimiento poblacional, lo que se ve reflejado en la información censal.

Por otro lado, Veracruz muestra ahora un incremento en lo concerniente a emigración, por lo que este factor afecta directamente las variables demográficas para dicho estado.

Sin embargo, y a pesar de la concentración o dispersión de los grupos que se pretendieron formar en esta tesis, una conclusión es clara, aunque conocida desde hace tiempo: El desarrollo de los estados de nuestro país no es uniforme y existen profundas diferencias, ya que mientras haya algunos que "estén bien", otros "están mal" y algunos incluso "están muy mal".

Conclusiones y Recomendaciones

Actualmente no existen en nuestro país condiciones que permitan un desarrollo homogéneo entre las diversas regiones del país. Aún en regiones específicas existen diferencias notables entre estados vecinos. Ejemplo de ello son Tabasco y Veracruz.

Asimismo, es importante reiterar que la aplicación de este tipo de técnicas puede tener una amplia gama de aplicaciones y no limitarse a algo parecido a la simple estadística descriptiva, que aunque muy ilustrativa y útil para la toma de

Se observa en este año, que los grupos son más dispersos y que de acuerdo a las variables seleccionadas se dan "mezclas" entre estados que aparentemente no tienen muchos factores comunes entre sí.

El hecho que en algunos análisis se muestre a Veracruz muy "cercano" al D.F. no se presentaba en 1990. En ese año, se identificó a Quintana Roo como una observación discrepante, con elementos que en ocasiones se acercaban al D.F. y en ocasiones a Oaxaca.

Ahora, una de las observaciones discrepantes fue precisamente Veracruz. Básicamente en lo referente a variables demográficas, se muestra muy cercano al D.F.

Una posible explicación a lo anterior reside en el hecho que ambos estados presentan las menores tasas de crecimiento poblacional. Es decir, que la población crece menos en estos dos estados que en los demás.

La explicación para tal acontecimiento muy seguramente es diferente para ambos casos, ya que en 1990 se captó al D.F. como uno de los estados con mayor inmigración, y su población se integraba por un mosaico diverso proveniente de prácticamente todas las regiones del país. Ante el crecimiento poblacional en años anteriores, la política de población en la capital del país tuvo que enfocarse a la reducción del crecimiento poblacional, lo que se ve reflejado en la información censal.

Por otro lado, Veracruz muestra ahora un incremento en lo concerniente a emigración, por lo que este factor afecta directamente las variables demográficas para dicho estado.

Sin embargo, y a pesar de la concentración o dispersión de los grupos que se pretendieron formar en esta tesis, una conclusión es clara, aunque conocida desde hace tiempo: El desarrollo de los estados de nuestro país no es uniforme y existen profundas diferencias, ya que mientras haya algunos que "estén bien", otros "están mal" y algunos incluso "están muy mal".

Conclusiones y Recomendaciones

Actualmente no existen en nuestro país condiciones que permitan un desarrollo homogéneo entre las diversas regiones del país. Aún en regiones específicas existen diferencias notables entre estados vecinos. Ejemplo de ello son Tabasco y Veracruz.

Asimismo, es importante reiterar que la aplicación de este tipo de técnicas puede tener una amplia gama de aplicaciones y no limitarse a algo parecido a la simple estadística descriptiva, que aunque muy ilustrativa y útil para la toma de

decisiones, es necesario ahondar en la información y obtener resultados que faciliten y coadyuven mejor a tomar dichas decisiones.

El simple hecho de identificar que no existe un comportamiento poblacional, económico y educacional uniforme en nuestro país puede ser de ayuda para la distribución de recursos, como se mencionó anteriormente.

Dentro del enfoque que se le pretendió dar a este trabajo se pueden generar, desarrollando las corridas con las variables adecuadas, las siguientes aplicaciones a partir de este trabajo:

En el ámbito del Gobierno Federal:

- ✦ Asignación de recursos económicos, educacionales o de desarrollo social a determinadas regiones identificadas con las técnicas aquí presentadas.
- ✦ Identificación de regiones turísticas a impulsar o a ser rescatadas.
- ✦ Identificación de zonas proclives a la corrupción.
- ✦ Determinar relaciones entre estados que ayuden a cuestiones de seguridad pública.
- ✦ Distribución de servicios.

En la Iniciativa privada:

- ✦ Estrategias de mercadotecnia por sectores poblacionales o por regiones geográficas.
- ✦ En los grandes corporativos se pueden identificar áreas determinantes en la productividad empresarial.
- ✦ Identificar gamas de productos que sean factibles de producirse, comercializarse o publicitarse en conjunto.

Es importante mencionar que antes de proceder a aplicar alguna de las técnicas del análisis multivariado se debe tener identificado el objetivo que se persigue. Asimismo, se debe conocer para qué sirve y para qué no cada una de dichas técnicas.

Una vez que se conoce qué es lo que se desea y cuál es la técnica adecuada para conseguirlo, la siguiente parte es la obtención de los datos y su disposición en el formato adecuado para aplicar la técnica.

A continuación se debe determinar cuál es el paquete estadístico que se utilizará. A lo largo de este trabajo se hizo uso de SPSS versión 10, el cual es altamente recomendable si se reúnen las características antes mencionadas.

Posteriormente se deben analizar los primeros resultados y determinar si proceden ajustes a los datos o a su formato y volver a ejecutar las corridas necesarias e incluso es posible combinar varias de las técnicas con los mismos datos.

Finalmente, sigue el proceso analítico de interpretar los resultados. Ante esto se tienen dos perspectivas. La primera consiste en obtener conclusiones directamente de los resultados que los paquetes estadísticos arrojen, lo cual podría ser válido dependiendo del tipo de información que se esté manejando.

El segundo enfoque, que es el más común, es la obtención de conclusiones incorporando otro tipo de información exógena, que indudablemente deba ser tomada en cuenta para estar en posibilidades de implantar acciones factibles a partir de el análisis multivariado. Por ejemplo, si pretendemos identificar una región geográfica para la asignación de desayunos escolares y los resultados nos indican que dicha región está conformada por Baja California Norte y Yucatán, sin duda no podremos llevar a cabo ninguna acción directa, sino que se tendrán que incorporar otro tipo de elementos para llegar a decisiones que sean aplicables a la realidad.

Apéndice

Otras aplicaciones

La ejemplificación que se dio en este trabajo del análisis multivariado, permite tener una idea de la infinidad de aplicaciones que se pueden presentar. A pesar de que se le dio énfasis a la agrupación de individuos, el tratamiento puede ser en base a la agrupación de variables.

Además, las técnicas aquí presentadas son solo una parte de las disponibles en el análisis multivariado, pues existen otras que por sí solas requieren textos completos para su explicación teórica y práctica, y más que por su complejidad, es por su amplio campo de aplicación y de posibles interpretaciones que se requieren espacios mayores para detallar las demás técnicas de este tipo de análisis.

Sin embargo, continuando con el énfasis del presente documento, y con la intención de "sacudir" un poco la visión de nuestras entidades federativas como individuos del estudio, presentaré dos casos que se analizarán con las técnicas aquí descritas.

Tasa de piratería de software a nivel mundial

Dentro de la página del INEGI se puede encontrar información interesante y digna de análisis, más allá de las simples estadísticas. Ejemplo de ello es la información que presentan con el rubro "*Tasa de piratería de software por países seleccionados, 1998-2003*".

Como apreciamos actualmente, la piratería es un fenómeno mundial y tan ampliamente difundido que hasta se obtienen estadísticas de ello.

Para iniciar el análisis, lo fundamental es la obtención de los datos. Y después de conseguir los datos que se analizarán, debe construirse la matriz que alimentará al paquete estadístico en el cual se aplicarán las técnicas estadísticas.

Los datos se encuentran por países seleccionados, en este caso se incluye México, y la disposición de datos es por año, de 1998 a 2003. De tal manera que la matriz se compone en cada renglón por los países objeto del estudio y las columnas son los años.

En este caso en particular, nuestras variables serán las tasas de piratería en cada uno de los años en estudio. A diferencia de nuestros ejemplos anteriores, aquí cada variable se refiere al mismo tópico, solo que en diferentes años, lo cual las hace diferentes y permite considerarlas para el estudio. A continuación se muestra la tabla tal y como la presenta el INEGI en su página de Internet.

Tasa de piratería de software por países seleccionados, 1998-2003						
Países seleccionados	1998	1999	2000	2001	2002	2003
Total mundial	38	36	37	40	39	36
América						
Argentina	82	82	58	62	62	71
Brasil	61	58	58	58	55	61
Canadá	40	41	38	38	39	36
Chile	53	51	49	51	51	63
Colombia	60	58	53	62	51	53
Estados Unidos de América	25	25	24	26	23	22
México	59	58	56	55	55	63
Panamá	70	66	64	61	65	69
Perú	64	63	61	60	60	68
Puerto Rico	49	48	46	47	43	46
Uruguay	72	70	68	63	60	67
Venezuela	62	60	58	56	64	72
Asia						
China	95	91	94	92	92	92
Corea	64	50	56	48	50	48
Japón	31	31	37	37	35	29
Malasia	73	71	68	70	68	63
Singapur	52	51	50	51	48	43
Europa						
Alemania	28	27	28	34	32	30
España	57	53	51	49	47	44
Francia	43	39	40	48	43	45
Italia	45	44	48	45	47	49
Países Bajos	45	44	40	39	36	
	29	26	28	25		29
Rusia	92	89	88	87	89	87
Suecia	38	35	35	31	29	27
Oceania						
Australia	33	32	33	27	32	31
Nueva Zelanda	32	31	28	28	24	23
NOTA:						
En años anteriores al 2003, el estudio define a la piratería de software como la cantidad de programas informáticos de uso empresarial instalados o reproducidos sin su correspondiente licencia de uso; sin embargo, el análisis del 2003 incluyó categorías no cubiertas en los estudios anteriores, como los sistemas operativos, software de consumo hogareño y software desarrollado localmente. Estas categorías adicionales duplicaron el universo de software cubierto en el estudio.						
* Para 2003, la cifra se refiere a Gran Bretaña.						
ND No disponible.						
FUENTE: BSA. <i>Eight Annual BSA Global Software Piracy Study</i> , June 2003. Para 2003, BSA-IDC. <i>Primer Estudio Anual Mundial de Piratería de Software</i> , Julio de 2004.						

Para la aplicación del análisis se omitirá el dato referente al total mundial. Primero aplicaremos un análisis de conglomerados, para ver el comportamiento general de los datos. Los resultados al aplicar la técnica son los siguientes.

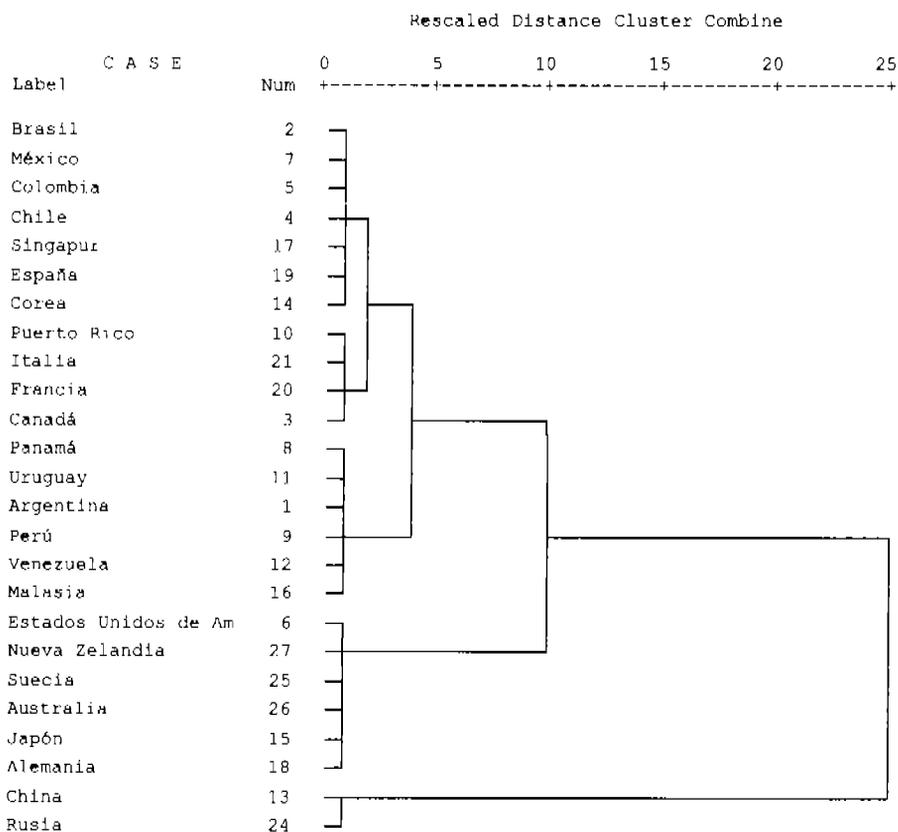
Case Processing Summary(a)

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
25	92.6	2	7.4	27	100.0

a Average Linkage (Between Groups)

***** H I E R A R C H I C A L C L U S T E R A N A L Y S I S *****

Dendrogram using Average Linkage (Between Groups)



Los resultados que se pueden obtener del dendograma si observamos que los datos proporcionados son de 1998 a 2003, es decir, que las conclusiones nos mostrarán el comportamiento de esos países a través de esos años y no únicamente en un momento en particular.

Como ya vimos, la intención es obtener grupos, y los primeros que se forman y muy claramente son:

Grupo 1: Brasil, México, Colombia, Chile, Singapur, España, Corea.

Grupo 2: Puerto Rico, Italia, Francia, Canadá.

Grupo 3: Panamá, Uruguay, Argentina, Perú, Venezuela, Malasia.

Grupo 4: Estados Unidos, Nueva Zelanda, Suecia, Australia, Japón, Alemania.

Grupo 5: China y Rusia.

En este caso, no nos iremos a continuar con el agrupamiento para formar conglomerados más sólidos, sino que lo dejaremos a este nivel de agregación.

Recordemos que los datos son acerca de la piratería de software y muy particularmente, al software empresarial, por lo que nos deja ver la nota del cuadro de datos.

El grupo 1 tal vez no tenga puntos a resaltar, pues la mayoría son elementos latinoamericanos y con economías muy similares. Sin embargo, resalta el hecho que se incluya a España en este grupo, pues se pensaría que un país con la economía más desarrollada no presentase este tipo de problemáticas dentro de su entorno empresarial.

El grupo 2, tal vez igual de homogéneo que el anterior, a excepción de Puerto Rico, lo cual solo podríamos explicar en esta primera aproximación por el hecho de ser una economía altamente vinculada a la estadounidense.

En el grupo 3 se observa nuevamente una homogeneidad, al igual que en el 4.

El grupo 5 es de los de comportamiento más alejado de los demás y presenta la particularidad de la coincidencia geográfica. Aunque ambas economías se encuentran actualmente con amplias posibilidades de crecimiento.

En este caso, a excepción del grupo 5, vemos que la situación geográfica no es tan influyente en la formación de grupos como lo esperamos que fuera para nuestras entidades federativas. Sin embargo, es interesante descubrir por estos métodos que las simples estadísticas como la tasa de piratería de software nos pueden mostrar similitudes en economías aparentemente distintas en su totalidad, tal y como lo pueden ser las de Europa con las de Asia y América.

A pesar que el análisis podría extenderse tanto como nos interesara, dejaremos el análisis de este ejemplo a este nivel, pues de las conclusiones anteriores se pueden extraer muchas hipótesis y generar un terreno fértil para Investigaciones de otro tipo.

Para concluir, presentaré un ejemplo adicional, que trata sobre el Gasto en tecnologías de información y comunicaciones para algunos países seleccionados, en 1995 y 2001.

Gasto en Tecnologías de Información

Antes de continuar con la presentación de los datos, y sin conocer los resultados del análisis, cabría esperar que los países con economías poderosas presentaran un comportamiento similar y que lo mismo ocurriera con las economías subdesarrolladas.

Las variables de este ejemplo son:

Gasto total en tecnologías de información y comunicaciones (Millones de dólares): Tanto para 1995 como para 2001, se presenta la información en millones de dólares invertidos en el rubro.

Gasto en tecnologías de información y comunicaciones como porcentaje del PIB: Es el monto que se invierte en este rubro, pero representado como porcentaje del PIB. Aquí cabría ver la significancia que presenta la conversión de valores netos a porcentajes, tal y como lo planteamos en algunos de los ejemplos precedentes.

Gasto en tecnologías de información y comunicaciones per cápita (dólares): Esta variable corresponde a la cantidad de dólares per cápita (por cabeza) que se invierten en los países seleccionados.

Si bien, para fines de agrupamiento, las variables anteriores podrían parecer a simple vista que representan lo mismo en cuanto a números, se aprecia que conceptualmente presentan grandes diferencias y con ello, podemos suponer que el problema de la multicolinealidad no debiera presentarse, aunque esto al igual que cualquier hipótesis que se plantee debe comprobarse.

Las variables que aquí se presentan pueden cuestionarse de muchas maneras, pues el hecho de que vengan presentadas por "pares" puede suponer algunos de los problemas que impiden la aplicación de técnicas estadísticas, en particular la violación de supuestos de homocedasticidad, linealidad y normalidad. Sin embargo, siempre depende del objetivo del estudio.

A continuación presentamos los datos tal y como se muestran (una vez más) en la página del INEGI.

Gasto en tecnologías de información y comunicaciones por países seleccionados, 1995 y 2001

Países seleccionados	Gasto total en tecnologías de información y comunicaciones (Millones de dólares)		Gasto en tecnologías de información y comunicaciones como porcentaje del PIB		Gasto en tecnologías de información y comunicaciones per cápita (dólares)	
	1995	2001	1995	2001	1995	2001
América						
Argentina	9 414	11 842	3.6	4.0	271	310
Brasil	18 882	50 031	2.7	6.3	121	287
Canadá	41 166	60 896	7.0	8.7	1 402	1 980
Chile	2 719	5 715	4.2	8.1	191	371
Colombia	2 983	10 434	3.2	12.0	85	231
Estados Unidos de América	557 252	812 835	7.5	5.6	2 119	2 924
México	10 619	19 211	3.7	3.2	113	196
Venezuela	2 724	4 943	3.5	4.0	128	199
Asia						
China	20 401	86 612	2.9	5.7	17	53
Corea	23 168	32 301	4.7	7.4	514	676
Japón	279 798	413 772	5.3	9.6	2 228	3 266
Malasia	4 438	6 325	5.0	6.6	221	282
Singapur	5 735	9 592	6.9	9.9	1 920	2 110
Europa						
Alemania	125 825	154 845	5.1	7.8	1 538	1 880
España	21 673	30 366	3.7	5.1	663	789
Francia	90 660	120 569	5.8	9.1	1 559	2 048
Italia	45 353	64 555	4.1	5.7	791	1 117
Países Bajos	28 207	37 145	6.3	9.3	1 091	2 327
Reino Unido	85 487	137 728	7.6	9.7	1 460	2 319
Rusia	6 188	9 908	1.8	3.3	42	88
Suecia	17 915	24 874	7.5	11.3	2 029	2 804
Oceania						
Australia	27 545	37 673	7.6	10.7	1 524	1 939
Nueva Zelanda	5 020	7 164	8.4	14.4	1 383	1 835

FUENTE: WB.

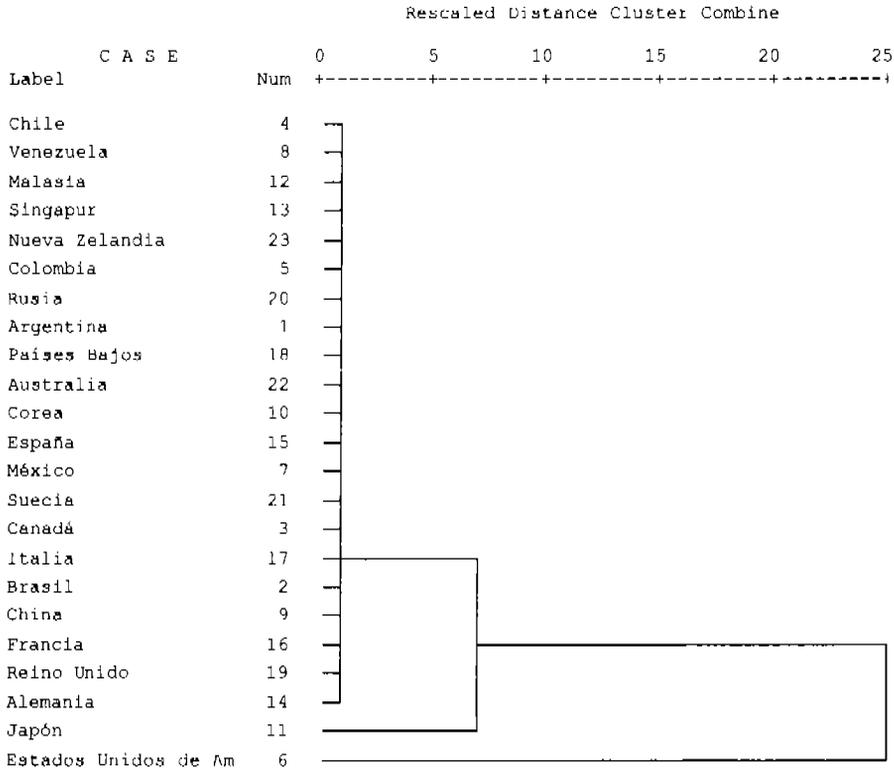
A simple vista, el país que podría "liderar" esta tabla debería ser Estados Unidos, y para ello, aplicaremos análisis de cluster a los datos para obtener el dendograma que ya hemos revisado en los ejemplos anteriores.

De dicho dendograma se observa claramente la disparidad que se tiene al comparar a los Estados Unidos con los demás países. Sin duda alguna es totalmente diferente en su comportamiento a los otros individuos objetos del análisis.

Caso también digno de mencionar es Japón, otro de los "diferentes" al resto de los países.

***** H I E R A R C H I C A L C L U S T E R A N A L Y S I S *****

Dendrogram using Average Linkage (Between Groups)



Si se tiene interés particular en alguno de los 23 países –en particular a nosotros México-, el dendrograma nos puede dar mucha idea de la situación en que se encuentra con respecto a los demás.

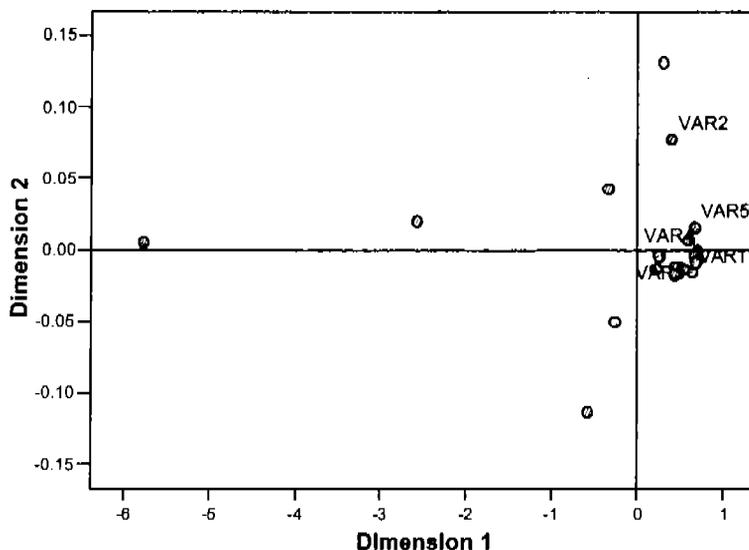
Sin embargo, este dendrograma nos muestra un comportamiento extremo, pues muestra tres grupos: Estados Unidos, Japón y el tercero conformado por 21 países.

En la primer gráfica el paquete no nos muestra completamente las etiquetas para identificar cada uno de los países en el espacio R2, sin embargo, con los datos que el mismo paquete genera se obtiene la segunda gráfica, que tal vez por su tamaño no permita apreciar adecuadamente la formación de grupos que se tiene, pero aún así se alcanzan a apreciar algunos detalles importantes.

Gráfica Original

Derived Stimulus Configuration

Euclidean distance model

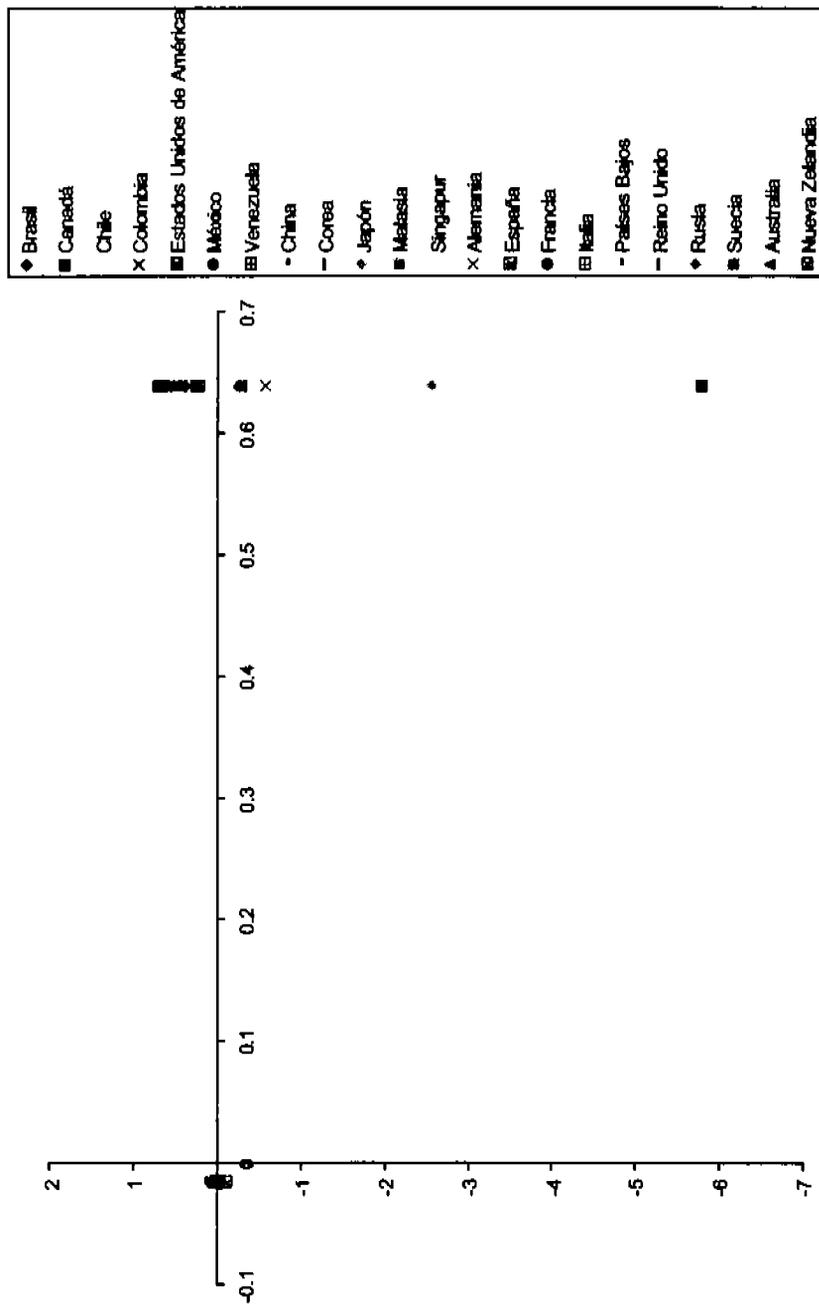


Como la gráfica anterior no nos permite mucho campo de acción, recurrimos a obtener la misma gráfica pero a través de Excel y se muestra más nítidamente cómo EUA se aleja del comportamiento estándar, lo cual indica que este país está muy por arriba que los demás en cuanto a inversión en Tecnologías de la Información (TI).

Japón, sin llegar a EUA, está muy cerca de su comportamiento y un poco más alejados Alemania, Reino Unido y Francia. El resto de los países realmente muestra un comportamiento muy similar, aunque dentro de estos se diferencian dos grandes grupos. Es fácil concluir que en uno se encuentran las economías desarrolladas y en otro las llamadas emergentes.

A partir de aquí se puede continuar con los análisis en los que se esté interesado y, reitero, dependerá del grado de interés que se tenga sobre un conjunto particular de datos, lo que permitirá aplicar técnicas más refinadas para el análisis. Desde luego, este documento como se mencionó al principio, pretende mostrar las bondades del análisis multivariado y sus aplicaciones prácticas con datos de nuestra realidad cotidiana.

Gráfica de Escalamiento en R2



Bibliografía

- 1)** Página de internet del INEGI : <http://www.inegi.gob.mx>
- 2)** *SPSS para Windows, Análisis Estadístico*. Ferrán Aranz Magdalena. Ed. Mc Graw Hill. 2001, España.
- 3)** *Análisis Estadístico Multivariante, Tesis de Maestría en Investigación de Operaciones, Carlos Barcia Portillo, DEPEI-UNAM, Diciembre del 2000, México D.F.*
- 4)** *Demos 2000, Carta demográfica sobre México, Número 13, ISSN 0187-7550. Revista anual, 2000, México*
- 5)** *Cluster Análisis for Applications, Michael R. Anderberg, Academic Press Inc, 1973, Londres.*
- 6)** *Applied Multivariate Analysis, Bernstein Ira, Springer-Verlag, 1988, Virginia, USA.*
- 7)** *Applied Multivariate Data Analysis, Everitt Brian & Dunn Graham, Ed. Edward Arnold, 1991, Londres.*