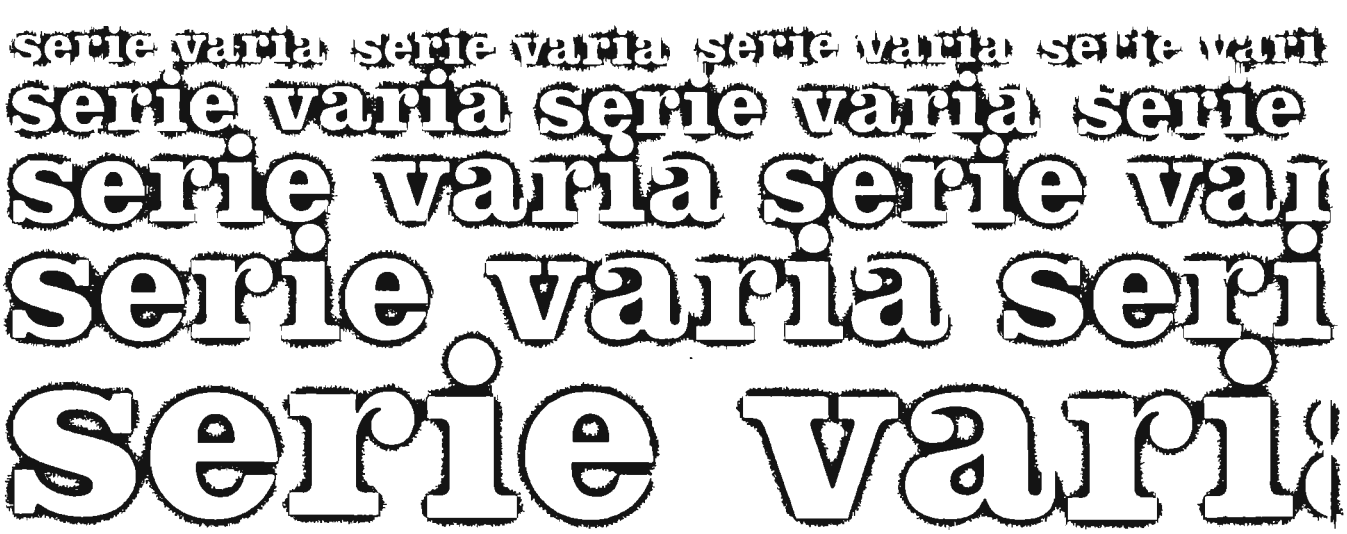
instituto de geografía

EL USO DE LA ESTADÍSTICA PARA LA CONSTRUCCIÓN DE CLASIFICACIONES Y REGIONALIZACIONES

IGNACIO KUNZ B.

Serie Varia T, 1, Núm 11
México
1988



EL USO DE LA ESTADÍSTICA PARA LA CONSTRUCCIÓN DE CLASIFICACIONES Y REGIONALIZACIONES

INSTITUTO DE GEOGRAFÍA

Serie Varia T. 1, Núm. 11 México 1988

EL USO DE LA ESTADÍSTICA PARA LA CONSTRUCCIÓN DE CLASIFICACIONES Y REGIONALIZACIONES

IGNACIO KUNZ B.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
MÉXICO 1988

Primera edición: 1988

INSTITUTO DE GEOGRAFÍA

Impreso y hecho en México

EL USO DE LA ESTADISTICA PARA LA CONSTRUCCION DE CLASIFICACIONES Y REGIONALIZACIONES

Ignacio Kunz B. *

RESUMEN

El presente trabajo está dirigido a investigadores sociales que se enfrentan a problemas de clasificación de objetos multivariados (tipologías, clasificaciones y regionalizaciones). Se estudia la aplicación de dos técnicas: el análisis de componentes principales y el análisis de cúmulos. No se tratan desde el punto de vista de las fórmulas y cálculos matemáticos, para lo cual existe abundante bibliografía y numersoso programas de cómputo. Por el contrario, se analizan las condiciones y características bajo las cuales se debe dar su aplicación, los problemas prácticos que aparecen comúnmente, la forma de evaluar e interpretar los resultados, sus limitaciones y el uso de algunas otras técnicas estadísticas que pueden enriquecer el análisis.

SUMMARY

The present work is addresed to social science reserchers who face the problem at the classification of multivariate objetcts (typology, clasification and regionalization). In this paper it is studied the appliance of two techniques: the principal component analysis and the cumulus analysis. This point of view does not deal with the formulae and mathematical calculus, for which there is a plantiful bibliography available and many computing programs. Quite the oppositive, in this work it is analyzed the conditions and characteristics in which these technics should have their appliance on, the practical problems that normally arise, the way to evaluate and interpreat the results, their restrains, and the use of some statistics elements that might be useful to enhance the appliance of these techniques.

INTRODUCCION

Este trabajo tiene su origen en tres aspectos principales. En primer lugar está la no utilización de la estadística en la investigación geográfica, ya sea por prejuicios, al identificar erróneamente el uso de herramientas formales con "la ciencia al servicio de las minorías" o, simplemente, por ignorancia no sólo del manejo de la técnica sino de su existencia.

En segundo lugar se encuentra el mal uso de la técnica estadística en la mayoría de los pocos casos en que se aplican dichas técnicas. Los errores no se dan tanto en los cálculos de que se requiere, sino en la incapacidad de concebir la solución estadística más adecuada al problema de investigación, en cuanto a las bases y los supuestos sobre los que se debe dar la aplicación, así como en la interpretación de los resultados y sus limitaciones. Muchos de los trabajos que "utilizan" la estadística en forma laxa pretenden, paradójicamente, basar su validez sólo en el uso de las técnicas cuantitativas utilizadas.

* Investigador del Instituto de Geografía de la UNAM.

En tercer lugar está el "Atlas Nacional de México", proyecto del Instituto de Geografía que requerirá el uso de técnicas estadísticas, sobre todo multivariadas, para clasificar objetos, realizar tipologías y construir regionalizaciones.

De esta manera, el artículo tiene como objetivo mostrar el uso de algunas técnicas estadísticas descriptivas para clasificación de objetos multivariados. Se tratarán: i) el tipo de problemas de investigación para los cuales resultan aplicables esas técnicas estadísticas; ii) dificultades comunes que se presentan en el proceso de aplicación, ya sean de índole teórico técnico o, bien, práctico. Estas últimas son de gran importancia ya que pueden limitar, sesgar y hasta detener la aplicación, y van a requerir, en la mayoría de los casos, de soluciones operativas sugeridas más por la experiencia que por el estudio de los aspectos teóricos de las herramientas estadísticas, pues tanto problemas como soluciones son independientes de ellas; iii) la forma en que se deben interpretar los resultados y las limitaciones de éstos; iv) las posibilidades existentes para la validación de los resultados.

A partir de lo anterior, es claro que el énfasis no se hace en cuanto a las características internas de las técnicas, ni se presentan éstas como una sucesión de fórmulas que se deben aplicar, lo cual se puede encontrar en numerosas fuentes bibliográficas y se ejecuta eficientemente por computadoras. Lo que se busca es ilustrar el adecuado manejo externo de las técnicas. El escrito está dirigido a científicos sociales que conocen de manera general las técnicas que se revisan, pero que no poseen suficiente experiencia en su aplicación, se busca que puedan saber cómo, cuándo y de qué manera usar técnicas estadísticas para clasificar, así como la manera de interpretarlas.

EL PROBLEMA A RESOLVER

El primer paso en cualquier investigación es la definición del problema. El desarrollo y los resultados dependen en gran medida de la claridad y precisión con que se haya definido el problema. También para el uso de la estadística es importante esta etapa, pues las técnicas estadísticas no son una solución mágica a todos los problemas, ni siquiera a todos los que suponen manejo de datos. Ciertos problemas de investigación podrán solucionarse mediante alguna técnica estadística: es fundamental la correspondencia exacta entre problema y solución, para lo que es necesario el planteamiento concreto y bien definido de aquél y el conocimiento y comprensión de la naturaleza de ésta.

En la introducción se manejaron dos tipos de objetivos: las regionalizaciones y las tipologías de las que, a pesar de compartir como problema de investigación algunos elementos, convendrá establecer ciertas diferencias. La regionalización implica un problema de clasificación particular de determinados objetos (unidades espaciales), mientras que la tipología supone una clasificación general a la cual podrán asignarse nuevos objetos. Así, la regionalización se da, necesariamente, por los objetos con que se trabaja; en cambio, la tipología puede ser independiente de los objetos. Hay independencia total cuando se construye exclusivamente a partir de un esquema teórico, y podrá existir cierta dependencia cuando los tipos resultantes han sido determinados no sólo teóricamente sino con un conjunto de objetos, pero se podrán asignar nuevos objetos no utilizados en la construcción. Además, en términos espaciales cada región tiene una definición específica y para muchos autores continua, mientras que los tipos podrán presentarse o reproducirse en espacios distintos; por tanto, los tipos no muestran continuidad espacial.

De esta manera, tanto la regionalización como la tipología suponen un problema de clasificación o formación de grupos homogéneos de objetos multivariados. Si se trabaja uno o dos atributos o variables no habrá necesidad de utilizar alguna técnica estadística para la agrupación, pues se está trabajando en una o dos dimensiones, según sea el caso; pero si se tiene un amplio conjunto de objetos caracterizados cada uno por numerosas variables (más de 10, por ejemplo) es imposible, sin ayuda de instrumentos estadísticos, identificar con precisión grupos homogéneos según distintos niveles de semejanza. Este es el problema de que se tratará: la identificación de niveles de semejanza de objetos multivariados y su clasificación en grupos que sean lo más homogéneos posible, lo que quiere decir que reúnan los objetos más parecidos en el conjunto total de variables y, así, construir regionalizaciones y tipologías.

EL TIPO DE INFORMACION

Cuando se habla de estadística multivariada se hace referencia a una multiplicidad de atributos que caracterizan a cierto número de objetos que, en este caso, serán unidades espaciales (entidades federativas). Dichos atributos o variables serán medidos por una serie de indicadores.

Las variables deben corresponder al problema de investigación planteado según ciertas categorías de análisis (González Casanova, p. 13 y ss.). Las variables no se obtienen por ocurrencia o imaginación, sino se desprenden de un marco teórico en el que se insertan las categorías de análisis que nos interesa evaluar a través de esas variables que, a su vez, serán concretadas o medidas por indicadores que deberán corresponder plenamente al sistema categoría-variable-indicador.

Los indicadores continuos y relativos (tasas, porcentajes, coeficientes) permiten mayor flexibilidad y precisión en cuanto a los cálculos, con ventaja sobre las variables no continuas, y ofrecen, al ser relativos, un elemento de referencia que facilita ciertas comparaciones. Hay variables que, siendo originalmente cualitativas u ordinales pueden traducirse en indicadores continuos. Por ejemplo, el sexo, que en sí es una variable dicotómica medida a veces como número total de hombres y mujeres, se puede ponderar mejor con el coeficiente de masculinidad que es un indicador continuo y relativo. Para ilustrar lo anterior mencionemos que, el estado de Quintana Roo, en 1980, era una entidad con pocos hombres en comparación con otros estados, no obstante mostraba el coeficiente de masculinidad más alto, lo que significa la entidad con mayor proporción de hombres respecto a mujeres (106 hombres por cada 100 mujeres, aproximadamente).

Si bien las variables continuas y relativas son atractivas en la aplicación de las técnicas aquí planteadas, el problema de investigación y su articulación con la posible solución determinará, a fin de cuentas, el tipo de variables que se usará.

En esta exposición se usará un ejemplo con 18 variables sobre aspectos sociodemográficos de los estados de la República (objetos). Se calcularon a partir del X Censo General de Población y Vivienda 1980, excepto las tasas de natalidad y mortalidad que se procesaron del Anuario Estadístico de los Estados Unidos Mexicanos 1983. Son variables continuas y relativas que se pueden agrupar en 7 dimensiones o aspectos que interesa analizar, pues esto significa parte del mínimo de conocimiento sobre aspectos sociodemográficos que se deben tener en cuenta para la planificación de la alimentación, educación, salud y seguridad social, empleo, vivienda y

desarrollo urbano, y desarrollo regional (Kunz, 1985).[1] Es evidente que para una planificación cuidadosa de estos aspectos se requiere de mucho más que el mínimo planteado y en una mayor escala que supondría usar como objetos de análisis los municipios. De cualquier manera, el ejemplo se utiliza únicamente para ilustrar la aplicación de las técnicas y en ningún momento se pretende hacer un análisis sociodemográfico.

Las dimensiones y sus variables son:

1. El volumen de la población medida como porcentaje de la población nacional.

2. La situación espacial en cuanto a concentración y dispersión de la población, medida mediante los porcentajes de población urbana (considerando así a la que viva en localidades mayores de 10 000 habitantes), y de población en localidades menores a 500 habitantes.

3. La estructura por edad y sexo, evaluada por 4 variables: el porcentaje de niños y jóvenes (entre 0-14 años); el porcentaje de población en edad de trabajar (entre 15 y 64 años); el porcentaje de ancianos o población postproductiva (mayores o iguales a 65 años) y, para el sexo, el coeficiente de masculinidad de la población.

4. La composición económica también analizada por 4 variables: el porcentaje de población activa, los porcentajes de población en los sectores primario y secundario respecto a la activa, y el coeficiente de masculinidad de la población activa, para conocer la participación de la mujer en el trabajo.

5. La composición educativa se trató a partir de dos variables, la primera con el fin de tener idea de la calidad de los recursos humanos actuales, medido por el porcentaje de mayores de 15 años que poseen algún grado de posprimaria, y para conocer el futuro de los recursos humanos se utilizó el porcentaje de población de niños entre 6 y 14 años que no asiste a la primaria.

6. Los procesos demográficos. Se usan cuatro variables: las tasas brutas de fecundidad y mortalidad, la proporción neta de migrantes respecto a la población total de la entidad (cuando la migración neta es negativa el resultado será negativo, y viceversa, a pesar de ser una proporción) y el crecimiento medio anual.

7. La composición étnica es medida por la población monolingüe, que es la sujeta a los mayores conflictos culturales.

Por otro lado, es importante cuidar los grados de confiabilidad de las fuentes de información, para esto existen muchas técnicas de evaluación y estrategia de ajuste que podrían ser material de un trabajo específico. Por ejemplo, el X Censo General de Población y Vivienda tiene un sesgo muy importante en los datos de población económicamente activa (Rendón, 1986), por lo que se decidió usar la ocupación en lugar de la rama de actividad para calcular los indicadores necesarios.

---

1 Kunz B., Ignacio 1985. Regionalización sociodemográfica del estado de Guanajuato. Tesis de maestría. División de Estudios de Posgrado de la Facultad de Filosofía y Letras. México. Aquí se trata con detalle el significado teórico de las variables utilizadas y la utilidad de su conocimiento para el Plan Nacional de Desarrollo 1983-1988.

Se tendrá que cuidar la pureza de la variable a fin de que refleje realmente la categoría de análisis que interesa y se vea lo menos afectada posible por otras variables, sobre todo cuando éstas se desconocen o no se pueden controlar. Esto se ve en el efecto que tiene la estructura de población sobre las tasas brutas de mortalidad y fecundidad: una estructura joven tiende a subestimar dichas tasas; en cambio, la esperanza de vida y la tasa neta de reproducción son indicadores que ilustran los mismos procesos, pero no se ven afectados por la estructura por edad.[2] Aquí al analizarse al mismo tiempo la variable de estructura se puede tener control sobre el efecto que produce.

La escala de significacia de las variables es otro aspecto a considerar. Hay fenómenos que se muestran sólo a ciertas escalas, así las variables que se usan para representarlos pueden perder significado si no corresponden al adecuado.

Cuando se presentan problemas de calidad pueden buscarse nuevos indicadores; si las dificultades están en la pureza o significancia se podrán buscar variables alternativas que reflejen de alguna manera las categorías de análisis. Así como no se deberán usar variables deficientes, tampoco se deberán eliminar variables necesarias según el planteamiento del problema, argumentando su mala calidad. Habrá que buscar solución a este tipo de problema. Hay que recordar que las variables que se usan en las técnicas estadísticas, y la misma técnica están en función del problema de investigación y no al contrario.

ESTADISTICAS BASICAS.

El análisis, por parte del investigador, de las estadísticas básicas (medidas de tendencia central, medidas de dispersión, sesgo, kurtosis, distribución de frecuencias, etc.) no es indispensable para la aplicación de muchas de las técnicas de estadísticas más complejas, a pesar de que la misma técnica en sus cálculos utilice elementos de las estadísticas básicas.

Sin embargo, es importante hacer este tipo de análisis con independencia del tipo de técnicas que se van a aplicar, pues ayuda a reseleccionar variables al detectar deficiencias de calidad, pureza y comportamiento estadístico (por ejemplo, en cuanto a sesgo) que no se vieron en el primer análisis; además, la comprensión de la estadística básica será un instrumento para la correcta interpretación de los resultados. A continuación se tratarán algunos aspectos del análisis de las estadísticas básicas.

---

2 Estos indicadores se pueden calcular de modo fácil estatalmente, lo mismo que se podría ajustar el sesgo de datos por rama de actividad, pero este trabajo sólo tiene por objeto ilustrar la técnica para regionalización y tipificación, y utiliza los datos demográficos de un proyecto de tipificación demográfica municipal en el cual no es posible realizar los ajustes por falta de información. Por eso se conservan los indicadores que originalmente se pensaron para aquel proyecto, excepto para la migración, pues mientras en este trabajo se utiliza la migración neta intercensal (76-80), en el proyecto sólo se puede utilizar el porcentaje de inmigrantes intercensales (76-80).

Las medidas de dispersión (varianza, desviación estándar y coeficiente de variación) determinan el grado de variación o heterogeneidad de cada variable. De esta manera pueden servir como criterio de selección al eliminar las variables que tienen una mínima variación, pues indican que existe homogeneidad en el atributo que caracterizan y, por tanto, no es necesario incluirlas en el cálculo. Esta exclusión tiene lugar en el procedimiento estadístico, no en el estudio; simplemente se parte de que todos los objetos originales, y, en consecuencia, todos los grupos que resulten serán homogéneos en dicho atributo. De cualquier manera, la clasificación será resultado de las variables que más varianza aporten. En el ejemplo que se aplica mostraron variaciones mínimas (medida con el coeficiente de variación) 5 variables: la población 0-14 (4.1%), la población 15-64 (3.8%), el coeficiente de masculinidad (2.2%), la población económicamente activa (6.0%) y la inasistencia a la primaria (6.9%), por lo que podrían considerarse variables con posibilidad de exclusión.

Las medidas de dispersión también se pueden usar para evaluar los resultados, comparando la variación original de las variables y su variación dentro de cada grupo. Es de esperarse una reducción significativa en la variación para que realmente tenga lugar una ganancia en cuanto a homogeneidad.

Los análisis de las distribuciones son importantes para evaluar el comportamiento de las variables y el significado de las medidas de tendencia central, sobre todo el de la media. Las técnicas que se proponen no requieren de distribuciones normales, pero, de cualquier manera, resulta conveniente no usar variables muy sesgadas. Cuando la distribución muestra mucha kurtosis y, sobre todo, un gran sesgo (cuando se disparan los valores de un objeto), y si, además. sucede en numerosas variables o en las de más varianza, la construcción de los grupos se puede ver muy afectada.

Es necesario, entonces, evaluar la distribución: si la distorsión se presenta en varias variables, por el comportamiento consistente de ciertos objetos, es conveniente aislar éstos, ya que al tomar regularmente valores extremos constituirán en sí mismos grupos o tipos específicos en la clasificación, y, por ser responsables del crecimiento de la varianza, el resto de los objetos quedará ditribuido en pocos grupos y, por tanto, los objetos, en su mayoría, estarán poco diferenciados.

Cuando es una variable la que muestra distorsión no por el comportamiento de un objeto que regularmente se dispersa, como en el caso anterior, sino por razones del propio indicador, podría señalarse alguna deficiencia en la información, o extrema sensibilidad del indicador a terceras variables o, simplemente, porque así es la realidad; en el primer caso convendría la refinación de la información o la sustitución del indicador; en el segundo y tercer casos se recomienda tener en cuenta la causa que provoca la distorsión así como las consecuencias que tiene sobre los resultados.

En el ejemplo de que se trata se esperaba que el Distrito Federal provocara sesgos significativos en las variables, y afectara los resultados. De hecho tuvo valores extremos en 10 de las 18 variables, representando hasta un 10% de la variación de dichas variables. Como consecuencia, la clasificación resultaba afectada por el comportamiento de dicha entidad; en cierto sentido los grupos quedaban constituidos en función del Distrito Federal.

En un segundo análisis, con exclusión del Distrito Federal aparecieron 3 entidades que provocaban sesgos en algunas variables: México, Quintana Roo y Zacatecas.

Por ejemplo, en el porcentaje de población respecto al total nacional, el estado de México provocó un sesgo (ver figura 1); al eliminarse esa entidad y recalcular las estadísticas básicas, el sesgo se redujo significativamente de 1.86 a 1.20 y la kurtosis de 6.84 a 4.36, que en curvas normales deben tender a 0 y 3, respectivamente. Si bien no llegan a ser distribuciones normales se mejoran en términos del efecto que pueden tener sobre la clasificación. La mejoría es clara al ver el número de frecuencias en el primer rango; con la separación del estado de México se pasó de 18 a 12 objetos (de un total de 31) en el primer rango. Por su parte, la desviación estándar pasa de 2.38 a 1.82.

La migración neta se veía muy sesgada positivamente y con una gran kurtosis por la presencia de Quintana Roo, incluso provocando la existencia de un rango vacío. Al excluirse este estado el sesgo pasó de 2.39 a 0.98, la kurtosis de 10.19 a 4.72 y la desviación estándar de 2.51 a 1.52, lo cual significaba mejoras importantes en la distribución (ver figura 1).

Otro ejemplo es sobre el coeficiente de masculinidad de la PEA, en el que originalmente había un sesgo de 1.8, una kurtosis de 8.1 y una desviación estándar de 50.3. La distorsión era causada, en parte, por Zacatecas que registró el mayor valor; al separarse esta entidad el sesgo bajó a 0.36, la kurtosis a 3.37 y la desviación estándar a 37.2, lo que significa que se aproximó mucho a la normalidad y que se redujo significativamente la variación.

Un último ejemplo es sobre el crecimiento, cuyos valores originales en sesgo y kurtosis eran de 2.25 y 9.01, pasando, con la exclusión del estado de México, a 0.65 y 3.13, mientras que la desviación estándar pasó de 1.6 a 0.88, lo que supone una mejoría significativa en la distribución (ver figura 1).

Estas 3 entidades afectaron de manera semejante otras variables que no se exponen por cuestión de espacio, pero su eliminación de los cálculos resultó convincente, en primer lugar, porque las distribuciones tendieron más hacia la normalidad provocando menores diferencias entre los rangos; en segundo lugar, hay una reducción en la variación (heterogeneidad); pero se puede afirmar que era una heterogeneidad artificial provocada en gran medida por un solo objeto, o sea, que dicho objeto provocaba gran parte de la varianza, de tal manera que al llevarse a cabo la clasificación la tendencia era a formar dos grupos, ese objeto por un lado y el resto de los objetos por otro, quedando éstos poco diferenciados y la construcción sesgada.

Estas distorsiones se pudieron confirmar en la construcción de los dendrogramas y en el escalamiento multidimensional en los que las entidades mencionadas tendían a formar cada una su propio grupo y a dejar poco diferenciado al resto del país. Así, si se quisieran obtener 5 regiones, una sería el Distrito Federal previamente excluido; probablemente 3 estarían constituidas por los 3 objetos aquí estudiados y la quinta por todo el resto del país. Por eso es conveniente excluirlas, al igual que el Distrito Federal, y realizar un nuevo análisis estadístico. Habrá que considerar que con esto se reduce la variación en todas las variables, llegando algunas a mostrar una gran homogeneidad en todos los objetos.

Por último se trata la estandarización, que es una operación que se realiza utilizando la media y la desviación estándar y que tiene como fin referir el conjunto de datos de cada una de las variables a su propia media, medidos en desviaciones estándar por arriba o por abajo de ésta. Así los valores de los datos son

Estadísticas básicas:

Con México, Quintana Roo y Zacatecas			Sin México, Quintana Roo y Zacatecas		
Porcentaje de población respecto al total nacional.					
Media	= 2.79		Media	= 2.62	
Desv. estándar	= 2.38		Desv. estándar	= 1.82	
Sesgo	= 1.86		Sesgo	= 1.20	
Kurtosis	= 6.84		Kurtosis	= 4.36	
Intervalos	Frecuencias		Intervalos	Frecuencias	
0.32 2.52	**********		0.32 1.86	**********	
2.52 4.71	*********		1.86 3.41	*********	
4.71 6.91	**		3.41 4.96	****	
6.91 9.11	*		4.96 6.51	*	
9.11 11.31	* (Méx)		6.51 8.06	**	
Porcentaje de migración neta					
Media	= 0.53		Media	= 0.12	
Desv. estándar	= 2.51		Desv. estándar	= 1.52	
Sesgo	= 2.39		Sesgo	= 0.98	
Kurtosis	=10.19		Kurtosis	= 4.72	
Intervalos	Frecuencias		Intervalos	Frecuencias	
-2.42 0.21	**********		-2.42 -0.96	*****	
0.21 2.86	**********		-0.96 0.50	**********	
2.86 5.50	**		0.50 1.96	*******	
5.50 8.14			1.96 3.43	*	
8.14 10.79	*(Q. Roo)		3.43 4.90	*	
Coeficiente de masculinidad de la PEA					
Media	= 297.54		Media	= 291.91	
Desv. estándar	= 50.33		Desv. estándar	37.23	
Sesgo	= 1.81		Sesgo	= 0.36	
Kurtosis	= 8.10		Kurtosis	= 3.37	
Intervalos	Frecuencias		Intervalos	Frecuencias	
218.0 272.2	*********		218.0 252.0	****	
272.2 326.4	**********		252.0 286.1	*********	
326.4 380.6	***		286.1 320.1	**********	
380.6 434.8	*		320.1 354.1	**	
434.8 489.0	*(Zac)		354.1 388.2	**	
Crecimiento de la población					
Media	= 3.56		Media	= 3.27	
Desv. estándar	= 1.61		Desv. estándar	= 0.88	
Sesgo	= 2.25		Sesgo	= 0.65	
Kurtosis	= 9.01		Kurtosis	= 3.13	
Intervalos	Frecuencias		Intervalos	Frecuencias	
1.63 3.27	**********		1.63 2.36	****	
3.27 4.92	**********		2.36 3.10	**********	
4.92 6.57	**		3.10 3.84	******	
6.57 8.22	*		3.84 4.58	*****	
8.22 9.87	*(Méx)		4.58 5.32	**	

Figura 1. Estadísticas básicas e histogramas de 4 variables.

independientes de las unidades de medida, con lo cual se pueden hacer cálculos y comparaciones. Si no se realiza la estandarización, la variación queda influida por la escala de la unidad de medidas que utiliza. Esta técnica siempre es conveniente y es indispensable para el análisis de cúmulos cuando no se usa una métrica ponderada.

El producto de la estandarización es una matriz de objetos-características-variables- (ver cuadro 1) equivalente a la matriz de datos originales. A partir de la matriz de datos estandarizados se realiza la aplicación de las técnicas.

Al igual que la estadística básica existen otras técnicas que, si bien no son indispensables para la construcción de los tipos y la regionalización, sí son un buen apoyo.

EL ANALISIS DE CORRELACION.

La correlación es una técnica que mide el grado de asociación entre pares de variables y, como la estadística básica, se utiliza como técnica de exploración y como apoyo interpretativo. Aunque la técnica de los componentes principales en que se centra la tipología parte de una matriz de correlación para sus cálculos (o de una matriz de varianza), no todos los investigadores realizan un análisis específico y profundo de esa matriz y se concretan a interpretar los resultados de la matriz variables-componentes. Aqui, por el contrario, se recomienda el estudio particular de las correlaciones por las razones expuestas al principio del párrafo.

La selección de variables, como se dijo anteriormente, debe obedecer a necesidades de la investigación. Es común que algunas de estas variables sean concomitantes, es decir, que muestran una asociación en su comportamiento estadístico, lo que significa que están correlacionadas. Puede ser que las variables se correlacionen por formar parte de un mismo proceso, porque existe algún tipo de relación causal o, simplemente, por coincidencia. Se dice, entonces, que se ubican en una misma dimensión. Se considera que el conjunto de multiples variables con que se trabaja es un espacio multidimensional en donde cada dimensión está dada por una variable; sin embargo, algunas de las variables pueden coincidir en ese espacio multidimensional, o sea, están correlacionadas. Cuando existe esa coincidencia entre un grupo, o al menos un par de variables, es posible eliminar algunas o una variables, según sea el caso, y conservar una de las variables correlacionadas que indirectamente informará sobre el comportamiento de las demás. Esto facilita los cálculos, pero se conservan las dimensiones de información que la investigación demanda. Para aplicar lo anterior es importante que la correlación sea muy clara y altamente significativa.

Lo que nunca se debe hacer es eliminar variables no correlacionadas, ya que no se podrán inferir a partir de las demás. Esto facilita aún más el trabajo, pues elimina números y al mismo tiempo dimensiones, lo que significa trabajar con menos información de la que originalmente se planteó como necesaria. La interpretación llega a ser simple pues se reduce artificialmente la complejidad de la realidad.

El análisis de correlación, además de ser útil en la selección de variables, es un instrumento de exploración del universo que se trabaja. Al conocer la asociación directa o inversa entre variables se obtienen elementos para la comprensión del fenómeno que se estudia; dichos elementos deberán corresponder a los resultados finales, con lo cual sirve, además, como elemento de validación.

ENTIDAD	NAL	URB	DISP	P.0-14	P.15-64	P. >=65	COE MASC	PEA	PEA I	PEA II	MASC PEA
AGS	-1.029	0.976	-0.509	0.925	-0.804	0.188	-1.546	-0.882	-0.984	1.123	-0.446
BC	-0.480	1.763	-1.692	-1.840	2.252	-1.503	-1.016	1.276	-1.711	0.791	-1.575
BCS	-1.282	0.377	0.227	-0.407	0.872	-1.335	2.510	0.198	-0.886	-0.185	1.199
CAMP	-1.111	0.343	-0.476	0.047	0.420	-1.329	0.243	-0.148	0.103	-0.058	0.763
COAH	-0.163	1.368	-1.024	-0.447	0.614	-0.279	-0.164	-0.705	-1.219	1.491	0.717
COL	-1.173	0.649	-0.881	0.042	0.097	-0.062	0.044	-0.498	-0.128	0.056	0.263
CHIS	0.276	-1.516	2.346	-0.173	-0.448	-1.938	1.656	1.885	2.271	-1.973	-0.747
CHIH	0.210	1.002	0.358	-1.201	1.222	-0.087	-0.659	0.592	-0.811	0.768	-0.652
DGO	-0.476	-0.487	0.898	1.352	-1.460	0.052	0.244	-1.247	0.042	-0.417	0.196
GTO	1.045	0.227	0.384	0.987	-1.021	0.497	-0.815	0.209	-0.515	0.509	-0.790
GRO	0.297	-1.019	0.995	0.891	-0.844	-0.039	0.021	1.184	1.075	-1.604	-2.019
HGO	-0.171	-1.471	1.382	0.624	-0.625	-0.419	0.826	0.275	0.575	-1.024	0.067
JAL	2.183	1.002	-0.534	-0.250	-0.402	1.012	-2.086	0.087	-0.898	0.973	-1.115
MICH	0.930	-0.514	0.628	0.742	-0.848	0.817	-1.077	-1.113	0.799	-0.688	0.342
MOR	-0.672	-0.413	-1.774	-0.674	0.626	0.298	-0.715	-0.074	-0.441	0.104	-0.712
NAY	-0.856	-0.446	-0.385	1.089	-1.094	0.700	0.896	-2.037	0.869	-0.779	2.634
NL	0.634	1.993	-1.247	-1.167	1.370	-0.495	0.026	-0.135	-1.805	2.327	-0.190
OAX	0.514	-1.751	1.124	-0.321	-0.082	0.932	-0.233	2.523	2.138	-1.860	-1.513
PUE	1.329	-0.597	-0.715	0.148	-0.569	0.746	1.197	0.068	0.933	-0.523	-0.333
QRO	-0.845	-0.498	0.555	1.613	-1.324	-0.293	0.057	-1.161	-0.059	1.368	1.267
SLP	-0.066	-0.532	1.381	0.793	-0.943	1.034	0.145	-0.257	0.349	-0.400	0.335
SIN	0.081	-0.216	0.446	1.614	-0.304	-0.803	1.394	-0.922	-0.101	-0.644	0.531
SON	-0.200	0.787	0.019	-1.056	1.345	-0.843	1.079	-0.130	-0.769	0.025	0.451
TAB	-0.575	-1.084	0.520	1.352	-0.699	-1.761	1.156	-0.870	0.707	-0.149	1.789
TAMS	0.143	1.268	-0.232	-1.218	1.203	0.244	-0.886	0.157	-1.021	0.866	-0.516
TLAX	-0.998	-1.051	-1.327	0.587	-0.904	1.438	0.157	-0.479	0.584	0.777	-0.191
VER	3.030	-0.561	0.592	-0.765	0.817	0.100	-0.101	0.714	0.698	-0.467	0.212
YUC	-0.575	0.400	-1.061	-2.286	1.531	2.490	0.040	1.490	0.208	-0.295	0.034

Cuadro 1. Matriz de datos estandarizados (28 objetos).
A partir de esta matriz se calcularon correlaciones. Componentes principales y cúmulos, por lo que sólo se incluyen los 28 objetos considerados en el análisis.

ENTIDAD	NO ASIST	POSTPRIM	NAT	MORT	MIG NETA	CREC	MONOLIN
AGS	1.006	0.109	0.207	0.024	0.847	1.277	-0.596
BC	-0.282	1.888	-1.854	-0.941	0.273	-0.238	-0.580
BCS	-0.984	1.374	0.406	-0.597	3.136	2.362	-0.564
CAMP	-1.375	0.021	-0.834	-1.161	1.145	2.299	-0.059
COAH	0.995	0.993	-0.141	-0.267	0.211	0.138	-0.624
COL	1.477	0.732	-0.714	0.387	1.465	0.469	-0.598
CHIS	2.104	-1.761	0.152	-0.005	-0.499	-0.455	2.934
CHIH	-0.476	0.459	-1.541	-0.020	-0.029	-1.236	-0.420
DGO	-0.563	-0.335	0.548	-0.889	-0.807	-1.093	-0.596
GTO	0.572	-1.061	-0.609	0.918	-0.481	-0.496	-0.623
GRO	0.368	-0.962	1.647	-0.489	-1.676	-0.527	1.155
HGO	-0.735	-0.933	0.938	1.628	-1.240	-0.748	1.212
JAL	-0.419	0.196	-0.108	0.027	0.376	-0.477	-0.601
MICH	1.597	-0.822	1.291	0.260	-0.607	-1.325	-0.453
MOR	-0.999	0.882	0.052	-0.428	0.571	1.285	-0.538
NAY	0.887	0.234	-0.513	-1.202	-0.181	-0.402	-0.428
NL	1.445	2.248	-1.450	-1.540	1.039	0.853	-0.638
OAX	-0.907	-1.765	1.115	2.543	-1.510	-1.898	2.713
PUE	0.821	-0.663	1.207	2.262	-0.616	-0.402	0.567
QRO	0.665	-0.435	0.738	0.937	1.045	1.176	-0.513
SLP	-0.405	-0.706	-0.596	-0.543	-0.864	-0.662	0.121
SIN	-0.424	0.324	0.304	-1.361	-0.276	0.672	-0.606
SON	-0.170	1.345	-1.335	-0.224	0.186	-0.025	-0.548
TAB	-1.586	-0.794	2.435	0.283	0.080	0.025	-0.495
TAMS	0.089	0.767	-0.834	-0.638	-0.099	-0.524	-0.619
TLAX	-0.718	-0.343	0.229	0.926	-0.272	-0.504	-0.491
VER	-0.285	-0.702	-0.364	-0.486	-0.536	0.268	0.177
YUC	-1.695	-0.288	-0.372	0.595	-0.681	0.189	1.712

Cuadro 1. Matriz de datos estandarizados. Continuación.

En el estudio de las entidades del país se calcularon los coeficientes de correlación de Pearson y se seleccionaron los mayores índices, que registraron un nivel de significancia descriptivo igual o menor a .001 que para el número de objetos con que se trabajó es un límite muy confiable, aproximadamente de 0.56 en valor absoluto. Por insuficiencias de espacio no se muestra la matriz de correlación completa, sólo se presenta un cuadro con las correlaciones significativas; las variables que no se incluyen no mostraron ninguna correlación importante, (ver cuadro2).

	DISP	15-64	PEA	PEA I	PEA II	POSPR	NAT	MIG	MONOL	MORT
URB	-0.62	0.63		-0.91	0.77	0.83	-0.72		-0.58	
DISP	1.00			0.61	-0.63	-0.69				
0-14		-0.94	-0.57				0.61			
15-64		1.00				0.66	-0.63			
PEA			1.00			-0.70			0.74	
PEA I				1.00	-0.84	-0.86	0.60	-0.57	0.73	
PEA II					1.00	0.66			-0.70	
MASC PEA			-0.70							
POSPRIM						1.00	-0.63		-0.64	-0.58
CRECIM						0.65		0.79		

Cuadro 2. Matriz simplificada de correlaciones. Sólo se incluyen correlaciones significativas. Las variables que no están presentes no tuvieron ninguna correlación.

Posteriormente se formaron grupos de variables asociadas por sus correlaciones; esto es, conjuntos de variables que registran entre sí correlaciones positivas o directas. Generalmente se forman dos grupos principales que, como se acaba de decir, muestran correlaciones positivas entre las variables que los constituyen y, al mismo tiempo, correlaciones negativas o inversas entre las variables de un grupo respecto al otro. Pueden también aparecer conjuntos o variables con pocas ligas o totalmente independientes de los dos principales (ver figura 2).

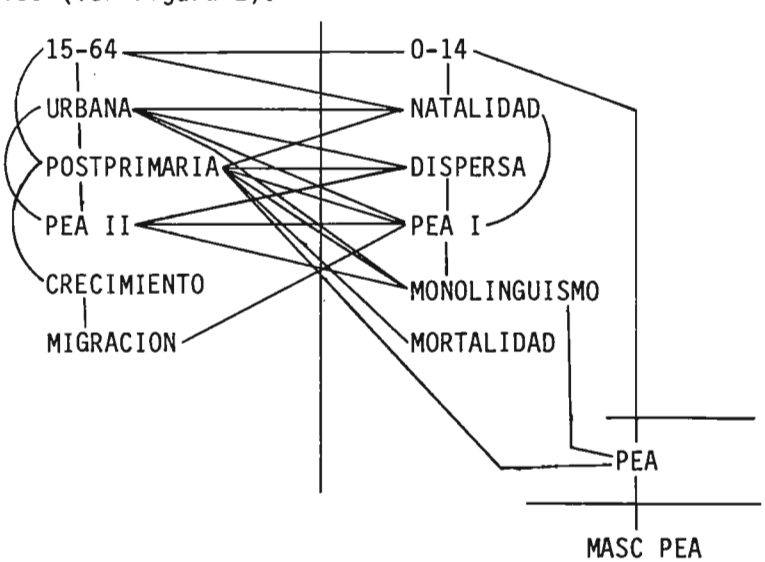


Figura 2. Esquema de correlaciones entre variables.

Como se podrá ver en la figura, se forman dos grandes grupos de variables con relaciones muy lógicas. En el primero se sitúa a la población urbana, a la población en edad de trabajar, a aquella que posee al menos un grado de posprimaria, a la población que participa en el sector secundario, al crecimiento y a la migración, todas ellas variables asociadas a condiciones de "desarrollo" socioeconómico de la sociedad; así, en lugares más urbanos es lógico esperar un mejor nivel de estudios, más importancia del sector industrial, lo que favorece la migración de población en edad de trabajar y, por ende, el crecimiento. En cambio, en el segundo grupo se presenta una situación contraria: dominio del sector primario, población dispersa, mayor natalidad y por ende, mayor proporción de niños y jóvenes, aunque esta también tiene origen en la menor importancia de la población en edad de trabajar que tiende a emigrar, mayor proporción de población indígena y mayor mortalidad. En realidad, estos dos conjuntos de variables forman una sola dimensión en sentidos contrarios; lugares con valores altos en las variables del primer grupo mostraran valores bajos en las variables del segundo, y viceversa. De esta manera representan las variables que en conjunto acarrearán más varianza y determinarán, en gran medida, la clasificación en el caso de la técnica de cúmulos, y constituirán la componente más importante en el caso del análisis de componentes principales. La población económicamente activa y su coeficiente de masculinidad no se asociaron de manera clara a ninguno de los grupos, por lo que se puede suponer representan otra dimensión. Las variables que no se incluyeron en el esquema son las que no registraron correlación significativa por ubicarse en dimensiones particulares, lo que quiere decir que son independientes del resto.

Se podría eliminar la población entre 0 y 14 años, pues está altamente correlacionada con la población entre 15 y 64 años de donde se podría inferir. De igual manera, se puede utilizar a la posprimaria como indicador de la población urbana y de la población en el sector primario con las que tiene altas correlaciones. La eliminación de las variables que mostraron mínima variación, como de aquellas que resultaron muy correlacionadas no afecta en nada los resultados, de aquí que pueda ser una medida interesante cuando se presentan problemas de manejo de volumen de información. Hay que recordar que dicha eliminación se debe dar sólo a nivel de los cálculos, pero las variables se deben incluir nuevamente durante la descripción y análisis de los resultados.

Es común que se presenten correlaciones signficativas que teóricamente no se esperaban, o bien, que no se presenten cuando sí se esperaban. En el primer caso se puede tratar de correlaciones espúreas, pero en ambas situaciones es recomendable el estudio de diagramas de dispersión en los que se puede describir el comportamiento de cada objeto, y así saber cuando algunos de ellos son responsables de alteraciones en la medida de correlación. En ocasiones hay objetos que por alguna razón particular no muestran la correlación y, sin embargo, pueden afectar el resultado general cuando tienen mucho peso; en otras palabras, son responsables de buena parte de la variación. De cualquier manera, el análisis de estos diagramas es importante, aun en el caso de correlaciones que se dieron y se esperaban, porque algunos objetos pueden no cumplir con el patrón; considerar estas excepciones hace más precisas las generalizaciones e inferencias.

COMPONENTES PRINCIPALES

La técnica de componentes principales se utiliza para la construcción de la tipología. Esta técnica busca la formación de familias de variables cuyo comportamiento sea concomitante, o sea, variables que estén correlacionadas. Esto se logra trasladando el sistema de referencia. En principio existe un espacio n dimensional determinado por las n variables con que se trabaja, pero algunas de estas variables pueden coincidir, en la misma dimensión, con otras variables, por lo que resulta conveniente considerar esas coincidencias y partir de un nuevo sistema de referencia representado por n componentes cada uno de los cuales es una dimensión en la que se agrupan la variación común entre variables; o sea, son familias de variables correlacionadas. La primera componente se ubica en una dimensión tal que reúne la mayor cantidad posible de varianza en una sola dimensión; la segunda, que se ubica en una dimensión ortogonal a la primera y, por tanto, es independiente de ella, se ubica en la siguiente dimensión que le permite reunir otra vez el máximo de la varianza restante, y así sucesivamente hasta completar las n dimensiones originales.

Como se podrá suponer, las primeras componentes explican la mayor parte de la variación, mientras que las siguientes van explicando cada vez menos, de tal manera que se pueden reducir las dimensiones del problema al considerar sólo las componentes que aportan más en la explicación de la varianza (ver cuadro 3).

Componente	% de varianza explicada	% acumulado de la varianza explicada
1	37.602	37.602
2	19.323	56.925
3	13.005	69.930
4	7.590	77.520
5	5.229	82.749
6	4.191	86.940
7	3.400	90.340
.	. .	. .
:	: :	: :
18	0.013	100.000

Cuadro 3. Distribución en porcentaje de varianza explicada por componente y de varianza explicada acumulada.

En el cuadro 3 se ve que los primeros 4 componentes reúnen el 77.52% de la variación explicada, y que desde el mismo cuarto componente las ganancias en variación explicada son poco importantes. Se podría trabajar sólo con los 3 primeros (69.3% de varianza) o los 4 primeros, pero antes de tomar la decisión es importante analizar el contenido de cada componente en la matriz de variables-componentes (Cuadro 4). Aquí sólo se presenta la primera parte de la matriz pues el resto carece de importancia para los fines de este trabajo.

Al igual que en la correlación, los valores que tienden a $\pm$ 1 indican asociación, directa o inversa (según sea el signo), mientras que los que tienden a 0 indican independencia. Así, en la primera componente se ubican de manera directa y en orden de la fuerza de su asociación: la PEA I, la población dispersa, el monolingüismo, la natalidad y la mortalidad, y en forma inversa la población con posprimaria, la población urbana, la PEA II, la migración, la población en edad de trabajar,

y el crecimiento. Como se podrá ver, lo anterior corresponde plenamente al esquema propuesto al estudiar las correlaciones. Podría afirmarse que es una dimensión en que se mide el grado de "desarrollo" de las entidades estudiadas, en términos de su desarrollo urbano y transformación de la composición económica, a lo que se asocian procesos como la migración y el crecimiento demográfico.

VARIABLES	COMPONENTES PRINCIPALES				
	1	2	3	4	5
P. Nacional	0.252	0.498	-0.286	-0.408	-0.101
P. Urbana	-0.908	0.248	-0.129	-0.152	0.009
P. Dispersa	0.725	-0.173	0.231	-0.420	-0.101
P. Niños	0.470	-0.737	-0.385	-0.094	0.089
P. Adultos	-0.630	0.558	0.465	0.055	-0.084
P. Ancianos	0.244	0.327	-0.564	0.485	-0.227
Coef. Masc.	0.130	-0.522	0.738	-0.043	-0.177
PEA	0.290	0.775	0.488	0.015	0.211
PEA I	0.921	-0.083	0.177	0.023	0.006
PEA II	-0.802	0.008	-0.407	0.044	0.136
Coef. Masc Pea	-0.144	-0.791	0.080	0.118	-0.282
No asistencia	-0.005	-0.061	-0.320	-0.706	0.360
Posprimaria	-0.943	0.021	0.092	-0.017	-0.068
Natalidad	0.688	-0.369	-0.095	0.206	0.250
Mortalidad	0.585	0.237	-0.217	0.398	0.412
Migración	-0.723	-0.378	0.223	0.071	0.394
Crecimiento	-0.565	-0.423	0.343	0.171	0.367
Monolingüismo	0.719	0.392	0.433	0.022	0.154

Cuadro 4. Valores de las cargas de cada variable en las 5 primeras componentes.

Por su parte, en la segunda componente se asocian de manera directa la PEA y la población en edad de trabajar, y, en forma inversa, el coeficiente de masculinidad de la PEA, el porcentaje de niños y jóvenes y el coeficiente de masculinidad general. Esta es una componente que trata sobre la composición por edad y sexo y de la participación en el trabajo también por edad y sexo. Así, las entidades que mayor población activa poseen también registran mayor proporción de población adulta y tienen mayor participación de la mujer en el trabajo (coeficiente de masculinidad bajo), lo que llega a reducir el coeficiente de masculinidad general.

La proporción de adultos se presentó de manera significativa en las dos primeras componentes y en ambas hay una clara posibilidad de explicación teórica; sin embargo, para términos de interpretación se considerará a esa variable en la segunda componente, pues ésta trata de la estructura por edad y sexo, en lo que la variable discutida juega un papel muy importante.

La tercera componente registra en forma directa al coeficiente de masculinidad y en forma inversa la proporción de población mayor o igual a 65 años. No obstante, la correlación entre ambas variables es sólo de -0.46; es decir, no es muy significativa, por lo que se puede pensar que su coincidencia en la tercera componente obedece más a una relación espúrea que a una relación real. Por esto, para fines de interpretación se considera el coeficiente de masculinidad en la segunda componente

en la que juega un papel importante por ser esta variable un indicador de la composición de la población por sexo, además de que en esa componente hay la posibilidad de interpretar claramente relaciones con el resto de las variables significativas. De cualquier manera, tanto la población adulta como el coeficiente de masculinidad fueron variables con una variación mínima (ver estadísticas básicas) por lo que no ofrecen riesgo en cuanto a alterar los resultados.

En la cuarta componente sólo se presenta la proporción de niños que no asiste a primaria. Los siguientes índices se encuentran muy alejados (% de la población nacional y población dispersa) y no parecen mostrar ninguna relación lógica con la no asistencia a primaria. Al analizar las correlaciones se verá que son totalmente independientes.

En estos cuatro componentes se interpretan todas las variables excepto una: el porcentaje de población respecto al total nacional, que es significativa hasta la séptima componente y que, de hecho, no mostró ninguna correlación alta. Se trata de una variable aparentemente independiente de los aspectos estudiados. Se dice aparentemente porque la independencia está asociada a la escala de análisis. Estatalmente sí existe independencia, pues la proporción de población no llega a intervenir en los procesos analizados con el esquema de variables aquí propuesto; sin embargo, municipalmente el volumen de población y, por tanto, la proporción de ésta respecto al total nacional estará íntimamente ligado al desarrollo urbano del municipio y se podrá esperar que muestre correlaciones importantes y se inserte de manera significativa en el primer componente. Por el momento, en la medida en que el ejemplo aquí tratado es estatal se excluirá dicha variable de la tipología.

Anteriormente se habló de la posibilidad de eliminar variables, algunas por su baja variabilidad y otras por su alta correlación. Realizando el análisis sin esas variables (ver páginas 6 y 13) los resultados fueron plenamente consistentes con lo aquí obtenido.

Después de reducir las 18 dimensiones originales a un número menor, según los componentes que se crea conveniente incluir (en este caso 4), el problema es determinar la ubicación o el valor de cada objeto en cada una de esas dimensiones. A estos valores se les llama scores, e igual que los datos estandarizados se presentan en desviaciones estándar por arriba o por abajo de la media. Como la tercera y cuarta componente sólo incluyen una variable (la dimensión de la componente se usa exclusivamente para interpretar la dimensión de una variable) es mejor utilizar los datos propios de esa variable en lugar de los scores, para obtener mejor caracterización, pues a pesar de que el resto de las variables sean poco significativas estadísticamente llegan a influir en los valores de los scores.

Para la presentación de este ejemplo se calcularon los scores de las 2 primeras componentes y se tomaron los datos estandarizados de la población mayor o igual a 65 años y del porcentaje de niños que no asiste a primaria (ver cuadro 5). Con los scores (o datos según sea el caso) de las entidades es posible formar rangos que agrupen a entidades semejantes en la dimensión o componente; a cada rango se le da una identificación (número o letra), de tal manera que las entidades se caracterizarán por una clave que es el conjunto de las identificaciones (una para cada dimensión); dichas claves corresponderán a un tipo.

Los 4 componentes se dividieron en 7,3,2 y 2 rangos, respectivamente, tratando de obtener una equivalencia entre el detalle de la división en rangos y la varianza

explicada por cada componente. En la primera componente se utilizaron números romanos del I al VII, de tal manera que el primer grupo mostrará los mayores indicadores de "desarrollo" (mayor proporción de población con posprimaria, de población urbana, de población en el sector II, etc., menor proporción de población en el sector I, menos porcentaje de población dispersa, menos proporción de monolingües, etc.), en cambio, el grupo opuesto, el VII, registra los menores índices de "desarrollo", o sea valores altos en variables que reflejan condiciones de atraso y bajos en variables que reflejan condiciones de progreso, exactamente al contrario del grupo I. Entre I y VII se establece un continuo asociado al nivel de "desarrollo" que nos refleja la componente (ver cuadro 5).

De manera análoga se trató la segunda componente. Se formaron tres grupos A, B y C en los que A tiene los menores valores de los coeficientes de masculinidad general y de la PEA, y los menores porcentajes de la población 0-14, al mismo tiempo que registra los porcentajes más altos de PEA y población entre 15 y 64 años, en tanto que C es el extremo opuesto con un patrón contrario al de A (ver cuadro 5).

Como se dijo anteriormente, la tercera y cuarta componente sólo se utilizarían para analizar una variable, por lo que se trabajaron los valores estandarizados de las propias variables, en lugar de los scores, ya que éstos dan la ubicación de los objetos en una dimensión en la que están involucradas otras variables que, a pesar de no ser significativas desde el punto de vista estadístico, si llegan a influir. En cambio, el uso directo de los valores de la variable permitió más pureza. De esta manera se dividió el porcentaje de población mayor de 65 años (variable que corresponde a la tercera componente) en dos rangos, 2 y 1, según que registren alta o baja proporción respecto a la media; y el porcentaje de no asistencia a primaria se dividió en b y a según que sea alta o baja, también respecto a la media, que en los datos estandarizados es 0 (ver cuadro 5).

Como se podrá ver, sólo coinciden en el mismo tipo Guanjuato y Michoacán, lo cual no tendría sentido, pues se conserva virtualmente el mismo detalle original y no existe ganancia en cuanto a generalidad. Esto se debe a la precisión con la que se realizó la tipología en relación con el reducido número de objetos. Existen dos formas de solucionar este aparente problema: en primer lugar, eliminar los dos últimos dígitos de la clave, teniendo en cuenta que se trata de atributos de baja variabilidad, de los que, como desde el principio se comentó, se podría prescindir. La segunda opción es reducir el número de rangos en que se dividieron los componentes. De hecho ésta es la alternativa más adecuada cuando no se pueden eliminar componentes o variables. En cualquier forma, el nivel de generalidad-detalle podrá variar en función de los objetivos que se persigan y del número de objetos con los que se trabaja. No existe un nivel óptimo determinado.

Después de determinar los tipos es importante caracterizar cada uno de ellos, para lo que resulta conveniente la construcción, para cada componente, de una tabla en la que las columnas son las variables significativas de la componente, los renglones los rangos que se determinaron, y las entradas los intervalos entre los valores máximo y mínimo de las variables originales (tasa, porcentajes, etc.) de los objetos que quedaron clasificados en cada rango de la componente (ver cuadro 6). Con ayuda en estas tablas se pueden caracterizar los objetos aislados en un principio.

Entidad	S C O R E S		DATOS ESTANDARIZADOS		
	Comp. 1	Comp. 2	%>=65	% no asist	TIPO
Ags	-4.453	-1.908	0.188	1.006	III B 2 b
BC	-11.437	5.814	-1.503	-0.282	I A 1 a
BCS	-7.372	-4.875	-1.335	-0.984	II C 1 a
Camp	-4.916	-2.786	-1.329	-1.375	III C 1 a
Coah	-7.213	-0.333	-0.279	0.995	II B 1 b
Col	-4.546	-1.581	-0.062	1.477	III B 1 b
Chis	11.911	0.891	-1.938	2.104	VII B 1 b
Chih	-4.226	4.193	-0.087	-0.476	III A 1 a
Dgo	3.506	-3.415	0.052	-0.563	V C 2 a
Gto	2.039	1.167	0.497	0.572	V B 2 b
Gro	9.668	1.527	-0.039	0.368	VI B 1 b
Hgo	9.290	-0.362	0.419	-0.735	VI B 2 a
Jal	-2.749	3.739	1.012	-0.419	III A 2 a
Mich	5.675	-1.086	0.817	1.597	V B 2 b
Mor	-4.793	0.718	0.298	-0.999	III B 2 a
Nay	0.630	-5.937	0.700	0.887	IV C 2 b
NL	-13.341	1.814	-0.495	1.445	I B 1 b
Oax	15.079	5.851	0.932	-0.907	VII A 2 a
Pue	6.005	1.966	0.746	0.821	VI B 2 b
Qro	0.174	-5.773	-0.293	0.665	IV C 1 b
SLP	4.252	-0.993	1.034	-0.405	V B 2 a
Sin	-0.410	-3.654	-0.803	-0.424	IV C 1 a
Son	-5.814	0.592	-0.843	-0.170	III B 1 a
Tab	4.452	-6.346	-1.761	-1.586	V C 1 a
Tams	-5.940	3.342	0.244	0.089	III A 2 b
Tlax	1.917	-1.044	1.438	-0.718	IV B 2 a
Ver	2.476	2.812	-0.100	-0.285	V A 1 a
Yuc	0.141	5.661	2.490	-1.695	IV A 2 a
Varianza	6.76	3.47	2.34 -	1.36	
No. rangos	7	3	2	2	
Rango I	<-10.1	A >2.1			
II	-10 a -6.1	B -2 a 2	1 < 0	a < 0	
III	-6 a -2.1	C <-2.1	2 > 0	b > 0	
IV	-2 a 2		estas dos columnas se tomaron		
V	2.1 a 6		de la matriz de datos estanda		
VI	6.1 a 10		rizados		
VII	>10				

Cuadro 5. Scores, datos estandarizados y tipos de las 28 entidades analizadas por componentes principales.

Así, el Distrito Federal (I' -II- A'2b) posee valores muy altos e incluso extremos en la mayoría de las variables del grupo I, excepto en la PEA II cuyo comportamiento es como el del grupo II. En la segunda componente todas sus variables son A, pero muestran valores extremos.

RANGOS digito 1	POSTPR en %	PEA I en %	URB en %	PEA II en %	DISP en %	MIGR * en %	MONOL en %	NAT por 1000 hab	MORT
I	>34	<12	>74	>25	<9	0-2	<1	<33	<6
II	29-34	12-20	53-74	19-30	9-19	0-5	<1	38-41	5-7
III	23-29	17-31	39-70	20-28	4-20	0-3	<3	32-40	5-7
IV	19-26	28-41	28-54	15-29	7-21	(-2)-(2)	<8	36-43	4-2
V	14-21	23-40	27-51	16-24	19-27	(-2)-(1)	<3	36-50	5-8
VI	14-19	36-48	21-37	10-18	11-27	(-3)-(0)	4-7	43-47	5-10
VII	<14	>48	<21	<10	>24	(-3)-(0)	>11	39-43	6-10
D.F.	46.9	0.9	100.0	22.8	0.0	-1.8	0.3	32.7	5.6
MEX.	29.1	14.4	62.4	32.3	4.8	4.0	0.6	28.8	6.4
Q.R.	22.4	28.1	48.2	15.7	18.6	10.8	6.8	34.9	4.2
ZAC.	13.7	48.9	26.4	16.1	28.6	-1.7	0.1	36.6	5.7

digito 2	MASC PEA coef.	PEA en %	P.0-14 en %	P.15-64 en %	C. MASC. coef.
A	234-300	32-37	39-44	51-57	95-100
B	218-318	30-35	41-46	50-55	96-103
C	299-389	28-33	43-47	49-54	99-104
D.F.	175.6	37.5	37.0	59.0	92.1
MEX.	265.8	31.9	44.6	52.5	98.6
Q.R.	295.5	35.1	45.3	52.4	106.1
ZAC.	489.0	26.4	48.2	47.2	97.5

digito 3	P.>=65 en %	digito 4	P. 6-14 NO ASISTE en %
1	<3.89	a	<27.0
2	>3.90	b	>27.1
D.F.	3.93	D.F.	27.8
MEX	2.70	MEX.	25.7
Q.R.	2.10	Q.R.	23.8
ZAC.	4.45	ZAC.	27.5

Cuadro 6. Caracterización de los tipos. Los límites superior e inferior de los rangos se aproximaron al entero inmediato superior e inferior, respectivamente.

* En la migración los porcentajes positivos se refieren a la inmigración neta y los negativos a la emigración neta.

El estado de México (II -I- B1a) es tipo II en la mayoría de las variables, a excepción del porcentaje de PEA II, de la población dispersa y de la tasa de natalidad en las que se comporta como grupo I. En cuanto al resto de la clave es B1a.

Quintana Roo (IV -I-III- C -A- 1a) es IV excepto en el porcentaje de migración en el que se comporta como I (con valores disparados incluso para este grupo) y en la tasa de natalidad en la que es III. En la segunda componente es C, con valores extremos en el coeficiente de masculinidad; es A en la proporción de PEA y es B en el coeficiente de masculinidad de la PEA, aunque con valores muy próximos a C.

Zacatecas (VII -V-VI- C2b), por su parte, es VII en las dos variables más significativas de la primera componente (porcentajes de posprimaria y de PEA I) y también en la proporción de población dispersa; es VI en el porcentaje de población urbana, y V en las restantes. En la segunda componente es C con valores disparados en todas las variables, excepto en el coeficiente de masculinidad en el que es B. El resto de la clave es 2b.

EL ANALISIS DE CUMULOS

El análisis de cúmulos se propone para la construcción de regionalizaciones, es una técnica estadística multivariada que permite construir clasificaciones jerárquicas de los objetos. El hecho de ser jerárquica facilita el manejo de distintos niveles en el continuo de generalidad detalle. Esto quiere decir que una misma clasificación puede ofrecer desde un mínimo de generalidad y máximo detalle al conservar todos los objetos originales por separado, hasta más generalidad y menos detalle al ir conjuntando todos los objetos en grupos hasta llegar a un solo conjunto.

La aplicación de la técnica supone dos etapas: la medida de la disimilaridad (distancia taxonómica -distancia en un espacio multidimensional-) entre objetos, y la construcción de la clasificación jerárquica de los objetos según las medidas anteriores, esto último representado gráficamente por un dendrograma (ver figura 3).

Hay varias formas de medir la disimilaridad (similiaridad) entre objetos; algunas (coeficiente de Gower, distancia de Mahalanobis, etc.) permiten ponderar las variables, lo que dependerá de las necesidades del trabajo, por ejemplo, para disimular la perturbación provocada por las diferencias de escala entre variables; otras medidas se usan para matrices de ausencia presencia (coeficiente de Jaccard, presencias comunes, presencias y ausencias comunes, etc.). Para el presente ejemplo se utilizaron la distancia euclidiana (medida de dismilaridad) que corresponde a la métrica de Minkowski de orden 2, que es la forma de medición de semejanzas entre objetos multivariados más extendida, y el coeficiente general de Gower sin ponderar (medida de similaridad). Ambos se calcularon a partir de una matriz de datos estandarizados (ver cuadro 1), para evitar los sesgos provocados por diferencias de escala entre variables.

Para conocer la forma de calcular las medidas de disimilaridad y de construir los dendrogramas se puede consultar: Anderberg, 1973; Espinoza, 1980; Johnson, 1980; Marriot, 1974; y Reyes, 1978.

Con las medidas de disimilaridad (similaridad) se debe construir una matriz simétrica en la cual tanto los renglones como las columnas son los objetos, y las entradas son las propias medidas; por tanto, la diagonal será 0, pues la disimilaridad de cada objeto consigo mismo es inexistente. A partir de la matriz se construyen

los dendrogramas, para lo que hay diversas técnicas. Para la regionalización se utilizaron 7 de las más comunes (disponibles en el programa ANALISIS/CUMULOS -Reyes, 1978-): conexión simple, conexión completa, promedios intergrupales, promedios intragrupales, promedios intragrupales sin la diagonal, promedios intragrupales ponderados, y el incremento a los promedios intragrupales ponderados.

Las medidas de disimilaridad que se utilicen dependen del tipo de matriz con la que se trabaje (matriz de objetos-características, matriz de presencia ausencia); sin embargo, para utilizar una de las distintas medidas que se pueden usar para los mismos tipos de matriz no existe teoría que establezca cuál es la más adecuada. Igual sucede con la técnica para la construcción del dendrograma. En ocasiones, un criterio para no considerar algunas medidas o técnicas es reconocer las deformaciones que provoca la aplicación de tales medidas en los datos originales; no obstante, esto se percibe sólo en casos extremos, por ejemplo, cuando un dendrograma se encadena. En este ejemplo tres de los dendrogramas mostraron serias deformaciones.

Para seleccionar el tipo de clasificación que se utilizaría de las 11 resultantes (2 técnicas de medida de disimilaridad por 7 técnicas de construcción de dendrogramas menos los tres deformados) se usó como criterio la minimización de la varianza interna de los grupos formados (Orford, 1976). Esto permite, además, determinar en qué nivel de generalidad-detalle del dendrograma formar los grupos. La razón reside en el objetivo del trabajo: construir una clasificación (que al llevarla al espacio sea una regionalización) en la que se agrupen entidades lo más semejantes posible para obtener regiones homogéneas. Entonces la suma de la varianza interna de los grupos (llamada W) deberá minimizarse (máxima homogeneidad), lo que, a su vez, supone que la suma de la varianza entre grupos se maximiza.

Ahora bien, en un mismo dendrograma la W variará al irse agregando los objetos; en la base del dendograma (figura 3) la W es cero, pues no se ha formado ningún grupo y, por tanto, no hay variación interna; los objetos se mantienen aislados y toda la varianza se da entre objetos, y conforme éstos se van agregando la W crece hasta llegar al máximo de varianza, o sea, cuando todos los objetos están agregados en un mismo grupo o región. Lo importante es determinar un nivel en el que teniendo el menor número de regiones posibles (máxima generalidad) se tenga, al mismo tiempo, el mínimo de W (máxima homogeneidad) (ver figura 4).

Se calculó por medio de un programa inédito, de M. Cortina, el valor de la W para los once dendogramas a niveles desde 6 a 10 regiones (con menos regiones se perdería mucho detalle, es decir se tendrían grupos de entidaes muy agregados y, por tanto, muy generales, y con más de 10 regiones -más las cuatro representadas por cada una de las entidades que se aislaron previamente- se tendría el extremo opuesto: demasiado detalle, pero con muy poca ganancia de generalidad), y se seleccionó la técnica de incrementos a los promedios intragrupales ponderados desarrollada a partir de una matriz de distancias euclidianas que minimiza la W a un nivel de 8 regiones o grupos que son, según el orden del dendrograma:

I	II	III	IV	V	VI	VII	VIII
Chih	BC	Yuc	Coah	Camp	Dgo	Mich	Gro
Tams	NL		Col	Mor	SLP	Pue	Hgo
Son			Ags	BCS	Sin	Gto	Oax
					Tlax	Jal	Chis
					Nay	Ver	
					Qro		
					Tab		

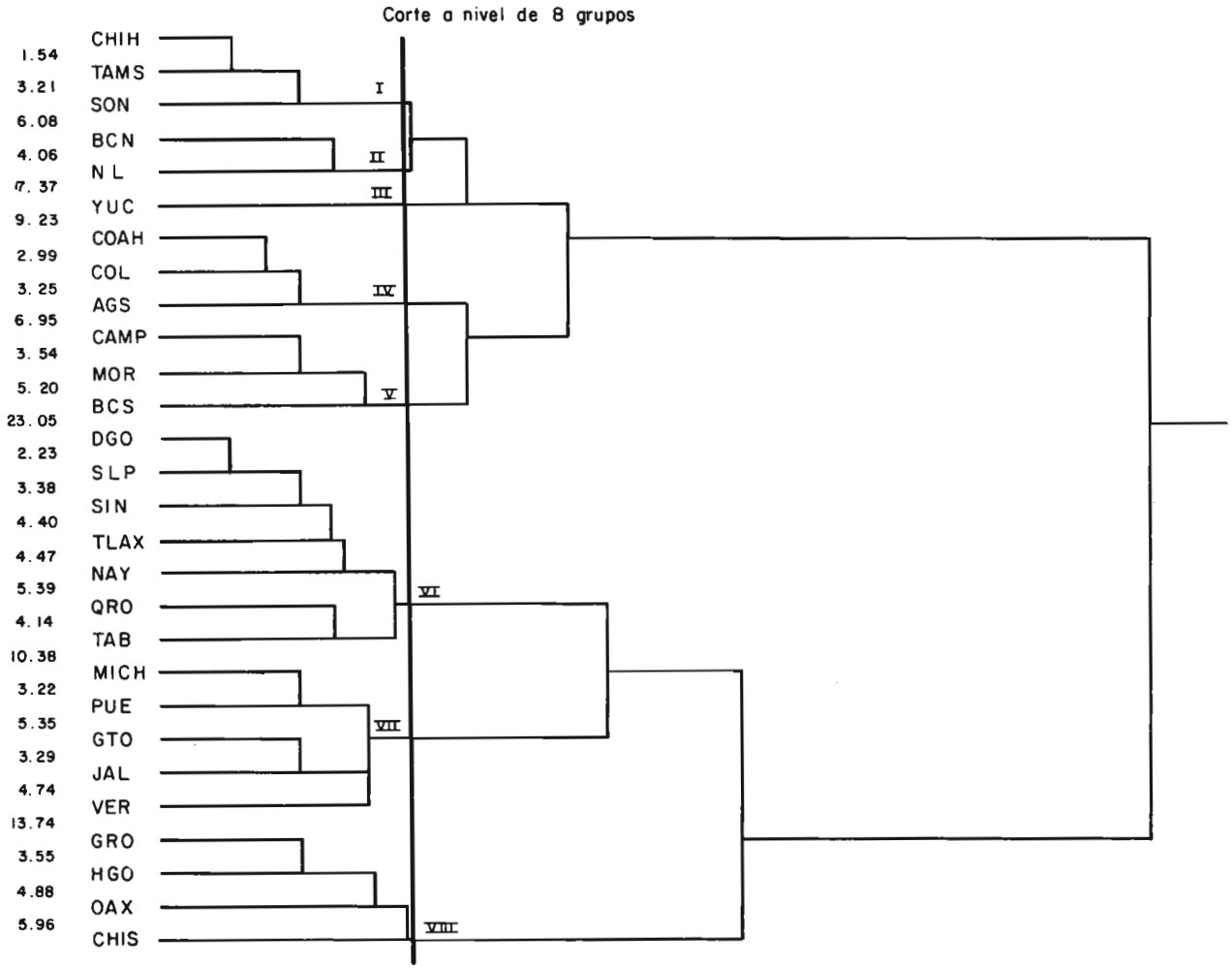


FIGURA 3 DENDROGRAMA CONSTRUIDO POR EL INCREMENTO A LOS PROMEDIOS INTRAGRUPALES PONDERADOS A PARTIR DE UNA MATRIZ DE DISTANCIAS.

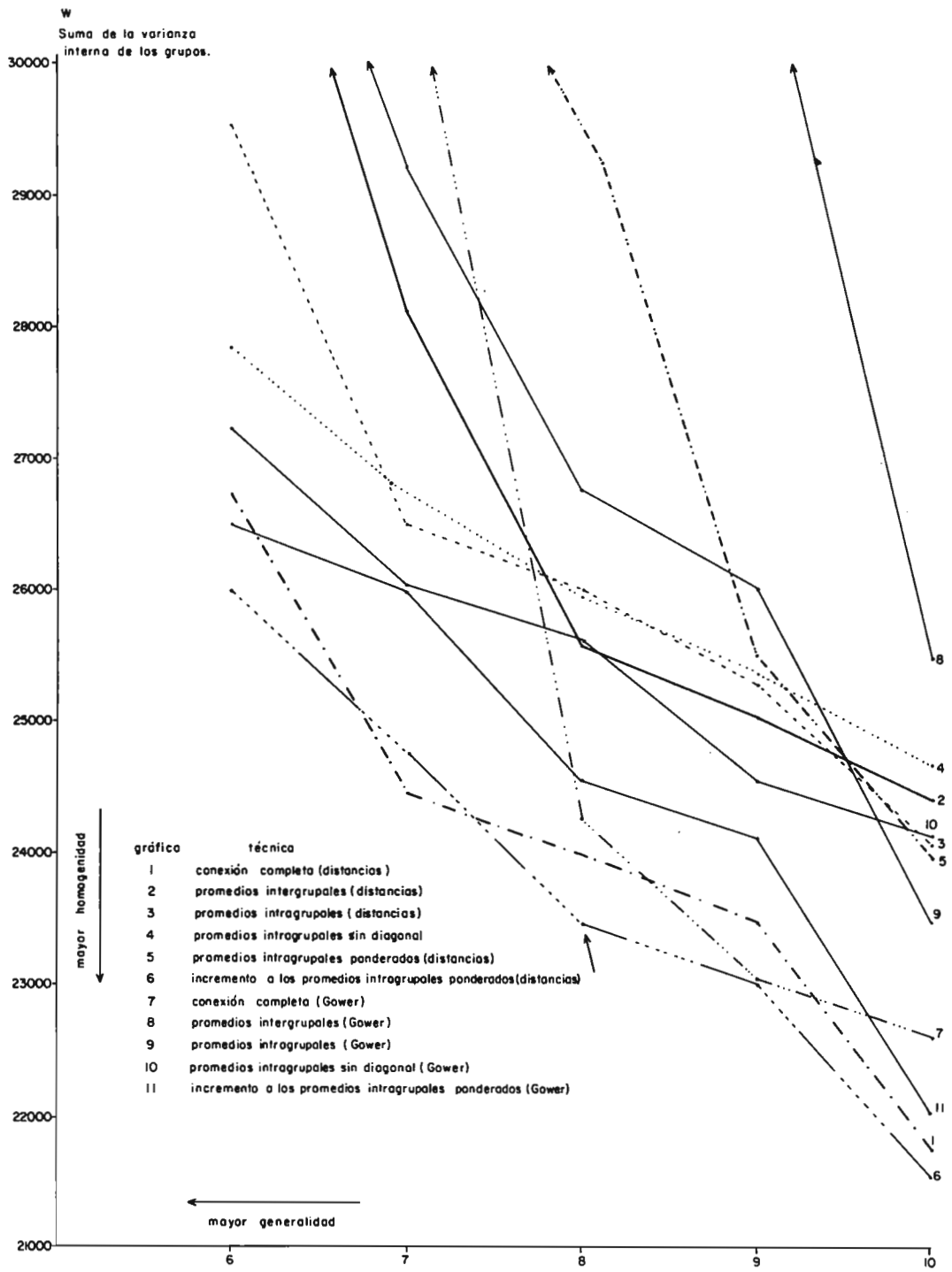


FIGURA 4. VALOR DE W (VARIANZA INTERNA) SEGUN DISTINTAS TECNICAS DE CLASIFICACION POR CUMULOS.

Además del Distrito Federal y los estados de México, Quintana Roo y Zacatecas que se comportan como regiones en sí mismos.

En la gráfica (figura 4) se verán valores de la W menores, pero a niveles de generalidad también menores. Se determinó la clasificación en el nivel de 8 grupos en el dendrograma porque es el punto donde se maximiza la generalidad (menos grupos) y se minimiza la W (mayor homogeneidad interna de los grupos formados).

Para algunos autores cada grupo constituye una región con independencia de su continuidad en el espacio; otros prefieren llamar tipos a los grupos y regiones a aquellos objetos que siendo del mismo tipo presentan continuidad espacial, de tal manera que distintas regiones pueden ser del mismo tipo.

El siguiente paso es la caracterización de cada tipo o región, para lo que se puede construir una tabla igual a la usada en el análisis de componentes principales (ver pág. 19). Con esa tabla se podrán identificar los valores límite para cada grupo en cada variable, lo que supone usar dos datos. Esto es útil, pero tiene el riesgo de que un objeto discrepante amplíe mucho el rango provocando la impresión de una variación alta. También resulta cómodo trabajar con los valores medios de una región, así un solo dato caracterizará el grupo; no obstante, pueden presentarse problemas en la representatividad de la media, ya que un mismo valor puede ser producto de distribuciones de datos muy distintos, por lo que es necesario utilizar una medida de dispersión y, en lo posible, el sesgo y la kurtosis, para evaluar la calidad de la media. Además, la dispersión de cada variable en los distintos grupos deberá compararse con la dispersión original de esa variable en el conjunto total de los datos. En la medida en que exista una reducción entre aquélla y ésta, se tendrá una ganancia de homogeneidad y la clasificación será eficaz en cuanto los valores medios regionales reflejan los objetos que forman la región. Lo más seguro es que dicha gananacia sólo se dé en algunas variables, respecto a las cuales la clasificación será confiable.

Existen dos técnicas multivariadas que también se pueden aplicar para la construcción de tipologías y regionalizaciones. El primero es K-promedios, que es un método de clasificación no jerárquico que determina cierto número de grupos homogéneos; el segundo es el escalamiento clásico multidimensional que, al igual que los componentes principales busca agrupar la mayor parte de la varianza en un número menor de dimensiones. La diferencia es que el poder de variación explicada por las primeras dimensiones del escalamiento es mayor que el de los primeros componentes, o sea, en dos o tres dimensiones se explica más en escalamiento que en componentes; sin embargo, no son interpretables en términos de las variables que se asocian a cada dimensión. Cuando se tiene la mayor parte de la variación explicada en dos dimensiones es posible graficar sobre ejes perpendiculares (uno para cada dimensión), de tal manera que los objetos que estén cercanos en la gráfica serán semejantes.

En el ejemplo se tuvo un 85.47% de variación en las dos primeras dimensiones. Para evaluar los resultados de las clasificaciones obtenidas se usó la gráfica del escalamiento multidimensional, sobre la cual se agruparon las entidades según los grupos formados por medio del análisis de componentes principales, con base en los dos primeros dígitos de la clave del tipo -dos primeras dimensiones- (figura 5); posteriormente se hizo lo mismo con los grupos resultantes del análisis de cúmulos (figura 6). Así se puede estimar la claridad de las clasificaciones obtenidas. En la medida que sea posible aislar o agrupar a los objetos de cada grupo sobre la gráfica del escalamiento, sin que se entrecrucen y, por tanto, sin que se pierda la

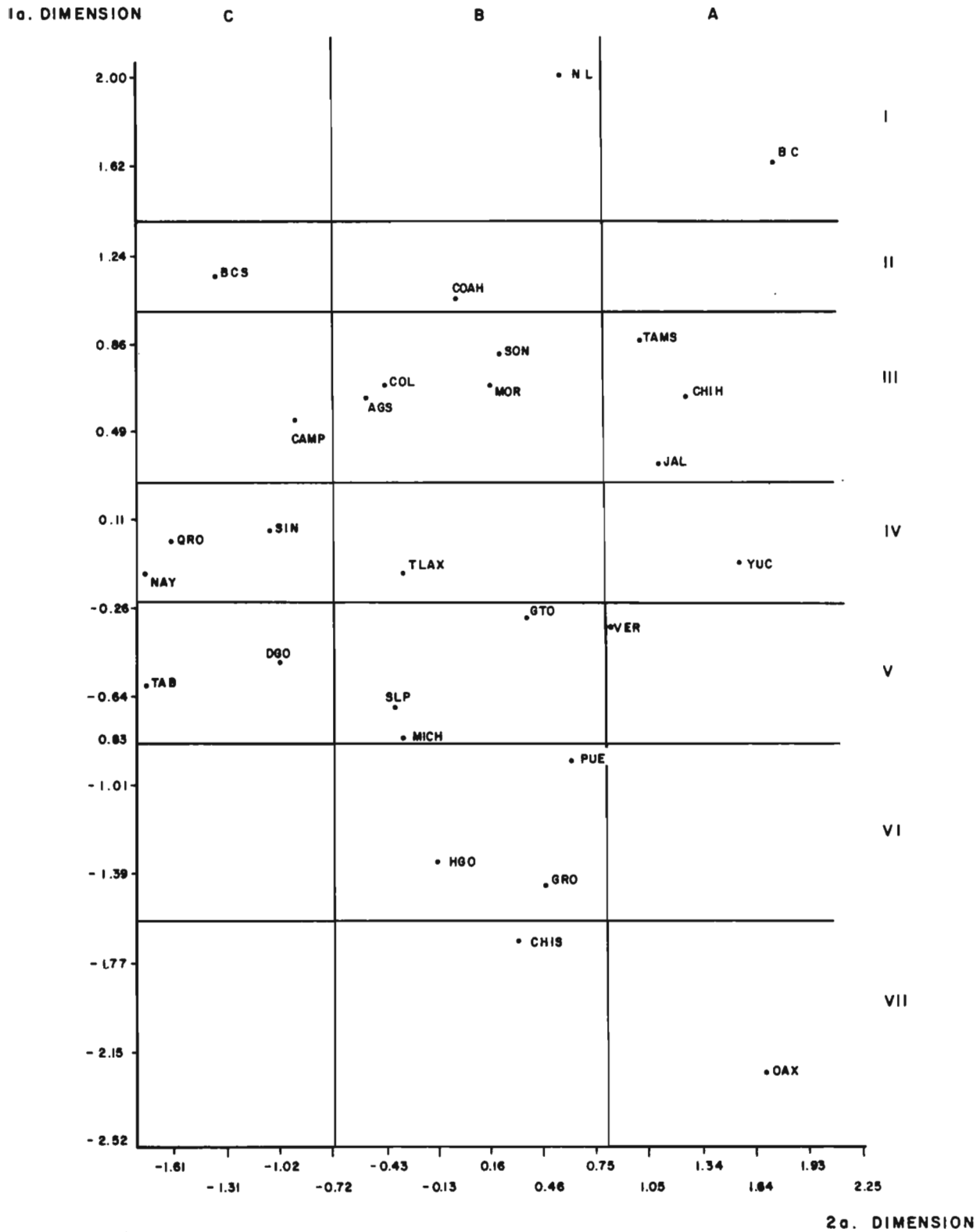
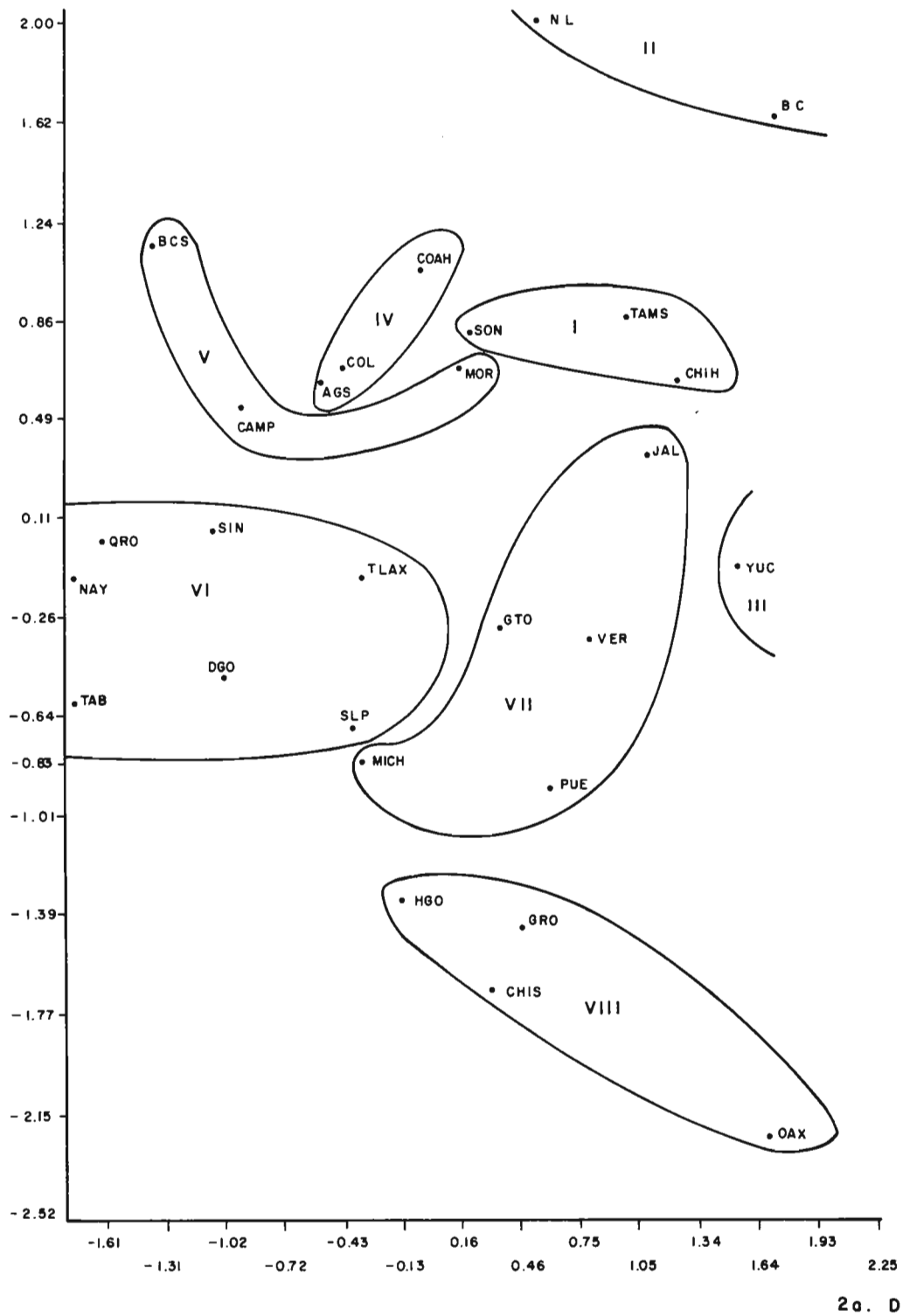


FIGURA 5. ESCALAMIENTO MULTIDIMENSIONAL COMPARADO CON LA CLASIFICACION DE COMPONENTES PRINCIPALES.

1a. DIMENSION



2a. DIMENSION

FIGURA 6. ESCALAMIENTO MULTIDIMENSIONAL COMPARADO CON LA CLASIFICACION DE CUMULOS.

continuidad del agrupamiento, se tiene una clasificación plausible.

En el caso de los componentes principales se verá que la agrupación se realiza claramente y que, además, existe coincidencia en las dimensiones, la primera componente coincide con el eje vertical del escalamiento y la segunda componente con el eje horizontal (figura 5), por lo que los ejes resultan interpretables en función de dicha coincidencia. En el análisis de cúmulos no se pueden considerar dimensiones como en el de componentes sino que se clasifica considerando el conjunto total de la variación, de aquí que los grupos resulten más "circulares" en la gráfica, y no son interpretables en función de las dimensiones ya mencionadas. No obstante, la definición de grupos también resultó clara, a excepción del estado de Morelos que aparentemente debería integrarse al grupo I o al IV; sin embargo, deben considerarse dos aspectos: en primer lugar, la pertenencia de Morelos a su grupo (V) puede darse por las dimensiones (por la variación de variables) no consideradas en la dos primeras dimensiones del análisis de componentes o del escalamiento; en segundo lugar, si la técnica de cúmulos se manejara a mayor generalidad, el grupo IV y V se integrarían.

Si bien hay ciertas diferencias entre las clasificaciones de componentes y cúmulos, ambas son válidas. Una razón son los distintos niveles con que se trabajaron las clasificaciones: mientras que con la técnica de componentes se obtuvieron 16 tipos o grupos, por medio de los cúmulos solamente se crearon 8 regiones o grupos (en ambos casos además de las cuatro entidades aisladas previamente). La segunda razón es que en el análisis de componentes no se consideraron todas las variables, en cambio en el análisis de cúmulos sí. No obstante, independientemente de las diferencias, los grupos que componen cada clasificación estarán descritos por un rango o por valores medios, quizá un conjunto de variables esté mejor representado en una de las clasificaciones y otro conjunto en la otra. Lo importante es determinar qué variables caracterizan adecuadamente a los grupos resultantes, en otras palabras, ¿en qué variables los grupos ganaron realmente en homogeneidad y en que variables se conserva la heterogeneidad original?, y esto se podrá evaluar comparando la dispersión (ver página 24).

CONCLUSIONES

1. La estadística en general, y las técnicas aquí revisadas, en particular, sólo son una herramienta en el trabajo del científico social.

2. Dicha herramienta sólo será útil si el problema de investigación supone el uso de alguna técnica -no todo los problemas suponen soluciones estadísticas-, si la técnica propuesta es la adecuada a ese problema y si la explicación e interpretación de los resultados de la técnica son los correctos.

3. Por tanto, el uso de técnicas estadísticas no son garantía de validez de ninguna investigación. En casos como éste, el papel de la estadística es básicamente descriptivo y no se debe esparar más que esto. La correcta interpretación y explicación de los fenómenos dependerá más de los recursos teóricos con que se cuente.

AGRADECIMIENTOS

Al Mtro. Mario Cortina B. por el procesamiento de la información, por la elaboración del programa para medir la varianza interna, y en general, por su amplia colaboración; al Sr. Carlos Jaso por la revisión de estilo, y a Lourdes M. Pérez Cardona por la mecanografía.

REFERENCIAS BIBLIOGRAFICAS

Anderberg, M. 1973. Cluster analysis for applications, Academic Press, London.

Espinoza, G. y A. López. 1980. Introducción a los métodos de clasificación jerárquica, IIMAS, UNAM, México.

González Casanova, Pablo. 1977. Las categorías del desarrollo económico y la investigación en ciencias sociales, UNAM, México.

Johnston, R.J. 1980. Multivariate statistical analysis in geography, Longman, London.

Kunz B., Ignacio. 1985. Regionalización sociodemográfica del estado de Guanajuato, Tesis de maestría, UNAM, México.

Marriot, F.H.C. 1974. Multivariate observations, Academic Press, London.

Orford, J.D. 1976. "Implementation of criteria for partitioning a dendrogram", Mathematica Geology 8, pp. 75-84.

Rendón, T. y Carlos Salas. 1986. "La población económicamente activa en el censo de 1980. Comentarios críticos y una propuesta de ajuste", Estudios Demográficos y Urbanos, Vol. 1, Núm. 2, pp. 291-309.

Reyes, L, G. Espinoza y A. López. 1980. ANALISIS/CUMULOS: un programa para clasificaciones jerárquicas, IIMAS, UNAM, México.

Secretaria de Programación y Presupuesto 1984, Anuario Estadístico de los Estados Unidos Mexicanos 1983, México.

Secretaria de Programación y Presupuesto. 1984. X Censo General de Población y Vivienda 1980, México.

El uso de la estadística para la construcción de clasificaciones y regionalizaciones, editado por la Dirección General de Publicaciones, se terminó de imprimir en la Imprenta Universitaria el 27 de julio de 1988.
La edición consta de 1000 ejemplares

El trabajo de esta publicación ha sido aprobado por la
Comisión Dictaminadora del Consejo Editorial
del Instituto de Geografía.

El contenido y forma de esta publicación
es responsabilidad del autor.

Para pedidos dirigirse a: Instituto de Geografía,
Ciudad Universitaria, 04510, México, D.F.

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO