Dr. John P. Cole

UNA INTRODUCCIÓN AL ESTUDIO DE MÉTODOS CUANTITATIVOS APLICABLES EN GEOGRAFÍA

INSTITUTO DE GEOGRAFÍA

Directora: Dra. Ma. Teresa Gutiérrez De MacGregor

Dr. John P. Cole

UNA INTRODUCCIÓN AL ESTUDIO DE MÉTODOS CUANTITATIVOS APLICABLES EN GEOGRAFÍA

Universidad Nacional Autónoma de México. *México, 1975*

Primera edición: 1975

Dirección General de Publicaciones

Impreso y hecho en México

Introducción

El uso de métodos cuantitativos en la geografía ha aumentado mucho desde el año de 1960. Es cada vez más importante que el geógrafo entienda los nuevos métodos que se aplican en la geografía, a fin de poder apreciar y criticar su uso en varios estudios y publicaciones, con el propósito de poder usarlos él mismo. Estos apuntes y ejercicios sobre geografía cuantitativa se publican para uso en los países de habla castellana. Muchos ejemplos se refieren a situaciones en México, pero pueden ilustrar la aplicación de métodos en cualquier ramo de la geografía y en cualquier parte del mundo.

El geógrafo recoge, presenta e interpreta datos acerca del mundo y de la distribución de fenómenos en su superficie. Los métodos cuantitativos se aplican en la recolección de información (por ejemplo, el uso de muestras), en la presentación de información (cartografía, proyecciones) y en la interpretación de los resultados (estadísticas diferenciales).

El mapa es una forma de representación de información en un contexto espacial. El mapa es un tipo o caso especial de gráfica, y su construcción comprende aspectos de geometría y trigonometría. Hasta ahora las matrices no se han usado mucho en la presentación y el procesamiento de datos geográficos. Su importancia será subrayada en esta publicación. Las matrices se manipulan con su propia álgebra. La computadora electrónica ya se aplica mucho en geografía. Es capaz de hacer cálculos aritméticos hasta cien mil o un millón de veces más rápido que una máquina calculadora convencional. Sin embargo, la computadora electrónica no tiene ningún valor para los geógrafos si no se cuenta con programas preparados específicamente para procesar sus datos y solucionar sus problemas. Con programas apropiados ofrece posibilidades enormes en el estudio de datos numéricos, lo que anteriormente no era posible.

En la ciencia, en general, y en muchos ramos de la geografía en particular, se buscan procesos, patrones que se repitan, generalidades, leyes que expliquen la realidad en forma concisa y precisa. En muchas situaciones conviene construir y estudiar modelos, simplificaciones de la realidad que ayudan a entender su funcionamiento. Muchos modelos usan conceptos, términos matemáticos y datos numéricos.

En el Instituto de Geografía, en octubre de 1972 se impartió un seminario intensivo, de un mes, sobre geografía cuantitativa, en el cual se trató de exponer las principales técnicas empleadas en geografía, sin que, por ello, se pretendiera abarcar completamente el campo, sobre todo en el caso de técnicas más complicadas tales como análisis de factores, modelos gravitatorios, análisis de sistemas, programación lineal, etcétera.

Se pensó, entonces, en hacer un texto introductorio, en español, de dichas técnicas, a fin de incorporar a ellas a los científicos sociales, y que sirviera de base en estudios más avanzados.

La publicación de este libro no hubiera sido posible sin el apoyo de la Dra. Ma. Teresa Gutiérrez de MacGregor, directora del Instituto de Geografía de la Universidad Nacional Autónoma de México. Agradezco al señor Carlos Jaso su asesoría en la redacción del trabajo, y a la doctora Silvana Levi de López la ayuda que prestó en su preparación.

Conceptos generales

1. LOS NÚMEROS EN LA GEOGRAFÍA

1. *Tipos de números*

Es importante reconocer varios tipos de números. Usamos la base 10 para representar varios tipos, pero los mismos existen con cualquier base.

Los números siguientes se usan para contar:

0, 1, 2, 3, 4, etc.

Se emplean para contar unidades u objetos distintos. Por ejemplo, es posible decir 50 ovejas o 22 caballos, pero no 50.3 ovejas o 22.91 caballos. Los números que se usan para contar son parte del sistema de números enteros.

Cada número positivo (3, 17) tiene su número negativo equivalente (−3, −17). Los números enteros son de tres tipos: positivos, negativos, y cero.

Es evidente, sin embargo, que hay otros números entre los enteros. Entre 3 y 4 hay 3.1, 3.15, 3.999 y muchos más (una infinidad). Los números con punto decimal se llaman números reales. Cada número entero tiene su equivalente real: 1 y 1.0, −23 y −23.0.

Hay dos tipos de números reales, el número racional y el número irracional. Es posible expresar un número racional, en forma de una razón o fracción, con dos números enteros: $\frac{1}{2}$, $\frac{2}{7}$, etc. En forma de decimales estas fracciones terminan o se repiten: $\frac{1}{2}$ es lo mismo que .500 y $\frac{2}{7}$ es .285714285714... El número π (razón de la circunferencia de un círculo con su diámetro) es un número irracional, como también $\sqrt{2}$ (la raíz cuadrada de dos).

En los datos geográficos raramente se usan más de 3 o 4 valores decimales. Es suficiente distinguir entre los números enteros y los números reales.

1.2 *Bases*

Estamos acostumbrados a un sistema de números que emplea diez dígitos (0, 1, 2, 3, 4, 5, 6, 7, 8, 9). En el sistema que usamos, la posición de un dígito en un numeral indica su valor o tamaño. En el número 575, por ejemplo, el primer dígito 5 vale cien veces más que el segundo dígito 5 (500 y 5). En realidad, el número 575 expresa en forma breve la cantidad siguiente: $5 \times 100 + 7 \times 10 + 5 \times 1$.

Se pueden expresar los números así:

1,000	100	10	1
(10^3)	(10^2)	(10^1)	(10^0)

$$5 \times 100 + 7 \times 10 + 5 \times 1$$

El dígito cero tiene una función especial

porque indica la ausencia de una cantidad. Por ejemplo, en el número 505, cero significa 0×10, la ausencia de 10.

Estamos tan acostumbrados al sistema de números con 10 dígitos o base 10, que nos es muy difícil concebir otro sistema. Sin embargo, es posible construir y aplicar sistemas con otros números dígitos. El sistema binario funciona exactamente como el sistema base 10, con solamente dos dígitos (UNO o 1, y CERO o 0). Tiene ciertas ventajas especiales. Sobre todo, es el único posible en las computadoras electrónicas. También ofrece un sistema numérico sencillo para la representación de información no cuantitativa; el símbolo 1 indica la presencia de un atributo, o *SÍ*, y el símbolo 0 (cero) indica la ausencia de un atributo, o *NO*.

Se necesita un fuerte salto mental para apreciar el funcionamiento del sistema binario. Representamos el número binario 10110 como ya representamos el número 575 base 10.

32	16	8	4	2	1	(base 10)
(2^5)	(2^4)	(2^3)	(2^2)	(2^1)	(2^0)	(base 10)
	1	0	1	1	0	(base 2)

$1 \times 16 \quad +0 \times 8 \quad +1 \times 4 \quad +1 \times 1 \quad +0 \times 1$

En el sistema binario el 1 indica presencia, el 0 indica ausencia.

El cuadro 1.2a indica algunos números en base 2 y sus equivalentes en base 10:

CUADRO 1.2a

8 4 2 1 (base 10) base 2	base 10
0 0 0 0	0
0 0 0 1	1
0 0 1 0	2
0 0 1 1	3
0 1 0 0	4
0 1 0 1	5
0 1 1 0	6
0 1 1 1	7
1 0 0 0	8
1 0 0 1	9
1 0 1 0	10

Ejercicio 1.2 Calcular lo siguiente (base 2). Ver la respuesta al final del capítulo.

(*i*)	(*ii*)	(*iii*)	(*iv*)
100	100	1111	1000
+ 10	+ 111	− 101	− 111
___	___	___	___

Hay otras bases, por ejemplo, base 8 (usando 8 dígitos; 0, 1, 2, 3, 4, 5, 6, 7), que también tiene aplicaciones en las computadoras.

Base 12 tiene dos símbolos más que base 10.

1.3 *Números grandes y aproximación*

a) Muchas veces se encuentran errores en la publicación de números grandes. Siempre hay que preguntarse si un número o el resultado de un cálculo es aproximadamente correcto. Por ejemplo, la superficie total de Brasil es de, más o menos, 8.000,000 km ^{2}, y su superficie en hectáreas es de cerca de 800.000,000. Sin embargo, ¿quién, leyendo que la superficie de Brasil es de 80.000,000 de hectáreas, no lo aceptaría? Los dos números son muy grandes. Muchas veces se pierde o se agrega un cero, y de vez en cuando más de uno.

b) La precisión superflua. Según el censo del año 1970, el Distrito Federal de México tenía 6.874,165 habitantes. Pero, ¿qué significa esta cifra? Hubo nacimientos y muertes el día del censo, de modo que no se puede dar un número exacto para todo el día. Además, es muy probable que el censo haya perdido cierta proporción de la población total. Se ha calculado que en el censo de Estados Unidos del año 1960, tres millones de personas no fueron enumeradas. Sería mejor decir que el Distrito Federal tenía 6.900,000 en el año 1970.

c) En ocasiones es útil expresar los números muy grandes en la forma siguiente:

1.000,000	es un millón o 10^6
50.000,000	5×10^7
3,700.000,000	3.7×10^9 (la población del mundo)
1.000,000.000,000	de dólares 10^{12} (la renta nacional de Estados Unidos en 1972).

Como Estados Unidos tiene aproximadamente 200.000,000 de habitantes o 2×10^8, la renta nacional por habitante era de 5,000 dólares por habitante al año (no 500 o 50,000).

1.4 *Matrices*

Con frecuencia es conveniente representar la información numérica en forma de una matriz. En general, se usan en la geografía matrices de dos dimensiones (con renglones y columnas). Sin embargo, es posible tener matrices de una dimensión o de más de dos dimensiones.

Una matriz de dos dimensiones tiene un determinado número de renglones y de columnas. El número de entradas de la matriz es igual al número de renglones multiplicado por el número de columnas. Cada entrada de la matriz debe contener solamente un valor numérico, aunque éste sea cero.

Una matriz con el mismo número de renglones que de columnas se denomina matriz cuadrada.

Cuando los valores en las entradas de la parte inferior izquierda reflejan los valores de la parte superior derecha, la matriz se denomina simétrica.

La matriz 1.4a contiene información sobre 5 ciudades de México. Para cada ciudad hay cuatro preguntas: ¿Tiene o no tiene la ciudad un atributo determinado? Por ejemplo, ¿tiene Guadalajara más de un millón de habitantes? La respuesta afirmativa se representa en la matriz con el número 1 (uno). A la pregunta ¿tiene Tijuana un puerto marítimo? la respuesta negativa se representa con el número 0 (cero).

□ *Atributos*

A 1) ¿Tiene la ciudad más de un millón de habitantes?
A 2) ¿Es un puerto?
A 3) ¿Es capital de la República?
A 4) ¿Tiene frontera con Estados Unidos?

MATRIZ 1.4a

	A1	*A2*	*A3*	*A4*
México	1	0	1	0
Guadalajara	1	0	0	0
Tijuana	0	0	0	1
Ciudad Juárez	0	0	0	1
Veracruz	0	1	0	0

El lector se dará cuenta de que los números de la matriz tienen la forma de los números binarios empleados en la sección 1.2. Cada ciudad tiene cierta combinación de uno y cero. En este ejemplo se encuentran 4 combinaciones, porque Tijuana y Ciudad Juárez tiene la misma combinación (0001). Pero hay 16 posibilidades (por ejemplo, 0000, 1111, etc.). El número de combinaciones es igual a 2^n, siendo n el número de atributos. Así, con 4 atributos tenemos 2^4 o $2 \times 2 \times 2 \times 2$ o 16.

Es posible transferir la información de la matriz anteriormente mencionada, a una matriz que establezca un índice de semejanza entre cada par de ciudades. Esta matriz tendría 5 renglones y 5 columnas. El método para el cálculo del índice de semejanza es el siguiente:

Se forma una matriz de 5 renglones por 5 columnas.

	Mexico	*Guadalajara*	*Tijuana*	*Ciudad Juárez*	*Veracruz*
México	4				
Guadalajara	3				
Tijuana	1				
Juárez	1				
Veracruz	1				

Se toman las ciudades por pares, y se les compara. Por ejemplo, México con Guadalajara:

1 0 1 0
1 0 0 0

tienen tres atributos iguales, *A1*, *A2* y *A4*. Son diferentes en *A3*.

Tienen, entonces, índice de semejanza de 3 con un máximo posible de semejanza completa de 4.

Otro ejemplo, México y Tijuana

1010
0001

tienen un índice de 1 porque poseen solamente el atributo *A2* en común (00).

En general, 1 y 1 o 0 y 0 indican semejanza, 1 y 0 o 0 y 1 indican diferencia.

Ejercicio 1.4 Completar la matriz.

Se puede dar a las características un valor según su importancia. Por ejemplo, sería posible dar doble importancia al atributo *A3* (capital nacional), contestando afirmativamente con 2 en vez de 1 y negativamente con 0, como antes. Esta operación da más peso a un atributo que a los demás.

Se puede afinar más la escala usando, por ejemplo, 0, 1 y 2 en vez de 0 y 1 o 0 y 2. Sería posible aplicar esta escala al primer atributo, clasificando las ciudades de la manera siguiente:

más de 5 millones	2 (México)
entre 1 millón y 5 millones	1 (Guadalajara)
menos de 1 millón	0 (las demás)

Es evidente que Tijuana y Ciudad Juárez son idénticas según los cuatro atributos o criterios usados. Sería posible comenzar una clasificación de las cinco ciudades poniendo en un grupo o una clase a estas dos ciudades, sin perder ninguna información. El par más semejante que sigue son México y Guadalajara, porque tienen un índice de 3. Sería posible unirlas en un grupo sin perder mucha información.

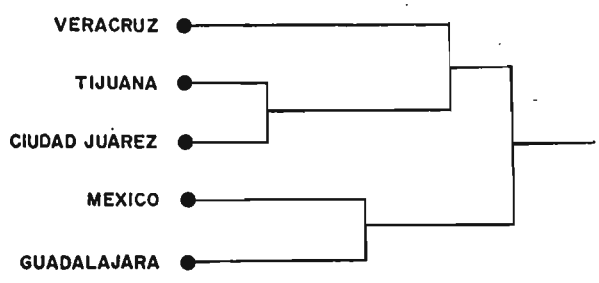


figura 1.4a

Entonces quedaríamos con 3 grupos de ciudades. Sería posible representar esta situación en forma gráfica (figura 1.4a).

Entre varias conclusiones que se pueden obtener analizando la matriz, se mencionan las siguientes:

1) Si dos ciudades son iguales entre sí, como Tijuana y Ciudad Juárez, cada una de ellas tiene, a su vez, un índice de semejanza igual con las demás.

2) Podemos calcular un índice de semejanza total entre cada ciudad y todas las demás sumando los valores en las columnas.

El resultado sería el siguiente:

México	10
Guadalajara	13
Tijuana	13
Ciudad Juárez	13
Veracruz	11

En este caso la ciudad de México es la menos típica del grupo. Si se toman en cuenta otros atributos el resultado podría ser diferente.

3) La matriz con índices de semejanza entre las ciudades es una matriz cuadrada porque tiene el mismo número de renglones que columnas.

4) Esta matriz es también simétrica porque los valores de las entradas de la parte

inferior izquierda reflejan los valores de la parte superior derecha.

1.5 *Gráficas*

La representación gráfica de la información es más conocida por el geógrafo, que el uso de las matrices. El mapa es un tipo particular de gráfica. Los dos ejes generalmente representan distancia hacia el este y distancia hacia el norte de algún punto determinado.

En la figura 1.5a se encuentran 5 lugares en un mapa. La distancia entre D y E, por ejemplo, es de 2 km. La distancia entre cualquier par de puntos en el mapa (o la gráfica) está relacionada con la diferencia de sus posiciones en el eje que representa la distancia al este y en el eje que representa la distancia al norte. Se puede calcular la distancia entre dos puntos, por ejemplo A y B, usando la fórmula siguiente:

$$D_{AB} = \sqrt{(A_E - B_E)^2 - (A_N - B_N)^2}$$

D_{AB} es la distancia en km entre A y B.

A_E y B_E son las distancias de A y B al este del punto de origen (0) de la gráfica.

A_N y B_N son sus distancias al norte.

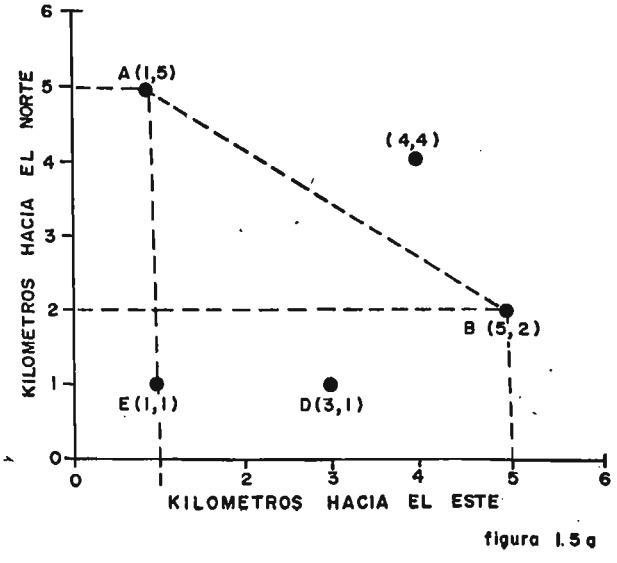


figura 1.5a

$$D_{AB} = \sqrt{(1-5)^2 - (5-2)^2} =$$

$$\sqrt{16+9} = \sqrt{25} = 5 \text{ km}$$

Para medir un número pequeño de distancias en un mapa, es suficiente usar una regla. Pero si se quisiera medir un número mayor de distancias, por ejemplo, entre cien puntos, tendrían que medirse 4,500 distancias diferentes. Para medir estas distancias sería más rápido elaborar un programa para una computadora electrónica, sobre la base solamente de 200 números; es decir, los valores este y norte de cada punto, por ejemplo, A (1, 5). El programa usaría la fórmula antes mencionada (basada en el teorema de Pitágoras), para hacer los cálculos.

Otra aplicación de la gráfica sería representar la relación entre dos o más columnas de datos numéricos. En el caso siguiente (cuadro 1.5a) se compara el porcentaje de población económicamente activa dedicada a la agricultura ($V1$) con el porcentaje de personas con ingresos menores de 500 pesos mensuales ($V2$) en 10 estados de la República Mexicana, en 1970.

Las columnas de datos se llaman variables.

CUADRO 1.5a

	V1	V2	V2(A)
1. Baja California	22 %	12 %	88 %
2. Coahuila	30	34	66
3. Chiapas	73	68	32
4. Distrito Federal	2	15	85
5. Guerrero	52	55	45
6. Nuevo León	17	23	77
7. Oaxaca	72	68	32
8. San Luis Potosí	53	58	42
9. Tlaxcala	55	58	42
10. Yucatán	55	65	35

Cada estado está representado en la gráfica por un punto. La posición de cada punto está determinada por su valor en variable 1 y en variable 2.

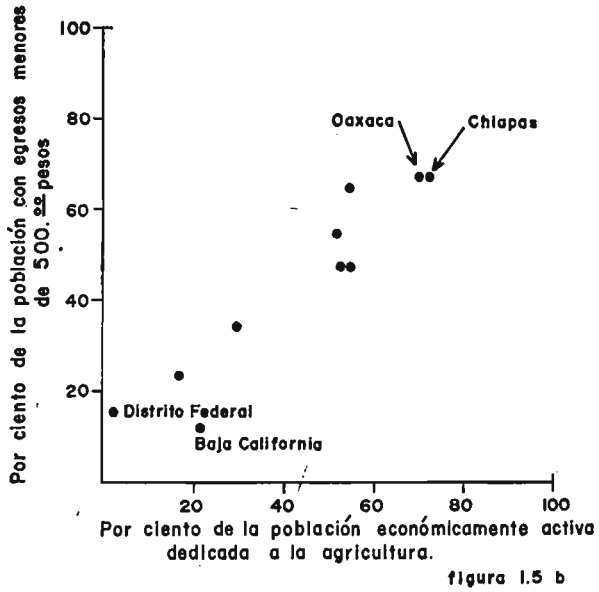


figura 1.5 b

En el caso del mapa, los ejes este y norte representan distancias en el espacio. En el caso de las dos variables, los ejes X ($V1$) y Y ($V2$) representan distancias entre los valores en las variables.

La distribución de los puntos en la situación ilustrada en la figura 1.5b, indica una fuerte correlación positiva entre las dos variables. Es posible calcular un índice numérico de correlación entre las variables (ver capítulo 5).

Si en vez de tomar en cuenta la población con ingreso bajo (menor de 500 pesos mensuales) se hubiera tomado en cuenta la población con ingresos mayores de 500 pesos mensuales, entonces tendríamos la variable $V2$ (A).

Ejercicio opcional: localizar los 10 estados en una gráfica, usando las variables $V1$ y $V2$ (A) (A = alternativa). La dirección de la distribución de los 10 puntos en la gráfica es diferente e indica una correlación negativa.

1.6 *Conclusiones*

1. Analice ahora una diferencia fundamental entre la gráfica y la matriz. Las escalas de la gráfica son continuas, de tal manera que en la superficie de la gráfica hay un número infinito de puntos. En el contexto geográfico es posible localizar con precisión en el mapa, o en la gráfica, cualquier fenómeno. En contraste, la matriz tiene un número finito de entradas, dependiendo del número de renglones y de columnas. En el contexto geográfico es posible poner una red con la forma de una matriz, sobre una superficie. Es posible escoger una red gruesa o fina, pero todos los lugares en cada cuadrado se encuentran en la misma posición.

CUADRO 1.6a

	A1	*A2*	*A3*
México	1	0	1
Guadalajara	1	0	0
Tijuana	0	0	0
Ciudad Juárez	0	0	0
Veracruz	0	1	0

2. Los ejes de una gráfica se intersectan en ángulos rectos (ortogonales). Por esta razón no es posible dibujar gráficas de más de tres dimensiones. Sin embargo, se pueden hacer cálculos matemáticos suponiendo que existen más de tres ejes que se intersectan en ángulos rectos. El cálculo explicado en la sección anterior, para medir la distancia entre dos puntos en una gráfica con dos ejes ortogonales, puede hacerse en tres dimensiones, aunque no sea posible representar la situación geométricamente.

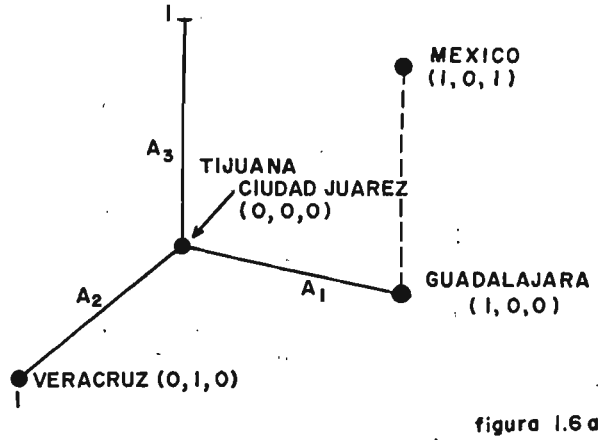


figura 1.6 a

Los datos del cuadro 1.6a pueden representarse tanto en forma de matriz como en forma gráfica. Según su índice (1 o 0) en la matriz, cada ciudad toma una posición en tres dimensiones (ver figura 1.6a). Sin embargo, no sería posible repetir esta operación geométrica o gráficamente en cuatro dimensiones, usando información para cuatro atributos, ya que no pueden representarse cuatro dimensiones. La matriz, entonces, es la forma capaz de contener información multivariada (con varias dimensiones).

3. Se pueden mencionar tres tipos de matrices aplicables en geografía.

a) Sector-región, que contiene información acerca de sectores u otros aspectos, sobre la base de divisiones regionales o territoriales.

b) Sector-sector, que contiene información sobre transacciones entre sectores de la economía.

c) Región-región, que contiene información sobre transacciones o distancias entre regiones.

Respuestas a los ejercicios.

EJERCICIO 1.2

(i) 10 (ii) 1011 ((iii) 1010 (iv) 1

EJERCICIO 1.4

	M	*G*	*T*	*C*	*V*
M	4	3	1	1	1
G	3	4	2	2	2
T	1	2	4	4	2
C	1	2	4	4	2
V	1	2	2	2	4

2. CONJUNTOS Y TOPOLOGÍA

2.1 *Los conjuntos*

La teoría de los conjuntos se considera como uno de los aspectos más importantes y básicos de la matemática moderna. En las publicaciones geográficas de algunos países se encuentran ya los conceptos y los símbolos de la teoría de los conjuntos. Por esta razón se incluye aquí una introducción breve a los símbolos y a las operaciones de dicha teoría.

Se mencionan también algunos ejemplos de su aplicación en la geografía. Como ilustración se usarán los datos en el cuadro 2.1a.

1. *Definición de un conjunto.* Es necesario definir con exactitud los miembros o elementos de un conjunto. Es posible hacer una lista de los elementos, representando cada uno de ellos con un símbolo apropiado, entre paréntesis especiales: { }.

CUADRO 2.1a (1 significa presencia)

	Iglesia	*Cine*	*Estación*	*Correo*	*Ayuntamiento*	*Garoje*	*Gasolinera*
A	1	0	0	1	0	0	0
B	1	0	1	0	1	0	0
C	1	0	0	1	0	0	0
D	1	0	0	1	1	1	1
E	1	0	0	0	0	0	0
F	1	0	0	1	1	1	1

En el cuadro 2.1a el conjunto de pueblos se representaría así: $\{A,B,C,D,D \text{ E } F\}$. Cuando el número de elementos es grande es preferible dar una definición exacta.

Se podría hablar, por ejemplo, del conjunto de los números enteros entre 0 y 100: 1, 2, . . . 99 o del conjunto de las ciudades de México con más de 100,000 habitantes en el año 1970.

Para no repetir la lista o la definición de un conjunto, muchas veces es conveniente referirse al conjunto con alguna letra o nombre. El conjunto de pueblos del cuadro 2.1a podría denominarse conjunto *P*.

2. *C* E *P* significa que el pueblo *C* es un elemento del conjunto *P*.

3. Cuando se está estudiando algún conjunto, se denomina *conjunto universal* a

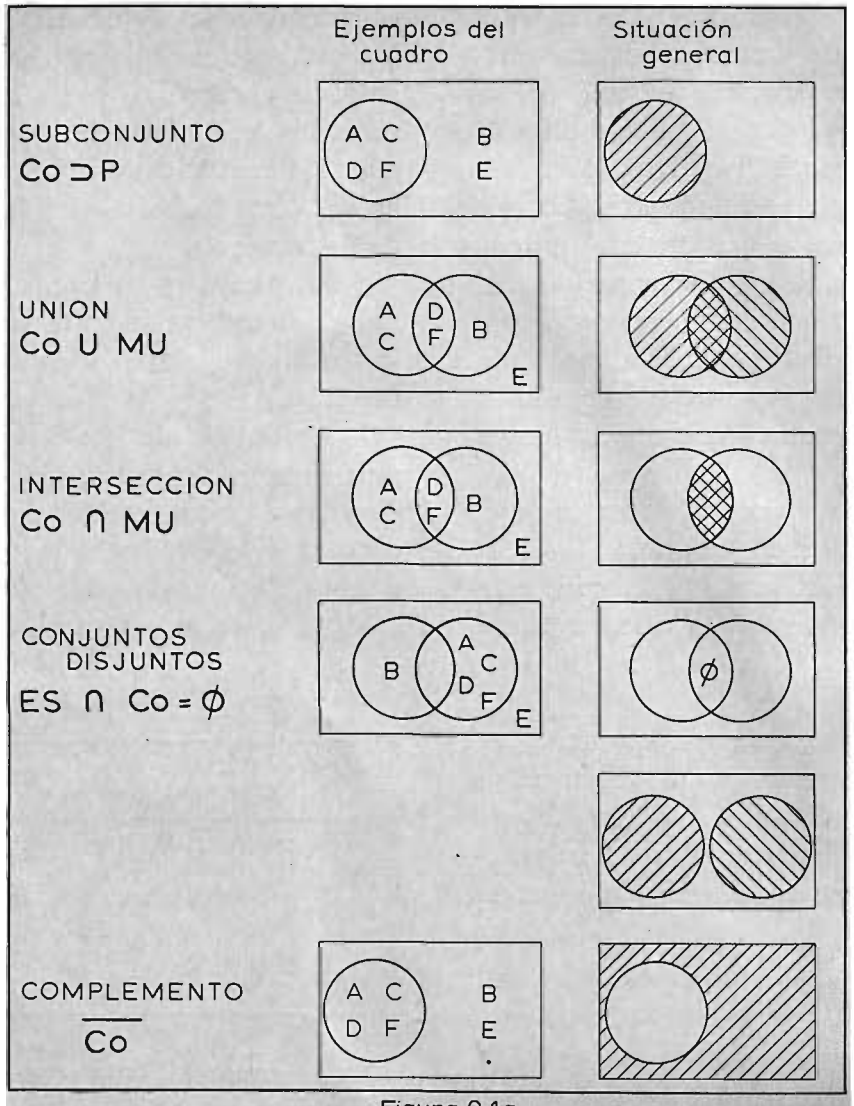


Figura 2 1a
(ver texto para una explicación)

éste o a la totalidad de elementos que se incluyen en él. El conjunto universal se representa por el símbolo *U* o *E*. En el cuadro 2.1a, los 6 pueblos forman el conjunto universal.

4. Cuando un conjunto no tiene elementos se le denomina conjunto vacío y se le representa por el símbolo ∅ o el símbolo { }.

En el cuadro, el conjunto de pueblos sin cine ejemplifica este tipo de conjunto.

Las situaciones y las operaciones de la teoría de los conjuntos se pueden representar por medio de los diagramas *Venn* (ver figura 2.1a).

5. *Subconjunto.* Un subconjunto, cuyo símbolo es ⊃, incluye una parte de los elementos del conjunto universal. Todos los elementos del subconjunto deben pertenecer al conjunto. Un caso particular de subconjunto sería cuando éste contiene los mismos elementos que el conjunto universal. El subconjunto de pueblos con iglesia contiene los seis elementos del conjunto universal.

En la figura 2.1a el subconjunto de pueblos con correo (*Co*), esto es, *A*, *C*, *D*, *F*, se

incluye en un círculo dentro del rectángulo que forma el límite de conjunto universal (*U*). Los pueblos sin correo se encuentran fuera del círculo.

6. *Unión.* La unión de dos * conjuntos incluye los elementos de ambos. La unión se representa en la figura 2.1a con el caso de los pueblos con correo (*Co*) o los pueblos con ayuntamiento (*Mu*), o con los dos atributos. El símbolo *U* significa la palabra *O*.

7. *Intersección.* La intersección de dos conjuntos incluye solamente los elementos que pertenecen a ambos conjuntos. La intersección se representa en la figura 2.1a con el caso de los pueblos con correo y con municipalidad (solamente *D* y *F*). El símbolo significa la palabra y. La unión es más amplia que la intersección.

8. *Conjuntos disjuntos.* Dos conjuntos son disjuntos si no tienen ningún elemento en común. Dos conjuntos disjuntos se distinguen por el símbolo ∅ (conjunto vacío), porque su intersección no tiene elementos. Por ejemplo, el subconjunto de pueblos con estación (pueblo *B*) no incluye ningún elemento del subconjunto de pueblos con correo (*A*, *C*, *D*, *F*), de manera que su intersección forma un subconjunto vacío, ∅.

9. *Complemento.* El complemento de un subconjunto incluye todos los elementos del conjunto universal que no están en el subconjunto especificado. Por ejemplo, el complemento del subconjunto de pueblos con correo (*Co*) incluye *B* y *E*. El complemento significa NO. El símbolo de NO es una raya colocada sobre la letra o las letras que denominan el subconjunto (por ejemplo $\overline{CO}$ o, también, $C\acute{O}$).

10. *Igualdad de conjuntos.* Dos conjuntos son iguales si tienen los mismos elementos, por ejemplo, {1, 3, 2 = 1, 2, 3}.

* Es posible considerar más de dos conjuntos al mismo tiempo, pero su representación por intermedio de diagramas Venn es difícil con más de 4 conjuntos.

11. *Equivalencia.* Dos conjuntos son equivalentes si hay correspondencia, uno a uno, entre todos sus elementos. El conjunto que forman los estados de México tiene una correspondencia uno a uno (símbolo <=>) con el conjunto formado por sus respectivas capitales.

EJERCICIOS

1-1	1-2	1-3	1-4	1-5	1-6
11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

Matriz 2.1a

La matriz 2.1a tiene todos los resultados posibles de combinaciones de un juego con dos dados. El conjunto universal tiene 36 elementos. El primer número en cada entrada de la matriz representa el resultado en el primer dado y el segundo representa el resultado en el segundo dado. Para entender mejor algunos aspectos de la teoría de los conjuntos, conteste a las preguntas siguientes:

1. ¿Cuáles son los elementos del *subconjunto* de resultados que suman 7?

2. ¿Cuáles son los elementos de la *unión* de resultados con 1 en el primer dado, o en el segundo dado (o en los dos)?

3. ¿Cuáles son los elementos de la *intersección* de resultados con 1 en el primer dado y en el segundo dado?

4. ¿Qué relación tiene el *subconjunto* con resultados que suman 7 con el *subconjunto* con resultados que suman 5?

5. ¿Cuántos elementos tiene el *complemento* del subconjunto de resultados que suman 7?

(Ver respuestas al final del capítulo).

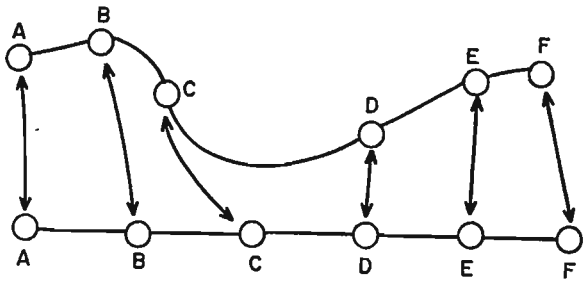


figura 2.2 a

2.2 *La topología*

Desde el siglo XIX se ha reconocido en la matemática, que es posible tener varios tipos de geometría, no solamente la euclidiana. La topología es una geometría menos exigente que la de Euclides.

En la topología no se conservan la distancia, la línea recta, la dirección y la orientación. En las transformaciones es necesario, sin embargo, conservar el orden y contigüidad.

En la figura 2.2a, el mapa del ferrocarril *A-F* no es equivalente, en la geometría euclidiana, a la línea recta *A-F*. La dirección del ferrocarril y las distancias entre las estaciones se han cambiado. En la topología, las dos líneas son equivalentes porque la transformación topológica que permite pasar de una a otra conserva el orden de las estaciones.

Las transformaciones topológicas de este tipo se usan mucho en la representación de redes de transporte, cuando se quiere representar claramente cierta información. El diagrama o mapa topológico del sistema del metro de México es un ejemplo. Cuando uno viaja en una línea del metro no importa la distancia entre pares de estaciones, sino el número de estaciones (con sus nombres) entre el principio y el fin del viaje.

De la misma manera, las tres figuras de la figura 2.2b son equivalentes del punto de vista topológico porque el orden *A B C D* se conserva en los tres casos; *X* queda dentro de la región, *Y* fuera, y la *intersección* de la línea *X Y*, con el límite de la región, se encuentra entre *B* y *C*.

Un aspecto de la topología, de interés geográfico, es el estudio de redes o el ramo que se llama teoría de las gráficas, el que no debe confundirse con las gráficas usadas en la sección 1.5.

El estudio de las redes en la topología usa los términos siguientes (ver figura 2.2c):

1. *NODO*: un punto (en la geografía podría ser un cruce de caminos, o una ciudad en la red de carreteras de un país). No tiene dimensiones.

2. *RAMO*: una línea que une dos NODOS (en la geografía podría ser una carretera o un río tributario, entre su fuente y su unión con otro río). Tiene una dimensión.

3. REGIÓN: una superficie limitada por ramos. Tiene dos dimensiones.

4. SENDERO: una ruta o un viaje entre dos NODOS en una red (por ejemplo, *A* y *C*).

5. *DIÁMETRO*: el sendero más largo que cruza la red; en este caso la red tiene un diámetro de 3 (*D-C-B-A*).

Las redes que se estudian en geografía pueden ser muy complejas, así sean redes

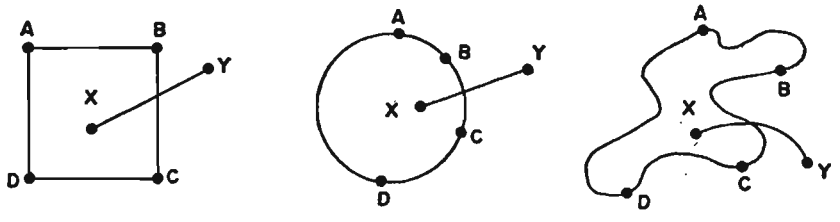


figura 2.2 b

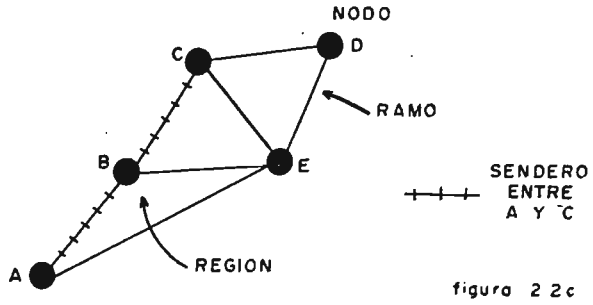


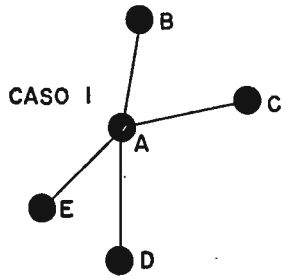
figura 2.2c

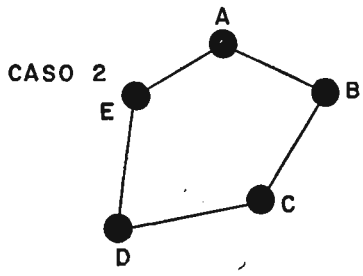
hidrológicas, de carreteras o de límites administrativos. Es útil sacar la información esencial del mapa convencional y representarla en forma topológica. Después se puede transferir la información de la red topológica a una matriz, de tal manera que se pueda estudiar numéricamente la información.

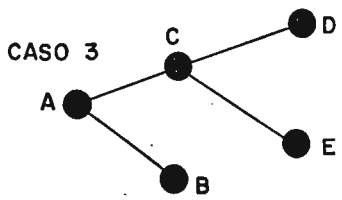
Los ejemplos de las figuras 2.2d ilustran el método de transferencia de información topológica (en forma de redes) a matrices.

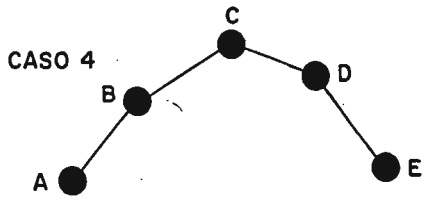
El caso 1 se representa en la matriz, el número 1 indica que entre dos NODOS existe un ramo que no pasa entre ningún otro nodo. Por ejemplo, *A-B* no pasa por otro nodo, pero *B-C* pasa por *A*. Ejercicio: represente en forma de matriz los casos 2, 3, 4 y 5.

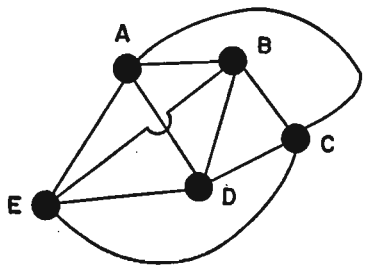
Los ejemplos de redes, en los casos 1-5 ilustran situaciones diferentes. En el primer caso la red está dominada por el











CASO 1

	A	B	C	D	E
A	0	1	1	1	1
B	1	0	0	0	0
C	1	0	0	0	0
D	1	0	0	0	0
E	1	0	0	0	0

figura 2.2 d

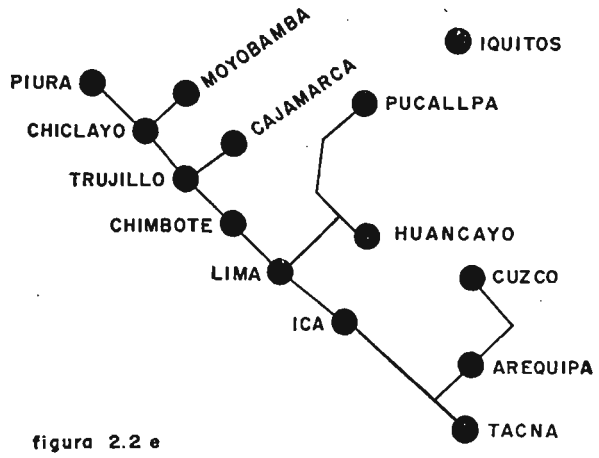


figura 2.2 e

nodo *A*. En el caso 2 cada nodo está vinculado con dos nodos. El caso 3 ejemplifica un tipo de red, el 'árbol'. Las redes hidrológicas que representan jerarquías administrativas son de este tipo. El número de ramos es igual al número de nodos menos 1. El caso 5 es una red completa. Se puede pasar entre cualquier par de nodos sin pasar por otro nodo. La matriz que representa la situación en el caso 5 contiene solamente valores de 1, con excepción de las entradas que representan el vínculo entre cada punto y él mismo. También en esta diagonal de la matriz sería justificable poner 1 en vez de 0, porque cada lugar tiene vínculos internos. El otro extremo sería una matriz exclusivamente con valores de 0. Representaría una situación con varios nodos, pero sin ramos que unieran a los nodos.

La figura 2.2e representa en forma muy sencilla la red de carreteras, y los nodos las ciudades principales de Perú. La capital, Lima, tiene una posición dominante en el sistema de carreteras. Iquitos, centro importante de la parte interior del país, no está vinculada con el resto del país.

La figura 2.2f ilustra dos redes del mismo tipo que la de las carreteras del Perú. En el primer caso, que corresponde a las rutas aéreas internacionales de AVIANCA (Colombia), es natural tener una red con varios ramas y sin 'regiones' importantes. En el caso de la red de comunicaciones de México se nota la posición dominante de México, la capital, y la falta de comunicaciones directas entre muchas ciudades importantes.

Es posible incluir en una matriz varios tipos de información acerca de la relación entre pares de nodos. La matriz que sigue incluye el número de ramos entre cada par de nodos en la red, caso 3, figura 2.2d.

	A	*B*	*C*	*D*	*E*
A	0	1	1	2	2
B	1	0	2	3	3
C	1	2	0	1	1
D	2	3	1	0	2
E	2	3	1	2	0

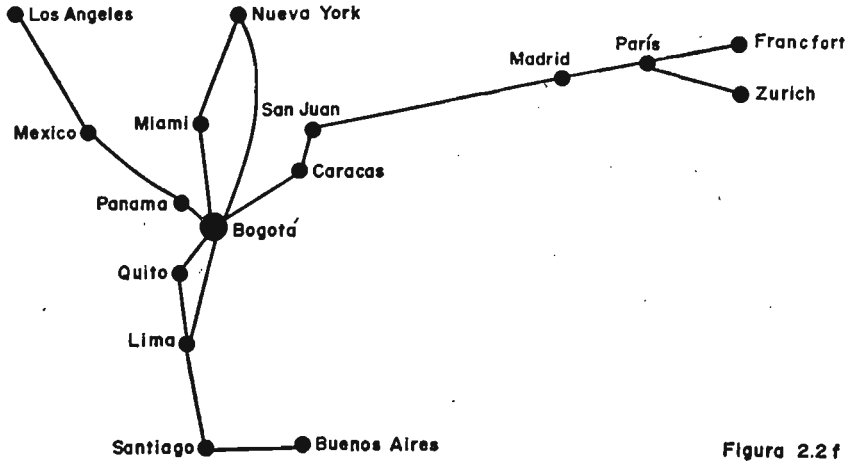
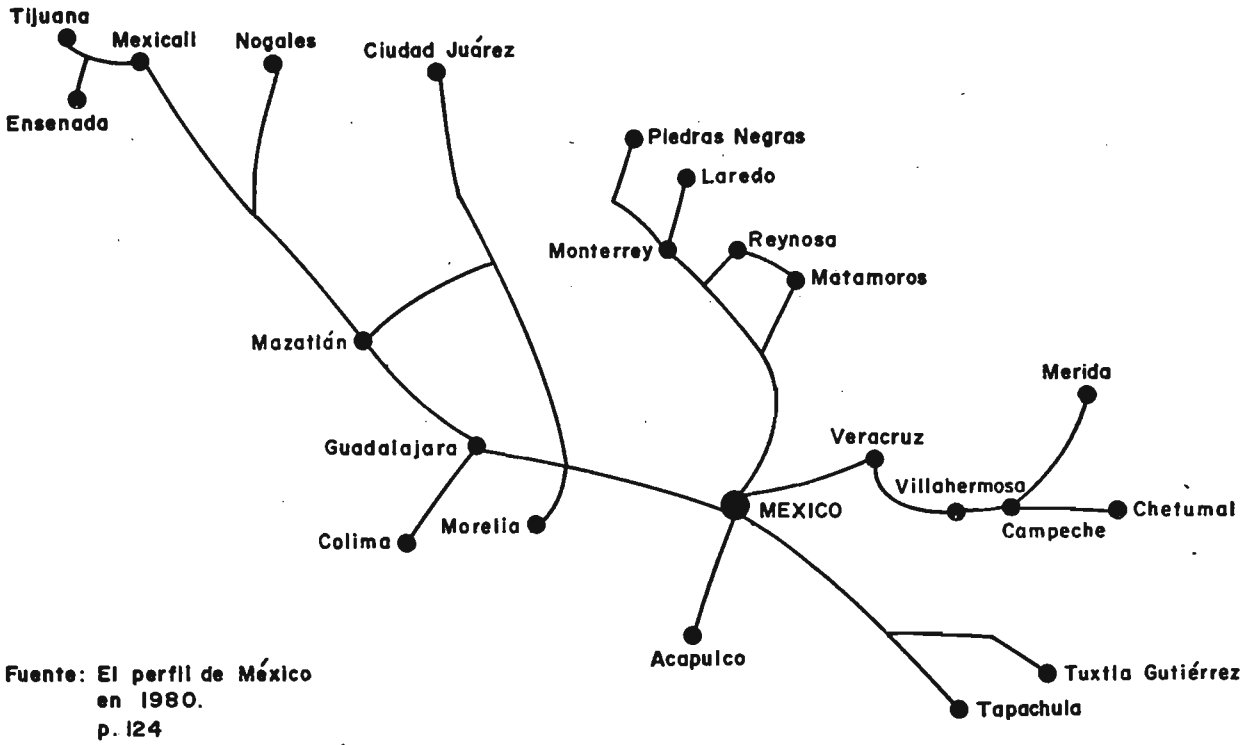


Figura 2.2 f



Fuente: El perfil de México en 1980. p. 124

figura 2.2 f (continuación)

Otra aplicación de la topología en la geografía es la construcción de mapas en los que el área es proporcional a la población o a otra variable y no a la superficie territorial. La construcción de este tipo de mapas es subjetiva, por lo que es posible obtener resultados diferentes. Es importante conservar la contigüidad.

La figura 2.2g muestra el método de construcción de un mapa topológico de las zonas administrativas de un país pequeño, con 5 divisiones.

La figura 2.2h muestra una transformación topológica de las entidades de México. Cada entidad está en proporción con el número de habitantes que tenía en el año 1960. En particular, se nota el gran tamaño del Distrito Federal y lo pequeño de los dos territorios.

POBLACION EN MILES
A, 100 , B, 100, C, 200 , D, 200, E, 400

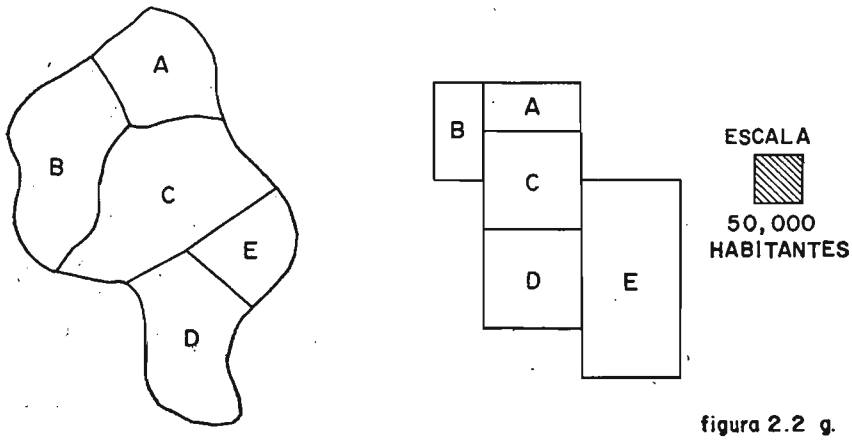


figura 2.2 g.

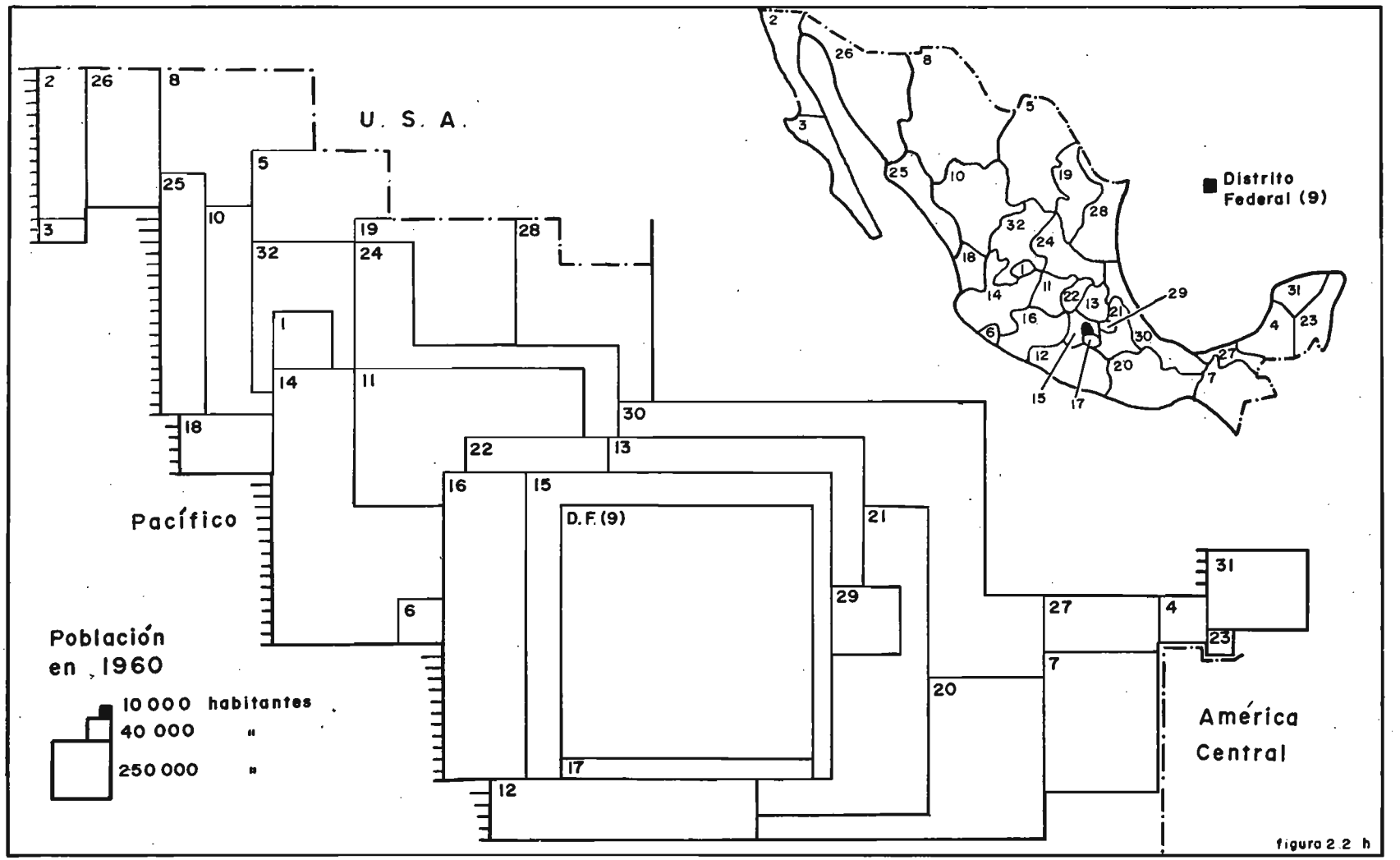

U. S. A.
Pacífico
D. F. (9)
América Central
Población en 1960
10 000 habitantes
40 000 "
250 000 "
Distrito Federal (9)

figura 2.2 h

Respuestas al ejercicio de la sección 2.1

1. 61, 52, 43, 34, 25, 16
2. 11, 21, 31, 41, 51, 61, 12, 13, 14, 15, 16,
3. 11
4. Son conjuntos disjuntos (sin elementos en su intersección).
5. 30.

Respuestas al ejercicio de la sección 2.2

CASO 2

	A	*B*	*C*	*D*	*E*
A	0	1	0	0	1
B	1	0	1	0	0
C	0	1	0	1	0
D	0	0	1	0	1
E	1	0	0	1	0

CASO 3

	A	*B*	*C*	*D*	*E*
A	0	1	1	0	0
B	1	0	0	0	0
C	1	0	0	1	1
D	0	0	1	0	0
E	0	0	1	0	0

CASO 4

	A	*B*	*C*	*D*	*E*
A	0	1	0	0	0
B	1	0	1	0	0
C	0	1	0	1	0
D	0	0	1	0	1
E	0	0	0	1	0

CASO 5

	A	*B*	*C*	*D*	*E*
A	0	1	1	1	1
B	1	0	1	1	1
C	1	1	0	1	1
D	1	1	1	0	1
E	1	1	1	1	0

3. APLICACIONES DE MÉTODOS NUMÉRICOS

3.1 *La curva Lorenz*

La curva Lorenz se usa con frecuencia en las ciencias sociales, para estudiar distribuciones. Da un índice numérico de desigualdad que permite comparar con precisión varias distribuciones. Los apuntes que siguen muestran con ejemplos los pasos del cálculo de los valores de las coordenadas necesarias para la representación gráfica de la curva.

En muchos países hay una distribución muy desigual (o muy poco equitativa) de ingresos entre los habitantes. En un país ficticio los ingresos se distribuyen de la manera siguiente:

Grupo	*Número de personas*	*Ingresos en dólares*	*Ingresos por habitante*
I	20,000	60.000,000	3,000
II	40,000	20.000,000	500
III	40,000	10.000,000	250
IV	100,000	10.000,000	100

Paso 1: Convertir los valores de las primeras dos columnas en porcentajes del total (200,000 habitantes y 100 millones de dólares):

Población (%)	*Ingresos* (%)
10	60
20	20
20	10
50	10

Es evidente que 10 % de la población recibe 60 % de los ingresos, mientras que, al otro extremo, 50 % recibe 10 %.

Paso 2: Sumar acumulativamente los valores que representan porcentajes:

10	60
30	80
50	90
100	100

Paso 3: Representar gráficamente, de la manera indicada en la figura 3.1a, con puntos, las posiciones de cada par de valores del paso 2, con ingresos en el eje horizontal y población en el eje vertical.

Paso 4: El coeficiente GINI es el índice que da numéricamente el nivel de desigualdad de la distribución. Hay que calcular el área del triángulo *ABC*,* y el área entre la diagonal *AB* y la curva que representa la distribución de ingresos de la población. Se divide el segundo valor entre el primero. El resultado siempre se encuentra entre 0 y 1. Un índice de 0 indica una distribución

* En este caso el área del triángulo ABC es 5 000 unidades que es la mitad del área total de la gráfica, la cual tiene dos lados de 100 unidades o porcentajes de cada uno.

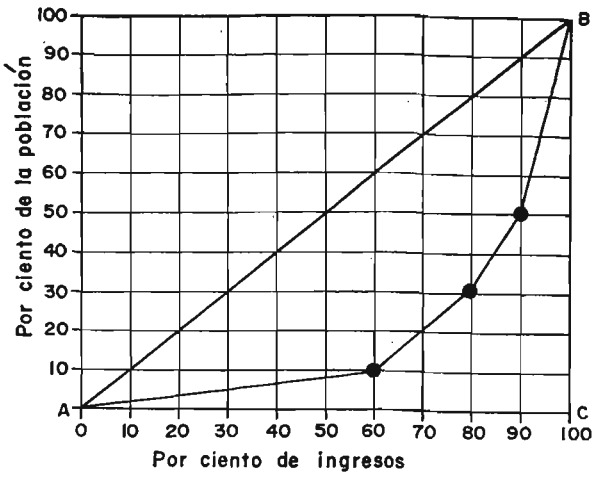


figura 3.1a

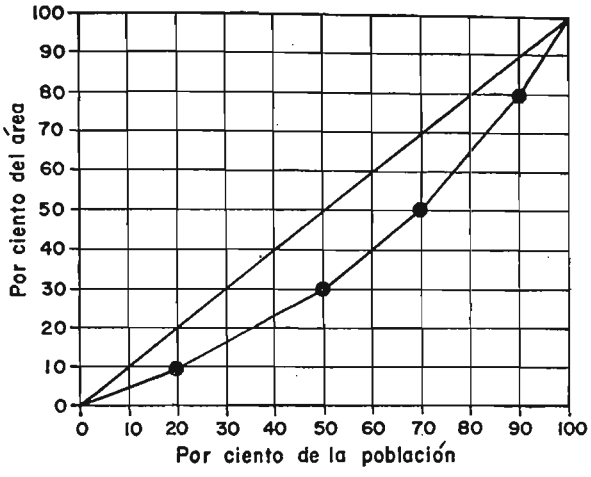


figura 3.1 b

completamente igual (20 % de la población con 20 % de los ingresos, 40 % de la población con 40 % de los ingresos, etc.). Un índice de 1 indica una concentración completa en el punto *C*. En el ejemplo, el área entre la diagonal y la curva ocupa aproximadamente 3,000 unidades de superficie. El coeficiente GINI viene a ser 0.6 (3,000/5,000).

Ejercicio 3.1 Con los datos siguientes, calcule y dibuje la curva de LORENZ de la distribución de población por área, en una región de 5 entidades territoriales, *A*, *B*, *C*, *D* y *E*.

	Población	*Área*	*Densidad*
A	200,000	5,000	40
B	300,000	10,000	30
C	100,000	10,000	10
D	200,000	10,000	20
E	200,000	15,000	13

Antes de comenzar el primer paso es necesario poner en orden descendente las entidades según su densidad de población. En este ejemplo *C* tiene que ser la última porque su densidad es menor que la de *D* y *E*.

En este ejemplo la escala que representa población ocupa el eje horizontal y la escala que representa área ocupa el eje vertical. Ver figura 3.1b.

La curva tiene muchas aplicaciones en geografía porque da un índice sucinto y preciso de las distribuciones. El segundo ejemplo ilustra su aplicación a una situación espacial, distribución de población sobre área. Sería posible comparar con varias curvas, en la misma gráfica, la distribución de la población en varios países (por ejemplo, Francia, España, Brasil y México). Sirve también para comparar la misma distribución (por ejemplo, la de la población de Venezuela) en varios años (por ejemplo, cada 10 años), para averiguar si se ha concentrado más en los últimos decenios por el crecimiento rápido de la población de ciertos estados.

El cálculo de la posición de la curva de LORENZ y de su coeficiente GINI necesita mucho tiempo cuando se consideran como 20 o 30 unidades en varios casos. Es posible hacer todos los cálculos con una computadora electrónica, dándole solamente la información inicial, como, por ejemplo, la población total y el área de cada entidad, en el segundo caso.

3.2 *Localización de una fábrica*

El método que sigue ilustra el uso de datos numéricos en una sencilla situación de localización. En un país ficticio hay

ocho ciudades. Las ciudades están conectadas por ferrocarril de la manera indicada en la figura 3.2a. Cerca de la ciudad *A* hay un depósito de carbón. Cerca de la ciudad *H* hay un depósito de mineral de hierro. El gobierno del país quiere establecer una fábrica de acero. La fábrica tiene que construirse en una de las ocho ciudades. Todo el acero se consumirá en la ciudad *E* (el mercado). El costo de transporte de una tonelada de carbón, hierro o acero por el sistema de ferrocarriles, está indicado en el mapa. En cada ramo de la red cuesta 1 dólar. Una tonelada de acero necesita dos toneladas de mineral de hierro y una tonelada de carbón. Tomando en cuenta estas cantidades y el costo de transporte de mineral de hierro y de carbón a la fábrica, y de acero de la fábrica al mercado, ¿cuál sería la posición con menores costos de transporte?

El cuadro 3.2a contiene toda la información necesaria para escoger la ciudad con costos de transporte mínimos. En la producción de una tonelada de acero hay que comparar los costos de transporte entre las ocho ciudades, las fuentes de carbón y hierro, y el mercado. Como la cantidad de hierro que precisa la producción del acero es dos veces más que la cantidad de carbón o de acero, los costos de transporte del hierro llegan a ser el doble de los otros costos de transporte.

CUADRO 3.2a

Ciudad	*Hierro*	*Carbón*	*Acero*	*Total*
A	8	0	4	12
B	6	1	3	10
C	6	2	2	10
D	8	3	1	12
E	10	4	0	14
F	4	2	3	9
G	2	3	4	9
H	0	4	5	9

Es evidente que, en esta situación, *F*, *G* y *H* tienen los costos más bajos, y que sería necesario usar otros criterios para decidir cuál sería la ciudad privilegiada.

Ejercicio 3.2 Calcule los costos de transporte y busque la ciudad o las ciudades con costos más bajos, considerando que del acero producido en la fábrica se destinará la mitad al mercado en *E* y la otra mitad en *G*. En el cálculo hay que mover ½ tonelada de acero a cada uno de dos mercados.

En realidad, las soluciones de los problemas de localización en geografía necesitan mucha más información que la usada en este ejemplo. Sin embargo, el ejemplo muestra la ventaja del uso de cálculos numéricos y de la aplicación de un método

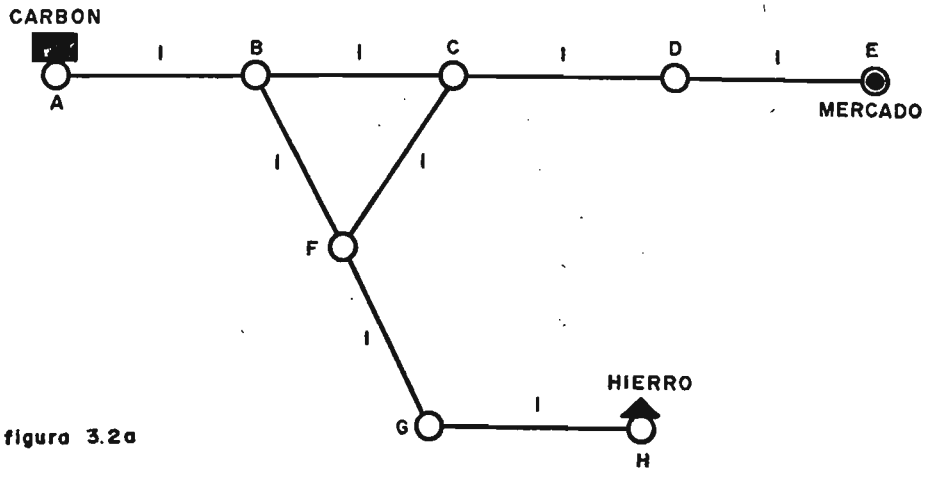


figura 3.2a

sistemático que tome en cuenta todas las posibilidades y llegue con certeza a la solución óptima.

3.3 *Tipos de distancia en geografía*

En la sección anterior, las distancias entre las ocho ciudades se calcularon según el *costo* de transporte de cargas. En geografía es importante reconocer varios tipos de distancia.

1. Distancia directa, línea recta entre dos lugares o un círculo grande en la superficie del globo.

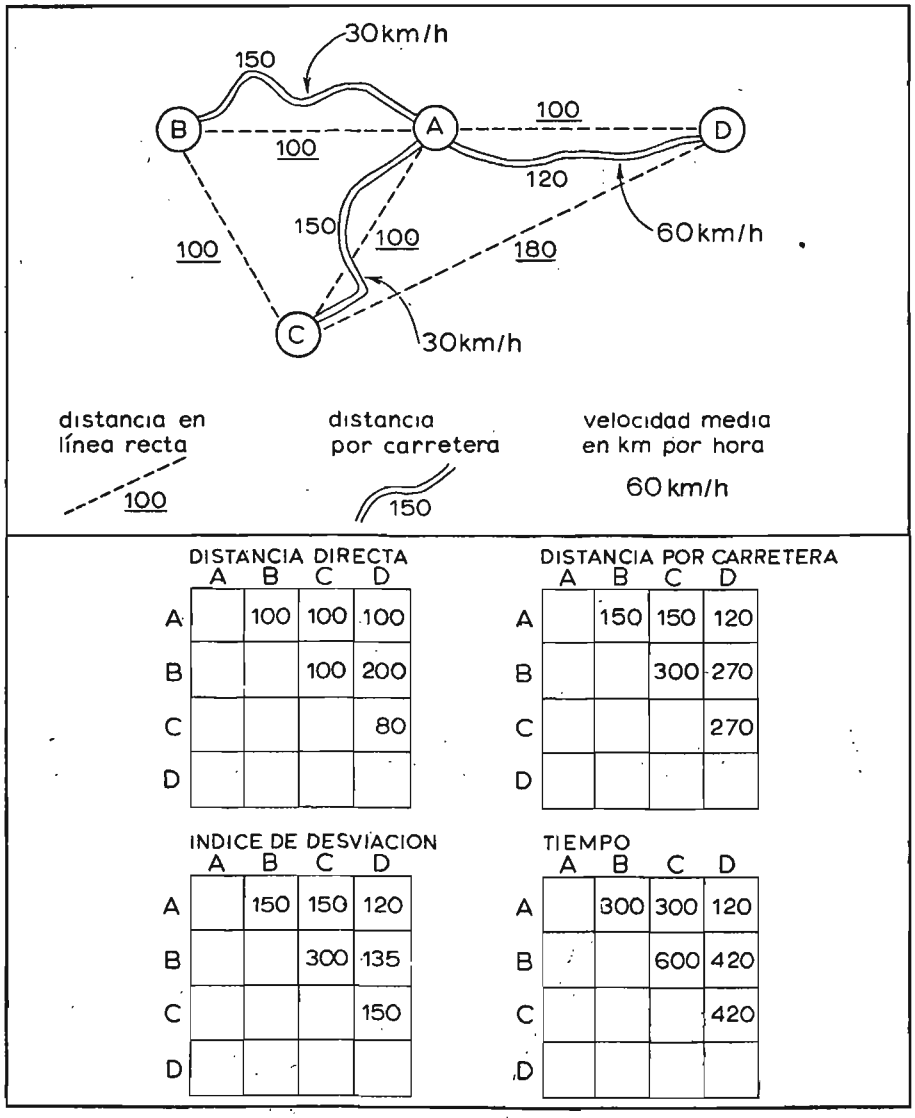
2. Distancia por carretera o por otra vía de comunicación entre dos lugares.

3. Distancia según la duración de un viaje entre dos lugares.

4. Distancia según el costo del movimiento de productos o de pasajeros entre dos lugares.

5. Distancia percibida. Por ejemplo, para un inglés, Australia está más cerca de Inglaterra, desde el punto de vista cultural, que Afganistán o Etiopía.

Es posible comparar la distancia en línea recta entre dos puntos, con la distancia por carretera, con la fórmula siguiente, calcu-



DISTANCIA DIRECTA

	A	B	C	D
A		100	100	100
B			100	200
C				80
D				

DISTANCIA POR CARRETERA

	A	B	C	D
A		150	150	120
B			300	270
C				270
D				

INDICE DE DESVIACION

	A	B	C	D
A		150	150	120
B			300	135
C				150
D				

TIEMPO

	A	B	C	D
A		300	300	120
B			600	420
C				420
D				

Figura 3.3a

lando un índice de desviación de la carretera.

$$\text{Índice} = \frac{\text{distancia por carretera} \times 100}{\text{distancia directa}}$$

En la República Mexicana hay muchos casos de índices muy altos de desviación. Por ejemplo, la distancia entre Tecpan (Guerrero) y Morelia es aproximadamente 820 km por carretera (pasando por la ciudad de México), mientras que la distancia en línea recta es menor de 300 km. El índice llega a ser aproximadamente 270, mientras que un índice de 100 representa una correspondencia exacta entre la distancia en línea recta y la distancia por carretera.

En la figura 3.3a hay cuatro ciudades. Sirven como ejemplo para ilustrar distancias en línea recta, por carretera y según el tiempo que dura su recorrido. Se puede afirmar que, hasta cierto punto, la distancia según el costo de transporte es el resultado de la influencia de los otros tipos de distancia.

3.4 *Potencial de población*

Para calcular el potencial de población de un lugar X de un país, se divide la población de todos los otros lugares entre la distancia entre cada uno de ellos y X, y se suman los resultados. Este índice da, en forma sucinta, la relación del lugar con la población total del país y permite la comparación entre varios lugares desde el punto de vista de su posición en un país o en una región.

La figura 3.4a muestra una región con 5 centros de población. Las distancias entre los centros están indicadas en el mapa, así como la población de los centros. Las distancias están en la matriz 3.4a. Es convencional incluir una distancia 'interna' para cada centro, es decir, la distancia media de los viajes dentro del estado en el cual está situado el centro. Se usa la mitad de la distancia del centro al lugar más próximo. La matriz 3.4b contiene los resultados de dividir la población entre la distancia.

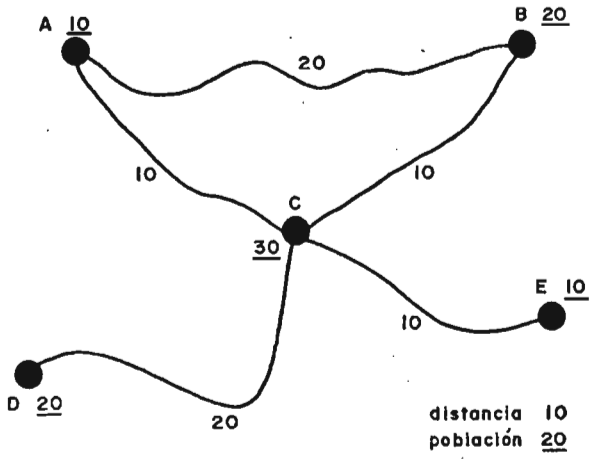


figura 3.4 a

MATRIZ 3.4a

	A	*B*	*C*	*D*	*E*
A	(5)	20	10	30	20
B	20	(10)	20	40	30
C	10	20	(5)	20	10
D	30	40	20	(10)	30
E	20	30	10	30	(5)

MATRIZ 3.4b

A	*B*	*C*	*D*	*E*	*Total*
2	1	3	.7	.5	7.2
.5	2	1.5	.5	.3	4.8
1	2	6	1	1	11.0

Las distancias 'internas' entre paréntesis.

El potencial de población (*PPc*) del centro C es la suma de los valores siguientes:

(P = población, D = distancia).

Así:

$$PPc = \frac{Pa}{Dc\text{–}a} + \frac{Pb}{Dc\text{–}b} + \frac{Pc}{Dc\text{–}c} + \frac{Pd}{Dc\text{–}d} + \frac{Pe}{Dc\text{–}e} = 11.0$$

El potencial de A es 7.2 y el de B es 4.8.

CUADRO 3.4a

Distrito Federal	2,415	San Luis Potosí	197	Aguascalientes	50
México	1,295	Tamaulipas	160	Sonora	46
Veracruz	671	Chiapas	138	Yucatán	42
Puebla	620	Coahuila	129	Colima	26
Jalisco	484	Zacatecas	126	Baja California	26
Michoacán	444	Tlaxcala	114	Campeche	15
Guanajuato	438	Querétaro	97	Baja California (T)	6
Hidalgo	297	Chihuahua	87	Quintana Roo	4
Guerrero	255	Sinaloa	83		
Oaxaca	249	Durango	82		
Morelos	204	Nayarit	56		
Nuevo León	199	Tabasco	52		

Ejercicio 3.3 Calcule el resto de los valores en la matriz 3.4b y el potencial de *D* y *E*.

El potencial de *C* es mayor que el de *A* y *B*, debido a su posición central y su población mayor. En una red mayor el potencial de los lugares centrales puede ser mucho mayor que el de los lugares periféricos, no solamente dos veces como entre *C* y *B* del ejemplo anterior.

El cálculo del potencial de población de cada estado de la República Mexicana dio los resultados siguientes, en el cuadro 3.4a. Hay que notar que los números son relativos, no absolutos. Se puede decir que el potencial del Distrito Federal es aproxima-

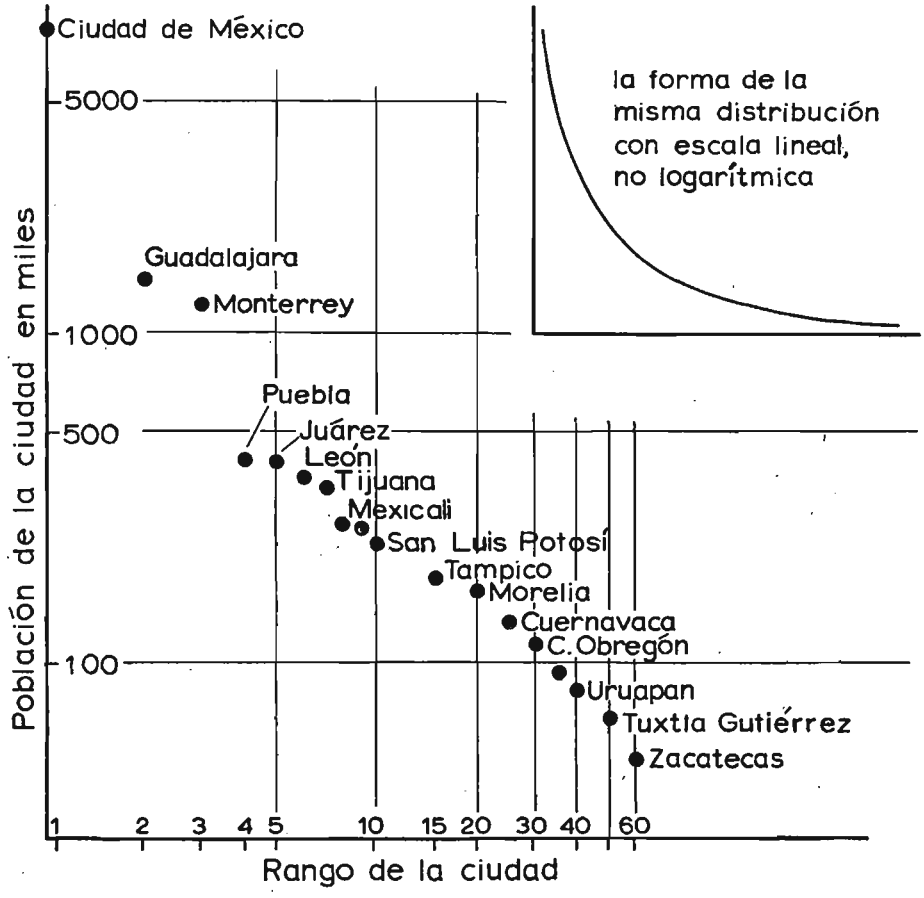


Figura 3. 5a

damente 600 veces el de Quintana Roo. La población de los estados, en el año de 1970, fue la base del cálculo. Las distancias son directas.

3.5 *Modelo de gravedad*

El modelo de gravedad tiene su origen en la física, pero en los últimos años ha sido muy usado por los geógrafos. La fórmula mide la interacción teórica entre dos lugares, por ejemplo, dos ciudades. La intensidad de la interacción está relacionada con la masa (por ejemplo, la población) de ambos lugares y la distancia entre ellos. La fórmula, en su forma más sencilla, es la siguiente:

$$Tij = \frac{Pi\,Pj}{Dij} K$$

T representa las transacciones teóricas entre lugares i y j.
Pi y Pj son las masas de los lugares.
Dij es la distancia entre ellos.

Es posible introducir una constante K, para reducir el resultado, si es un número demasiado grande. (K podría ser, por ejemplo, 10^{-5}).

En general, se usa la distancia al cuadrado entre los dos lugares, como en el modelo de la física, no la distancia misma. Sin embargo, no es obligatorio usar el exponente 2, y a veces otro exponente expresa la situación real en forma más precisa.

3.6 *La curva de Zipf* (ver figura 3.5a)

Se ha observado que muchas distribuciones de datos numéricos tienen forma

CUADRO 3.6a. LAS CIUDADES DE MÉXICO CON MÁS DE 50,000 HABITANTES EN EL AÑO 1970

Ciudad	*Habitantes*	*Ciudad*	*Habitantes*
1. Ciudad de México	8.362,670	31. Toluca	114,078
2. Guadalajara	1.455,824	32. Querétaro	112,993
3. Monterrey	1.213,479	33. Villahermosa	99,565
4. Puebla	413,269	34. Oaxaca	99,509
5. Ciudad Juárez	407,370	35. Orizaba	92,517
6. León	364,990	36. Ciudad Madero	90,830
7. Tijuana	327,400	37. Tepic	87,540
8. Mexicali	267,356	38. Ciudad Victoria	83,897
9. Chihuahua	257,027	39. Pachuca	83,892
10. San Lus Potosí	230,039	40. Uruapan	82,677
11. Torreón	223,104	41. Celaya	79,977
12. Heroica Veracruz	214,072	42. Gómez Palacio	79,650
13. Mérida	212,097	43. Heroica Caborca	78,495
14. Aguascalientes	181,277	44. Monclova	78,134
15. Tampico	179,584	45. Ensenada	77,687
16. Hermosillo	176,596	46. Coatzacoalcos	69,753
17. Acapulco	174,378	47. Campeche	69,506
18. Culiacán	167,956	48. Minatitlán	68,397
19. Saltillo	161,114	49. Los Mochis	67,953
20. Morelia	161,048	50. Tuxtla Gutiérrez	66,851
21. Durango de Victoria	150,541	51. Salamanca	61,039
22. Nuevo Laredo	148,867	52. Tapachula	60,620
23. Heroica Matamoros	137,749	53. Colima	58,450
24. Reinosa	137,383	54. Zamora de Hidalgo	57,775
25. Cuernavaca	134,117	55. Hidalgo del Parral	57,619
26. Jalapa	122,377	56. Heroica Guaymas	57,492
27. Poza Rica	120,462	57. Delicias	52,446
28. Mazatlán	119,553	58. Heroica Nogales	52,108
29. Irapuato	116,651	59. Ciudad Mante	51,247
30. Ciudad Obregón	114,407	60. Zacatecas	50,251

asimétrica. En una gráfica la distribución tiene la forma de una curva. Con papel gráfico logarítmico doble, la curva se aproxima a una línea recta. Esta característica es típica de la distribución de la población de las ciudades de muchas regiones del mundo. El cuadro 3.6a incluye todas las ciudades de México con más de 50,000 habitantes en el año 1970. El eje vertical representa la población de las ciudades, el eje horizontal el rango de cada ciudad según su población, esto es, su posición en una escala ordinal. Teóricamente, la distribución de los valores sigue una línea recta. En el caso de México el tamaño excepcional de la capital influye en la forma de la distribución que se caracteriza, en particular, por la diferencia marcada entre la ciudad de México y las dos ciudades que le siguen, y entre éstas y Puebla, la cuarta ciudad. Como la escala horizontal disminuye de la izquierda a la derecha, en la gráfica no es posible marcar todas las ciudades. La curva de Zipf sirve, sobre todo, para distinguir anomalías en la distribución de la población de las 8-12 ciudades más grandes de la región. Sería posible, también, comparar las ciudades de México en 1970 con las mismas en años anteriores (por ejemplo, 1950, 1930, 1910).

Respuesta: 3.1

CUADRO 3.1a

	Porcentaje de población	*Área*	*Porcentaje de población*	*Área*
A	20	10	20	10
B	30	20	50	30
D	20	20	70	50
E	20	30	90	80
C	10	20	100	100

Respuesta 3.2

CUADRO 3.2b

Ciudad	*Hierro*	*Carbón*	*Acero aE*	*Acero aG*	*Total*
A	8	0	2	1½	11½
B	6	1	1½	1	9½
C	6	2	1	1	10
D	8	3	½	1½	13
E	10	4	0	2	16
F	4	2	1½	½	8
G	2	3	2	0	7
H	0	4	2½	½	7

Resulta que, ahora, solamente *G* y *H* tienen los costos mínimos.

Respuesta 3.3

	A	*B*	*C*	*D*	*E*
D	.3	.7	1.5	2	.3
E	.5	1	3	.7	2

Elementos de estadística

4. ESTADÍSTICA DESCRIPTIVA Y MUESTRAS

4.1 *Tipos de datos*

Se reconocen en la estadística tres tipos de datos: nominales, ordinales y con intervalos.

El término *datos nominales* se refiere a la información sobre objetos que no están en escala cuantitativa. Por ejemplo, una persona de 30 años de edad es mayor que una de 20; no se puede decir, sin embargo, que rojo es mayor que verde; éstas son cualidades. De la misma manera, no se puede decir que una de las marcas de coche siguientes: Volkswagen, Renault o Datsun sea mayor o menor que las otras.

Salta a la vista esta característica cuando se quiere colocar las tres marcas de coche en la escala de un histograma para representar la frecuencia de venta de coches de cada marca, como, por ejemplo, en Veracruz en 1970: Datsun 731, Renault 455, Volkswagen 1,645. En este caso la escala horizontal no es una escala cuantitativa.

Los *datos ordinales* usan una escala graduada. Los casos o las observaciones se ponen en el orden de la escala de los números enteros positivos, representando cada número la posición en la escala ordinal.

En ciertas competencias es costumbre poner a los concursantes en orden, sin preocuparse del intervalo que hay entre ellos: primero, segundo o tercer lugar. Hay un intervalo exactamente de uno en cada caso, excepto en el caso de un empate.

La escala que se usa más comúnmente para las medidas utiliza *datos con intervalos*. En este tipo de escala hay una distancia conocida que separa a dos números cualesquiera, por ejemplo: la precipitación en centímetros: 25.3, 18.7, 19.4, 22.8, 23.9. La conversión de estos valores a milímetros no cambiaría la forma básica de los datos: 253, 187, 194, 228, 239 subsisten como datos con intervalos. No es forzoso usar decimales en una escala con intervalos.

De los tres tipos de datos, cada uno tiene ventajas y desventajas, características y pruebas propias.

Solamente los datos con intervalos pueden tener media y desviación estándar. Una prueba que se aplica mucho con los datos con intervalos es la de Pearson (ver capítulo 5, sección 3).

Los datos ordinales tienen mediana (equivalente a la media) y cuartiles (parecidos a la desviación estándar). Una prueba que se aplica mucho con los datos ordinales es la de Spearman (ver capítulo 5, sección 2).

Los datos nominales tienen clase modal como medida de tendencia central. Una prueba que se aplica con los datos nominales es la prueba chi-cuadrado (ver capítulo 5, sección 4).

4.2 *Estadística descriptiva*

Media aritmética. La media aritmética

o promedio de un conjunto de números se calcula de la manera siguiente: se suman los valores y la suma se divide entre el número (N) de los valores.

En la estadística se usan muchos símbolos. Muchas veces la letra x se refiere a cada elemento de un conjunto de datos numéricos: $x_1, x_2, x_3 \ldots x_n$, siendo x_i cualquiera de los datos x.

La suma de los valores x_i se escribe así: Σx_i.

Ejemplo:

$$51 + 26 + 23 + 28 + 34 + 18 = 180 \; \Sigma x_i = 180 \; (N = 6).$$

$$\text{Media } (M \text{ o } \overline{X}) \; \overline{X} = 180/6 = 30.$$

Mediana. Esta medida se emplea solamente con datos ordinales. Sin embargo, es posible calcular la mediana en los datos con intervalos. La mediana es el número que divide todo el conjunto de números en partes iguales. Por ejemplo: 1,2,3,4,5,6,7. La mediana es 4 en este caso.

Cuando el número de observaciones es par, la mediana es la media de los dos números del centro de la serie. En el ejemplo que sigue es 4.5 (la media entre 4 y 5): 1,2,3,4 (4.5), 5,6,7,8.

La mediana del siguiente conjunto de números con intervalos es 3.1 porque es el que divide a la serie en dos partes iguales; pero no es la media. 1.7, 1.8, 2.4, 2.6, 3.1, 3.6, 3.8, 4.0, 4.3.

Clase modal. La clase modal es la que tiene mayor frecuencia de casos u observaciones. En la figura 4.1a es la clase Volkswagen, porque es ésta la que tiene un mayor número de casos. La clase modal se destaca cuando los datos se representan en forma de polígono de frecuencias o de un histograma.

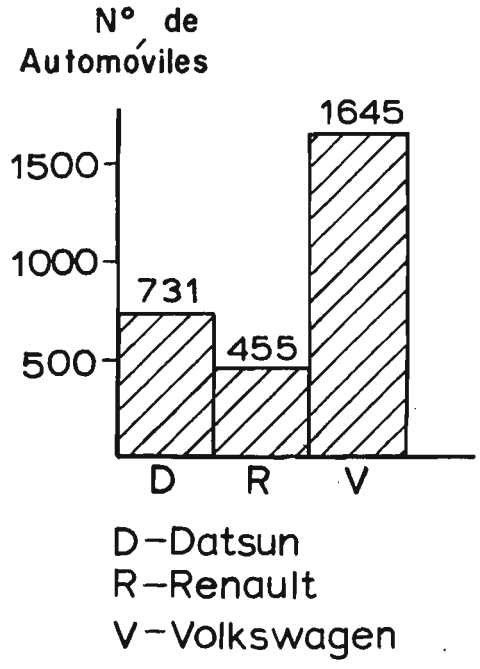


Figura 4 1a

La clase modal se aplica también a datos con intervalos, como se ve en el ejemplo siguiente: Los valores 92, 96, 97, 98, 99, 100, 100, 100, 102, 104, 104, 107, 111 representan índices de inteligencia de varios niños. El problema es agrupar las observaciones en clases, para deducir la cantidad de casos sin perder mucha información. Los límites inferior y superior de las clases no deben coincidir con los valores de las observaciones.

Generalmente todos los intervalos entre los límites de las clases deben ser iguales. En el cuadro 2a, 92.5, 97.5, 102.5 y 107.5 son los límites de las clases, la diferencia entre el límite superior e inferior, en este caso, es de 5.

Cuando hay muchas observaciones conviene agruparlas en clases. Es posible calcular la media aritmética, M, de los datos agrupados, de la manera siguiente:

Paso 1. Identificar los límites superior e inferior de las clases (columna 1, cuadro 4.1a).

Paso 2. Calcular el punto medio de la clase x, sumando los límites superior e inferior de la clase y, dividiendo entre dos; por ejemplo, el punto medio de la clase 92.5 + 97.5 es: $(92.5 + 97.5)/2 = 95$.

CUADRO 4.2a

Columna 1	*Columna 2*	*Columna 3*	*Columna 4*
Clases	x	f	fx
< 92.5	90	1	90
92.5– 97.5	95	2	190
97.5–102.5	100	7	700
102.5–107.5	105	3	315
> 107.5	110	1	110
		$\Sigma f = 14$	$\Sigma fx = 1{,}405$

$M = \Sigma fx/\Sigma f = 1{,}405/14 = 100.4$

Paso 3. Contar el número de casos, f, en cada clase y escribir el resultado en la columna 3.

Paso 4. Multiplicar cada valor x por su valor f correspondiente y poner los resultados en la columna 4.

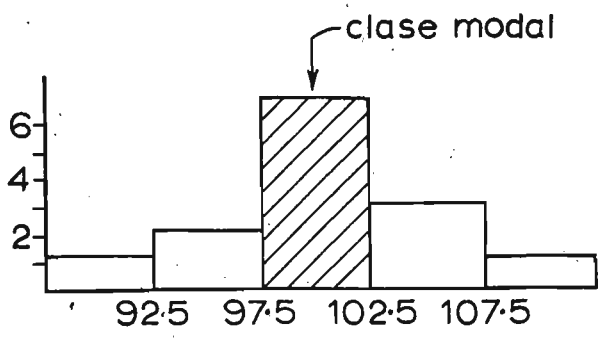
Paso 5.Sumar la columna 3 y la 4 y dividir la suma de la columna 4, Σfx, entre la suma de la columna tres, Σf o N.

Los cálculos se hacen más rápidamente con datos agrupados, pero se pierde siempre un poco de la información inicial.

La clase modal, la de mayor frecuencia, es la clase de 97.5-102.5 con una frecuencia de 7 casos (ver figura 4.1b).

Desviación estándar

Esta medida da una idea de la forma de una distribución y de su dispersión alrededor de la media. Además, el cálculo de la desviación estándar facilita la comparación de distribuciones con valores iniciales muy diferentes. Es posible convertir cualquier serie de datos con intervalo, en la misma forma: con una media de cero y una desviación estándar de uno (símbolo σ, sigma minúscula).



(datos del cuadro 4 2 a)

Figura 4.1b

Una de las fórmulas para el cálculo de la desviación estándar es la siguiente:

$$\sigma = \sqrt{\frac{\Sigma (X - M)^2}{N}}$$

donde X = número de observaciones
M = la media
N = número de observaciones

Ejemplo: calcular la desviación estándar de los números siguientes ($N = 10$):

5, 8, 3, 2, 7, 9, 8, 2, 2, 4.

Paso 1. Sumar los números $\sum_{i=1}^{N} X_i = 50$.

Paso 2. Calcular M (la media) $50/N = 5$.

Paso 3. Calcular $(x - M)$ sustrayendo la media de cada valor en turno:

0, 3, −2, −3, 2, 4, 3, −3, −3, −1.

Paso 4. Calcular $(x - M)^2$:

0, 9, 4, 9, 4, 16, 9, 9, 9, 1.

Paso 5. Calcular $\Sigma (x - M)^2$, esto es, 70.

Paso 6. La varianza es $70/N = 7$.

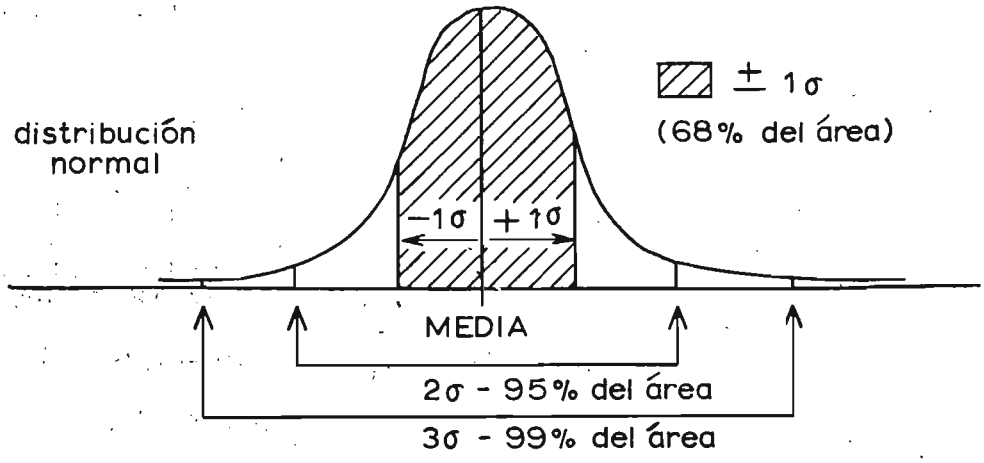


Figura 4.1c

Paso 7. σ es la raíz cuadrada de 7, aproximadamente 2.64.

Cuando muchas observaciones están distribuidas en forma de curva *normal* (ver figura 4.1c), las distancias de una desviación estándar a la izquierda y a la derecha de la media, incluyen aproximadamente 68 % del área de abajo de la curva, dos desviaciones estándar incluyen 95 %, y tres desviaciones estándar incluyen 99 %.

Muchas cosas naturales (por ejemplo, la estatura de las personas) tienen valores que se distribuyen en forma de una curva normal, con muchas observaciones cerca de la media. Sin embargo, hay muchas otras formas de distribución. En geografía se encuentran muchos datos en forma de distribución asimétrica, como, por ejemplo, la distribución de ingresos por persona (muchos sueldos altos). Ver la figura 4.1d.

A veces se encuentran distribuciones con dos clases modales. En tales casos, la media no se encuentra en la parte más alta o densa de la distribución.

La estadística convencional trabaja con distribuciones de datos en una dimensión, en una línea. Es posible transferir conceptos y métodos de la estadística de una dimensión, a distribuciones en dos dimensiones. Los mapas muestran distribuciones en dos dimensiones y la mayor parte de los estudios geográficos se ocupan de situaciones espaciales.

Es posible calcular la MEDIA de una distribución en dos dimensiones.

La MEDIA de una distribución en una dimensión es el punto de equilibrio de varios pesos en una línea (ver figura 4.1f). De la misma manera, la media o centro de gravedad de una distribución en dos di-

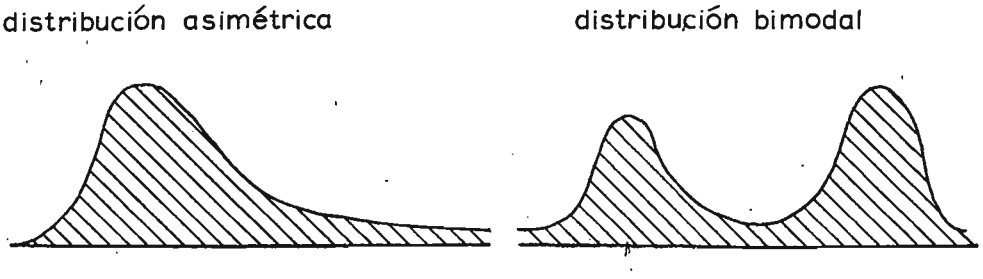


Figura 4.1d

Figura 4.1e

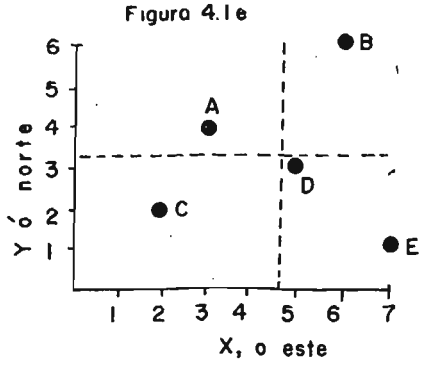
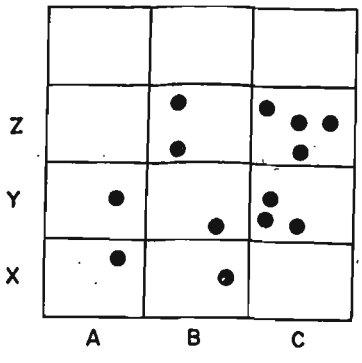
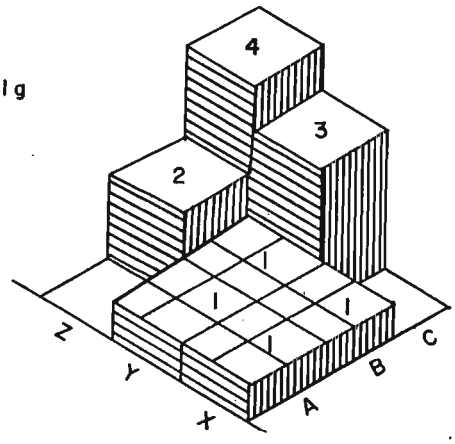


Figura 4.1f

Figura 4.1g





mensiones es el punto en que la distribución se encontraría en estado de equilibrio. Para los Estados Unidos se ha calculado la media de la distribución de la población, cada diez años, desde fines del siglo XVIII. Después de la independencia se encontraba cerca de Washington (la capital). Entre 1800 y 1960 se movió hacia el oeste, hasta el estado de Illinois. También en la Unión Soviética se ha calculado la media de la población.

La figura 4.1e indica el método de cálculo de la media de una distribución en dos dimensiones. Hay que considerar, primero, la distribución desde el punto de vista del eje x, o Este, y después desde el punto de vista del eje y, o Norte. Cada punto (representando, por ejemplo, 100,000 personas) tiene un valor equivalente a su distancia al este y al norte de 0, el punto de origen de la gráfica.

Calcular la media en la dirección Este (x) y en la dirección Norte (y) ($N=5$ puntos).

	Distancia x	*Distancia* y
A	3	4
B	6	6
C	2	2
D	5	3
E	7	1
	$\Sigma x = 23$	$\Sigma y = 16$

Media, dirección x $23/5 = 4.6$
Media, dirección y $16/5 = 3.2$

Las dos líneas cruzan en la media de la distribución, en dos dimensiones.

Cuando hay muchos puntos es posible dividir las dos partes (x, y) en clases y hacer los cálculos con los puntos agrupados.

Es posible, también, adaptar el concepto de histograma y de la clase modal a una distribución en tres dimensiones (ver figura 4.1g). Es preciso sobreponer una red cuadrada a una distribución de puntos y contar el número de puntos en cada cuadrado. Hay que representar la distribución,

de nuevo, con un diagrama que representa tres dimensiones, y levantar columnas sobre cada cuadrado, proporcionales en altura al número de puntos que representan. Es un método un poco complicado y tiene el defecto de que las columnas altas pueden esconder columnas más bajas (no pasa en el ejemplo). Sin embargo, es una manera de representar, en forma dramática, discrepancias y contrastes en la densidad de la población y en otros fenómenos geográficos.

4.3 *La prueba 't'*

La prueba 't' tiene aplicación en todas las ramas de la geografía. La función de esta prueba es averiguar si con un determinado coeficiente de confianza dos conjuntos de observaciones podrían venir de la misma población, como se verá en los dos ejemplos siguientes:

1. Dos muestras de la estatura de hombres, una de Suecia, la otra de Portugal, dan los resultados siguientes, en pulgadas.

Suecia (muestra de 12): 72, 71, 69, 73, 72, 74, 70, 68, 66, 73, 70, 71.

Portugal (muestra de 10): 70, 66, 63, 64, 66, 68, 65, 69, 65, 66.

Es evidente que la estatura media de los portugueses es menor que la de los suecos. Sin embargo, es posible que, por casualidad, la primera muestra haya incluido más hombres altos que la segunda y que, en realidad, no haya diferencia en la estatura media de los suecos y de los portugueses. En otras palabras, se pregunta: ¿Es posible que los elementos de los dos conjuntos vengan de la misma población total, o hay una diferencia significativa entre las dos muestras y, en consecuencia, entre las dos poblaciones totales?

2. En el centro de México se construyó un embalse y se creó un lago artificial. Hipótesis: ¿Hubo un cambio significativo en la temperatura media de la zona del rededor del lago, después de la creación de éste? Las observaciones meteorológicas incluyen datos sobre la temperatura durante un periodo de 13 años anteriores a la construcción del embalse y 15 años después de su construcción (temperaturas °C).

Antes: 18.5, 17.6, 19.2, 19.5, 18.8, 17.5, 17.8, 18.2, 18.8, 17.9, 19.0, 18.0, 18.8.

Después: 17.2, 16.8, 17.5, 17.3, 18.6, 18.9, 16.9, 18.3, 18.2, 17.0, 16.8, 16.6, 18.5, 17.4, 18.8.

El promedio de las temperaturas después de la construcción del embalse es menor que el del periodo anterior. A pesar de esto la diferencia no es muy marcada, y dadas las fluctuaciones de la temperatura de año en año, podrían ser variaciones aleatorias.

En la prueba 't' se comparan las observaciones de dos muestras y el índice permite una conclusión, en términos de coeficientes de confianza, acerca de la relación entre los dos conjuntos de datos. Hay dos métodos para calcular el índice 't'. Se explica uno de los métodos.

Los datos

Muestra X: 4,4,5,6,6.
Muestra Y: 6,7,9,5,8,7.

Paso 1. Calcular el número (N) de valores en cada muestra, la media de las dos muestras y la varianza de las dos muestras (ver la sección 4.2).

Los valores de N son 5 (N_x) y 6 (N_y).

X	$X-\overline{X}$	$(X-\overline{X})^2$	Y	$Y-\overline{Y}$	$(Y-\overline{Y})^2$
4	−1	1	6	−1	1
4	−1	1	7	0	0
5	0	0	9	+2	4
6	+1	1	5	−2	4
6	+1	1	8	+1	1
25		4	7	0	0
			42		10

$\overline{X}$ (la media de X) $= 25/N_x = 5.$

$\overline{Y}$ (la media de Y) $= 42/N_y = 7.$

S_x^2 (la varianza de X) $= 4/N_x = 0.8.$

S_y^2 (la varianza de Y) $= 10/N_y = 1.66.$

Paso 2. Calcular la mejor estimación de la varianza de las dos muestras ($\hat{\sigma}^2$). Se usa la fórmula siguiente:

$$\hat{\sigma}^2 = \frac{N_x S_x^2 + N_y S_y^2}{N_x + N_y - 2} =$$

$$= \frac{5 \times 0.8 - 6 \times 1.66}{5 - 6 - 2} = \frac{14}{9} = 1.55$$

calcular $\hat{\sigma}$, esto es, la mejor estimación de la desviación estándar es la raíz cuadrada de $\hat{\sigma}^2$, esto es $\sqrt{1.55} = 1.25$.

Paso 3. Calcular $\hat{\sigma}_w$

$$\hat{\sigma}_w = \hat{\sigma} \sqrt{\frac{1}{N_x} + \frac{1}{N_y}} =$$

$$= 1.25 \sqrt{1/5 + 1/6} = 1.25 \sqrt{0.36} =$$

$$= 1.25 \times 0.6 = 0.75.$$

Paso 4. Calcular la diferencia absoluta entre las medias

$$X - Y = 5 - 7 = 2.$$

Paso 5. $t =$ diferencia entre las medias dividida entre $\hat{\sigma}_w = 2/0.75 = 2.66$.

Paso 6. Referirse a las tablas de valores de 't' con grados de libertad igual a $N_x + N_y - 2$; en este ejemplo, $5 + 7 - 2$, esto es, 10.

El resultado 2.66 queda entre los coeficientes de confianza de 95 % (1.81) y 99 % (2.76).

4.4 *Las muestras*

Esta sección no trata de la teoría de las muestras, sólo consiste en un ejercicio práctico que expone algunos problemas y dificultades en la manera de escoger una muestra y de interpretar los resultados. El problema es elegir 50 casas entre una *población* total de 200 casas en un barrio de una ciudad, visitar las 50 casas, hacer preguntas e interpretar los resultados.

En realidad, la función de una muestra es estimar, con la información de la muestra, ciertas características de la población total. Por ejemplo, si uno fuera a visitar 100 casas en un barrio de una ciudad y encontrara 60 con televisor, sería posible decir que 60 % de todas la casas (posiblemente miles) del barrio, tenían televisor. Desgraciadamente, con datos de una muestra de 100 no se puede estimar con tanta exactitud una característica de la población total, porque es poco probable que la muestra sea exactamente representativa de la población total.

En el ejercicio que sigue, el cuadro 4.4b tiene información sobre la población total de 200 casas. En realidad, si uno ya tuviera toda la información deseada, no sería necesario tomar una muestra. Además, una población de 200 elementos es tan pequeña, que sería mejor visitar todas las casas. Es necesario tomar en cuenta estas limitaciones del ejercicio.

Paso 1. Usando el mapa (figura 4.3a), escoger una muestra de 50 casas de la población de 200. La muestra debe ser representativa de toda la zona. No conviene, por ejemplo, incluir solamente las casas de la zona obrera. Posiblemente podría visitarse cada cuarta casa, pero sería mejor usar una lista de números aleatorios y escoger una muestra al azar.

Paso 2. Usando el cuadro 4.4a, identifi-

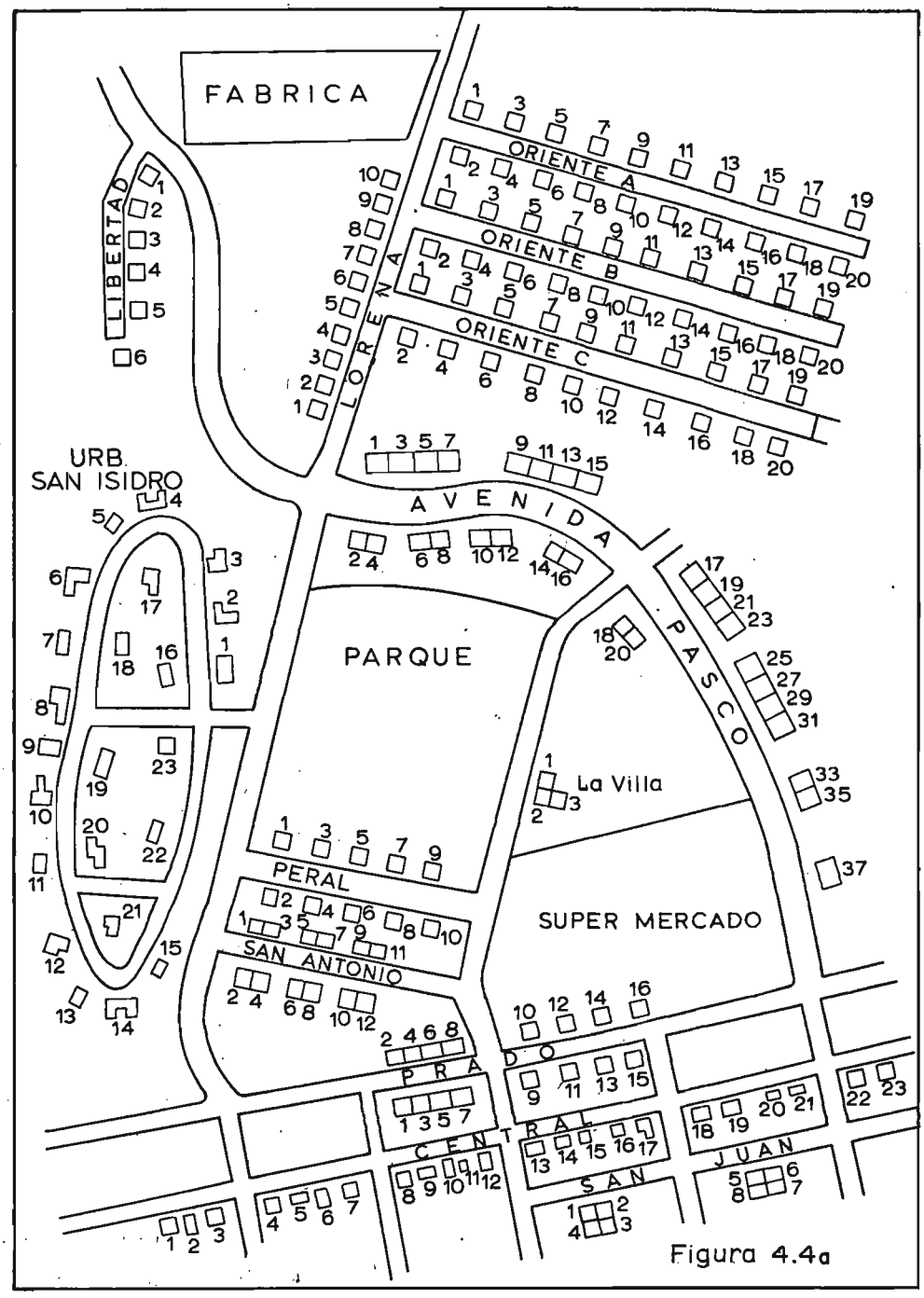


Figura 4.4a

car las 50 casas de su muestra, según columna II), su posición y su número en la calle. Identificar II y I, su número o posición en la lista completa de todas las casas, de 1 a 200. Por ejemplo, la casa Nº 15 de la calle Oriente C tiene el número 55 en la lista 4.4b.

Paso 3. A la familia de cada casa de su muestra hacerle las cuatro preguntas siguientes:

1. ¿La familia ha emigrado recientemente a este barrio, de otra entidad del país?
2. ¿Tiene televisor?

3. ¿Tiene coche?
4. ¿Vota por el Partido Democrático Nacional?

En este ejercicio es suficiente referirse al cuadro 4.4b. Aquí se encuentran las respuestas de *todas* las familias.

Por ejemplo, el dueño de la casa número 15 de la calle Oriente C, esto es, casa 55 en la lista completa del cuadro 4.4b, contesta así:

1. ¿Son inmigrantes? NO.
2. ¿Tienen televisor? NO.
3. ¿Tienen coche? NO.
4. ¿Votan por el Partido Democrático? Sí.

Sumar las respuestas positivas a cada pregunta.

Una muestra hecha por el autor dio los resultados siguientes (entre 50 casas):

1. Inmigrantes	3/50
2. Televisor	43/50
3. Coche	23/50
4. Partido Democrático	17/50

Paso 4. Estimar la proporción de casas en la población total, con cada uno de los cuatro atributos:

1. Inmigrastes	3/50	6 %
2. Televisor	43/50	86 %
3. Coche	23/50	46 %
4. Partido Democrático	17/50	34 %

Paso 5. Comparar las estimaciones de la muestra con la situación en la población total, la cual se conoce *en este ejercicio*, pero en la realidad no se conocería. Las proporciones son las siguientes:

1. Inmigrantes	15/200	7.5 %
2. Televisor	160/200	80.0 %
3. Coche	100/200	50.0 %
4. Partido Democrático	70/200	35.0 %

La muestra en el paso 4 dio una estimación buena de la proporción que vota por el partido D, pero una estimación excesiva de la proporción con televisor y una estimación demasiado baja de la proporción con coche. El "error" en la estimación es resultado de fuerzas aleatorias que impiden escoger una muestra perfecta.

Es evidente, entonces, que se necesita algún índice que dé confiabilidad a la estimación de la proporción.

En ejercicio en clase se obtuvieron 15 muestras diferentes:

Conclusiones

1. La confiabilidad de una muestra, o su capacidad de estimar características de una población total, depende mucho *del número de casos* de la muestra y poco de la proporción de la muestra en la población total.

No es necesario incluir en una muestra, por ejemplo, 5 % o 10 % de la población total. Además, en general, no vale la pena tener más de 500-600 casos en la muestra. Muchas veces no se conoce el número de casos en la población total (por ejemplo, granos de arena en una playa), de modo que no sería *posible* saber si incluye, por ejemplo, 1 % o 2 % de tal población.

2. La confiabilidad de la estimación depende, hasta cierto punto, de la naturaleza de la pregunta y de la proporción de respuestas afirmativas o negativas. En los ejemplos anteriores, la proporción verdadera de inmigrantes (7.5 %) fue estimada entre 2 % (muestra 6) y 12 % (muestra 9). Es evidente que en este caso, por la rareza de inmigrantes, la muestra no fue verdaderamente útil. En el caso de los coches, las estimaciones de la proporción verdadera (50 %) varían entre 38 % y 66 %.

	Pregunta				*Porcentajes*			
	1 (IN)	*2 (TEL)*	*3 (COCHE)*	*4 (P.D.)*	*1*	*2*	*3*	*4*
1	3	42	23	20	6	84	46	40
2	4	41	24	18	12	82	48	36
3	2	44	24	18	4	88	48	36
4	4	41	28	15	8	82	56	30
5	5	40	20	13	10	80	40	26
6	1	47	24	23	2	94	48	46
7	4	44	29	22	8	88	58	44
8	3	45	29	26	6	90	58	52
9	6	38	19	18	12	76	38	36
10	3	45	28	21	6	90	56	42
11	3	42	27	16	6	84	54	32
12	3	43	33	19	6	86	66	38
13	3	38	24	15	6	76	48	30
14	4	41	28	19	8	82	56	38
15	6	40	23	19	12	80	46	38
					Porcentajes verdaderos en la población total			
					7.5	80	50	35

3. Existen tablas que indican los tamaños de las muestras según la confiabilidad deseada. Por ejemplo, para tener una con coeficientes de confianza de 95 %, que no varíe más que 5 % del valor verdadero de la población total, se necesitan los números siguientes:

Una muestra de 278 para una población total estimada de 1000;

Una muestra de 370 para una población estimada de 10,000;

Una muestra de 383 con población de 100,000.

(Las tablas pueden encontrarse en una publicación americana: *Tables for Statisticians*, College Outline Series, Barnes and Noble, New York, 1964, pp. 145-152.)

Lo siguiente ejemplifica el método para calcular la confianza de la estimación de una muestra.

En una muestra de 400, 160 personas declararon que votan por el Partido Democrático. Se estima, entonces, que 40 % (160/400) de la población total vota por el Partido Laborista.

Paso 1. Calcular p (punto estimativo de π, proporción con la población total):

$$p = 160/400 = 0.4.$$

Paso 2. Calcular la desviación estándar:

$$(\hat{\sigma}p) = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{0.4 \times 0.6}{400}} = 0.0245$$

Paso 3. La tabla de límites de confiabilidad indica que el límite de confiabilidad al nivel de 95 % es 1.96.

El resultado varía:

$$p \pm 1.96 \times \hat{\sigma}\,p = 0.4 \pm 1.96 \times 0.0245$$
$$= 0.4 \pm 0.048$$

esto es, entre 0.352 y 0.448.

Paso 4. Convirtiendo las proporciones a la muestra se puede afirmar, con una confiabilidad de 95 %, que la proporción de la población que vota por el Partido Democrático se encuentra entre 142 — (160)—179, o entre 35.5 % y 45 % de la población total.

CUADRO 4.4a. LOS NÚMEROS DE LAS CASAS EN LAS CALLES

I. *Número de la casa en el cuadro 4.4b.*
II. *Número de la casa en el cuadro 4.4a.*

I	*II*	*I*	*II*	*I*	*II*	*I*	*II*
Oriente A		*Oriente C*		*Central*		*Av. Pasco*	
1	1	51	11	101	1	151	12
2	3	52	12	102	2	152	14
3	5	53	13	103	3	153	16
4	7	54	14	104	4	154	18
5	9	55	15	105	5	155	20
6	11	56	16	106	6	*Urb. S. Isidro*	
7	13	57	17	107	7	156	1
8	15	58	18	108	8	157	2
9	17	59	19	109	9	158	3
10	19	60	20	110	10	159	4
11	2	*Lorena*		111	11	160	5
12	4	61	1	112	12	161	6
13	6	62	2	113	13	162	7
14	8	63	3	114	14	163	8
15	10	64	4	115	15	164	9
16	12	65	5	116	16	165	10
17	14	66	6	117	17	166	11
18	16	67	7	118	18	167	12
19	18	68	8	119	19	168	13
20	20	69	9	120	20	169	14
Oriente B		70	10	121	21	170	15
21	1	*Libertad*		122	22	171	16
22	2	71	1	123	23	172	17
23	3	72	2	*La Villa*		173	18
24	4	73	3	124	1	174	19
25	5	74	4	125	2	175	20
26	6	75	5	126	3	176	21
27	7	76	6	*Av. Pasco*		177	22
28	8	*Prado*		127	1	178	23
29	9	77	1	128	3	*Peral*	
30	10	78	3	129	5	179	1
31	11	79	5	130	7	180	3
32	12	80	7	131	9	181	5
33	13	81	9	132	11	182	7
34	14	82	11	133	13	183	9
35	15	83	13	134	15	184	2
36	16	84	15	135	17	185	4
37	17	85	2	136	19	186	6
38	18	86	4	137	21	187	8
39	19	87	6	138	23	188	10
40	20	88	8	139	25	*San Antonio*	
Oriente C		89	10	140	27	189	1
41	1	90	12	141	29	190	3
42	2	91	14	142	31	191	5
43	3	92	16	143	33	192	7
44	4	*San Juan*		144	35	193	9
45	5	93	1	145	37	194	11
46	6	94	2	146	2	195	2
47	7	95	3	147	4	196	4
48	8	96	4	148	6	197	6
49	9	97	5	149	8	198	8
50	10	98	6	150	10	199	10
		99	7			200	12
		100	8				

CUADRO 4.4b. LA SITUACIÓN VERDADERA

Los números 1-200 corresponden a las 200 casas del mapa (figura 4.4a)
I - Inmigrante, T - Televisor, C - Coche, P - Partido Demócrata Nacional.
1 Significa presencia de, o sí.
0 Significa ausencia de, o no.

	I	*T*	*C*	*P*		*I*	*T*	*C*	*P*		*I*	*T*	*C*	*P*		*I*	*T*	*C*	*P*
1	0	1	0	0	51	0	1	1	0	101	0	1	1	0	151	0	1	0	1
2	0	1	0	0	52	0	0	0	0	102	0	1	0	1	152	1	1	0	1
3	0	0	0	0	53	0	1	0	0	103	0	1	1	1	153	1	0	0	0
4	0	0	1	0	54	0	1	0	0	104	0	1	1	1	154	0	1	0	1
5	0	1	0	0	55	0	0	0	1	105	0	1	1	0	155	0	1	0	0
6	0	1	0	0	56	0	1	0	1	106	0	1	1	1	156	0	1	1	0
7	0	1	0	0	57	0	1	0	1	107	0	0	0	0	157	0	1	1	0
8	0	1	0	0	58	0	1	0	0	108	0	1	1	1	158	0	1	1	0
9	0	0	1	0	59	0	0	0	1	109	0	1	0	0	159	0	1	1	0
10	0	1	0	0	60	0	0	0	1	110	1	1	1	1	160	0	1	1	0
11	0	1	0	0	61	0	1	0	0	111	0	1	1	0	161	0	1	1	0
12	0	1	1	0	62	0	1	0	0	112	0	1	1	1	162	0	1	1	0
13	0	0	1	0	63	0	1	1	0	113	0	1	1	0	163	0	1	1	0
14	0	1	0	0	64	0	0	0	0	114	0	0	0	0	164	0	1	1	0
15	0	1	0	0	65	0	1	0	0	115	0	1	1	1	165	0	1	1	0
16	0	0	0	0	66	0	1	0	1	116	0	1	1	0	166	0	0	1	1
17	0	1	0	0	67	1	0	0	0	117	0	1	0	1	167	0	1	1	0
18	0	1	0	0	68	1	1	1	0	118	0	1	1	0	168	0	1	1	0
19	0	0	0	0	69	0	1	0	1	119	0	1	1	0	169	0	1	1	0
20	0	0	0	0	70	0	1	1	1	120	0	1	1	0	170	0	1	1	0
21	0	1	0	0	71	0	1	1	0	121	0	1	1	0	171	0	1	1	0
22	0	1	1	0	72	0	1	1	1	122	0	0	1	0	172	0	1	1	0
23	0	1	0	0	73	0	1	0	1	123	0	1	1	0	173	0	1	1	0
24	0	0	0	0	74	0	1	0	0	124	0	1	1	1	174	0	1	1	0
25	0	1	0	0	75	0	1	0	1	125	0	1	1	1	175	0	1	1	0
26	0	1	0	0	76	0	1	1	1	126	0	1	0	1	176	0	1	1	0
27	0	1	0	0	77	0	1	0	1	127	0	1	1	1	177	0	1	1	0
28	0	0	0	0	78	0	1	1	1	128	1	1	1	0	178	0	1	1	0
29	0	0	0	0	79	0	1	1	1	129	0	1	0	1	179	0	1	1	1
30	0	1	0	0	80	0	1	1	0	130	0	0	0	1	180	0	1	1	1
31	0	1	0	0	81	0	1	0	0	131	0	1	1	1	181	0	1	0	0
32	0	1	1	0	82	0	1	0	1	132	0	0	1	1	182	0	0	1	1
33	0	1	0	0	83	0	1	1	1	133	0	1	1	1	183	0	1	1	0
34	0	0	0	0	84	0	1	0	1	134	0	1	0	1	184	0	1	1	0
35	0	1	0	0	85	0	1	0	1	135	1	0	1	0	185	0	1	1	1
36	0	1	0	0	86	0	1	0	0	136	1	1	0	0	186	0	1	1	0
37	0	1	0	0	87	0	0	1	0	137	0	0	1	1	187	0	0	0	1
38	0	1	1	0	88	0	1	1	1	138	1	0	0	0	188	0	1	1	1
39	0	0	0	0	89	0	1	1	0	139	0	1	0	1	189	0	1	1	0
40	0	1	0	0	90	0	1	1	1	140	1	1	0	0	190	0	0	1	1
41	0	0	1	0	91	0	1	0	0	141	1	0	0	0	191	0	1	1	0
42	0	0	0	0	92	0	1	0	1	142	1	1	1	0	192	0	1	0	1
43	0	1	0	0	93	0	1	1	1	143	0	1	1	1	193	0	1	1	1
44	0	1	0	0	94	0	1	0	1	144	1	1	0	0	194	0	1	1	1
45	0	0	0	0	95	0	1	1	1	145	1	1	0	0	195	0	1	0	0
46	0	1	0	0	96	0	1	1	0	146	1	1	0	0	196	0	0	1	1
47	0	1	0	0	97	0	1	0	1	147	1	0	1	0	197	0	1	1	1
48	0	0	1	0	98	0	1	1	1	148	0	0	1	1	198	1	1	1	0
49	0	1	0	0	99	0	1	0	0	149	0	1	1	1	199	0	1	1	0
50	0	1	0	0	100	0	1	0	1	150	0	1	0	1	200	0	1	0	1

5. TRES PRUEBAS ESTADÍSTICAS

5.1 *Introducción*

Se incluyen en este capítulo tres pruebas estadísticas que, en términos generales, dan un índice (o coeficiente) de correlación entre dos variables.

La prueba *chi-cuadrado* (X^2) usa datos nominales, la prueba corficiente de correlación gradual de Spearman emplea datos ordinales, y la prueba de Pearson para el coeficiente de correlación producto momento utiliza datos con intervalos. Se presenta primero la prueba Spearman (sección 5.2) porque de las tres es la más fácil de entender; le sigue la prueba Pearson, ya que tiene mucho en común con la de Spearman (sección 5.3).

Después de efectuar el cálculo de un índice de correlación, es esencial averiguar si el resultado pudiera haberse obtenido fácilmente, por casualidad. Convencionalmente se usan dos coeficientes de confianza, 95 % y 99 %; el segundo límite es más exigente.

En la prueba de Spearman un coeficiente (*RHO*) alrededor de cero significa que no hay correlación entre dos variables. Un índice de 1.0 significa una correlación positiva perfecta, mientras que uno de −1.0 significa una correlación negativa perfecta. Un índice de ±0.3 o ±0.5 puede obtenerse por casualidad si en el estudio no se usaron suficientes casos. Por ejemplo, el resultado de la aplicación de la prueba dio un índice de +0.6, siendo el coeficiente de confianza +0.4. Esto quiere decir que la correlación de +0.6 es significativa, con 95 % de confianza, porque un índice mayor que dicho límite se obtendría por casualidad, menos de una vez en cien (probabilidad de menos de 1/100).

En cada una de las secciones siguientes hay un cuadro con los límites de confianza apropiados para la prueba que se describe.

5.2 *El coeficiente de correlación gradual*

Una de las pruebas estadísticas más fáciles de entender y de calcular es la de Spearman, que se aplica a datos ordinales o en rango. Esta prueba puede usarse con datos en orden creciente o decreciente (por ejemplo, las posiciones de varios competidores según dos jueces examinadores) o con datos en forma de intervalos. Los datos con intervalos tienen que transformarse en datos ordinales.

En el ejemplo que sigue hay ocho estaciones meteorológicas y los datos de altitud y de precipitación para cada estación, por lo que hay ocho pares de observaciones. Se puede aplicar la prueba de Spearman a estos datos, para saber si hay correlación entre las dos variables (ver cuadro 5.2a).

La fórmula de la prueba de Spearman es:

$$RHO = 1 - \frac{6 \Sigma D^2}{N^3 - N}.$$

RHO es el coeficiente de correlación.
D es la diferencia de cada lugar (de *A* a *H*) entre su posición en la primera variable y su posición en la segunda.
N es el número de casos, en este ejemplo las ocho estaciones meteorológicas. 1 y 6 en la fórmula son constantes.

Paso 1. Como los datos están con intervalos, es preciso sustituir cada valor, en las dos variables, con un número entero según el rango del valor. *H*, el lugar más alto, viene primero; *G* es el segundo, *D* el tercero.

Para obtener los datos de rango se ordenan los datos iniciales, de mayor a menor, dando al primero valor 1, al siguiente 2, y así sucesivamente hasta terminar con la serie.

CUADRO 5.2a

Datos iniciales

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Altitud (metros)	300	350	200	450	150	250	600	700
Lluvia (cm) anual	71	73	62	100	60	75	162	145

Datos en rango

Altitud	5	4	7	3	8	6	2	1
Lluvia	6	5	7	3	8	4	1	2

Paso 2. Calcular la diferencia, *D*, entre cada par de valores (altitud-lluvia) de los datos en rango, y elevarlos al cuadrado, D^2.

Cálculo de la diferencia *

D	−1	−1	0	0	0	2	1	−1
D^2	1	1	0	0	0	4	1	1

Paso 3. Sumar los ocho valores de D^2.

* El resultado de elevar al cuadrado un número negativo da uno positivo; la operación de multiplicar dos números negativos da como producto un número positivo; por ejemplo, −3 por −3 = +9.

La suma, esto es, D^2, es igual a 8 en este ejemplo.

Paso 4. Calcular *RHO*:

$$RHO = 1 - \frac{6 \times 8}{512 - 8}$$

$$RHO = 1 - \frac{48}{504}$$

$$= 1 - .095$$

$$= +0.905$$

Paso 5. Referirse al cuadro 5.2b para ver los coeficientes de confianza del *RHO* de Spearman. *N* es el número de *casos* estudiados, en este ejemplo las 8 estaciones meteorológicas. En la aplicación de la prueba de Spearman hay que tomar en cuenta lo siguiente:

1. La prueba de Spearman es una prueba *no-paramétrica*. Eso significa que las variables no tienen la forma de una distribución normal, porque no tienen ni media, ni desviación estándar, que son las propiedades fundamentales de los datos paramétricos.

2. El valor *RHO* de la prueba siempre se encuentra entre −1 y +1.

Se puede decir: $-1 \leqslant RHO \leqslant +1$.

Los coeficientes alrededor de 0 (cero) indican falta de correlación. Los coeficientes cercanos a +1 indican una correlación alta positiva, y los valores cercanos a −1 una correlación alta negativa.

3. El cuadro 5.2b da los coeficientes de confianza de Spearman.

La posición del límite depende del número de casos (*N*). A medida que *N* aumenta, el límite se acerca a cero.

Sin embargo, aun con un número grande de *N* (por ejemplo, 100) es posible, de vez en cuando, obtener un índice de correlación de ±0.2 o 0.3, por casualidad.

4. En ciertas situaciones dos casos pue-

den tener el mismo rango. Hay que modificar sus posiciones. Por ejemplo, un juez ha puesto a dos concursantes en la tercera posición:

1,2,3= ,3=,4,5...

En este caso es necesario conservar correctamente la secuencia de números enteros:

1,2,3.5,3.5,5,6...

3.5 y 3.5 ocupando las posiciones 3 y 4.

Cuando hay 3 valores iguales se hace así:

1,2,3= ,3= ,3= 4,5

tiene que ser:

1,2,4,4,4,6,7.

La regla general de este cálculo es dar a todos los casos con rango igual (por ejemplo 5=, 5=,5=) la *media* de las posiciones que habrían tenido si hubieran tenido *rangos diferentes* (esto es, 5,6,7, siendo seis la media).

Otros ejemplos. A la pregunta: Escriba cuál de las seis ciudades preferiría si tuviera que vivir en una de ellas, cuatro personas (*A*, *B* *C* y *X* dieron las opiniones expresadas en la tabla que sigue:

	A	*B*	*C*	*X*
Guadalajara	1	1	6	3
Monterrey	2	2	5	1
Tijuana	3	3	4	6
Mérida	4	4	3	5
Veracruz	5	5	2	4
Acapulco	6	6	1	2

Ejercicio 5.2a

Calcule el coeficiente de Spearman entre:

A y *B*, *A* y *C*, *A* y *X*

usando los pasos explicados anteriormente.

Por los resultados es evidente que *A* y *B* están completamente de acuerdo en sus preferencias, mientras que *A* y *C* tienen preferencias opuestas. Los criterios de *D* no corresponden de ninguna manera con los criterios de *A*; no hay ninguna correlación significativa.

Ejercicio 5.2b

Teniendo los números 1-6 en el orden siguiente: 1,2,3,4,5,6, ¿cuál es la probabilidad de obtener el mismo orden, sacando los números 1-6 al azar?

Ejercicio 5.2c

Datos del Censo de población de México del año 1970.

Variable 1 (V1): porcentaje de la población económicamente activa en la agricultura.

Variable 2 (V2): porcentaje de personas con ingresos menores de 500 pesos mensuales. Ponga en orden los valores de las dos variables y calcule el coeficiente de correlación de Spearman.

Estado	*V1*	*V2*
Baja California	22	12
Coahuila	30	34
Chiapas	73	69*
Distrito Federal	2	15
Guerrero	52	55
Nuevo León	17	23
Oaxaca	72	68
San Luis Potosí	53	59*
Tlaxcala	55	58
Yucatán	54*	65

* Estos valores eran 55, 68 y 58. Se modificaron en este ejercicio, para facilitar el cálculo eliminando la necesidad de que tuvieran rangos iguales.

CUADRO 5.2b. Coeficientes de confianza de Spearman

N (*Número de casos*)	.05 (*95% de confiabilidad*)	.01 (*99% de confiabilidad*)
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746
12	.506	.712
14	.456	.645
16	.425	.601
18	.399	.564
20	.377	.534
22	.359	.508
24	.346	.485
26	.329	.465
28	.317	.448
30	.306	.432

(Fuente: S. Siegel, *Non-Parametric Statistics for the Behavioral Sciences*, p. 284).

5.3 *La prueba de Pearson*

El coeficiente de correlación 'r' de la prueba Pearson, *producto momento*, se aplica a datos con intervalos. Se supone que la distribución de valores de las dos variables tiene aproximadamente la forma de una curva normal, con muchos valores cerca del valor de la media. La prueba de Pearson es más exacta que la prueba de Spearman porque conserva los intervalos entre los valores observados, mientras que la prueba de Spearman pierde esta información debido a su forma de rangos, con intervalos de 1 entre cada observación. La prueba de Pearson necesita más cálculos que la prueba de Spearman. Sin embargo, tienen ciertos pasos en común. Hay varias fórmulas para el cálculo de la prueba de Pearson. Aquí se considera solamente una, el método más conveniente cuando se escribe un programa para computadora.

En el ejemplo que sigue se usan las mismas variables que se usaron en el ejercicio 5.2c (sección 5.2).

Para reducir los cálculos, los valores originales se expresan en 'por 10' en vez de 'por cien'.

V1: *agricultura* (*X*) *V2*: *ingresos bajos* (*Y*)

	V1 (*X*)	*V2* (*Y*)	X^2	Y^2	*XY*
Baja California	2	1	4	1	2
Coahuila	3	3	9	9	9
Chiapas	7	7	49	49	49
Distrito Federal	0	2	0	4	0
Guerrero	5	6	25	36	30
Nuevo León	2	2	4	4	4
Oaxaca	7	7	49	49	49
San Luis Potosí	5	6	36	36	30
Tlaxcala	6	6	36	36	36
Yucatán	6	7	36	49	42
	43	47	238	273	251

Nótese: *X* y *Y* no son valores *en rango*.

Paso 1. Escriba los valores de las variables *X* y *Y* en dos columnas.

Paso 2. Calcule los valores de X^2, Y^2 y *XY*.

Paso 3. Con estos datos se puede calcular la fórmula 'r' de Pearson:

$$R = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$R = \frac{10 \times 251 - 43 \times 47}{\sqrt{(10\times 237 - 43\times 43)(10\times 273 - 47\times 47)}}$$

$$R = \frac{2{,}510 - 2{,}021}{\sqrt{(2{,}370 - 1{,}849)(2{,}730 - 2{,}209)}}$$

$$\frac{489}{\sqrt{521 \times 521}}$$

$$R = +.94$$

Nótese que la prueba de Spearman dio un coeficiente de +0.93 para los mismos datos (ejercicio 5.2a), índice menor debido a la pérdida de información en el proceso de transformar los datos con intervalos en datos ordinales (1,2,3...10).

El cuadro 5.3a da los coeficientes de confianza de Pearson. El coeficiente de correlación de Pearson siempre se encuentra en-

tre −1 y +1; un resultado alrededor de cero significa que no hay correlación.

5.4 *La prueba chi-cuadrado* (X^2)

La prueba X^2 (X es una letra griega) difiere de las pruebas Spearman y Pearson en los aspectos siguientes:

1. Usa datos nominales y frecuencias (*pero no porcentajes*). La palabra frecuencia, en estadística significa 'número de' o 'cantidad de' casos específicos o discretos, por ejemplo: 45 estudiantes, 100 camiones o 22 ciudades.

2. Los datos para el cálculo se presentan no en forma de dos renglones o columnas, sino en forma de una tabla o matriz de contingencias con M renglones y N columnas.

3. El coeficiente X^2 se encuentra entre 0 y un número muy grande; tiende al infinito.

El ejemplo que sigue ilustra una situación en la cual conviene aplicar la prueba X^2.

En un estudio sobre choques de camiones de carga en las carreteras, en una muestra se han notado las frecuencias (cantidades) siguientes:

Camiones marca *A*, con defecto mecánico,	13
Camiones marca *A*, sin defecto mecánico,	7
Camiones marca *B*, con defecto mecánico,	7
Camiones marca *B*, sin defecto mecánico,	13

Esta información puede expresarse en forma de una tabla de contingencias, más concisamente:

	Marca A	*Marca B*	*Total*
Con defecto	13	7	20
Sin defecto	7	13	20
Total	20	20	40

Nótese que, en este caso, el número de camiones de marca A es igual al número de camiones de marca B y que el número de camiones con defecto, de la marca A, es igual al número de camiones, sin defecto, de la marca B. En general, esta simetría no se encuentra en las tablas de contingencias.

Se puede preguntar, entonces, ¿qué justificación hay para suponer que mecánicamente la marca B es superior a la marca A? Hay que tener en cuenta que 40 camiones forman una muestra muy pequeña, y que posiblemente hay 50,000 camiones de cada marca en el país. La prueba X^2 da un índice que permite consultar un cuadro de coeficientes de confianza. De esta manera se puede saber si, en la muestra, la preponderancia de camiones marca A, con defecto, se hubiera obtenido por casualidad.

Con los resultados siguientes se concluye, intuitivamente, que no hay diferencia entre las dos marcas:

	Marca A	*Marca B*
Con defecto	10	10
Sin defecto	10	10

Al contrario, sería inevitable concluir que la marca A es inferior a la marca B, con los resultados siguientes:

	Marca A	*Marca B*
Con defecto	19	1
Sin defecto	1	19

Para calcular el índice X^2 es necesario, primero, calcular valores esperados, para compararlos con los valores observados.

Si se considera por el momento el caso de una moneda, se sabe que echándola al aire cae en un lado o en el otro. Hay dos resultados posibles, por ejemplo: en México, sol y águila. Teóricamente, la probabilidad de obtener sol ($P(S)$) es igual a la probabilidad de obtener águila ($P(A)$). Echando una moneda 40 veces, uno esperaría obtener igual número de sol y de águi-

la, esto es, 20 y 20. En realidad, una prueba no daría siempre esta combinación teórica. Se obtendría 22 y 18 o 24 y 16, por casualidad, fácilmente.

Sería mucho más difícil obtener 35 y 5 (o 5 y 35).

El ejemplo geográfico que sigue ilustra la prueba X^2.

El estudio de 40 ejidos de una región dio la información siguiente:

13 ejidos de la llanura tienen tractor
7 ejidos de la llanura no tienen tractor
7 ejidos de la zona montañosa tienen tractor
13 ejidos de la zona montañosa no tienen tractor.

La tabla siguiente presenta la información:

	Ejidos en	
	Llanura	*Montaña*
Sin tractor	7	13
Con tractor	13	17

Se puede afirmar que la preponderancia de mecanización en la llanura es significativa, o hay que concluir que, debido al reducido número (40) de casos de la muestra sería fácil obtener las proporciones de la tabla, por casualidad. Si no hubiera ninguna relación (o correlación) entre la forma del terreno y el progreso de la mecanización, sería justificado y estadísticamente correcto esperar igual proporción de ejidos con tractor, en la llanura que en la montaña, así como, también, de ejidos sin tractor. Los valores 'esperados' o 'teóricos' serían, entonces, 10, 10, 10 y 10.

La prueba X^2 se calcula con la fórmula siguiente:

$$X^2 = \Sigma \frac{(0 - E)^2}{E}$$

Σ significa sumar todos los valores que siguen.

0 es un valor observado.
E es el valor correspondiente, esperado.

En el ejemplo de los 40 ejidos la tabla de contingencias se expresaría así:

013 / E 10	07 / E 10
07 / E 10	013 / E 10

Es mejor hacer el cálculo paso a paso, de la manera siguiente:

0	E	$(0-E)$	$(0-E)^2$	$\frac{(0-E)^2}{E}$
13	10	3	9	0.9
7	10	−3	9	0.9
7	10	−3	9	0.9
13	10	3	9	0.9
				3.6

Resultado: $X^2 = \Sigma \frac{(0-E)^2}{E} = 3.6$

Hay que consultar el cuadro de los coeficientes de confianza de X^2 (cuadro 5.4a), para averiguar si el resultado es significativo, esto es, si la relación entre terreno y mecanización aparente, en los datos observados, se hubiera obtenido, fácilmente, por casualidad. Cuando la tabla de contingencia tiene únicamente 2 renglones y 2 columnas, posee solamente un grado de libertad.* Como el valor de 3.6 es menor que el coeficiente de confianza de 95 % (0.05), esto es, 3.84, hay que aceptar que

* Un ejemplo ayudará a entender el concepto:

	Tipo A	*Tipo B*	
CON X	13		20
SIN X			20
	20	20	40

Como el número de cada categoría (A y B, con X y sin X) ya se conoce, una vez que se sabe que hay 13 en la primera posición de la tabla, los otros valores siguen inevitablemente.

el resultado se encuentra en la zona de valores que se obtendrían, por casualidad, más de una vez en 20 (probabilidad de 1/20 o 0.05).

Ejercicio 5.4a

Calcule los índices de chi-cuadrado para las situaciones siguientes: el valor E es igual a 10 en los dos ejemplos:

(*i*)	*Zona lluviosa*	*Zona seca*
Cultivan algodón	3	17
No cultivan algodón	17	3

(*ii*)	*Zona lluviosa*	*Zona seca*
Cultivan maíz	10	10
No cultivan maíz	10	10

En muchas muestras el número total de casos en un renglón de una columna *no es* igual al número de casos en otro renglón o en otra columna. Se presenta, ahora, el método general para calcular los valores *esperados*.

R significa renglón, $R1$ el primer renglón, $R2$ el segundo.
C significa columna, $C1$ la primera columna, $C2$ la segunda.
RT significa total de casos en el renglón.
CT significa total de casos en la columna.
GT significa gran total (suma de los valores de RT o de CT).

	$C1$	$C2$	
$R1$	$R1C1$	$R1C2$	$RT1$
$R2$	$R2C1$	$R2C2$	$RT2$
	$CT1$	$CT2$	GT

Ejemplo: valores observados.

	$C1$	$C2$	RT
$R1$	21	37	58
$R2$	31	11	42
CT	52	48	100 (GT)

Los valores esperados se calculan así:

	$C1$	$C2$
$R1$	$(RT1 \times CT1)/GT$	$(RT1 \times CT2)/GT$
$R2$	$(RT2 \times CT1)/GT$	$(RT2 \times CT2)/GT$

Usando el ejemplo:

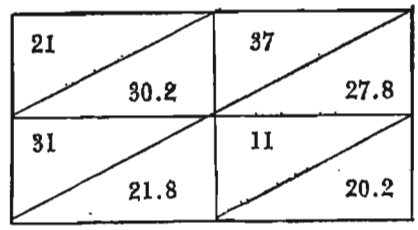
Paso 1. Presentar los datos para el cálculo

$\frac{58 \times 52}{100}$	$\frac{58 \times 58}{100}$	58
$\frac{42 \times 52}{100}$	$\frac{42 \times 48}{100}$	42
52	48	100

Paso 2. Calcular los valores esperados.

	$C1$	$C2$
$R1$	30.2	27.8
$R2$	21.8	20.2

Paso 3 (opcional): comparar valores de 0 con valores de E.



Paso 4. Calcular X:

$R1C1$	21	30.2	−9.2	84.5	2.8
$R1C2$	37	27.8	−9.2	84.5	3.0
$R2C1$	31	21.8	−9.2	84.5	3.9
$R2C2$	11	10.2	−9.2	84.5	4.2
					13.9

$X^2 = 13.9$, un índice significativo al nivel de confianza, de 95 % (6.635).

Ejercicio 5.4b Muchas veces la tabla de contingencias para la prueba X^2 tiene más de dos renglones o de columnas.

El ejemplo que sigue tiene tres columnas.

Calcule el índice de X^2, usando los pasos del ejemplo anterior. La muestra tiene 70 casos.

	Ejidos		
	Montaña	*Colina*	*Llanura*
Con tractor	5	8	22
Sin tractor	15	12	8

Los grados de libertad de una tabla se calculan de la manera siguiente:

El número de renglones menos 1, multiplicado por el número de columnas menos 1.

Por ejemplo: una tabla con 4 renglones y 3 columnas tiene 6 grados de libertad:

$$(NR - 1 \times NC - 1 \text{ o } 3 \times 2)$$

Interpretación del índice de chi-cuadrado:

Ejemplo 1:

$X^2 = 3.1$, con 2 grados de libertad. 3.1 es menor que 5.991, de modo que no es significativo.

Ejemplo 2:

$X^2 = 16.5$, tabla con 4 grados de libertad. 16.5 es mayor que 13.277, de modo que es significativo con un grado de confiabilidad de 99 %.

Ejemplo 3:

$X^2 = 8.9$, tabla con 3 grados de libertad. 8.9 es mayor que 7.815, de modo que es significativo con grado de confiabilidad de 95 %, pero es menor que 11.345, con grado de 99 %.

Conclusiones: La prueba de X^2 es muy útil en la interpretación de los resultados de encuestas, sobre todo cuando hay respues Sí o NO.

Se usa cuando la información se divide en clases o categorías, como, por ejemplo, camiones marca *A* y *B*. Tiene la ventaja, también, de poder usar datos con intervalos modificados. La escala continua de observaciones de la cantidad de lluvias anuales en varios lugares puede dividirse en dos o tres categorías: lluvioso, seco; o lluvioso, seco, árido, poniendo los valores siguientes (lluvia en cm) en clases: 80, 65, 49, 46, 42, 20, 12, 10. *Lluvias*: 80, 65; *seco*: 49, 46, 42; *árido*: 20, 12, 10.

Dos puntos finales:

1. Recuerde que se usan solamente *frecuencias* y *números enteros* (los valores esperados pueden ser *reales*).
2. En la tabla todas las entradas deben tener por lo menos 5 casos.

Respuestas a los ejercicios de la sección 5.2a

A y *B*: $RHO = +1.0$
A y *C*: $RHO = -1.0$
A y *X*: $RHO = +0.08$

El cálculo *A* y *X* es el siguiente:

A	1	2	3	4	5	6	
X	3	1	6	5	4	2	
D	−2	1	−3	−1	1	4	
D^2	4	1	9	1	1	16	$\Sigma D^2 = 32$

$$RHO = 1 - \frac{6 \times 32}{216 - 6} = 1 - \frac{192}{210}$$

$$= 1 - 0.92 = +0.08$$

Ejercicio 5.2b

720

Ejercicio 5.2c

En este ejemplo las dos variables se escriben en dos columnas, no en dos renglones como en los otros dos ejemplos. N=10.

	V1	*V2*	*D*	D^2
Baja California	8	10	−2	4
Coahuila	7	7	0	0
Chiapas	1	1	0	0
Distrito Federal	10	9	1	1
Guerrero	6	6	0	0
Nuevo León	9	8	1	1
Oaxaca	2	2	0	0
San Luis Potosí	5	4	1	1
Tlaxcala	3	5	−2	4
Yucatán	4	3	1	1

$\Sigma D^2 = 12$

$$RHO = 1 - \frac{6 \times 12}{1{,}000 - 10} = 1 - \frac{72}{990}$$

$$= 1 - 0.07 = +0.93$$

El resultado es un coeficiente de correlación positivo, muy alto y significativo, el coeficiente de confianza con $N = 10$, siendo +0.746 del grado 99 %.

Respuestas a los ejercicios de la sección 5.4

5.4a (*i*) $X^2 = 19.6$, un índice significativo a nivel de confiabilidad de 99 %.

(*ii*) $X^2 = 0.0$, un índice que se obtendría fácilmente por casualidad y que, efectivamente, representa el índice más bajo de la escala de X^2.

5.4b

	O	*E*	$(O-E)$	$(O-E)^2$	$\frac{(O-E^2)}{E}$
R1C1	5	10	5	25	2.5
R1C2	8	10	2	4	.4
R1C3	22	15	7	49	3.3
R2C1	15	10	5	25	2.5
R2C2	12	10	2	4	.4
R2Cs	18	15	7	49	3.3

$X^2 = 12.4$. Con 2 grados de libertad, se encuentra más allá del valor 9.210 (ver cuadro 5.4a), de modo que la relación entre terreno y mecanización (tractores) es significativo.

6. ANÁLISIS DE FACTORES Y CLASIFICACIÓN MULTIVARIADA

Este capítulo se divide en dos secciones. La primera describe los pasos del análisis de factores, y la segunda un ejemplo de la clasificación sobre una base de muchas variables. El análisis de factores no debe confundirse con el análisis factorial. El análisis de *componentes principales* es un caso especial de análisis de factores.

6.1 *Análisis de factores*

La aplicación del análisis de factores ha aumentado en los últimos diez años en varias ramas de la ciencia (psicología, geología, arqueología, por ejemplo). Desde el año 1960, aproximadamente, se ha aplicado en la geografía, siendo los primeros en hacerlo algunos geógrafos de los Estados Unidos, en particular B. J. Berry. En muchas publicaciones sobre geografía, ya sea física o humana, se encuentra el uso de este método. Es necesario, por lo menos, entender los conceptos, los pasos matemáticos y estadísticos y los defectos del análisis de factores.

En la geografía, para correlacionar dos o más distribuciones se usan varios métodos, entre los cuales se encuentran los siguientes:

1. Localizar dos o más variables (por ejemplo, cantidad de lluvia, altitud, temperaturas, cultivo de maíz) en una serie de mapas, y compararlas.
2. Representar gráficamente la relación entre dos o, más difícilmente, tres variables (ver capítulo 1).

3. Calcular un índice de correlación entre un par de variables (ver capítulo 5).

4. Identificar familias de variables semejantes, y agruparlas. Esto es la función del análisis de factores.

En términos generales, se trata de una matriz de datos iniciales con determinado número (N) de casos (por ejemplo, estaciones meteorológicas, cuencas fluviales, ciudades, estados, zonas dentro de una ciudad, fábricas) y con determinado número (M) de variables. El método de análisis de factores busca correlaciones entre las variables. Si hay intercorrelación entre las variables, las expresa en forma de un número menor de factores o componentes principales. De esta manera se reduce el número inicial de variables a algunos factores o ejes que representen familias de variables intercorrelacionados.

En este estudio se usa como ejemplo una matriz de datos demográficos y económicos de las 32 entidades de México. Se usaron datos preliminares del censo de 1970. Cuando se hizo el estudio, los datos disponibles eran limitados.

Las 11 variables usadas fueron las que se encuentran en el cuadro 6.1a.

En el cuadro 6.1b se encuentran la media y la desviación estándar de las variables.

Las variables tomadas en cuenta son las siguientes:

1. Densidad de la población, habitantes por kilómetro cuadrado.
2. Aumento de la población, 1950-1970 (1950 = 100).
3. Proporción de hombres a mujeres (mujeres = 100).
4. El porcentaje de personas ocupadas en la agricultura, con respecto a la población total.
5. El porcentaje de personas ocupadas en la industria de transformación, con respecto a la población total.
6. Personas ocupadas en los servicios, como porcentaje de la población total.
7. Población de 10 años y más, *alfabeta*, como porcentaje de la población total.
8. Personas con alguna instrucción, como porcentaje de la población total.
9. Viviendas que disponen de agua entubada, como porcentaje del número total de viviendas.
10. Viviendas con más de un cuarto, como porcentaje del número total de viviendas.
11. Personas con ingreso mensual de más de 500 pesos, como porcentaje de la población total.

Las características de la matriz del cuadro 6.1a son las siguientes:

1. El número de observaciones es mayor que el número de variables.
2. La distribución de los valores de las variables es aproximadamente normal. Cuando el número de casos es muy grande (mayor de 100), la condición de normalidad tiene menor importancia. Para que los datos que no tienen una distribución normal la tengan, hay métodos para convertirlos usando, por ejemplo, la transformación logarítmica.
3. Se convirtieron los valores absolutos en relativos, con el fin de evitar las dificultades que se presentan en el manejo de datos debido a las variaciones en el tamaño de las entidades.

Los pasos del análisis de factores son los siguientes:

Paso 1. Calcular el coeficiente 'r' de correlación * entre cada par de variables (cuadro 6.1a) y poner los resultados en una matriz de índices de correlación (cuadro 6.1c). La matriz tiene índices de 1.00

* Ver capítulo 5.3, la correlación de Pearson.

CUADRO 6.1a

Estados	*Variables*										
	1	2	3	4	5	6	7	8	9	10	11
1. Aguascalientes	61	180	98	9	4	4	56	80	79	75	88
2. Baja California	12	384	100	6	5	6	59	79	67	76	97
3. Baja California (T)	2	210	105	9	2	5	59	80	67	67	95
4. Campeche	5	207	101	13	4	4	52	73	48	51	86
5. Coahuila	7	155	102	8	5	5	59	81	74	71	91
6. Colima	45	215	101	12	2	5	54	75	82	45	91
7. Chiapas	21	172	102	19	1	2	38	57	38	39	82
8. Chihuahua	6	191	102	9	3	5	67	80	66	70	92
9. Distrito Federal	4,550	225	93	1	10	10	64	86	96	71	95
10. Durango	8	149	104	13	1	3	56	78	53	69	88
11. Guanajuato	73	171	101	12	4	3	42	61	56	64	88
12. Guerrero	25	174	100	15	2	2	36	55	38	39	87
13. Hidalgo	57	140	101	15	3	2	41	62	48	54	84
14. Jalisco	41	189	98	9	6	5	54	74	66	72	90
15. México	179	275	102	8	6	4	48	69	63	62	91
16. Michoacán	39	163	101	14	2	2	44	62	53	55	88
17. Morelos	125	226	99	12	4	5	50	71	68	55	87
18. Nayarit	20	187	103	16	2	3	52	74	47	48	92
19. Nuevo León	26	229	101	5	9	6	61	85	81	64	93
20. Oaxaca	23	153	98	19	2	1	39	59	35	41	83
21. Puebla	74	154	99	15	4	3	45	66	48	53	84
22. Querétaro	41	181	100	13	3	3	40	62	52	54	87
23. Quintana Roo	2	326	108	15	2	3	49	74	40	60	86
24. San Luis Potosí	20	150	102	14	3	3	47	68	46	56	85
25. Sinaloa	22	199	104	14	2	4	52	75	51	53	94
26. Sonora	6	215	101	10	3	5	59	79	68	74	96
27. Tabasco	31	212	103	15	2	3	49	72	34	44	86
28. Tamaulipas	18	203	99	9	3	5	58	79	67	59	91
29. Tlaxcala	108	148	103	14	4	2	51	75	50	56	86
30. Veracruz	52	188	101	14	2	3	47	67	51	53	86
31. Yucatán	17	147	100	15	3	3	52	70	42	51	83
32. Zacatecas	13	143	100	15	1	2	52	75	43	68	87

en la diagonal principal, pues representan la correlación de cada variable consigo misma.

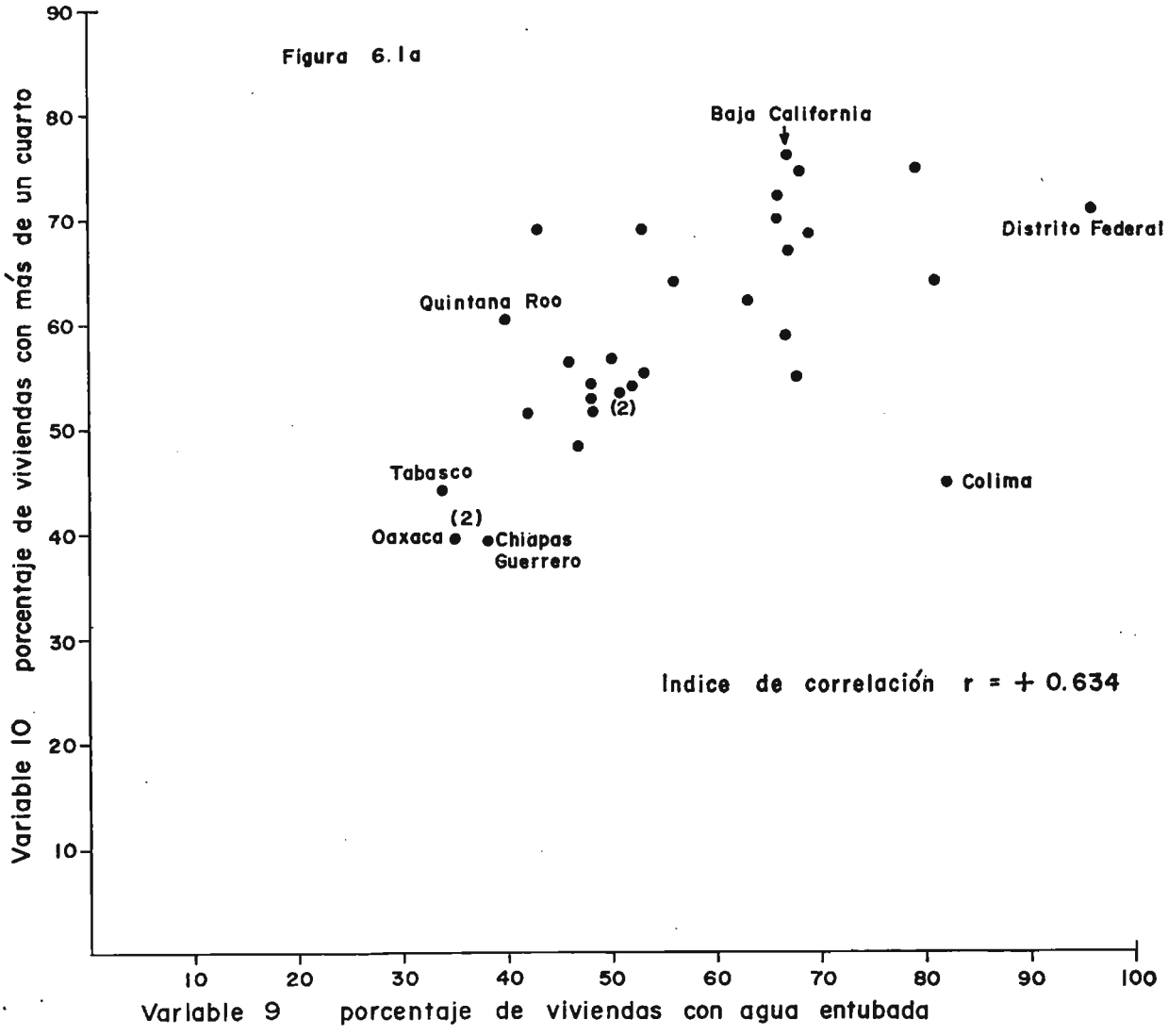
Se trata de una matriz cuadrada (número igual de renglones y columnas) y simétrica. Por esta última razón es suficiente usar únicamente la parte inferior izquierda de la matriz (por ejemplo, la correlación entre variables 5 y 8 es igual a la correlación entre variables 8 y 5).

Tomando como ejemplo el índice de correlación entre las variables 9 (viviendas con agua entubada) y 10 (viviendas con más de un cuarto), se encuentra en la matriz 6.1c, renglón 10 con columna 9, un coeficiente de +0.634, es decir, una correlación relativamente alta y positiva entre las dos variables. Este índice es significativo al nivel de confianza de 99 % (ver cuadro 5.3a). En la figura 6.1 está representada gráficamente esta relación.

CUADRO 6.1b

Media y desviación estándar de las 11 variables

Variable	*Media*	*Desviación*
1	179.0	786.0
2	105*7	52.0
3	101.0	2.5
4	12.0	3.9
5	3.4	2.0
6	3.8	1.7
7	51.0	7.7
8	72.0	8.0
9	56.8	15.1
10	58.1	10.7
11	88.7	4.0



Se pueden analizar los datos de la matriz 6.1c de la manera siguiente:

(*i*) Encontrar el índice de correlación más alto; en este caso, el de +0.961 entre variables 7 (alfabetismo) y 8 (instrucción). Esta correlación muy alta es de esperarse porque las dos variables expresan aproximadamente la misma cosa. Se puede representar esta relación en un diagrama (ver figura 6.1b).

(*ii*) Encontrar el índice de correlación más alto después del primero (paso (i)), en este caso el de −0.902 entre las variables 4 (agricultura) y 6 (servicios). Esta correlación negativa muy alta es de esperarse también porque, en general, los servicios están relativamente menos desarrollados en las zonas agrícolas que en otras zonas. Esta relación también se representa en el diagrama.

(*iii*) Estos pasos se repiten hasta que todas las variables con correlación mayor a un valor determinado estén representadas en el diagrama. En este ejemplo se han incluido todas las variables con correlación mayor de ±0.70 con, por lo menos, una variable diferente. Se puede identificar una familia de variables que reflejan el desarrollo. Las otras tres variables no tienen mucha relación con esta familia.

Paso 2. Si se analiza la matriz de índi-

CUADRO 6.1c

Matriz de índices de correlación *r* (la prueba producto-momento de Pearson) entre cada par de variables

Variable											
1	1.000										
2	0.103	1.000									
3	−0.572	0.151	1.000								
4	−0.514	−0.446	0.361	1.000							
5	0.593	0.296	−0.522	−0.797	1.000						
6	0.650	0.473	−0.394	−0.902	0.732	1.000					
7	0.292	0.313	−0.061	−0.750	0.431	0.766	1.000				
8	0.303	0.340	−0.009	−0.729	0.454	0.742	0.961	1.000			
9	0.471	0.287	−0.411	−0.891	0.679	0.862	0.710	0.693	1.000		
10	0.208	0.203	−0.122	−0.744	0.451	0.567	0.711	0.696	0.634	1.000	
11	0.275	0.519	−0.034	−0.761	0.404	0.742	0.722	0.694	0.700	0.608	1.000
Variable	1	2	3	4	5	6	7	8	9	10	11

ces de correlación, es evidente que en este estudio de México hay mucha intercorrelación entre la mayor parte de las variables. Matemáticamente es posible agrupar variables altamente intercorrelacionadas en sus factores.[1]

[1] Hay que distinguir aquí entre el uso de la palabra factor en un sentido técnico y su uso general en el sentido de influencia.

En el álgebra de matrices los factores son *eigenvectores*. Cada *eigenvector* tiene un tamaño determinado. Cuando se tienen once variables, como es el caso en este estudio, hay once *eigenvectores*. El largo de los *eigenvectores* depende del grado de correlación entre las variables. Como hay mucha intercorrelación en este estudio, el

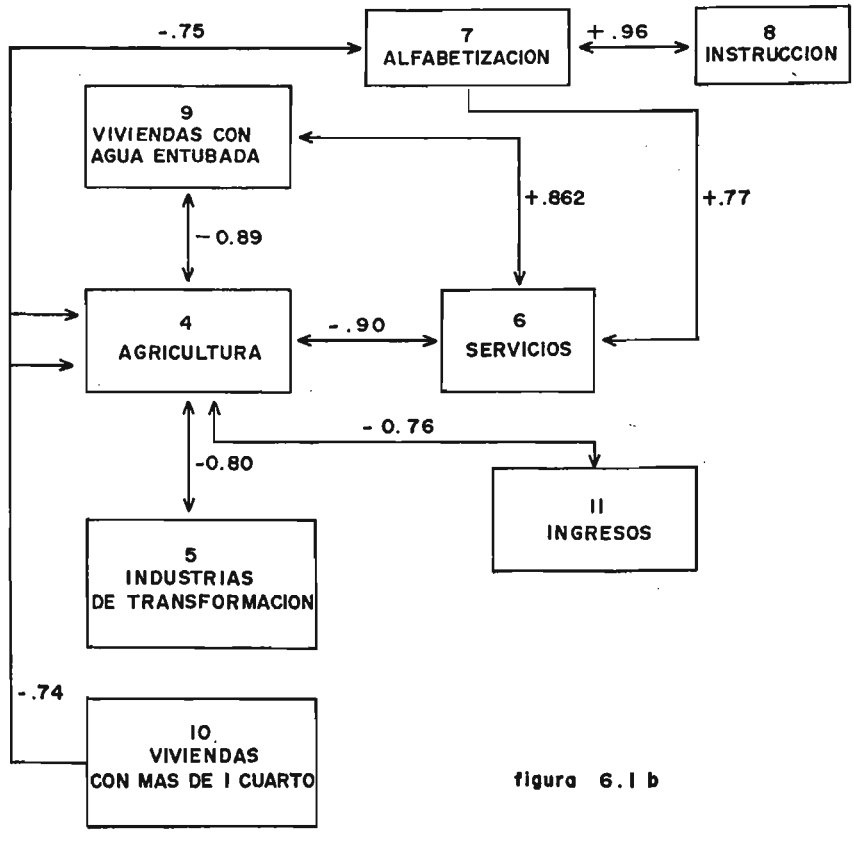


figura 6.1 b

primer *eigenvvector* es muy largo: tiene un *eigenvalor* de 6.498 en un total de 11.0 (ver cuadro 6.1d). Se considera que el número de unidades de variación es igual al número de variables, de manera que el primer factor o *eigenvector* representa 59.1 por ciento de la variación total (6.498 entre 11.00). El segundo factor (ver cuadro 6.1d) tiene un *eigenvalor* de 1.797, el cual representa 16.3 por ciento de la variación total. Los otros factores tienen menor tamaño e importancia. Generalmente los factores con un *eigenvalor* menor de 1.00 se excluyen de la parte siguiente de este estudio, porque tienen menos peso que cada una de las variables solas.

CUADRO 6.1d

Componente	*Eigenvalor*	*Porcentaje de la variación total*	*Porcentaje acumulado*
I	6.498	59.1	59.1
II	1.797	16.3	75.4
III	0.900	8.2	83.6
IV	0.532	4.8	88.4
V	0.386	3.5	91.9
VI	0.328	3.0	94.9
VII	0.274	2.5	97.4
VIII	0.162	1.5	98.9
IX	0.063	0.6	99.5
X	0.036	0.3	99.8
XI	0.024	0.2	100.0
Total	11.000		

En este estudio, entonces, ha sido posible concentrar 75 por ciento de la variación total de 11 variables (y 11 unidades de variación) en dos factores; una reducción notable, con una pérdida de información de solamente 25 %. Esto ilustra una de las ventajas del análisis de factores: su capacidad de economizar y de concentrar la información.

Se puede apreciar mejor el significado de los *eigenvectores* si se consideran las tres matrices de índices de correlación en el cuadro 6.1e. La matriz *A* representa la situación que se encuentra en el estudio de México, pero en forma reducida (4 variables en vez de 11). Estas correlaciones darían *eigenvectores* con valores aproximadamente de: 2.5, 1.0, 0.3 y 0.2 (total, 4).

En el caso de la matriz *B*, los *eigenvectores* tendrían los valores:

4.0 0.0 0.0 0.0 (total, 4)

Y en el caso de la matriz *C*, los siguientes:

1.0 1.0 1.0 1.0 (total, 4)

En general, el grado de contraste entre el tamaño de los *eigenvectores* depende del nivel de intercorrelación (o covariación) en la matriz de índices de correlación.

Una explicación geométrica o gráfica de los *eigenvectores* o 'ejes' puede ayudar también al entendimiento de lo que son. En la figura 6.1c hay dos ejemplos de correlaciones entre dos variables. En el primer caso hay una correlación alta entre las dos variables y un eje o *eigenvector* es mucho más largo que el otro. En el segundo caso

CUADRO 6.1e

	1.0	+.7	+.9	+.8		1.0	1.0	1.0	1.0
	+.7	1.0	+.6	+.9		1.0	1.0	1.0	1.0
A	+.9	+.6	1.0	+.9	*B*	1.0	1.0	1.0	1.0
	+.8	+.9	+.9	1.0		1.0	1.0	1.0	1.0
	1.0	0.0	0.0	0.0					
	0.0	1.0	0.0	0.0					
C	0.0	0.0	1.0	0.0					
	0.0	0.0	0.0	1.0					

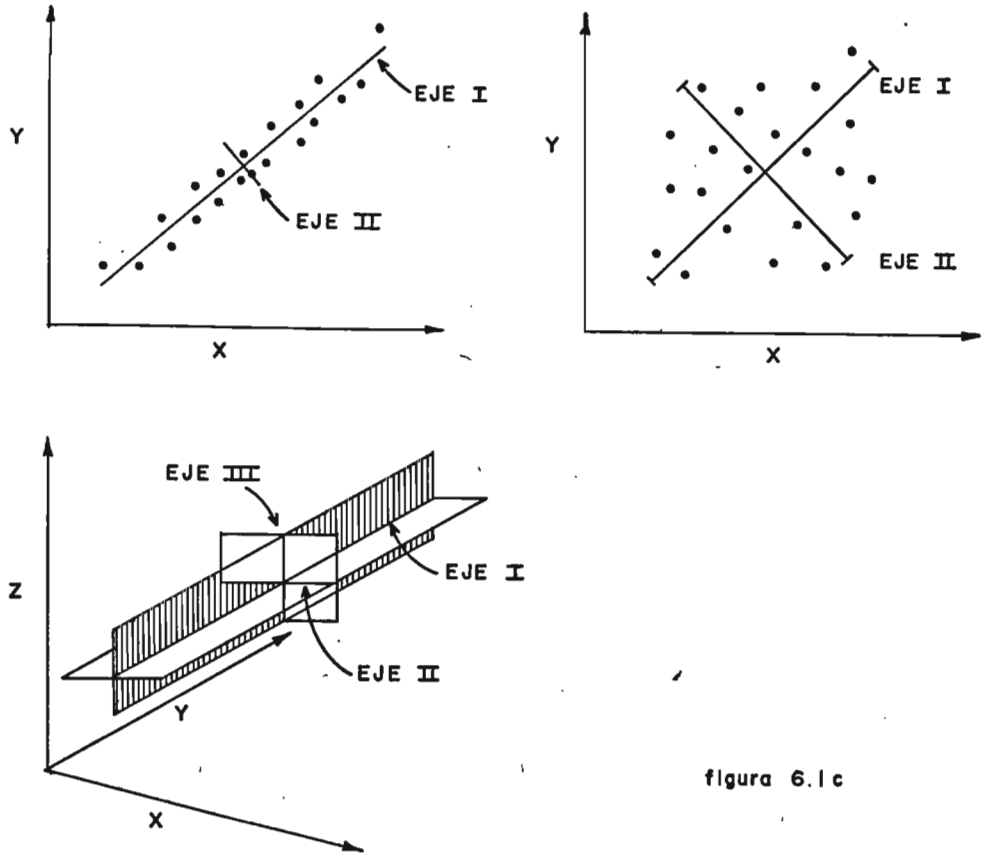


figura 6.1c

hay poca correlación entre las variables y ejes del mismo tamaño. También se representa en el diagrama el caso de correlaciones altas entre tres variables.

Paso 3. Los *eigenvectores* o factores tienen relación con las variables. En el estudio actual de México, las variables 4 a 11 tienen una correlación alta con el primer factor (factor 1). Las variables 1 (densidad de población) y 3 (proporción entre hombres y mujeres) tienen correlaciones altas con el factor 11. La variable 2 (aumento de la población) no se asocia mucho con ninguno de los dos factores.

En el cuadro 6.1f se encuentran los índices de correlación de cada variable con cada uno de los dos factores. Un índice de 1.00 indicaría una correlación perfecta con el factor, lo que significaría que toda la relación de la variable se representa en el factor. En este ejemplo, la variable 4 (agricultura) coincide casi exactamente con el primer factor. Un índice negativo bajo (hacia −1.00) indica también una asociación fuerte con el factor, pero en el otro sentido (variables 5 a 11 con el factor I). Se notará que el signo de la variable 4 (agricultura) es positivo en el factor I y que los signos de todas las otras variables, desde 5 a 11, son negativos. Esto refleja la correlación negativa o contraria entre agricultura y las otras variables, desde 5 a 11, en la matriz de índices de correlación (cuadro 6.1c).

Es posible cambiar la posición de los *eigenvectores* en relación a las variables, por una rotación. El resultado del tipo de rotación VARIMAX se encuentra en la segunda parte (ii) del cuadro 6.1f. Se puede decir que tiene el efecto de considerar la misma situación desde otro punto de vista. A veces la rotación representa más nítidamente la existencia de familias de variables.

CUADRO 6.1f

(i) *Factores 'principal axis'*

Variables	*Factores*		*Factores: Porcentaje de variación*	
	I	*II*	*I*	*II*
1. Dens. Pob.	−0.57	−0.61	32	37
2. Aum. Pob.	−0.46	0.37	21	14
3. Hom.-Muj.	0.35	0.83	12	69
4. Agric.	0.96	0.05	93	0
5. Indust.	−0.75	−0.42	57	17
6. Serv.	−0.95	−0.11	90	1
7. Alfab.	−0.85	0.33	72	11
8. Instr.	−0.84	0.34	70	12
9. Viv. agua	−0.70	−0.10	81	1
10. Viv. 2+cuartos	−0.75	0.25	56	6
11. Ingr. + 500 P	−0.81	0.33	65	11

(ii) *Factores 'varimax rotation'*

Variables	*Factores*		*Factores: Porcentaje de variación*	
	I	*II*	*I*	*II*
1. Dens. Pob.	−0.16	−0.82	2	67
2. Aum. Pob.	−0.59	0.07	35	0
3. Hom.-Muj.	−0.14	0.89	2	79
4. Agric.	0.79	0.55	63	30
5. Indust.	−0.42	−0.75	18	56
6. Serv.	−0.76	−0.60	56	36
7. Alfab.	−0.89	−0.17	80	3
8. Instr.	−0.89	−0.15	80	2
9. Viv. Agua	−0.71	−0.56	50	32
10. Viv. 2+cuartos	−0.77	−0.19	59	3
11. Ingr. + 500 P	−0.86	−0.14	74	2

Paso 4. Para tener una idea más clara de la asociación (o falta de asociación) entre las variables y los factores, se pueden modificar los índices de correlación de la manera siguiente:

Multiplicar el índice por sí mismo y multiplicar el resultado por 100. Ejemplo: $0.57 \times 0.57 \times 100 = 32$.

Se puede decir que 32 % de la variación total de la primera variable se 'explica' en el primer factor. En el caso de la variable 4, 93 % de su variación total se concentra en el primer factor.

De esta manera se puede tener un índice numérico sucinto y preciso del grado de asociación de grupos de variables con sus familias. En geografía es muy raro encontrar correlaciones perfectas entre las variables. Es más común encontrar tendencias.

Paso 5. Los factores tienen dos propiedades importantes. En primer lugar, son ortogonales (se cruzan en ángulos rectos), aun cuando hay más de 3 factores. Esto significa que no hay ninguna correlación entre dos factores.[1] En segundo lugar, es posible dar a cada uno de los casos originales (las entidades de México en este

[1] Existen, sin embargo, métodos para sacar factores oblicuos, que no son ortogonales.

CUADRO 6.1g

Entidad	*I*	*II*	*Entidad*	*I*	*II*
1. Aguascalientes	−0.53	−0.62	17. Morelos	0.12	0.57
2. Baja California	−2.00	0.25	18. Nayarit	−0.04	0.85
3. Baja California (T)	−1.50	1.07	19. Nuevo León	−1.44	−0.71
4. Campeche	0.21	0.03	20. Oaxaca	2.04	−0.31
5. Coahuila	−1.00	−0.10	21. Puebla	1.11	−0.54
6. Colima	−0.37	0.04	22. Querétaro	0.91	−0.27
7. Chiapas	1.75	0.36	23 Quintana Roo	−0.60	1.97
8. Chihuahua	−1.29	0.44	24. San Luis Potosí	0.64	0.17
9. Distrito Federal	−0.92	−4.49	25. Sinaloa	−0.49	0.97
10. Durango	−0.47	1.03	26. Sonora	−1.30	0.36
11. Guanajuato	0.59	−0.25	27. Tabasco	0.13	0.77
12. Guerrero	1.65	−0.13	28. Tamaulipas	−0.61	−0.23
13. Hidalgo	1.23	−0.12	29. Tlaxcala	0.20	0.36
14. Jalisco	−0.36	−0.87	30. Vercruz	0.58	0.17
15. México	−0.49	−0.12	31. Yucatán	0.06	−0.10
16. Michoacán	0.70	0.14	32. Zacatecas	0.20	0.45

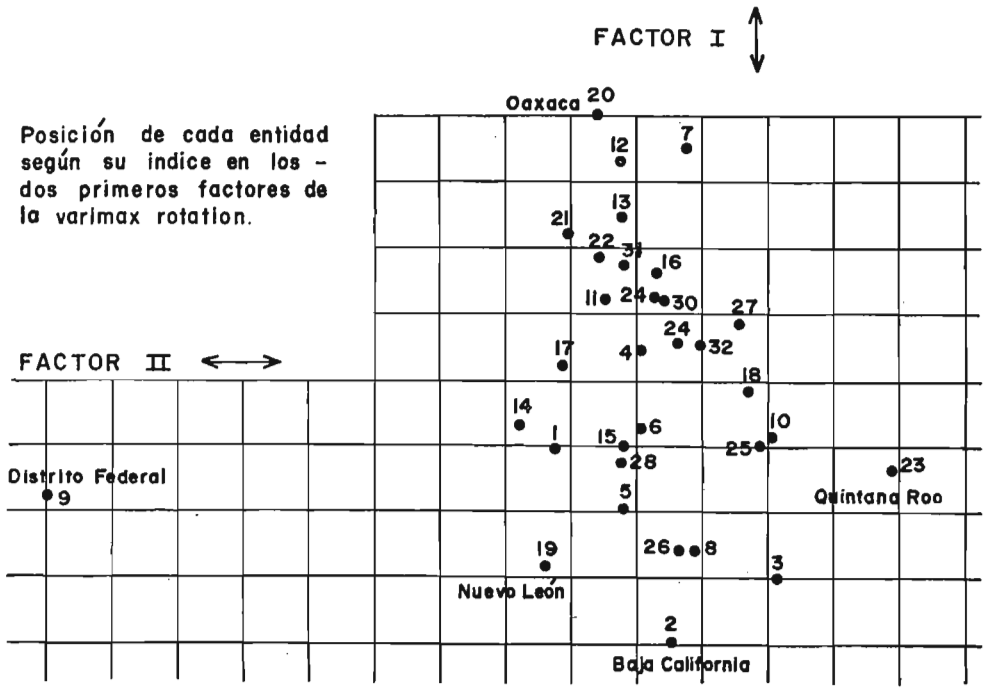


Figura 6.1 d

ejemplo) una posición en cada factor, la cual indica su índice en relación con la nueva familia de variables.

En el cuadro 6.1g las 32 entidades de México están en dos escalas: las de los factores I (desarrollo económico y social) y II (densidad de población y sexo). Los datos de cada columna tienen media de 0 (cero) y desviación estándar de 1. Los signos sirven para indicar la posición de cada entidad en las escalas, pero en el caso del primer factor los signos negativos no significan bajo nivel de desarrollo; su significado es exclusivamente matemático, porque ni el método ni la computadora 'entienden' si la agricultura o los servicios significan alto nivel de desarrollo en la vida mexicana.

Es posible representar gráficamente los dos factores en el cuadro 6.1g. En la figura 6.1d las 32 entidades tienen posiciones según sus índices en los factores. La posición del Distrito Federal lejos de las otras entidades se debe, sobre todo, a su densidad de población extremadamente alta en comparación con los estados (ver cuadro 6.1a, variable 1).

6.2 *La clasificación*

La clasificación se usa para reducir un gran número de elementos a un número menor de grupos o clases (o subconjuntos). Se emplean varios criterios. Uno de los métodos más importantes de la clasificación viene del concepto de agrupar elementos semejantes en determinado número de clases 'homogéneas', minimizando, al mismo tiempo, la pérdida de detalle o de información. Entre un número dado o inicial de elementos hay un par de elementos más semejantes que ningún otro par. Estos dos elementos se agrupan en una clase, de modo que el número total de elementos se reduce de uno. Después se ponen los dos elementos más semejantes que quedan. La primera clase, con sus dos elementos, se considera como un solo elemento nuevo. Este concepto fundamental de la clasificación se ilustra con dos ejemplos. Los dos

se clasifican en la base de más de dos variables y, por esto, se llaman métodos multivariados.

Ejemplo 1. En el primer capítulo (sección 1.6) se incluye el ejemplo de una matriz de índices de semejanza. Esta idea se extiende en el ejemplo actual. Hay 5 municipios (*A, B, C, D, E*). Se usan seis características (*A1, A2, A3, A4, A5,A6*) para clasificarlos (ver cuadro 6.2a).

CUADRO 6.2a

Municipio	*A1*	*A3*	*A4*	*A5*	*A2*	*A6*
A	0	0	0	0	1	1
B	0	0	0	0	1	0
C	0	0	0	0	0	1
D	1	1	1	0	0	0
E	1	1	1	1	0	0

A1. ¿Es una zona montañosa?
Sí = 1, No = 0
A2. ¿Es una zona lluviosa?
Sí = 1, No = 0
A3. ¿Es una zona fría?
Sí = 1, No = 0
A4. ¿Tiene población indígena?
Sí = 1, No = 0
A5. ¿Tiene centro urbano?
Sí = 1, No = 0
A6. ¿Predomina la agricultura?
Sí = 1, No = 0

Cuando el número de casos (municipios) y de variables no es grande, es fácil ver cuáles son los pares más semejantes. Por ejemplo, *A* y *B* se distinguen sólo por ser diferentes según el atributo *A6* (agricultura). Cuando hay mayor número de datos no es tan fácil identificar en el cuadro los pares más semejantes. Es posible calcular un índice de semejanza entre cada par de municipios. El método que sigue puede ser modificado.

Paso 1. Preparar una matriz con un número de renglones y de columnas igual al número de casos, para los resultados de los cálculos que siguen. En este ejemplo tiene 5 renglones y 5 columnas (ver cuadro 6.2b).

CUADRO 6.2b

	A	*B*	*C*	*D*	*E*
A	6				
B	5	6			
C	5	4	6		
D	1	2	2	6	
E	0	1	1	5	6

Paso 2. Comparar, por turno, cada par de municipios; por ejemplo, *A* con *B*. Estos municipios 'están de acuerdo' en 5 de los atributos. Se puede considerar, entonces, que tienen un índice de semejanza de 5/6, siendo 6 el máximo posible. Cuando los dos municipios tienen 1 y 1 o 0 y 0, están de acuerdo; cuando tienen 1 y 0 o 0 y 1, pierden un punto. En la matriz, la diagonal principial tiene valores de 6, el índice de semejanza entre cada municipio y él mismo.

Paso 3. Identificar los pares más semejantes (ver figura 6.2a).

En este ejemplo es fácil identificar dos grupos:

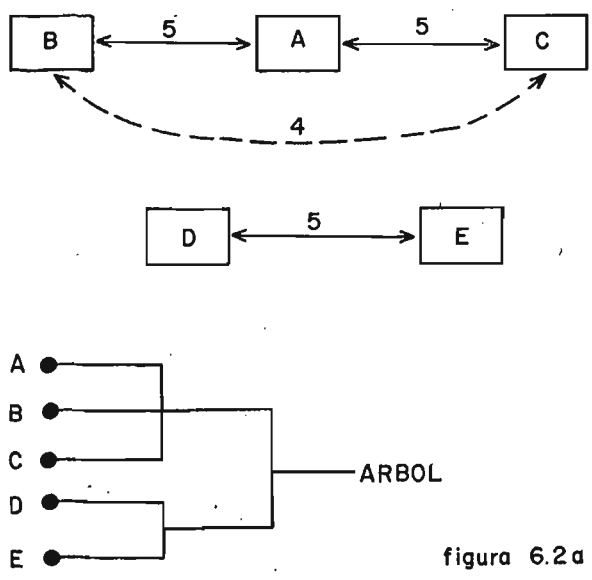


figura 6.2a

1. *A, B, C*
2. *D, E*

Se puede representar por dos grupos, en forma de un 'árbol'. Uniendo los municipios de esta manera se redúce el número de elementos de 5 a 2, con pérdida de poca información, 1 punto entre *D* y *E*, 4 puntos entre *A, B* y *C*. (*A* con *B*1, *A* con *C*1, *B* con *C*2).

Generalmente los grupos no se distinguen tan fácilmente. Se puede usar la computadora electrónica para hacer los cálculos y separar los grupos.

Es posible que en una clasificación se desee dar más énfasis a ciertos atributos que a otros. Un método para lograr pesos diferentes es aumentar la diferencia entre los extremos 0 y 1 en el caso de los atributos más importantes.

Se supone que *A*1 y *A*5 tienen más peso que los otros atributos.

*A*1. ¿Es una zona montañosa?
Sí = 3, No = 0
*A*5. ¿Tiene centro urbano?
Sí = 2, No = 0

Los otros atributos tienen extremos de 1 (Sí) y 0 (No). El cuadro inicial (6.2a) tiene que modificarse (ver cuadro 6.2c). Además, el índice máximo de semejanza ahora viene a ser 9, no 6. Los nuevos índices de semejanza se encuentran en el cuadro 6.2d. Los resultados han cambiado un poco las relaciones entre los municipios (ver figura 6.2b).

CUADRO 6.2c

Municipio	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*
A	0	0	0	0	2	1
B	0	0	0	0	2	0
C	0	0	0	0	0	1
D	3	1	1	0	0	0
E	3	1	1	1	0	0

CUADRO 6.2d

	A	*B*	*C*	*D*	*E*
A	9				
B	8	9			
C	7	6	9		
D	1	2	2	9	
E	0	1	1	8	9

Es posible modificar el método todavía más. Por ejemplo, en vez de preguntar si tiene o no centro urbano, se puede contestar de la manera siguiente a tres preguntas:

¿Tiene centro urbano grande? Sí = 2
¿Tiene centro urbano pequeño? Sí = 1
¿Tiene algún centro urbano? No = 0

Se puede introducir la idea de una escala con grande, mediano, pequeño, o muy importante, importante, sin importancia. En este caso el atributo, matemáticamente, tiene un peso de 2 (entre 0 y 2). Podría ser preciso aumentar 'en compensación' el peso de otros atributos, con sólo 0 y 1.

Ejemplo 2. En éste se usan los resultados de la aplicación del análisis de factores a los estados de México. En el cuadro 6.1g se encuentran las posiciones de las 32 enti-

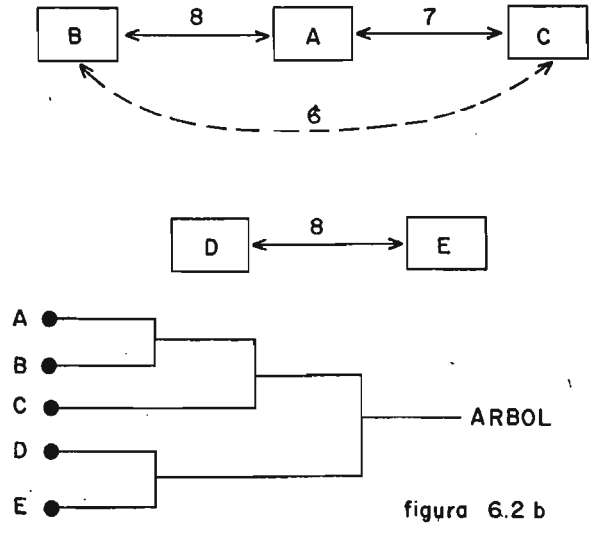


figura 6.2 b

dades de México en dos escalas, 0 factores. Cada factor representa ya un grupo de variables. En el cuadro 6.1f (ii) se indican las variables más asociadas con cada factor. Dado que los dos factores son ortogonales (cruzan en ángulos rectos) es posible usarlos como dos ejes, en una gráfica, y el valor de cada estado en los dos factores indica su posición en el 'espacio' de la gráfica. La distancia entre cada par de estados mide también su semejanza.

El par de estados más parecidos, según los datos usados en el estudio (económicos y demográficos), indican que Veracruz (30) y San Luis Potosí (24) son los más parecidos. Sus posiciones en los factores son las siguientes:

	Factor I	*Factor II*
Veracruz	0.58	0.17
San Luis Potosí	0.64	0.17

Son muy semejantes porque en la matriz de datos iniciales (cuadro 6.1a) tiene valores muy parecidos en casi todas las columnas. El par que sigue (ver cuadro 6.2c) son Sinaloa (25) y Durango (10). En la gráfica (figura 6.1c) están muy cerca el uno al otro.

El proceso de identificar pares de estados continúa hasta que todos los estados formen un solo grupo. Muchas veces un estado se asocia con un grupo existente, no con otro elemento individual. Por ejemplo, en el diagrama (figura 6.2c) que representa las etapas en el agrupamiento de los estados, Michoacán se une con un grupo existente, Veracruz y San Luis Potosí.

El árbol de la figura 6.2a representa, en forma de una red de tipo especial, la clasificación en grupos homogéneos de 32 elementos. Hay que tener en cuenta que el uso de otros datos iniciales (cuadro 6.1a) daría resultados diferentes. Por ejemplo,

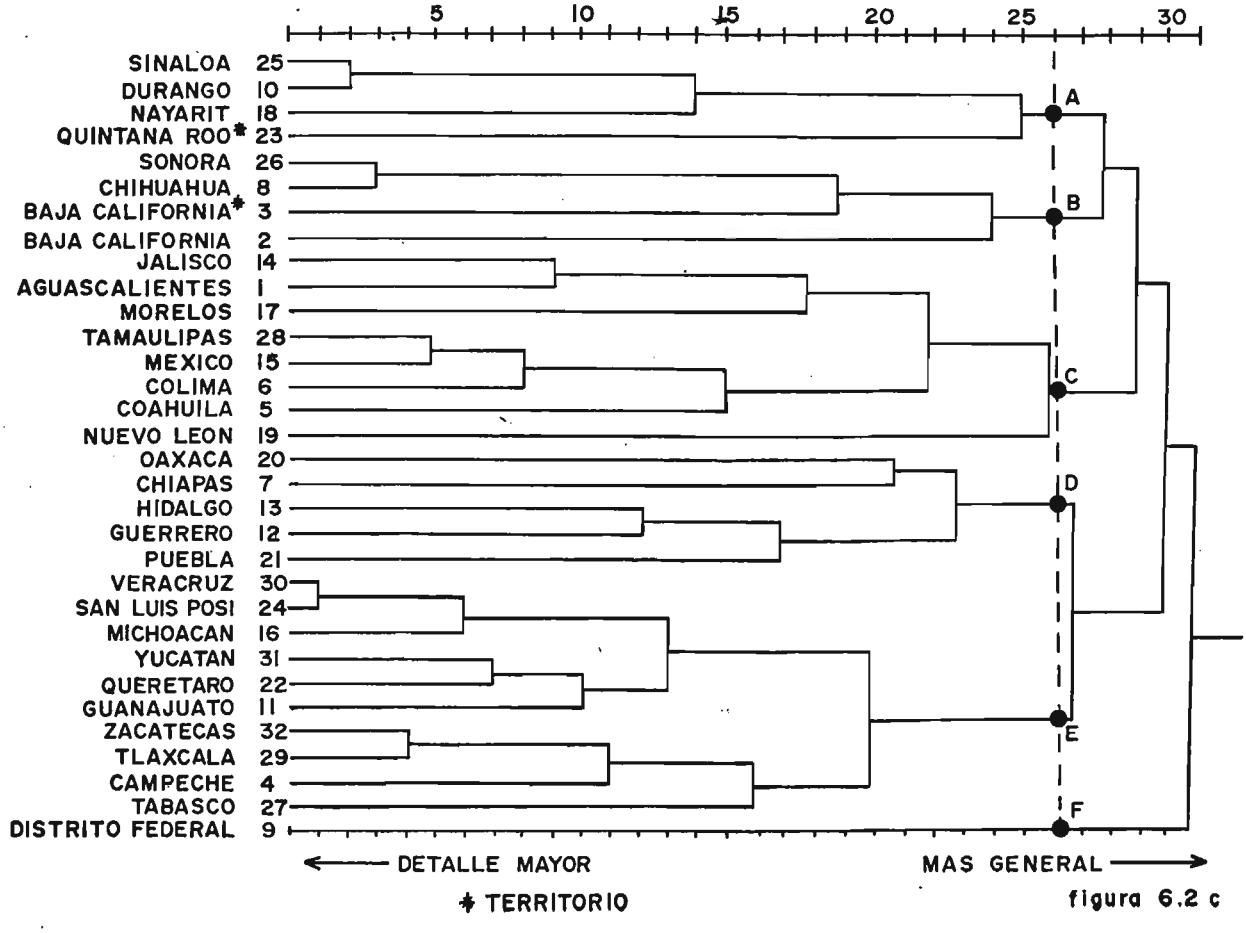


figura 6.2 c

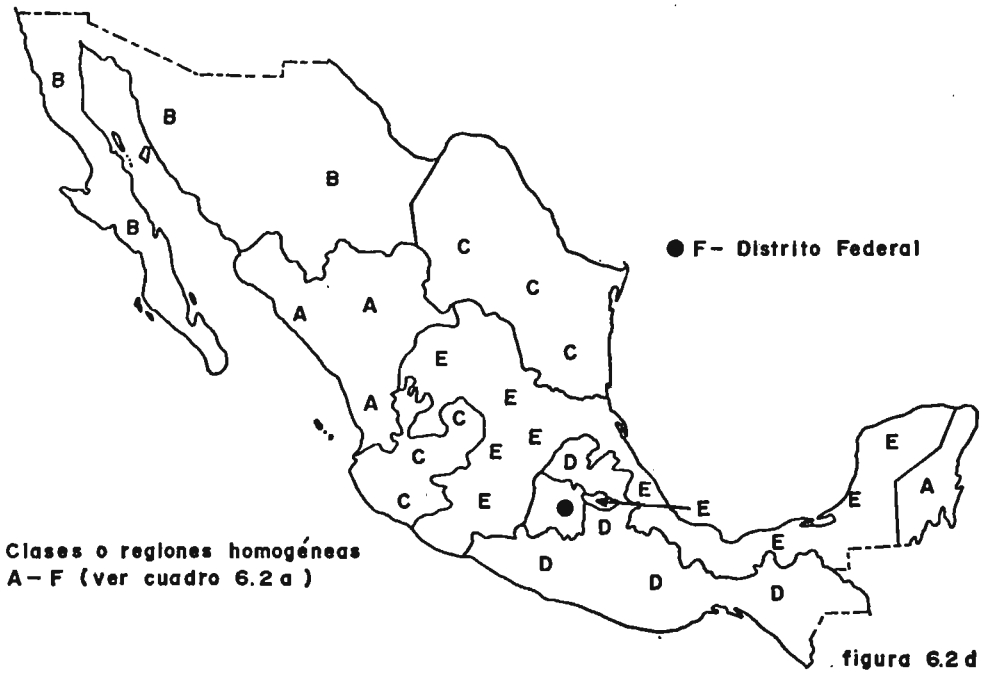


figura 6.2 d

la inclusión de datos físicos (cantidad de lluvia, altitud, vegetación) reduciría la proximidad de Veracruz y San Luis Potosí. La exclusión de la variable densidad de población movería al Distrito Federal más cerca de las otras entidades.

Es posible cortar el árbol en cualquier etapa de la clasificación. Sin embargo, hay poca ventaja en un corte que deje 26 o 24 regiones, y ninguna ventaja en un corte en que el árbol tenga solamente un ramo. En general, hay una 'zona' en la escala en la que la pérdida de información en el proceso de agrupamiento no ha sido muy notable, mientras que el número de elementos se ha reducido mucho. En este ejemplo sería conveniente tener entre 8 y 5 regiones.

Muchas veces se notan diferencias muy marcadas en el número de elementos (estados) en cada región homogénea. Por ejemplo, el corte de seis grupos (ver *A-F* en la figura 6.2c) da un grupo (*E*) con 10 elementos (Veracruz-Tabasco) y un grupo (*F*) sólo con un elemento (D. F.).

Otro aspecto notable del método usado es que muchas veces incluye en la misma región homogénea entidades que no son contiguas (ver figura 6.2d). Por ejemplo, la región *A* contiene 3 estados contiguos, Sinaloa, Durango y Nayarit, pero incluye también Quintana Roo, al otro extremo del país. Durante el proceso de agrupamiento de los elementos iniciales en regiones, es posible incluir como condición que dos entidades se unan solamente si son contiguas. Lo mismo pasaría cuando se une una entidad con un grupo existente.

7. PROGRAMACIÓN LINEAL Y TEORÍA DE LOS JUEGOS

La programación lineal y la teoría de los juegos son dos métodos semejantes, pero con objetivos y conceptos un poco diferentes. El primero minimiza el uso de recursos o maximiza la ganancia. El segundo estudia estrategias entre jugadores y la forma de decisiones. En este capítulo se consideran con ejemplos sencillos los aspectos básicos de los dos métodos. En realidad, muchas veces es preciso usar más variables que en los ejemplos que siguen, y

hacer cálculos con computadora electrónica

7.1 *Programación lineal*

La programación lineal es un medio para encontrar soluciones óptimas a problemas prácticos. Se considera que en geografía es posible aplicar programación lineal a dos tipos de situación o problema:

a) Problema de producción.
b) Problema de transporte.

a) *Problema de producción*

El dueño de una granja tiene 18 hectáreas. Puede cultivar solamente patatas o lechugas (en el problema sería posible incluir otros cultivos, pero complicaría los cálculos). Se quiere saber, dadas ciertas condiciones, cuántas hectáreas conviene cultivar con patatas y cuántas con lechugas. Las condiciones o límites son los siguientes:

1. El costo de producción por hectárea (incluso mano de obra) es de 5 libras por hectárea de lechugas y 3 por hectárea de patatas. Puede gastar hasta 60 libras en total.

2. Se necesitan 3 días-hombre por hectárea con lechugas y uno por hectárea con patatas. Se dispone de un máximo de 30 días-hombre.

3. Como tiene solamente 18 hectáreas de superficie total, la superficie máxima que puede cultivar con patatas y lechugas es 18.

4. No puede sembrar menos de 0 hectáreas, ni con patatas, ni con lechugas.

Al buscar la solución óptima, esto es, la combinación de patatas y lechugas que simultáneamente satisface todas las condiciones y maximiza su ganancia, debe tenerse en cuenta lo que se llama la función objetiva. La ganancia (P) es una función de lo que gana en términos de dinero por hectárea de lechugas y de patatas: 12 libras por hectárea de lechugas y 8 por hectárea de patatas.

Cuando hay solamente dos variables es posible solucionar el problema gráficamente, con la aplicación de desigualdades.

$x = y$ es una ecuación o igualdad. Representada gráficamente, es una línea.

$x \geqslant y$ es una igualdad. En la gráfica es una región, la parte de la gráfica en donde los valores de x son iguales a los valores correspondientes a y, o mayores.

Las ecuaciones de este problema son las siguientes (L significa lechugas, P patatas):

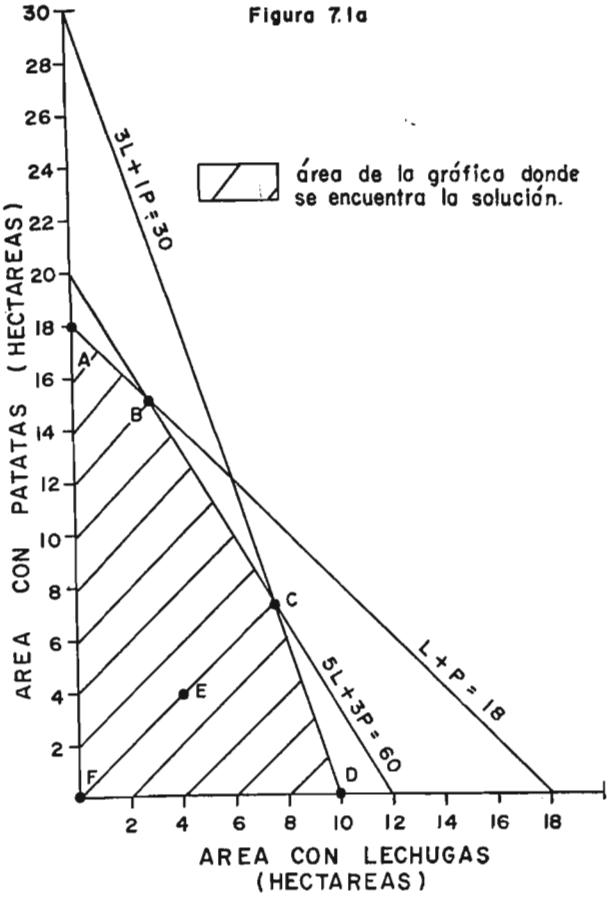
1. $5 \times L + 3 \times P \leqslant 60$
   (Condición 1: no puede gastar más de 60 libras)
2. $3 \times L + 1 \times P \leqslant 30$
   (Condición 2: tiene solamente 30 horas)
3. $L + P \leqslant 18$
   (Condición 3: tiene solamente 18 hectáreas)
4. $L \geqslant 0$ y $P \geqslant 0$.

Estas desigualdades se representan en la gráfica como ecuaciones. Las ecuaciones de las restas forman los límites de regiones. La solución del problema tiene que encontrarse en cierta región de la gráfica.

Las escalas de los ejes de la gráfica (ver figura 7.1a) miden hectáreas de lechugas (eje horizontal o x) y de patatas (eje vertical o y).

Las tres ecuaciones limitan el área donde se encuentra la solución. Tiene la forma de un polígono irregular. Es suficiente comparar los resultados que se obtienen en cada uno de los puntos A, B, C y D en la gráfica.

Resulta que al dueño de la granja le conviene sembrar 3 hectáreas con lechugas y 15 con patatas. Hay muchas otras soluciones posibles (por ejemplo E, la cual da $4 \times 12 + 4 \times 8 = 80$ libras, y F, la cual da 0 porque no siembra nada).



Ganancia

Punto	*Lechugas (12)*	*Patatas (8)*	*Total*
A	0 × 12	18 × 8	144 libras
B	3 × 12	15 × 8	156 „
C	7.5 × 12	7 × 8	146 „
D	10 × 12	0 × 8	120 „

Ejercicio 7.1a (la solución se encuentra al final del capítulo).

Usando el mismo método, solucione el problema siguiente:

El dueño de una fábrica de muebles produce dos tipos de sillas, *A* y *B*. Los límites o las condiciones son los siguientes:

1. El costo de materiales es 4 libras por silla tipo *A* y 5 libras por silla tipo *B*. Puede gastar hasta 200 libras.

2. El trabajo en horas-sombre viene a ser 8 horas por silla tipo *A* y 5 horas por tipo *B*. Tiene hasta 320 horas disponibles.

3. Tiene contratos para vender, *por lo menos* (límite mínimo, *no* máximo), 15 sillas de tipo *A* y 10 sillas de tipo *B*. Atención: estos límites no forman una ecuación como en el ejemplo anterior, sino dos líneas, una vertical y otra horizontal.

4. Dadas las condiciones anteriores, no es preciso estipular que el número de sillas deba exceder 0.

Calcule las ecuaciones, represéntelas gráficamente y busque la solución que maximice las ganancias.

La función objetiva, o la ganancia por silla producida, es la siguiente:

Tipo *A*: 1.75 libras
Tipo *B*: 1.50 libras

b) *Problema de transporte*
(ver figura 7.1b)

La presentación de este problema tiene forma distinta de la del problema de producción.

Una compañía que tiène dos fábricas (*A* y *B*) y dos almacenes (1 y 2) tiene que transportar todos los productos de las dos fábricas a los dos almacenes, con un mínimo de costos de transporte.

Las distancias son los costos de transporte, en pesos, por unidades de producción de la fábrica. Las condiciones siguientes tienen que cumplirse:

Fábrica *A* produce 7 unidades
Fábrica *B* produce 4 unidades
Almacén 1 necesita 6 unidades
Almacén 2 necesita 5 unidades

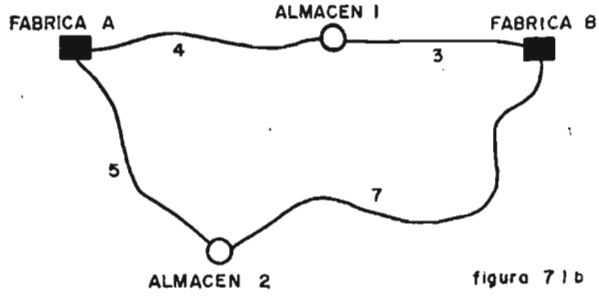


figura 7.1b

CUADRO 7.1a

Desde:	*Costos de transporte A:* *Almacén 1*	*Almacén 2*	*Producción de la fábrica*
Fábrica *A*	4	5	4
Fábrica *B*	3	7	7
Necesidad del almacén	6	6	Total 11

Toda la información necesaria para solucionar el problema se coloca en una matriz (ver cuadro 7.1a).

Atención: no confundir en la matriz los costos de transporte con las unidades de producción.

Se pueden expresar las relaciones entre las fábricas y los almacenes en forma de ecuaciones (F = fábrica, A = almacén):

$$F_A \text{ a } A_1 + F_A \text{ a } A_2 = 4$$

(Esto significa que el número total de unidades de producción que se puede mandar de la fábrica F_A a los dos almacenes es 4).

De la misma manera:

$$F_B \text{ a } A_1 + F_B \text{ a } A_2 = 7$$

Además:

$$F_A \text{ a } A_1 + F_B \text{ a } A_1 = 6$$

(Porque el almacén A_1 necesita un número total de 6 unidades de las dos fábricas).

De la misma manera:

$$F_A \text{ a } A_2 + F_B \text{ a } A_2 = 5$$

Cuando no hay más que dos fábricas y dos almacenes, es fácil encontrar la solución sin cálculos elaborados. En este caso conviene mandar toda la producción (4) de la fábrica *A* al almacén 2, seis unidades de la fábrica *B* al almacén 1, y una unidad de la fábrica *B* al almacén 2. Se pueden examinar las alternativas de la manera siguiente:

CUADRO 7.1b

(*U - unidades enviadas C - costo T - total*)

	Alternativa 1 U	C	T	*Alternativa 2* U	C	T	*Alternativa 3* U	C	T
FA-A1				4	× 4	= 16	2	× 4	= 8
FA-A2	4	× 5	= 20				2	× 5	= 10
FB-A1	6	× 3	= 18	2	× 3	= 6	4	× 3	= 12
FB-A2	1	× 7	= 7	5	× 7	= 35	3	× 7	= 21
	Total		45	Total		57	Total		51

Hay varias soluciones posibles, pero se puede averiguar, en este caso, si la alternativa 1 es la óptima o no lo es. Cuando hay muchas fábricas y almacenes (o, en general, puntos de origen y de destino de movimientos de carga) es posible considerar algunas alternativas y compararlas, pero no se pueden examinar todas las soluciones posibles, de modo que no se puede estar seguro si se ha encontrado la solución óptima.

Felizmente existen varios métodos que evitan la necesidad de calcular todas las alternativas. Sin embargo, se necesita la aplicación de la computadora electrónica para hacer los cálculos.

Ejercicio (ver figura 7.1c). Hay dos minas de carbón y 4 estaciones termoeléctricas. Cada mina de carbón es capaz de producir 2.5 millones de toneladas al año y cada estación termoeléctrica consume un millón de toneladas al año.

En este ejemplo la capacidad de producción excede la demanda del mercado, de modo que hay que determinar no solamente la dirección del movimiento del carbón, sino el nivel de producción de las dos minas.

La información necesaria para la solución del problema se encuentra en el cuadro 7.1c. Las posiciones de las minas y

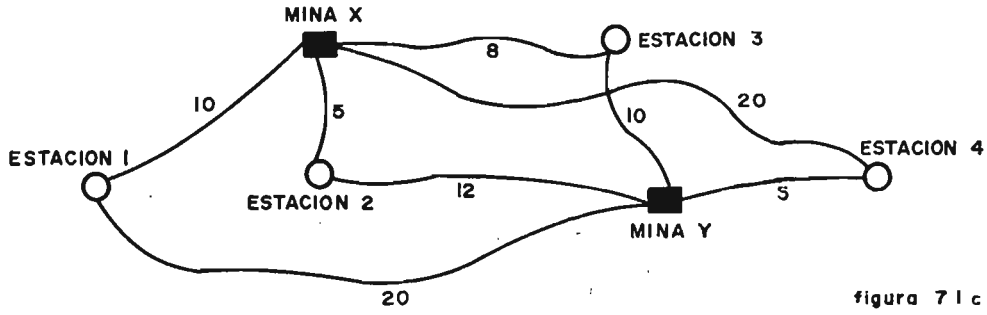


figura 7 1 c

estaciones se encuentran en la figura 7.1d. Calcule la solución que minimice los costos de transporte.

CUADRO 7.1c

(distancias en kilómetros)

	Estaciones 1	2	3	4	*Producción máxima*
Minas *x*	10	5	8	20	2.5 millones de ton
y	20	12	10	5	2.5 millones de ton
Demanda (millones de toneladas)	1.0	1.0	1.0	1.0	

7.2 *Teoría de los juegos*

La teoría de los juegos se ha aplicado en varios estudios geográficos en los sesentas. Uno de los primeros artículos publicados fue el de P. R. Gould, "Man Against His Environment: a Game Theoretic Framework", *Annals of the Association of American Geographers*. Vol. 53, Nº 3, septiembre, 1963, pp. 290-297.

Gould sugirió el caso de los cultivadores de Ghana, los cuales tienen varias posibles combinaciones de cultivos (maíz, arroz, frijol). No saben cómo serán las condiciones meteorológicas cuando crezcan y maduren sus cultivos. Ciertas condiciones favorecen algún tipo de cultivo, otras condiciones favorecen otro. Tienen que adoptar una estrategia para afrontar todas las estrategias posibles (varias condiciones meteorológicas, lluvia, calor) del oponente, el otro jugador, esto es, el clima. De los conceptos de la teoría de los juegos se han desarrollado muchos juegos geográficos, modelos o simulaciones de la realidad que ayudan en el estudio de las situaciones geográficas y en las investigaciones.

Se ilustra con cuatro ejemplos la teoría de los juegos.

Ejemplo 1

Dos fábricas (*A* y *B*) producen el mismo artículo y lo venden en el mismo mercado. En el futuro inmediato (por ejemplo, el año entrante), hay cuatro alternativas que resultan de dos estrategias posibles para *A* y dos para *B*.

a) *A* reduce su precio
B no reduce su precio
b) *A* reduce su precio
B reduce su precio
c) *A* no reduce su precio
B no reduce su precio
d) *A* no reduce su precio
B reduce su precio

Se puede considerar la situación desde el punto de vista de *A* o de *B*. *A* prefiere la primera alternativa a todas las otras porque le daría la oportunidad de aumentar sus ventas (al costo de *B*). *B*, al contrario, prefiere la alternativa *d*) por la misma razón. La segunda preferencia de los dos es la alternativa *c*), porque es preferible a la alternativa *b*). Con *c*), los dos, por lo menos mantienen la situación actual, mien-

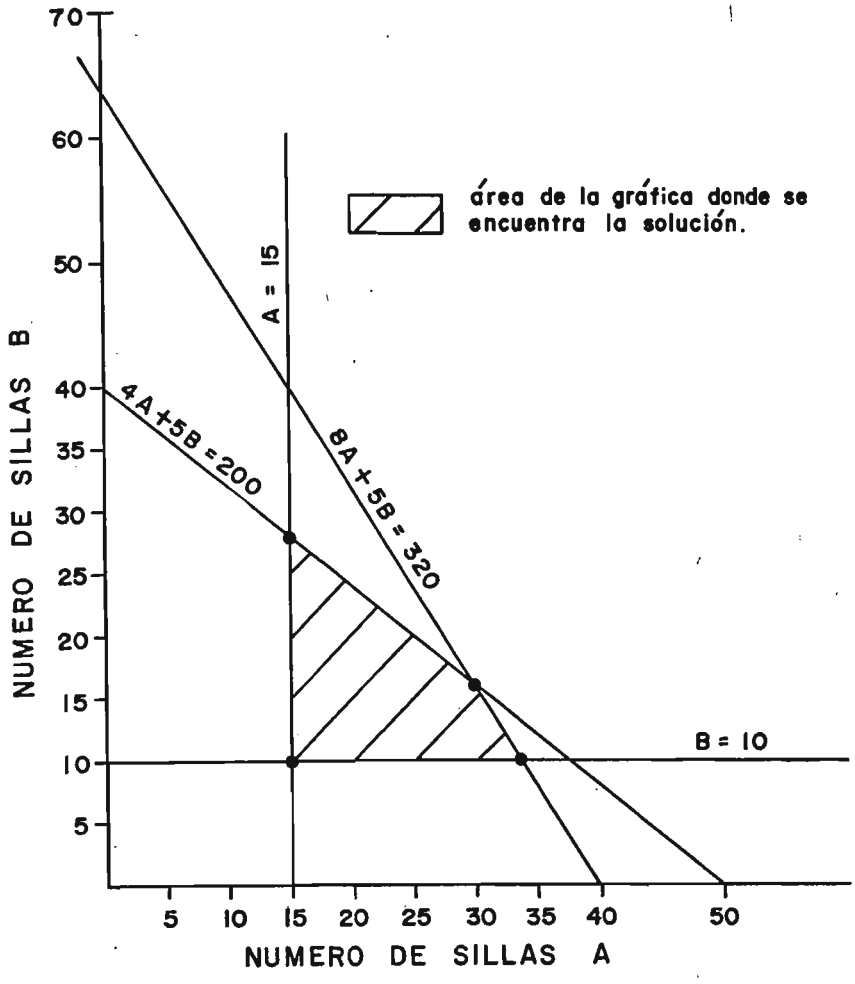


figura 7.1 d

tras que con *b*) los dos pierden dinero por la reducción bilateral de sus precios.

Es posible representar las estrategias y las preferencias de *A* y *B* en forma de una matriz (ver cuadro 7.2a). Las estrategias de *A* forman los renglones de la matriz, las estrategias de *B* las columnas.

4 es la preferencia más alta.

1 es la preferencia más baja.

Las preferencias de *A* y de *B* se apuntan en cada entrada de la matriz, y el total de sus preferencias se encuentra en el centro. Los dos jugadores *A* y *B* usan una escala 4, 3, 2, 1 para ordenar sus preferencias.

Es interesante notar que, aunque la preferencia máxima de *A* (alternativa *a*) no coincida con la preferencia máxima de *B* (alternativa *d*), la estrategia que más conviene a los dos juntos (alternativa *c*) tiene un total de 6 (3 + 3), el cual supera a *a* (4 + 1), y a *d* (4 + 1 también).

En conclusión, se puede decir que la tabulación de las preferencias y el estudio numérico de las estrategias ayuda a aclarar una situación, aunque al principio los tér-

CUADRO 7.2a

	BB				
AA	*no reduce su precio*			*reduce su precio*	
reduce su precio	*a*) Total 5	1		*b*) Total 4	2
	4			2	
no reduce su precio	*c*) Total 6	3		*d*) Total 5	4
	3			1	

minos y los conceptos presenten dificultades.

Ejemplo 2

El conflicto entre Israel y Egipto (y otros países árabes) ha durado más de dos decenios. Es cuestión del reconocimiento del Estado de Israel por Egipto. Al mismo tiempo entra la cuestión del Canal de Suez que, cuando funciona, es una fuente de ingresos para Egipto. Naturalmente, el problema tiene más aspectos, pero en términos básicos, como en el ejemplo anterior, se pueden identificar 4 estrategias.

Se supone que los dos países dan más importancia a la cuestión del reconocimiento de Israel que a la cuestión del canal.

a) Egipto no reconoce a Israel, el canal funciona.
b) Egipto no reconoce a Israel, el canal no funciona.
c) Egipto reconoce a Israel, el canal funciona.
d) Egipto reconoce a Israel, el canal no funciona.

La situación *a*) es la que existía hasta el año 1956 y de nuevo entre 1957 y 1967. El canal no funcionó en el periodo 1956-57 y se cerró otra vez en el año 1967.

Se pueden representar las preferencias, en la forma de una tabla, de la manera siguiente:

Orden de preferencia

Alternativa	*Egipto*	*Israel*	*Total*
a)	4	2	6
b)	3	1	4
c)	2	4	6
d)	1	3	4

Egipto prefiere la alternativa *a*) a todas las otras, mientras que Israel prefiere la alternativa *c*). Sin embargo, la situación *b*) ha durado 6 años. Se puede explicar la existencia de una situación que no conviene a ninguno de los dos jugadores, de esta manera: Egipto prefiere perder los ingresos del canal a la alternativa de reconocer a Israel [prefiere *b*) a c), 3 puntos a 2]. Además, si varía de *b*) a *c*) concediendo el reconocimiento de Israel y ganando la rehabilitación del canal, cambia de la situación de menor preferencia de Israel [Israel da 1 punto a *b*)] a la situación preferible para Israel [*c*) vale 4 puntos]. De esta manera Egipto no sólo perdería un punto en el juego, sino cedería 3 puntos a su enemigo.

Ejemplo 3

El dueño de una hacienda puede escoger entre 5 cultivos: maíz, trigo, alfalfa, algodón y melones. Tiene que decidir qué combinación de cultivos le conviene en su hacienda (la cual tiene 10 campos de igual tamaño), dadas ciertas condiciones meteorológicas de su región. Él sabe, por experiencia, que hay 4 tipos de condiciones). No sabe qué tipo le tocará durante el periodo de crecimiento y maduración de sus plantas, pero sabe que tienen frecuencias distintas. Además, sabe cuánto vale la cosecha de cada cultivo según las condicioies meteorológicas que pueden experimentarse. Se puede aplicar la teoría de los juegos para calcular qué combinación de cultivos le conviene más (maximice su ganancia) dadas las probabilidades de cada conjunto de condiciones meteorológicas.

Condiciones del modelo

1. *Tiempo*

Caliente/lluvioso	2 años en 6
Caliente/seco	1 año en 6
Frío/lluvioso	2 años en 6
Frío/seco	1 año en 6

2. *Valor de producción por campo, de cada cultivo, según el tiempo*:

	Caliente lluvioso	*Caliente seco*	*Frío lluvioso*	*Frío seco*
Maíz	8	5	3	1
Trigo	1	3	6	8
Alfalfa	6	4	5	3
Algodón	4	10	1	6
Melón	6	10	1	3

3. *Dado* 1 o 2: caliente/lluvioso
3: caliente/seco
4 o 5: frío/lluvioso
6: frío/seco

El uso del modelo

Se supone que cada año el dueño o gerente de la granja decide qué plantas va a cultivar. Por ejemplo:

4 campos con maíz
1 campo con trigo
2 campos con alfalfa
2 campos con algodón
1 campo con melones

¿Qué pasa si resulta que viene un verano frío y lluvioso? Se puede calcular el valor de su producción de la manera siguiente:

	Campos	*Valor por campo*	*Valor total*
Maíz	4	3	12
Trigo	1	6	6
Alfalfa	2	5	10
Algodón	2	1	2
Melón	1	1	1
			31

Prácticamente pierde toda su cosecha de algodón; el maíz rinde poco. Si hubiera cultivado exclusivamente algodón, habría sufrido un desastre. Sin embargo, le conviene cultivar algodón porque en otras condiciones (caliente/seco) la cosecha vale mucho.

Es posible usar el modelo para simular una serie de años. Se usa un dado para determinar el tiempo, pero *después* de la decisión sobre la combinación de cultivos. Se pueden apuntar las decisiones y calcular los resultados en una tabla (cuadro 7.2b).

El dado representa la estrategia del tiempo. Ejemplo (ver columna 1 en el cuadro 7.2b).

(*i*) Cada cultivo ocupa el número de campos indicado.
(*ii*) Este año el resultado del dado es 1, lo que significa caliente/lluvioso.
(*iii*) Multiplicar el número de campos por el valor de la producción.

Según caliente/lluvioso: maíz 3×8 (24), trigo 2×1 (2), alfalfa 2×6 (12), algodón 2×4 (8), melón 1×6 (6). Poner los resultados en la columna 1, valor de la producción.

Repitiendo los pasos año por año es posible estudiar las fluctuaciones de la fortuna del hacendado, como consecuencia del "juego" entre su estrategia (los cultivos que escoge) y la estrategia del ambiente físico.

Usando conceptos y métodos de la teoría de los juegos es posible calcular la estrategia que, a largo plazo (decenios), asegure al hacendado el ingreso promedio más alto posible. Hay que tener en cuenta, sin embargo, que los precios cambian año por año; que la calidad de las semillas y de las plantas puede mejorarse y que las condiciones climáticas también son capaces de cambiar dentro de un periodo de 50 años. Es posible complicar y adaptar el modelo, para tener en cuenta muchos factores de este tipo.

Ejemplo 4

Se pueden adaptar las ideas de la teoría de los juegos a varias ramas de la geogra-

CUADRO 7.2b

	Campos con cada cultivo						*Valor de la producción*					
	1	*2*	*3*	*4*	*5*	*6*	*I*	*II*	*III*	*IV*	*V*	*VI*
Maíz	3						24					
Trigo	2						2					
Alfalfa	2						12					
Algodón	2						18					
Melones	1						6					
Total	10	10	10	10	10	10	62					

fía. El ejemplo siguiente ilustra su aplicación en la geografía urbana.

Hay muchos incendios en la ciudad ficticia de San Luis del Pozo Alto. La municipalidad ha comprado dos carros de bomberos. Hay que decidir dónde localizarlos en la ciudad, para *minimizar* la distancia (o el tiempo) entre la estación de bomberos y cualquier incendio. Sería posible construir dos estaciones en diferentes lugares o construir una estación y tener los dos carros en el mismo sitio.

Es posible simular la situación y experimentar con varias posiciones antes de decidir finalmente dónde invertir el dinero en la construcción de las estaciones. Las distancias en la ciudad de San Luis han sido calculadas en unidades de tiempo (por ejemplo, 30 segundos por unidad). Se necesita cierto periodo de tiempo para moverse de un lugar a otro, incluso el movimiento por las calles y la demora delante de los semáforos. En el centro comercial la velocidad se reduce mucho. En la peri-

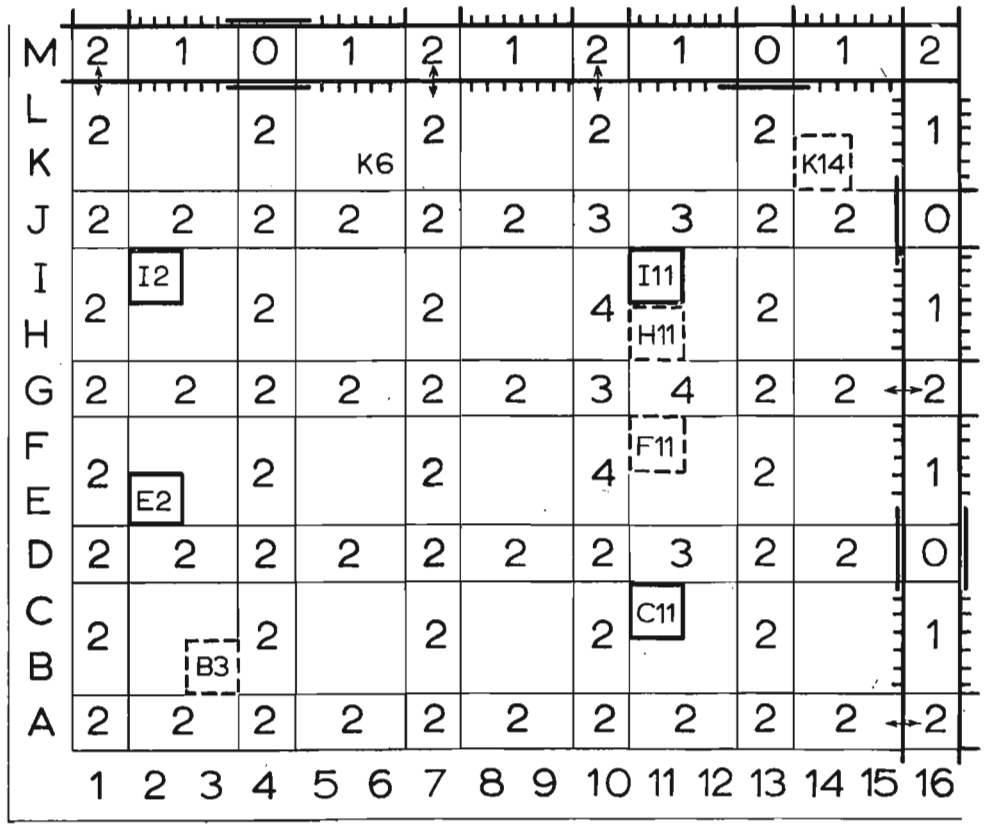


Figura 7.2a

feria, una vía rápida facilita mucho el movimiento.

En la figura 7.2a se representa en forma diagramática la ciudad de San Luis. Los números de las calles representan el lapso estimado en unidades de tiempo necesarias para moverse de un lugar a otro. Se usan números y letras para localizar las calles y los edificios en la ciudad, los números en la dirección oeste-este y las letras en la dirección sur a norte. Por ejemplo, el viaje del edificio *B*3 al edificio *B*11 dura 2 + 2 + 2 + 2 + 2 + 2 + 2 + 3, esto es, 17 unidades de tiempo.

En la ciudad, ciertos lugares tienen serio riesgo de incendio:

*I*2, depósito de madera
*E*2, depósito de petróleo
*I*11, tienda grande
*C*11, fábrica de muebles

Se calcula que, en el caso de un incendio, sería preciso mandar los dos carros de bomberos dentro de un máximo de 10 unidades de tiempo.

Otros riesgos se encuentran en *B*3, *F*11, *M*11, *K*14. Necesitan, por lo menos, un carro de bomberos dentro de un máximo de 10 unidades de tiempo.

Todos los otros espacios (por ejemplo, *E*9, *K*6) necesitan un carro dentro de un máximo de 12 unidades de tiempo.

Problema: ¿Es posible localizar una o dos estaciones de bomberos para satisfacer las condiciones? ¿Es posible simular la situación colocando las estaciones de bomberos en determinados lugares y, usando números aleatorios, representar la ocurrencia de los incendios en la ciudad? Es posible mover los carros de bomberos por las calles y calcular el tiempo necesario.

En la forma de un juego sería posible tener dos jugadores: el primero decidiría la posición de sus estaciones. El otro representaría, al azar, los incendios. El primer jugador tendría que mover sus carros a los incendios, perdiendo puntos cuando no llegase dentro del límite máximo establecido.

La programación lineal y la teoría de los juegos son dos ejemplos de una serie de métodos que pertenecen a investigación de operaciones. Otros métodos se han aplicado en geografía. El método del análisis del sendero principal ayuda en el planeamiento de los proyectos complejos. La teoría de las colas tiene aplicaciones en muchos aspectos de la geografía.

La teoría de las colas estudia las relaciones entre algún servicio y los clientes que lo usan. El servicio dura un periodo determinado de tiempo. Si los clientes llegan con regularidad y la duración del servicio es menor que el intervalo entre la llegada de cada cliente, no se forman colas. Muchas veces los clientes llegan de manera fortuita. Es fácil simular el proceso con auxilio de una tabla de números aleatorios. La teoría de las colas ayuda en la optimización del uso del servicio y la minimización del tiempo perdido por la porción de clientes que esperan en cola. Cuando los clientes llegan de manera variable se ha observado que no se forman colas notables cuando el intervalo medio entre las llegadas de los clientes es el doble de la duración del servicio. En este caso, sin embargo, el servicio usa solamente 50 % del tiempo de la persona que lo ofrece. Cuando el servicio usa 80 % de ese tiempo, se forman colas muy grandes. Los remedios posibles incluyen la implantación de otro servicio, la reducción de la duración del servicio y el traslado de los clientes a otro lugar en dode se ofrece el mismo servicio: voluntariamente o por imposición de un límite al tamaño de la cola. Las colas se forman en varias situaciones de interés geográfico, entre las cuales se hallan las siguientes:

1. Congestión en las calles de una ciudad.

2. La llegada de aviones a un aeropuerto con mucho tráfico. A veces tienen que esperar turno para aterrizar y otras tienen que dirigirse a otro aeropuerto.

3. El agua que pasa por varios afluentes fluviales, que se reúnen en una zona donde hay peligro de inundaciones. Si simultáneamente llueve excesivamente en las cuencas de todos los afluentes, una cantidad muy grande de agua lleva a la parte crítica del río. Como el agua no es capaz de esperar (a menos que se haya construido un embalse para contenerla), no forma cola; pasa por el 'servicio' * rápidamente e inunda una zona cercana al río.

4. Se ha sugerido que la vida es un tipo de servicio; éste dura toda la vida. Hay una cola de niños que todavía no han nacido; si el número de nacimientos excede al de defunciones, es preciso aumentar las facilidades del servicio o pasar una parte de los clientes nuevos, los nacimientos, a otro sitio.

Solución al ejercicio 7.1.a

Las desigualdades son las siguientes:

$$4A + 5B \leqslant 200$$
$$8A + 5B \leqslant 320$$
$$A \geqslant 15$$
$$B \geqslant 10$$

La solución que maximice la ganancia, dadas las condiciones de la situación, es la siguiente. Ver también la figura 7.1d.

30 sillas tipo A	$= 30 \times 1.75 =$	52.5 libras
16 sillas tipo B	$= 16 \times 1.5 =$	24 „
	TOTAL	76.5 „

* La palabra servicio se usa aquí en un sentido muy general. Puede ser, por ejemplo, una estación de gasolina, la caja de un supermercado, o la pista de aterrizaje de un aeropuerto.

Solución al ejercicio 7.1b

Cantidades movidas

	A 1	2	3	4	
					Producción total
De *x*	1	1	0.5	0	2.5
y	0	0	0.5	1.0	1.5

Costos totales de transporte

x-1	$1 \times 10 =$	10
x-2	$1 \times 5 =$	5
x-3	$0.5 \times 8 =$	4
y-3	$0.5 \times 10 =$	5
y-4	$1 \times 5 =$	5
	TOTAL	29

8. ANÁLISIS DE SUPERFICIES DE TENDENCIA

8.1 *El estudio de difusión*

La conciencia de la forma en que muchos fenómenos, poblaciones, vegetales, invenciones, enfermedades, ideas, con el tiempo se difunden sobre un área, desde hace mucho debe haber sido parte de la experiencia del hombre. El estudio concienzudo, formal y científico de los patrones de difusión, por los geógrafos, es relativamente reciente. Quizá el primer artículo de un geógrafo, sobre el tema fue el de T. Hagerstrand: "The propagation of innovation waves".*

El concepto básico es que, al pasar el tiempo muchas cosas se difunden en un área. En un momento determinado muchas distribuciones espaciales aparentemente estáticas pueden estar en proceso de cambio. Han sido menos extensivas en el pasado y se volverán más en el futuro.

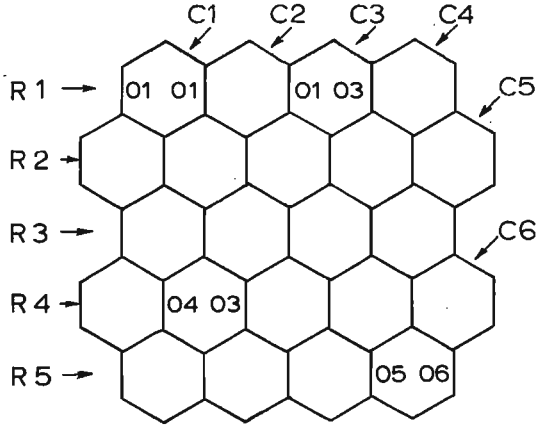
* Lund Series in Geography, Ser. B., Human Geography, Nº 4, 1952, Lund, Suecia.

En una escala relativamente pequeña, la difusión de una enfermedad de los pies y la boca, que afectara al ganado desde un lugar conocido, Oswestry, hasta cientos de ranchos del centro de Inglaterra, en 1967 y 1968, ejemplifica un proceso de difusión cuyo principio, desarrollo y conclusión pueden estudiarse con mucha precisión. A escala mundial, la difusión de la alfabetización a todas las clases sociales de una comunidad, y horizontalmente en especial de las ciudades a las zonas rurales es ejemplo de un proceso de difusión más complejo que, sin duda, continuará todavía por muchos decenios.

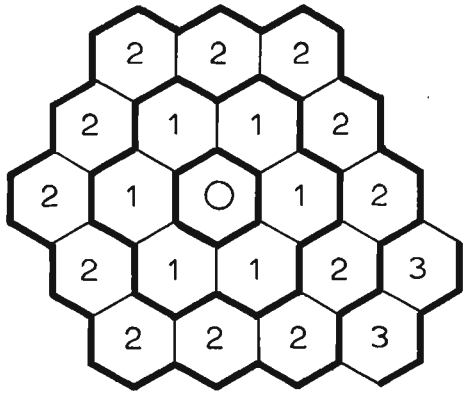
El propósito de esta sección es proveer ejemplos prácticos, todos ficticios, de dos tipos de difusión sobre áreas. Por conveniencia se llaman mecanísticas y probabilísticas.

Los términos. La figura 8.1a muestra los términos utilizados para manejar un diseño de hexágonos. Los hexágonos se utilizan, en vez de los cuadrados o triángulos, porque tienen seis lados, cada uno de los cuales coincide con un lado del hexágono adjunto. En un cuadrado hay cuatro lados y cuatro vértices, de manera que la proximidad de los cuadrados en la primera coro-

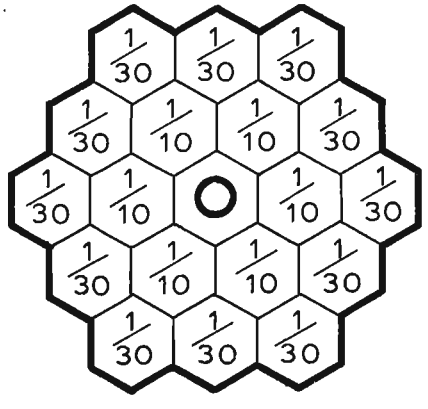
Sistema de referencia que se sugiere para especificar los exágonos en el sistema. R es el equivalente al renglón en una matriz y C a la columna. Cuando se usan 10 o más renglones o columnas se antepone un 0 a los dígitos entre 1 y 9.



Cada exágono tiene 6 exágonos adyacentes. Se le llama la primera corona. La segunda y la tercera coronas se muestran también en la figura. La numeración de la corona se da con relación al exágono original (0).



Probabilidad de información.



Probabilidad de información expresada en la forma de números aleatorios (pares de dígitos). No deben tomarse en cuenta 00 y 91-99.

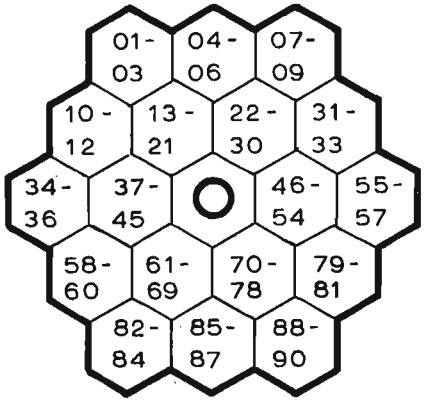


figura 8.1a

na es de dos tipos: cuatro contiguos en los lados y cuatro en los vértices. Cualquier hexágono de la red puede localizarse en forma única mediante un par ordenado de números, pero si las líneas se hacen horizontales, esto es, de izquierda a derecha, a lo ancho de la página, las columnas no están en ángulos rectos, sino a 60°. Para juegos y modelos del tipo que describimos aquí, es conveniente dividir el espacio en los compartimientos discretos formados por hexágonos (o cuadrados) en vez de considerarlo como continuo. En forma análoga, el tiempo se divide en componentes, tales como días o semanas.

La noción de una corona de hexágonos se muestra también en la figura 8.1a. Cualquier corona consiste en una combinación de hexágonos única, como la corona alrededor de un determinado lugar de origen (0). Nótese que un lugar de origen puede tener parte de su primera corona y otras coronas en el mar.

Difusión mecánistica. La figura 8.1b puede utilizarse como base para experimentar con una difusión mecanística. Para no marcar el mapa se puede usar papel transparente. Supóngase que hay un rancho con productos lácteos, en cada hexágono que no está ocupado por una montaña o un pantano. Hay en total 150 hexágonos libres y, por tanto, ranchos.

Surge una enfermedad en un rancho dado, el rancho de origen. En la primera semana todos los ranchos de los hexágonos contiguos al rancho de origen se contaminan, esto es, todos los ranchos de la primera corona. La semana siguiente todos los ranchos contaminados, viejos y nuevos, se ven afectados, y así sucesivamente.

Calcule y tabule el número de ranchos que se ven afectados cada semana, y el total contaminado al final de 8 semanas, si la enfermedad empieza (1) en 01-10 (esto es, en la península Wayward), (2) en 10-10 (un rancho muy central), (3) en 13-14.

Ejemplos de tabulaciones del lugar (1), en 01-10 sería como sigue:

	Semana	*Valor acumulado*
Comienzo	1	1
Final 1ª semana	3	4
2ª	1	5
3ª	2	7
4ª	2	9

y así sucesivamente.

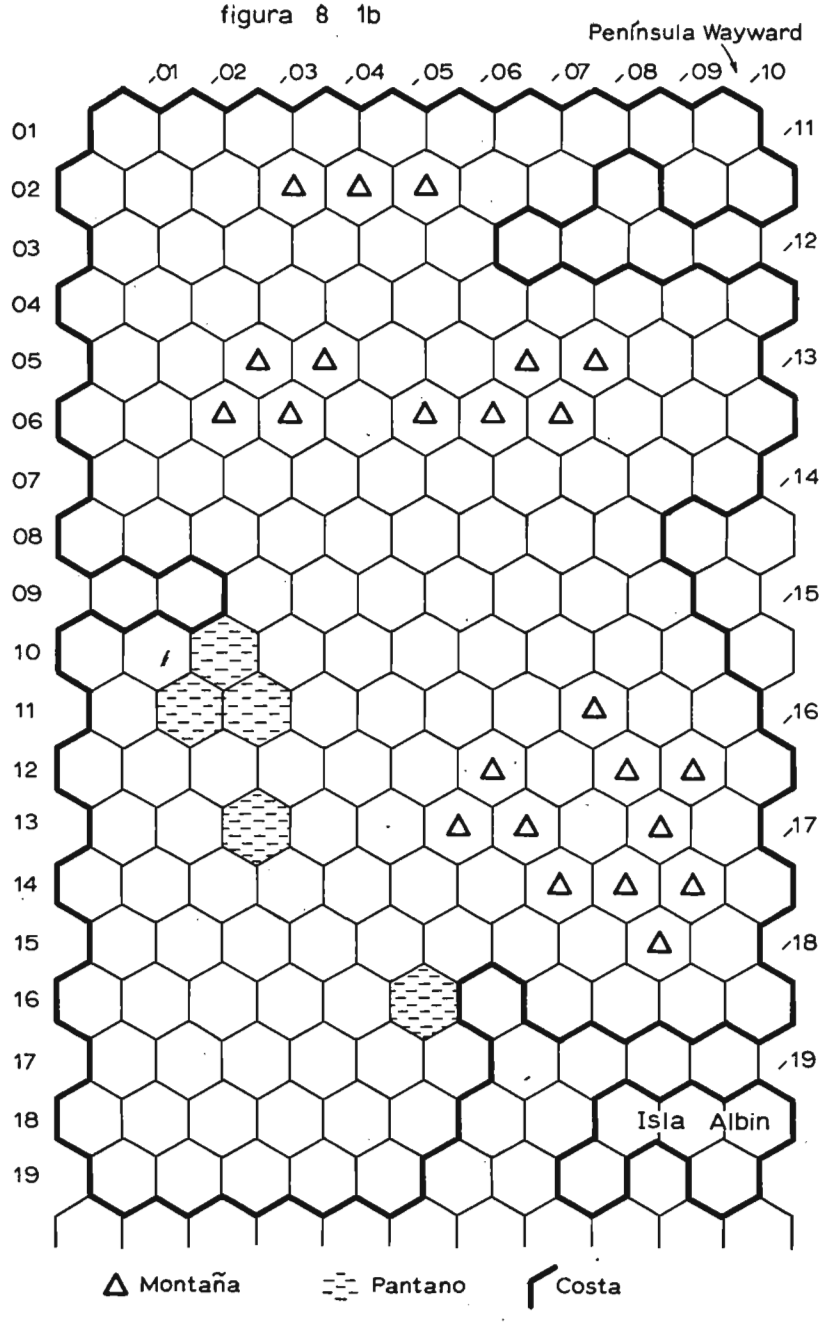
Puesto que no hay ranchos en las montañas, en los pantanos o en el mar, éstos no pueden contaminarse; actúan como barreras. Utilice un papel transparente y marque con diferente color la difusión de cada uno de los tres ranchos en turno.

Difusión probabilística. La figura 8.1b puede utilizarse como base para este ejercicio (use papel transparente). En contraste con la difusión mecanística la probabilística no es mecánica. La idea de la difusión de una innovación de un lugar inicial hacia afuera es la misma, pero se agrega un elemento de incertidumbre.

Supóngase que alguna idea nueva, tal como un nuevo tipo de semilla, fertilizante, o implemento agrícola, se introduce en un rancho o pueblo, ya sea por casualidad o deliberadamente, por ejemplo, por un experto de las Naciones Unidas. La innovación se conoce gradualmente y se adopta o rechaza por más y más ranchos o pueblos. No es posible decir cuál aceptará la innovación en un momento determinado. El ejercicio siguiente es un modelo estocástico de tipo muy sencillo. Muestra el tipo de resultado que pudiera obtenerse por medio de esa difusión al azar. Son posibles muchos resultados diferentes.

Para este ejercicio se supondrá que cada uno de los 150 hexágonos sin montaña o pantano, del mapa de la figura 8.1b con-

figura 8 1b



tiene un pueblo. Al pueblo en 10-10 se le informa de una nueva técnica, y adopta, al principio, la difusión. Este pueblo, marcado 0 (origen) en la figura 8.1c es elegible para comunicarlo a un nuevo pueblo, durante un periodo de tiempo (o turno) 1. A qué nuevo poblado se difunde la información depende de la oportunidad, pero existe una gran probabilidad de que esté en la primera corona, y es seguro que estará en la 1ª o en la 2ª corona. Esta regla ha sido elaborada por el autor para este juego particular y podría modificarse. ¿Cómo se determina el pueblo informado?

Use la tabla de números aleatorios que se proporciona más adelante. Empiece en algún lugar alejado del principio y use pares de dígitos sucesivos (ejemplo, 17, 53), pero no use el número 00 o los números de 91 a 99. En los diagramas inferiores de

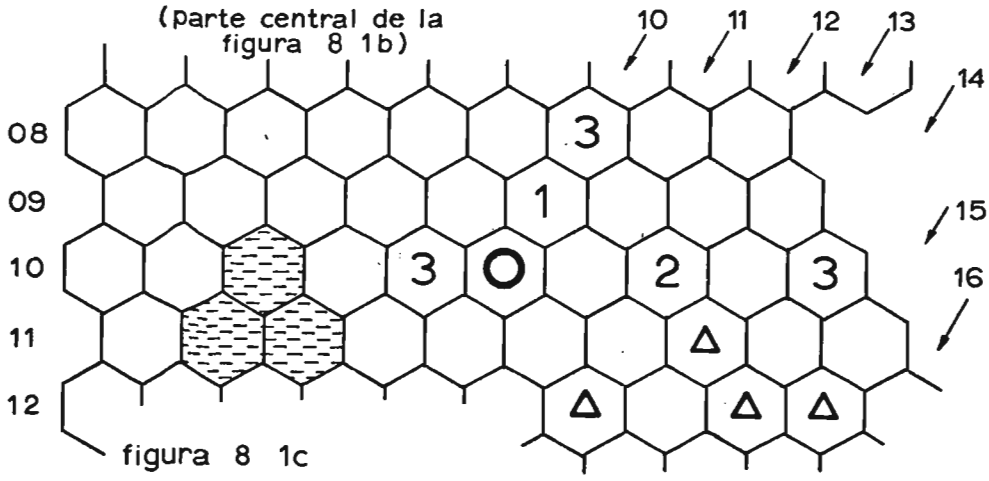


figura 8 1c

la figura 8.1a se ilustran las ideas de la red que se usan para decidir aleatoriamente cuál pueblo se informa. Imagine el 0 (centro) de la red izquierda colocada en su pueblo innovador 10-10. Hay una probabilidad de ser informado de 1/10 para cada uno de los seis pueblos de la corona interior y una probabilidad de 1/30 para cada uno de los doce pueblos de la segunda corona. Note que la probabilidad total es uno. Los números aleatorios de 01 a 90 se colocan en grupos de tres en la corona exterior y de nueve en la interior, para reproducir las proporciones de las probabilidades indicadas. Los números se muestran en la red de la derecha en la figura 2.1a. Esto se usará para simular la difusión. Se imaginará la colocación del 0 en cada pueblo, al ser éste elegible para comunicar la noticia.

Ejemplo satisfecho con resultados tabulados. Consultar el mapa de la figura 8.1c. Nótese que se desperdicia una información que cae en un hexágono con una montaña, un pantano o el mar; lo mismo pasa si cae en un pueblo que ya la conoce. Los números del cuadro 8.1a son los que se obtuvieron. Use los suyos propios.

En el ejemplo, el primer número aleatorio que se encontró fue el 29. Éste se halla en el pueblo localizado inmediatamente al noroeste de 0. Marque un 1 en ese pueblo. Final del periodo de tiempo 1. Anote el resultado en el cuadro 8.1a.

Tanto el pueblo en 10-10 que ya estaba informado, como el nuevo en 09-10, son elegibles para informar. Es más conveniente (aunque no obligatorio) seguir sucesivamente los renglones a través del mapa, de oeste a este, tomando cada vez un pueblo que ya está informado.

De la tabla de números aleatorios tome un nuevo número para cada pueblo. El número obtenido para el pueblo en 09-10 fue 81. Éste cae en un pueblo en la segunda corona de 09-10; es decir, 10-12. Anote 2

CUADRO 8.1a

	Periodo de tiempo		*Periodo de tiempo acumulado*	
	comunicado	*perdido*	*comunicado*	*perdido*
Comienzo	1	0	1	0
Fin del periodo 1	1	0	2	0
Fin del periodo 2	1	1	3	1
Fin del periodo 3	3	0	6	1
y así sucesivamente				

aquí. Para el pueblo 10-10 se obtiene el número 88. Éste se localiza en la montaña en 12-12 y se pierde. Anote los resultados en la tabla y luego proceda al periodo de tiempo 3. Salen los tres números siguientes: 22 para el pueblo 09-10 (marque un 3 en 08-10), 39 para el pueblo 10-10 (marque un 3 en 10-09) y 56 para el pueblo 10-12 (marque un 3 en 10-14). Anote los resultados en la tabla. Continúe por 8-10 periodos de tiempo. Si tiene ocasión, vuelva a empezar con otro pueblo de origen. Compare la proporción de informaciones provechosas con las perdidas. ¿Cuáles modificaciones puede usted sugerir para el proceso? ¿Qué situaciones reales se podrían probar en esta forma?

8.2 *Análisis de superficies de tendencia* *

El análisis de superficies de tendencia es un método para caracterizar distribuciones de puntos en tres dimensiones, con superficies. La *primera* superficie es plana, la *segunda* curva y las superficies de órdenes más altos son más complejas. Los puntos o lugares se localizan en tres dimensiones según sus posiciones en tres direcciones o ejes ortogonales.

1. La distancia en una dirección oeste-este (eje x) del punto de origen.
2. La distancia en una dirección sur-norte (eje y) del mismo punto de origen.
3. El índice del valor observado de cada punto o lugar en el mapa, por ejemplo, la precipitación, la densidad de población.

En muchas distribuciones geográficas los valores observados en una variable, para un número determinado de lugares, revelan tendencia a disminuir o aumentar de un lado de una región a otro. Por ejemplo, en el caso del norte de Portugal, la densidad de la población rural disminuye, de manera general, de oeste a este. Sin embargo, se notan excepciones a esta tendencia general. En la zona occidental se encuentran ciertas áreas con densidades relativamente bajas, en el este algunas áreas con densidades relativamente altas, y en el centro de la región, en la zona del cultivo de la vid, densidades muy altas.

La función del análisis de superficies de tendencia es calcular los valores numéricos de una serie de superficies que caracterizan la tendencia general de los valores observados que cambian de una parte a otra de la región. Se calculan los 'residuos', la distancia de cada punto, cada superficie, la diferencia entre el valor observado y el valor esperado según la superficie.

La superficie de tendencia es un tipo de valor medio en tres dimensiones. Se ajusta a una distribución de puntos de una manera que minimiza la distancia total entre los puntos y la superficie. La idea se ilustra diagramáticamente en la figura 8.2a (*a*), (*b*) y (*c*).

En el caso de una serie de valores en una escala lineal, el valor medio corresponde a un punto central de la escala [diagrama (*a*)]. Cuando los valores se localizan en dos dimensiones, su distribución se caracteriza con una línea [diagrama (*b*)]. Continuando la analogía, la superficie en tres dimensiones es equivalente a la línea en tres dimensiones. Como la superficie tiene el carácter de valor medio, algunos puntos se encuentran debajo de ella y otros sobre [diagrama (*c*)].

El valor residuo de cada punto indica su posición, negativa o positiva, en relación a la superficie que caracteriza la tendencia general [diagrama (*d*)]. Cuando el valor actual u observado del punto lo deja debajo del nivel de la superficie esperada [punto 3 en diagrama (*d*)], el residuo tiene un valor positivo porque hay una diferencia positiva entre el valor esperado y el valor observado. El valor esperado es la altura perpendicular de la superficie al pun-

* Inglés: *Trend Surface Analysis.*

80 30	23 64	67 96	21 33	36 90	03 91	69 33	90 13	34 48	02 19
61 29	89 61	32 08	12 62	26 08	42 00	31 73	31 30	30 61	34 11
23 33	61 01	02 21	11 81	51 32	36 10	23 74	50 31	90 11	73 52
94 21	32 92	93 50	72 67	23 20	74 59	30 30	48 66	75 32	27 97
87 61	92 69	01 60	28 79	74 76	86 06	39 29	73 85	03 27	50 57
37 56	19 18	03 42	86 03	85 74	44 81	86 45	71 16	13 52	35 56
64 86	66 31	55 04	88 40	10 30	84 38	06 13	58 83	62 04	63 52
22 69	58 45	49 23	09 81	98 84	05 04	75 99	27 70	72 79	32 19
23 22	14 22	64 90	10 26	74 23	53 91	27 73	78 19	92 43	68 10
42 38	59 64	72 96	46 57	89 67	22 81	94 56	69 84	18 31	06 39
17 18	01 34	10 98	37 48	93 86	88 59	69 53	78 86	37 26	85 48
39 45	69 53	94 89	58 97	29 33	29 19	50 94	80 57	31 99	38 91
43 18	11 42	56 19	48 44	45 02	84 29	01 78	65 77	76 84	88 85
59 44	06 45	68 55	16 65	66 13	38 00	95 76	50 67	67 65	18 83
01 50	34 32	38 00	37 57	47 82	66 59	19 50	87 14	35 59	79 47
79 14	60 35	47 95	90 71	31 03	85 37	38 70	34 16	64 55	66 49
01 56	63 68	80 26	14 97	23 88	59 22	82 39	70 83	48 34	46 48
25 76	18 71	29 25	15 51	92 96	01 01	28 18	03 35	11 10	27 84
23 52	10 83	45 06	49 85	35 45	84 08	81 13	52 57	21 23	67 02
91 64	08 64	25 74	16 10	97 31	10 27	24 48	89 06	42 81	29 10
80 86	07 27	26 70	08 65	85 20	31 23	28 99	39 63	32 03	71 91
31 71	37 60	95 60	94 95	54 45	27 97	03 67	30 54	86 04	12 41
05 83	50 36	09 04	39 15	66 55	80 36	39 71	24 10	62 22	21 53
98 70	02 90	30 63	62 59	26 04	97 20	00 91	28 80	40 23	09 91
82 79	35 45	64 53	93 24	86 55	48 72	18 57	05 79	20 09	31 46
37 52	49 55	40 65	27 61	08 59	91 25	26 18	95 04	98 20	99 52
48 16	69 65	69 02	08 83	08 83	68 37	00 96	13 59	12 16	17 93
50 43	06 59	56 53	30 61	40 21	29 06	49 60	90 38	31 43	19 25
89 31	62 79	45 73	71 72	77 11	28 80	72 35	75 77	24 72	98 43
63 29	90 61	86 39	07 38	38 85	77 06	10 23	30 84	07 95	30 76
71 68	93 94	08 72	36 27	85 89	40 59	83 37	93 85	73 97	84 05
05 06	96 63	58 24	05 95	56 64	77 53	85 64	15 95	03 91	59 03
03 35	58 95	46 44	25 70	31 66	01 05	44 44	62 91	36 31	45 04
13 04	57 67	74 77	53 35	93 51	82 83	27 38	63 16	04 48	75 23
49 96	43 94	56 04	02 79	55 78	01 44	75 26	85 54	01 81	32 82
24 36	24 08	44 77	57 07	54 41	04 56	09 44	30 58	25 45	37 56
55 19	97 20	01 11	47 45	79 79	06 72	12 81	86 97	54 09	06 53
02 28	54 60	28 35	32 04	36 74	51 63	96 90	04 13	30 43	10 14
90 50	13 78	22 20	37 56	97 95	49 95	91 15	52 73	12 93	78 94
33 71	32 43	29 58	47 38	39 96	67 51	64 47	49 91	64 58	93 07
70 58	28 49	54 32	97 70	27 81	64 69	71 52	02 56	61 37	04 58
09 68	96 10	57 78	85 00	89 81	98 30	19 40	76 28	62 99	99 83
19 36	60 85	35 04	12 87	83 88	66 54	32 00	30 20	05 30	42 63
04 75	44 49	64 26	51 46	80 50	53 91	00 55	67 36	68 66	08 29
79 83	32 39	46 77	56 83	42 21	60 03	14 47	07 01	66 85	49 22
80 99	42 43	05 58	54 41	98 05	54 39	34 42	97 47	38 35	59 40
48 83	64 99	86 94	48 78	79 20	62 23	56 45	92 65	56 36	83 02
28 45	35 85	22 20	13 01	73 96	70 05	84 50	68 59	96 58	16 63
52 07	63 15	82 30	66 23	14 26	66 61	17 80	41 97	40 27	24 80
39 14	52 18	35 87	48 55	48 81	03 11	26 99	03 80	08 86	50 42

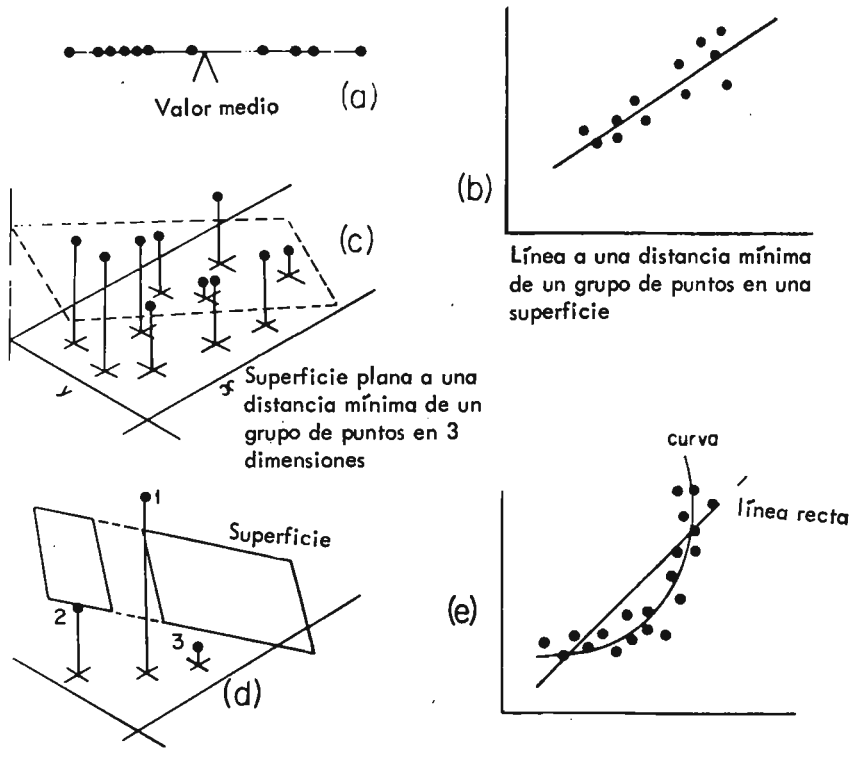
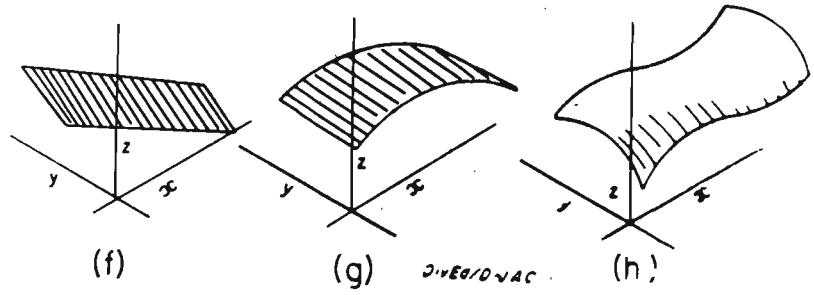


Figura 8.2a



to. Cuando el valor observado es mayor que el valor esperado [punto 1 en el diagrama (*d*)], se apunta un valor negativo. Los puntos que tienen residuos altos representan anomalías, casos que no conforman con la tendencia general. Son útiles porque indican los lugares que merecen investigaciones más detalladas.

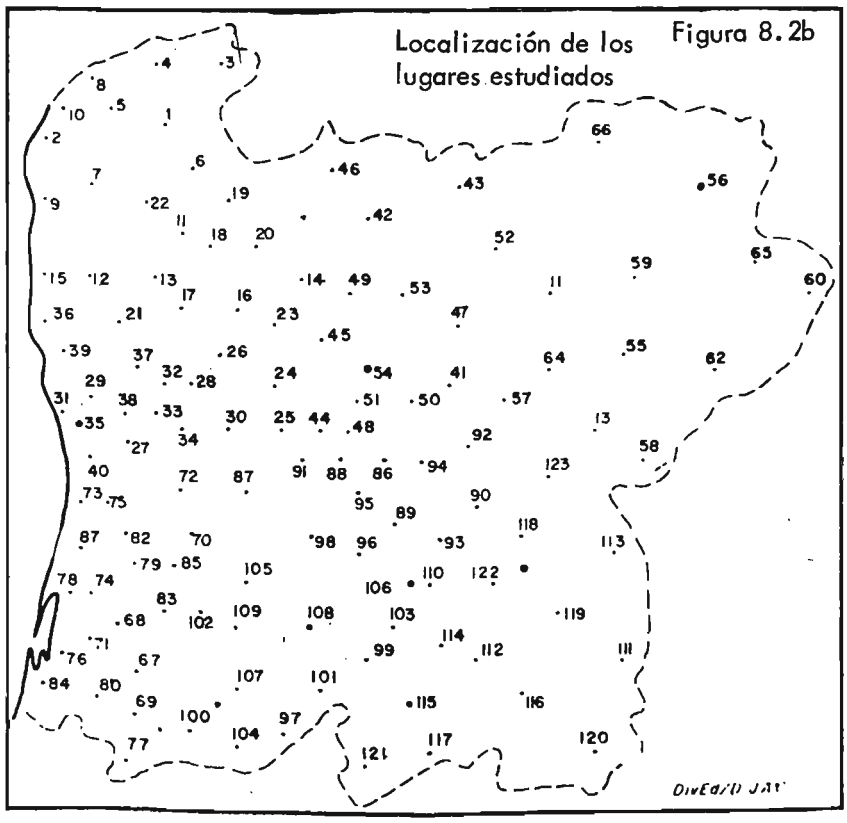
Generalmente, la primera superficie, la plana [ver diagrama (*f*)], se inclina del lado de la región que tiene los más altos valores observados al que tiene los valores más bajos, pero cuando los valores son iguales en toda la región, la superficie no se inclina. Como la primera superficie no es capaz de ajustarse exactamente a una distribución irregular de puntos en tres dimensiones, se calcula un índice que expresa el grado en que la superficie describe la distribución en forma de un porcentaje de 'explicación' de la variación total. En el caso de correspondencia exacta entre la superficie y la distribución de los puntos, la superficie explicaría 100 % de la distribución. Si, al contrario, los puntos tuviesen una distribución completamente fortuita, la superficie explicaría apenas 5-10 % de la variación total.

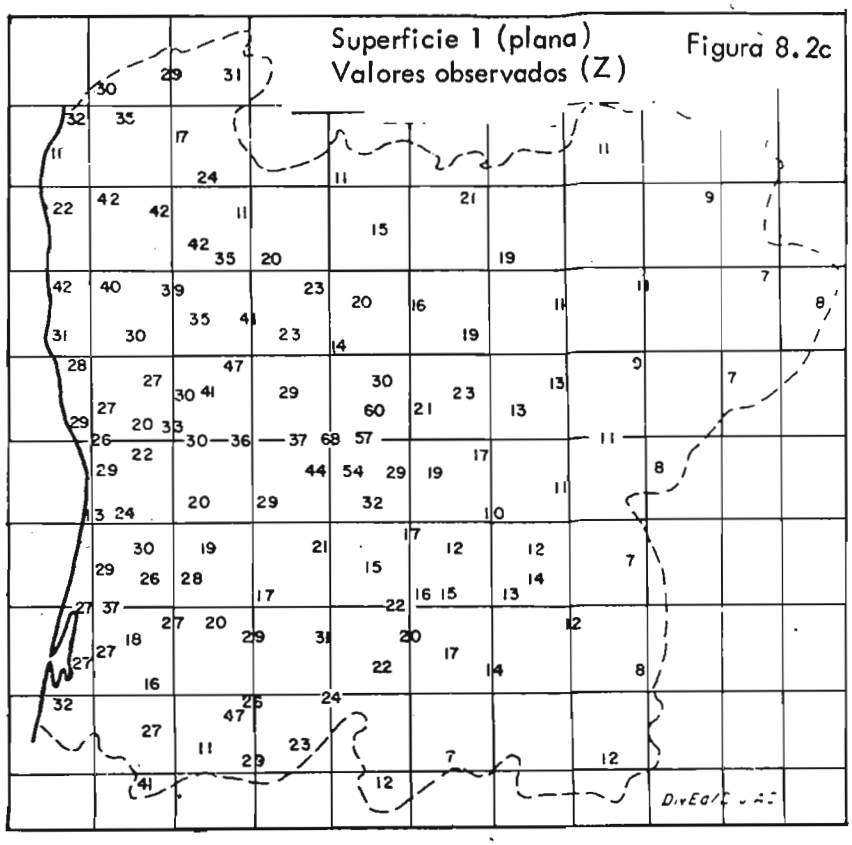
Es posible modificar la forma plana inicial de la superficie hasta una forma curva [diagrama (*g*)] que caracteriza con más precisión la distribución de los puntos en tres dimensiones. De la misma manera general, en el caso de una distribución de puntos en dos dimensiones, es posible que, en ciertas circunstancias [diagrama (*e*)], una curva caracterice mejor la distribución que una línea recta. Frecuentemente la aplicación de la segunda superficie disminuye notablemente los residuos y aumenta el porcentaje de la explicación. Las superficies más complejas se ajustan cada vez más exactamente a la distribución de puntos. Se llaman superficies cuadrada (segunda), cúbica (tercera) [diagrama (*h*)], cuarta, etcétera.

Las superficies de orden más elevado exigen cada vez más cálculos, y hasta los cálculos de la superficie plana son tan númerosos que necesitan el uso de computadoras electrónicas.

El análisis de superficies de tendencia se ilustra con el ejemplo de la densidad de la población rural en 123 consejos (unidades administrativas) en Portugal. La figura 8.2b muestra la posición de los consejos en el norte de Portugal y los datos del cuadro 8.2a la posición de cada consejo hacia el este (*x*) y hacia el norte (*y*) del punto de origen de una red escogida para el estudio (ver figura 8.2c). La tercera columna (*z*) contiene los valores observados de los consejos, esto es, la densidad de la población rural de cada consejo. En la cuarta columna se encuentran los valores esperados de la densidad de población según la primera superficie, la plana. Los residuos (columna 5) son las diferencias entre los valores observados y los valores esperados; se expresan en la forma de porcentajes (columna 6) en relación a los valores observados.

En la figura 8.2c se encuentran las densidades de la población rural de todos los




Superficie 1 (plana)
Valores observados (Z)
Figura 8.2c

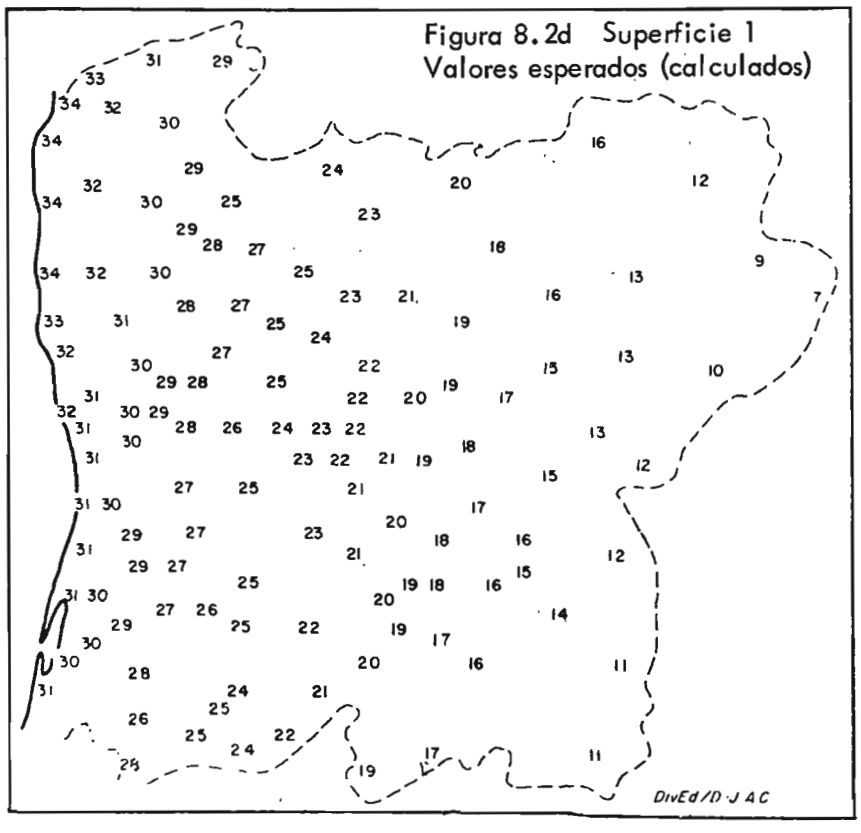
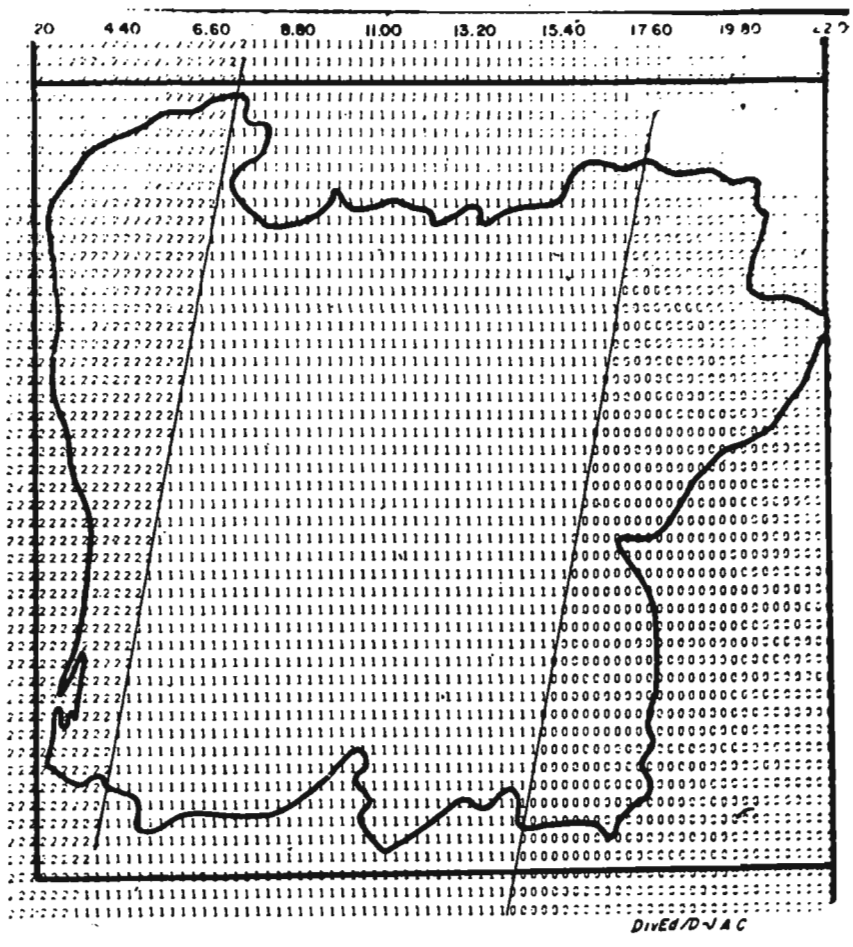


Figura 8.2d Superficie 1
Valores esperados (calculados)

Figura 8.2e Inclinación de la primera superficie



consejos. Se nota en este mapa una disminución general de los valores, del oeste al este. En el centro del área en estudio se hallan los consejos con las densidades más altas de toda la región, como, por ejemplo, consejo número 44, Mesão Frio, con una densidad de 68 personas por kilómetro cuadrado.

La figura 8.2d muestra los valores de la tendencia de la superficie de primer grado. Esta superficie explica apenas 28.5% de la variación total de la distribución de los puntos. En la figura 8.2e se encuentra la superficie plana desde la cual se calculan los valores esperados, los valores teóricos de los consejos si coincidieran exactamente con la superficie.

Se encuentran en la figura 8.2f el valor residuo de cada consejo, la diferencia entre su valor observado y su valor esperado en la primera superficie. Los valores residuos son positivos cuando los valores observados son inferiores a los valores esperados. Por ejemplo, en el caso del consejo número 2 (Caminha), el valor observado es 11 (densidad de la población rural actual), en tanto que el valor esperado es aproximadamente 35, debido a su posición en el oeste de la región en donde la tendencia es encontrar densidades altas; el valor residuo es 24. Lo contrario se nota en el caso del consejo número 44 (Mesão Frio) cuya densidad es muy alta (68 habitantes por kilómetro cuadrado); tiene un valor esperado de solamente 23.4 y un valor residuo muy grande, de 44.6.

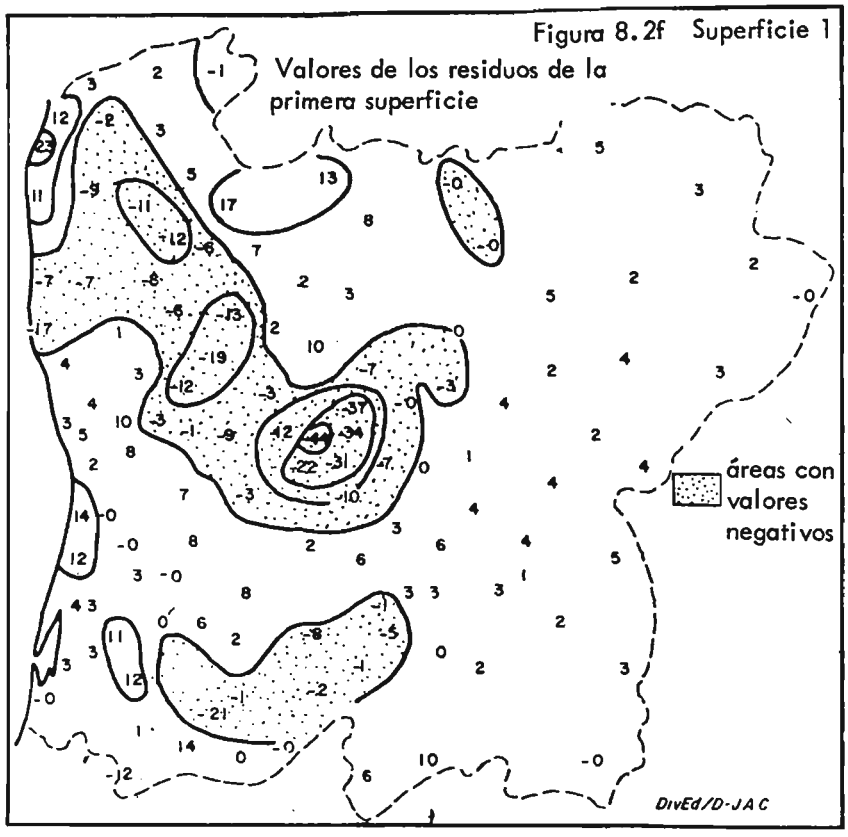


Figura 8.2f Superficie 1

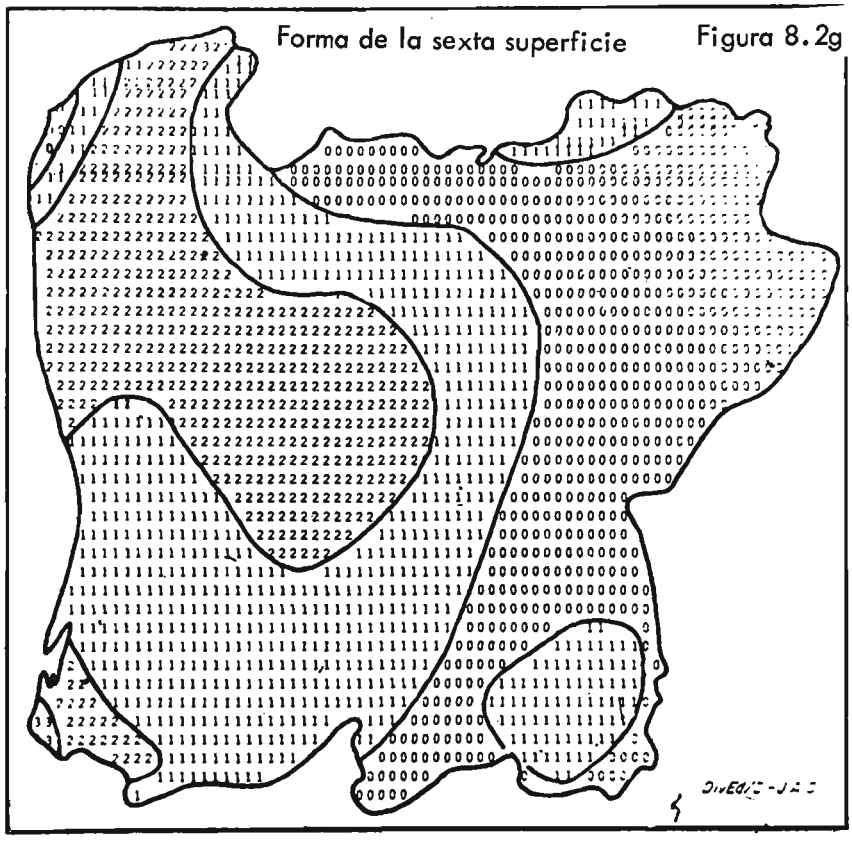


Figura 8.2g

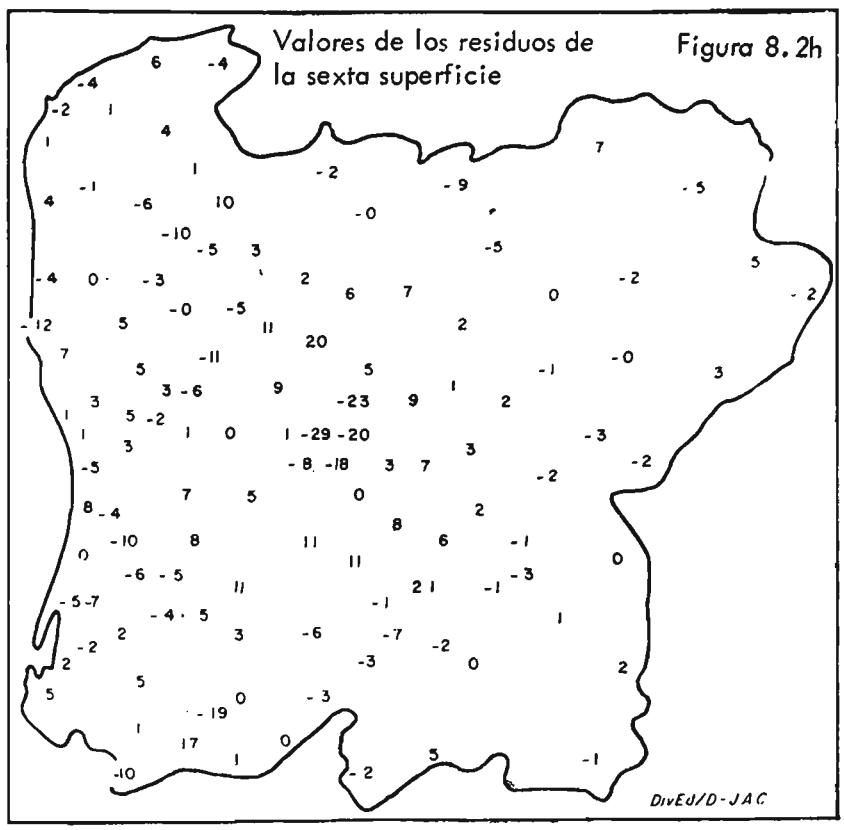
Interpretación de los resultados

La primera superficie de tendencia de primer grado es un plano inclinado de oeste-noroeste a este-sureste, lo cual indica que los valores de la densidad de la población rural en el norte de Portugal tienden a disminuir en esa dirección. La inclinación de la superficie revela cerca de la costa una zona de densidades altas; la zona es más ancha al norte de la ciudad de Porto (consejo número 35) que al sur. En la zona oriental del norte de Portugal, en la zona montañosa de las altiplanicies, cerca de la frontera con España se encuentran densidades muy bajas. Las excepciones más notorias de la tendencia general se encuentran en el medio Douro en donde, en la figura 8.2f, destaca la zona de producción del vino oporto.

En el estudio actual se calcularon seis superficies. En la figura 8.2g se representa la superficie de sexto grado, una figura compleja. Esta superficie ya explica 61.7 % de la variación total. Los residuos de la superficie se hallan en la figura 8.2h. Muchas anomalías se han reducido o eliminado, pero la zona del medio Douro destaca todavía como una región muy distinta.

Conclusiones

1. El análisis de superficies de tendencia tiene aplicaciones así en la geografía física como en la geografía humana.
2. Da una descripción general de una distribución complicada.
3. Revela anomalías en una distribución, indicando dónde vale la pena investigar más detalladamente.
4. Su función principal es permitir predicciones de una muestra de puntos a todos los puntos de una región. Por ejemplo, se pueden considerar las estaciones meteorológicas de una región como una selección o muestra de todos los puntos de la super-



ficie de la Tierra, porque en cada punto cae lluvia y se experimentan cambios de temperatura. De la muestra de puntos con sus valores observados se calculan las distintas superficies de tendencias. De los valores esperados de las superficies se estima el valor de cualquier lugar, con determinado coeficiente de confianza.

Sistemas en geografía

9. SISTEMAS EN GEOGRAFÍA

9.1 *El concepto de un sistema*

El concepto de sistema no es nuevo; sin embargo, se ha desarrollado mucho, desde la Segunda Guerra Mundial, el uso del término y el análisis de sistemas, una nueva rama de la ciencia aplicada.

La Society for General Systems Research publica un anuario desde el año 1955. El concepto de sistema se ha usado en ciertos estudios sobre geomorfología, meteorología y geografía humana.

En términos muy generales, se considera sistema a un conjunto de elementos distintos. Los elementos están interrelacionados y el sistema funciona como una entidad. Un cambio en un elemento del sistema influye en todos los elementos o, por lo menos, en una parte considerable de ellos. Muchas veces la representación diagramática de un fenómeno (la atmósfera, una máquina, la economía), en forma de sistema, ayuda al estudio y la apreciación de su funcionamiento.

Desde el decenio de los cincuentas se ha desarrollado un vocabulario de términos técnicos referentes a los sistemas. Por ejemplo: se reconocen sistemas abiertos y sistemas cerrados. El abierto es susceptible de tener influencias de fuera. Se habla de los límites de un sistema y del ambiente en el cual se encuentra. Hay niveles de sistema según sus capacidades de adaptación al ambiente. Un reloj es un sistema puramente mecánico, mientras que un animal, sistema orgánico, es capaz de adaptarse a variaciones de las condiciones del ambiente. Por ejemplo, las constantes fisiológicas del cuerpo de un mamífero se mantienen a determinado nivel; esta propiedad se llama homeostasia. Ciertos sistemas son capaces de aprender, se "retroalimentan" con sus acciones, lo que les ayuda a mantener control sobre lo que hacen. Se habla también de varios estados, siendo los más importantes un estado de equilibrio y un estado de desequilibrio. Los términos clave de la teoría de los sistemas se definen más ampliamente en un artículo de O. R. Young, aparecido en *Yearbook of the Society for General Systems Research*, Volume IX, 1964, p. 61.

Esta sección consiste en comentarios sobre algunos diagramas que ilustran unos sistemas sencillos. Hasta cierto punto, un sistema es un modelo; sobre todo, es una simplificación de un fenómeno observado o conocido en el mundo actual. Incluye solamente los elementos esenciales para revelar sus funciones y características básicas.

Ejemplo 1 (ver figura 9.1a)

Un sistema de calefacción en una casa particular. En este ejemplo, el gas en com-

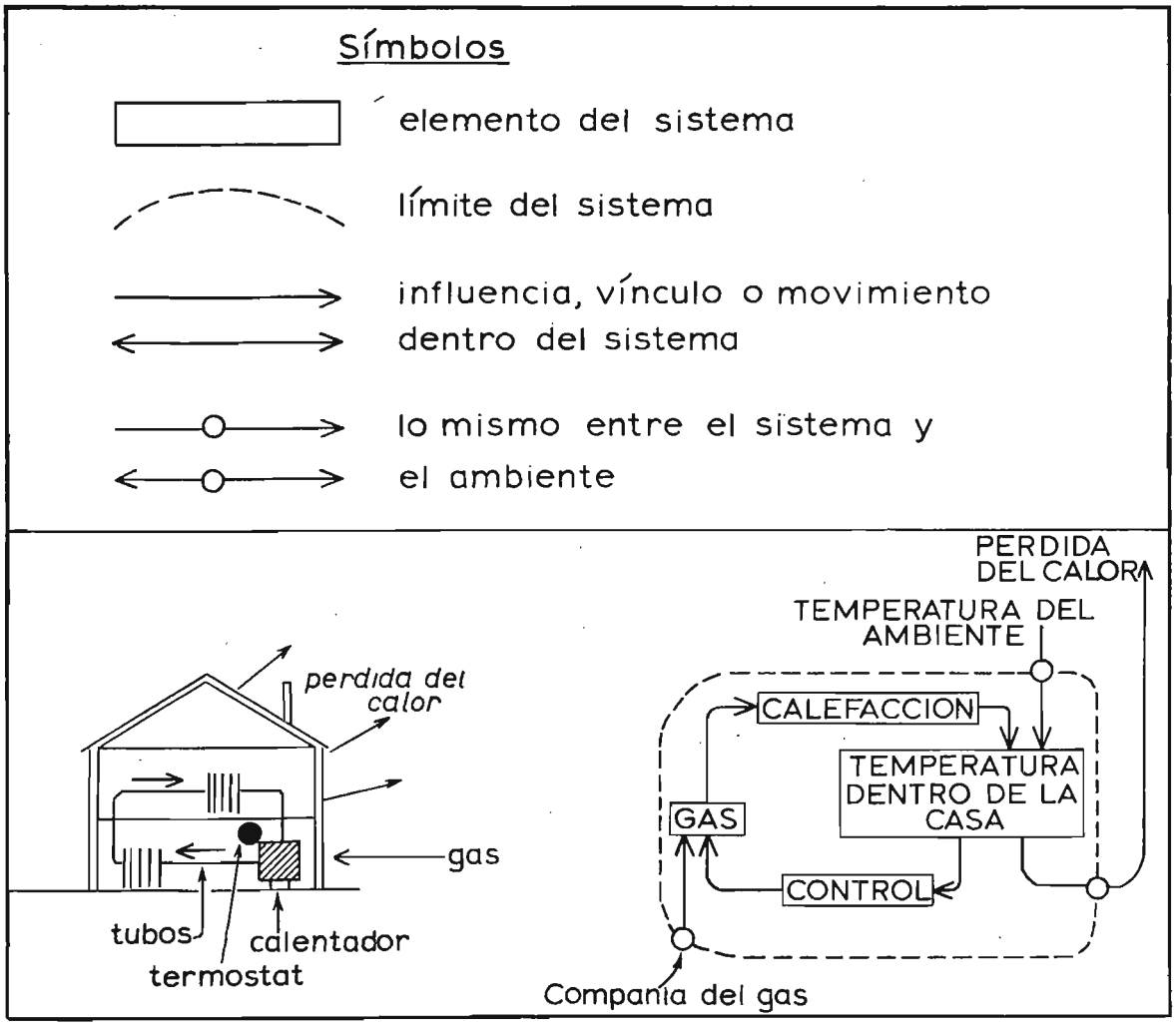


Figura 9.1a

bustión calienta el agua de un calentador y el agua pasa por tubos a los cuartos de la casa. La temperatura del agua y de los tubos afecta la temperatura interior de la casa que constantemente pierde calor por el techo, las ventanas y las paredes. El control del sistema es el termostato que regula constantemente la temperatura de la casa; cuando cae abajo de un nivel determinado, el termostato 'pasa la información' al calentador, el cual se enciende. Cuando la temperatura vuelve al nivel deseado, se corta el abastecimiento de gas.

En el sistema hay tres controles. Los cambios de la temperatura del ambiente influyen en la temperatura interior de la casa; variaría mucho si no fuera por el segundo control, el termostato. Hay otro control, la compañía que vende el gas, la que permite su uso solamente cuando se ha pagado.

En el sistema de calefacción, un cambio en un elemento del sistema afecta a los otros. Un cambio en el ambiente externo afecta al sistema interno, y un aumento de la temperatura dentro de la casa afecta al ambiente externo, aunque en escala muy limitada.

Ejemplo 2 (ver figura 9.1b)

La representación más sencilla posible

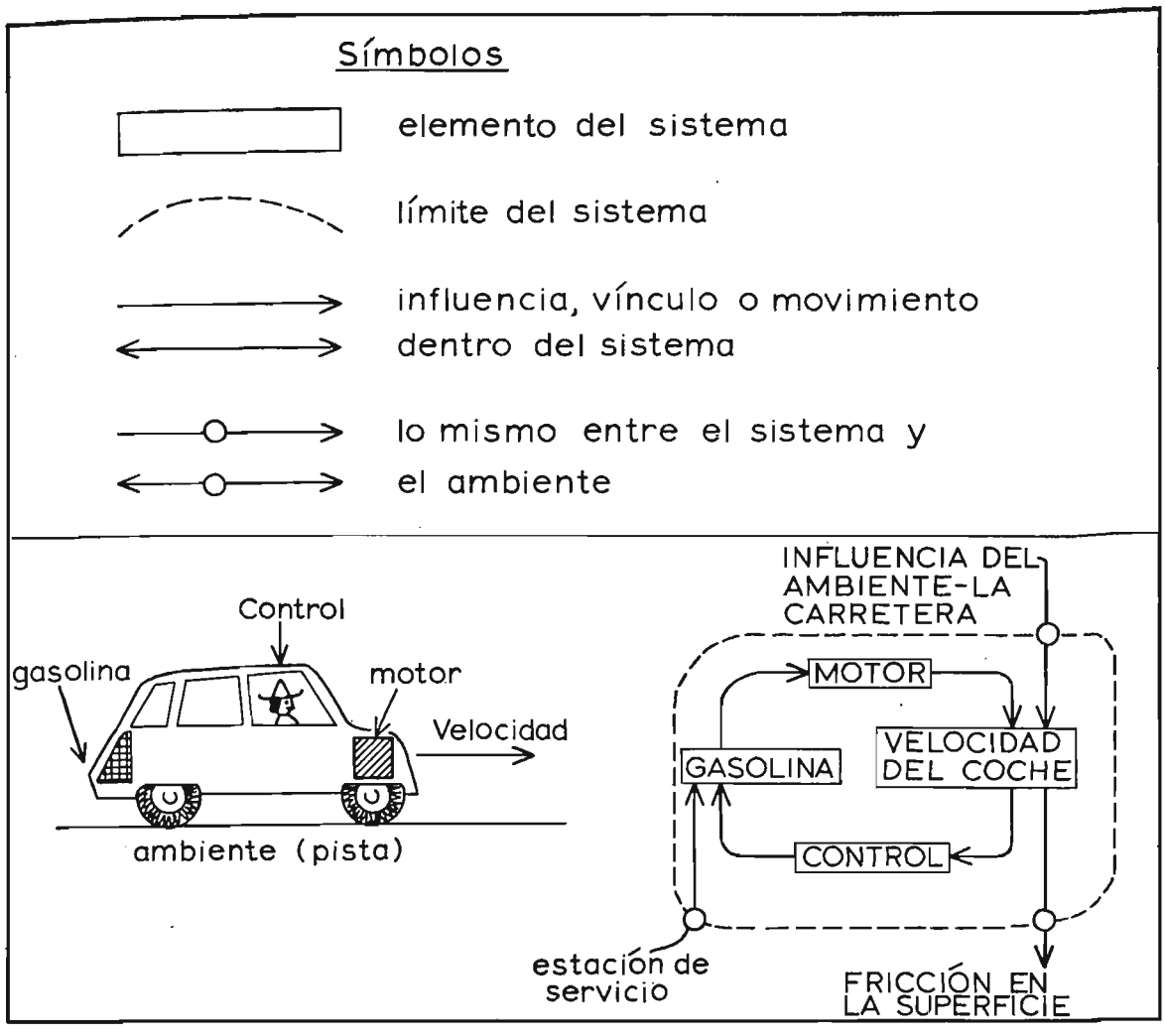


Figura 9.1b

de la función de un coche y de sus elementos fundamentales, en forma de sistema, demuestra semejanza notable con el sistema de calefacción, figura 9.1a. El control (el chofer) mantiene determinada velocidad aumentando o disminuyendo el consumo de gasolina por el motor, según las condiciones (declives, curvas) de la carretera que sigue y el ambiente en que viaja.

Ejemplo 3 (ver figura 9.c)

Muchas situaciones geográficas son muy complejas. En términos de sistema, tienen muchos elementos. Se presentan en la figura 9.1c los elementos básicos de la vida y el funcionamiento de una comunidad en los Andes del Perú. En el pueblo de Chaquicocha viven como 600 personas en una superficie de 600 hectáreas, a una altitud de 3,500 metros. Cultivan casi exclusivamente cebada y patatas y tienen varios tipos de ganado. Se puede dar en palabras una descripción del pueblo. Una descripción en forma de sistema da idea más clara y sucinta de los elementos y de las relaciones entre sí.

En el pueblo de Chaquicocha ciertos cambios en el ambiente afectan la vida entera de la comunidad. Cuando hay heladas las patatas sufren y el consumo humano de alimentos se reduce. Cuando no llueve

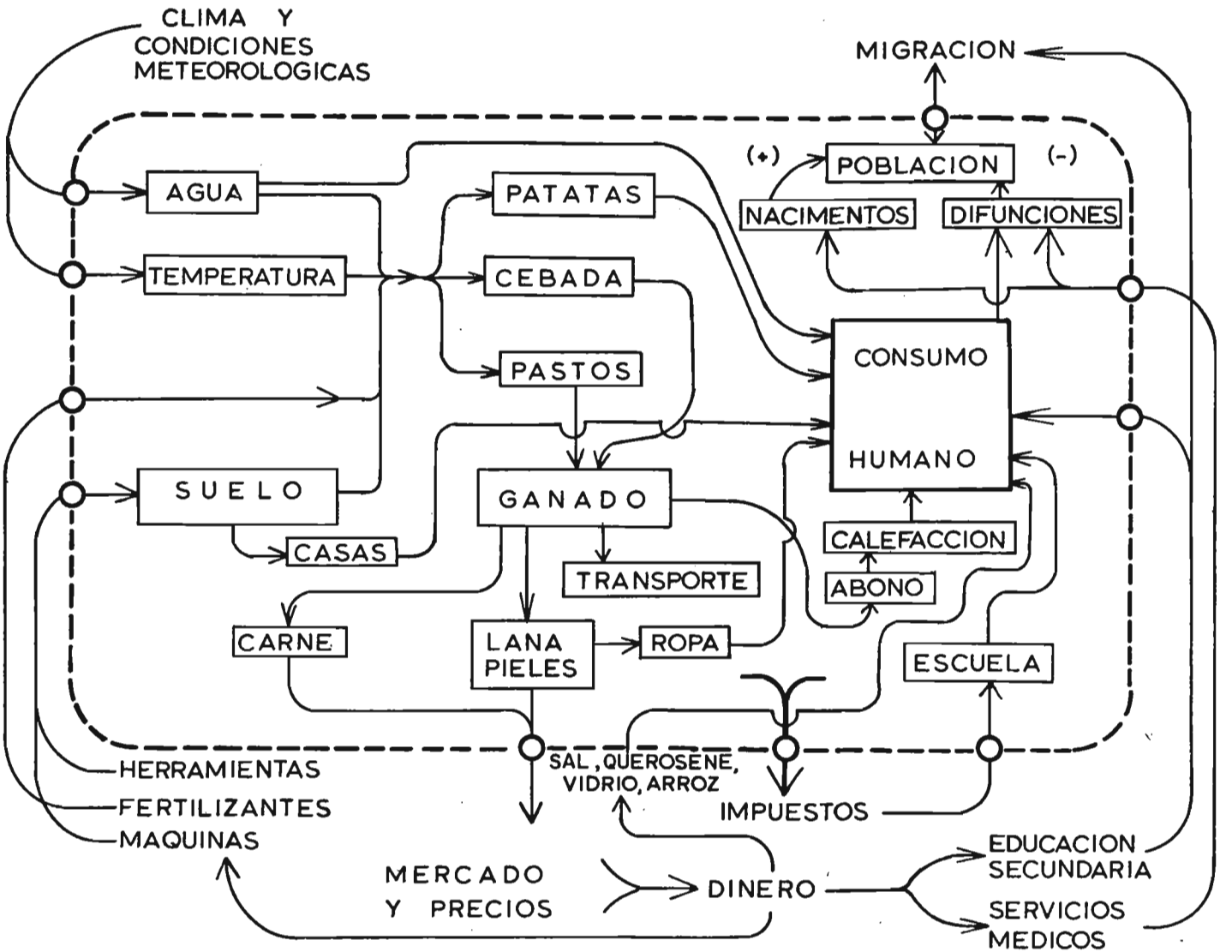

CLIMA Y
CONDICIONES
METEOROLOGICAS
MIGRACION
(+)
POBLACION
(-)
NACIMENTOS
DIFUNCIONES
AGUA
PATATAS
TEMPERATURA
CEBADA
PASTOS
CONSUMO
HUMANO
SUELO
GANADO
CASAS
TRANSPORTE
CALEFACCION
ABONO
CARNE
LANA
PIELES
ROPA
ESCUELA
HERRAMIENTAS
FERTILIZANTES
MAQUINAS
SAL, QUEROSENE,
VIDRIO, ARROZ
IMPUESTOS
MERCADO
Y PRECIOS
DINERO
EDUCACION
SECUNDARIA
SERVICIOS
MEDICOS

Figura 9.1c

mucho los pastos rinden poco, el ganado sufre y la única fuente de productos, que se venden fuera del pueblo, disminuye. Si la población crece excesivamente aumentan la migración o las defunciones. El nivel de vida en Chaquicocha es igual, aproximadamente, al consumo de productos y de servicios dividido entre el número de habitantes. ¿Es posible introducir cambios en el sistema, que aumentarían el nivel de vida? Sería posible reducir el número de habitantes o comprar y aplicar, de manera efectiva, más fertilizantes. Se pueden seguir las flechas del sistema estudiando las repercusiones de cambios en los elementos internos del sistema y en las influencias externas. Posiblemente el aspecto más difícil de entender es la mentalidad de los habitantes, sobre todo de las personas que toman las decisiones que afectan la vida de la comunidad.

Ejemplo 4

Si se necesitan 20 elementos o variables en el estudio de un pueblo de 600 habitantes, ¿cuántos elementos serían precisos para un estudio del mundo, con sus problemas actuales y prospectos futuros? Por lo menos 100, según los autores de un libro que

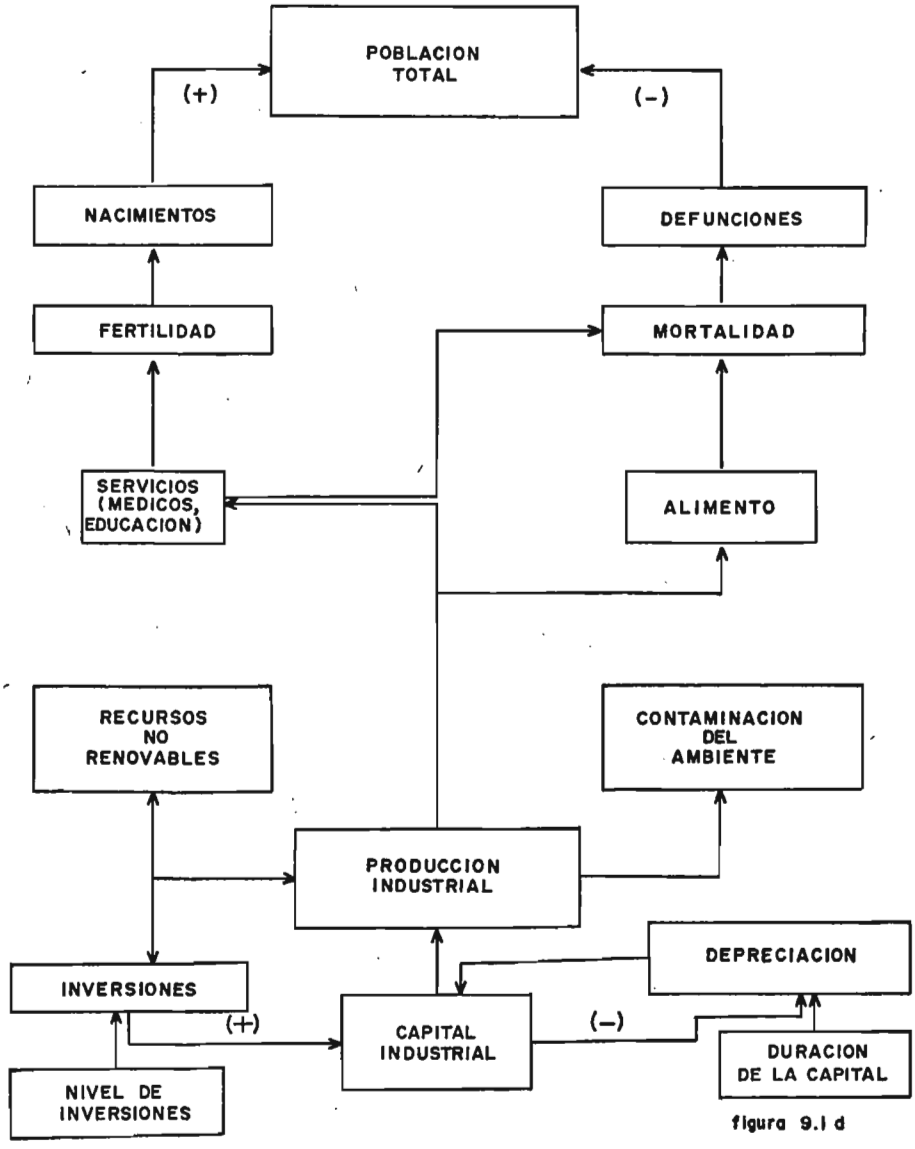


figura 9.1 d

trata de ese tema. En el libro *The Limits to Growth* * se reproduce un modelo o sistema que incluye los elementos fundamentales y las interacciones que forman la base de la vida humana.

En el mismo libro se subraya la necesidad de reducir el número de variables y de interacciones en el sistema mundial, para poder manejar los procesos más importantes y, simulando cambios durante un periodo de tiempo, plantear varias futuras alternativas. Su diagrama básico se reproduce en la figura 9.1d, con modificaciones. Se seleccionaron, en particular, los elementos siguientes:

1. Nacimientos.
2. Defunciones.
3. Población total.
4. Consumo de los recursos naturales.
5. Nivel de producción agrícola e industrial.
6. Nivel de servicios.
7. Contaminación del ambiente.

Es interesante notar dos diferencias fundamentales entre el sistema mundial (figura 9.1d) y los otros sistemas (9.1a-9.1c). En primer lugar, *el sistema* mundial no tiene límite. Eso significa que (salvo la influencia de la energía del Sol) no hay interacción entre este sistema y su ambiente. En segundo, en el sistema no hay ningún control central.

Ejemplo 5

¿Es posible considerar a un país como un sistema? Tendría que ser un sistema abierto, porque ningún país vive completamente aislado del resto del mundo. En el caso de muchos países actualmente subdesarrollados, son precisamente las influencias de otras partes del mundo las que penetran el sistema y producen cambios.

En este sistema se supone que las innovaciones tecnológicas provenientes de los países industriales producen cambios en un país subdesarrollado:

1. La introducción de métodos nuevos en la agricultura (fertilizantes, máquinas, semillas) aumentan la producción por persona y por hectárea.
2. Se hacen disponibles capital y mano de obra para el desarrollo de sectores de la economía agrícola.
3. El mejoramiento del abastecimiento de alimentos y la introducción de servicios médicos reducen la morbilidad y las defunciones.
4. La población aumenta.
5. Como las industrias de transformación y los servicios tienden a concentrarse en centros de población mayores, aumenta la población urbana.
6. Con la especialización regional disminuye la independencia económica de las comunidades rurales.
7. Aumenta el movimiento de carga, el comercio interregional y la red de ferrocarriles y de carreteras.
8. Con la especialización aumenta la diferenciación regional. Crece la migración de las zonas rurales a las ciudades y de regiones pobres a regiones afluentes.
9. La educación y el conocimiento de métodos para el control de la natalidad baja la fertilidad y reduce la discrepancia entre nacimientos y defunciones.

Desgraciadamente no es tan fácil representar esta cadena de relaciones y reacciones en forma de diagrama. Una tentativa se encuentra en la figura 9.1e.

* D. H. Meadows, *The Limits to Growth* (Earth Island Ltd., London, 1972).

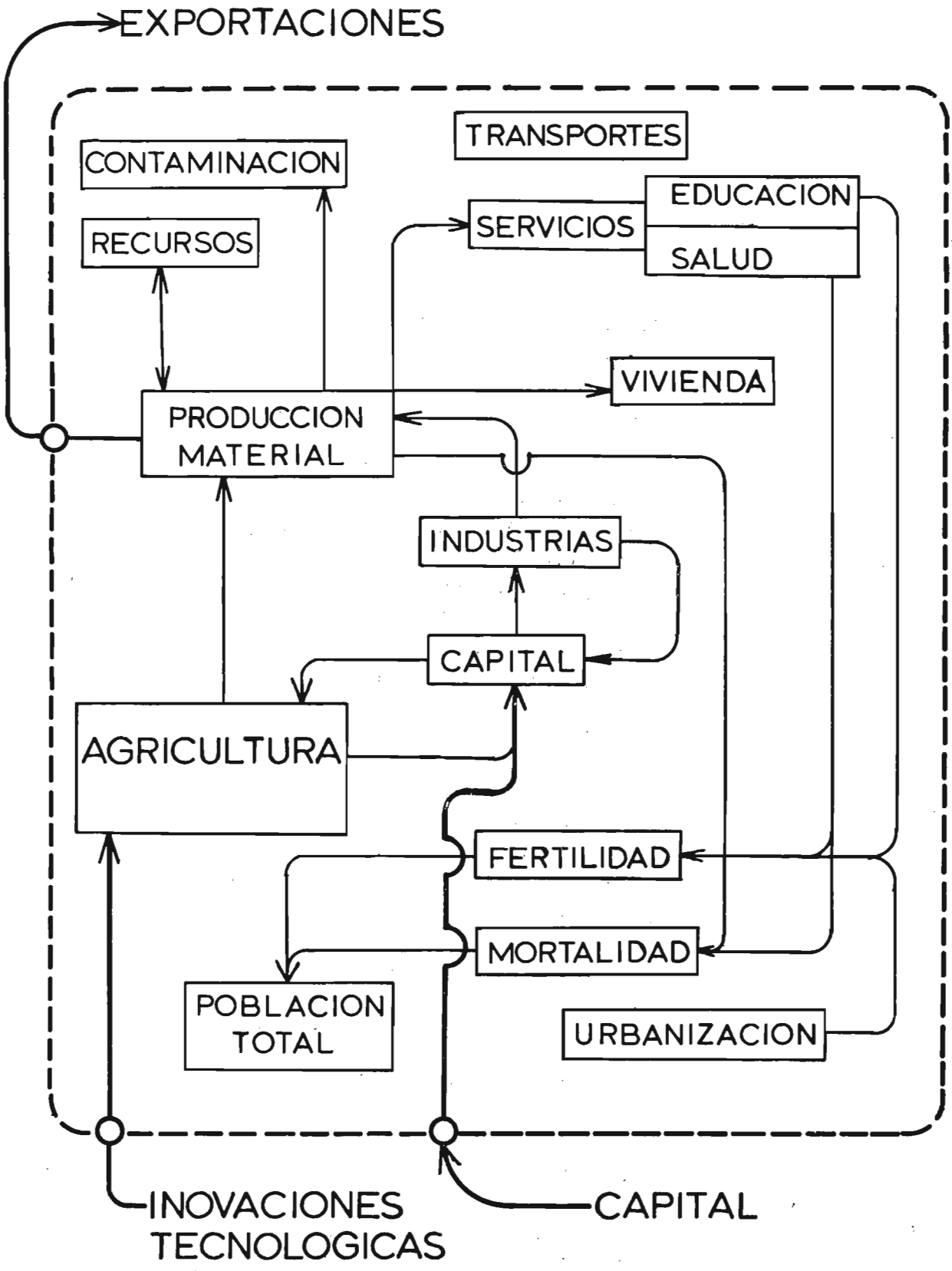

EXPORTACIONES
TRANSPORTES
CONTAMINACION
EDUCACION
SERVICIOS
SALUD
RECURSOS
VIVIENDA
PRODUCCION
MATERIAL
INDUSTRIAS
CAPITAL
AGRICULTURA
FERTILIDAD
MORTALIDAD
POBLACION
TOTAL
URBANIZACION
INOVACIONES
TECNOLOGICAS
CAPITAL

Figura 9.1.e

Bibliografía

ABLER R., J. ADAMS y P. GOULD, Spatial Organization, the geographer's view of the world. Prentice Hall International Editions, Londres, 1972.

BERGAMINI D., Mathematics, Time-Life International (Holanda), N. V., 1965.

CHORLEY, R. J. y HAGGET P. (Eda), Models in Geography, Methven, Londres, 1967.

COLE, J. P., "Mathematics and Geography". *Geography,* Vol. IV, Parte 2, abril, 1969, pp. 152-164. Inglaterra.

COLE, J. P. y Mather, "México 1970: Estudio geográfico usando análisis de factores", *Boletín del Instituto de Geografía,* Vol. V, UNAM. México, 1974.

COLE, J. P. y BENYON, N. J., New Ways in Geography, Blackwell, Oxford, 1968.

COLE, J. P. y KING, C. A. M., Quantitative Geography: Techniques and Theories, Wiley, Londres, 1968.

DUNCAN, O. D. y otros, Statistical Geography, Frec. Press of Glencoe, Londres, 1961.

FREUDENTHAL, H., Mathematics Observed, World University Library, Weidenfeld and Nicholson, Londres, 1967.

FRUCHTER, B., Introduction to Factor Analysis, Van Nostrand, Princeton, Nueva Jersey, 1954.

FUCHS, W. R., Modern Mathematics, Modern Science Series, Weidenfeld and Nicholson, Londres, 1967.

GOLDBERG, S., Probability an introduction, Prentice Hall, Englewoods Cliffs, Nueva Jersey, 1960.

GOULD, P. R., "Man against his environment, A game theoretic framework", *Annals of the Association of American Geographers,* Vol. 53, Nº 3, septiembre, 1963, pp. 290-297.

GREGORY, S., Statistical Methods and the Geographer, Longmans, Londres, 1963.

HÄGERSTRAND, Torsten, "The propagation of innovation waves", *Lund Studies in Geography,* Ser. B. Human Geography Nº 4. The Royal University of Lund, Suecia.

HAGGETT, P. Locational analysis in Human Geography, Arnold, Londres, 1965.
LEWIS, W. D., Mathematics Makes Sense, Heinemann, Londres, 1966.
ROSENTHAL, E. B., Understanding the New Mathematics, Souvenir Press, Londres, 1966.

Índice

Una introducción al estudio de métodos cuantitativos aplicables en geografía, editado por la Dirección General de Publicaciones, se terminó de imprimir en los talleres de GRÁFICA PANAMERICANA, S. DE R. L., Parroquia 911, México 12, D. F., el día 15 de diciembre de 1975.

Se tiraron 2 000 ejemplares.