LA SELECCION DE VARIABLES EN EL ANALISIS MULTIVARIANTE

José Cortizo Alvarez

(Universidad de León)

1. Introducción

El objetivo principal de esta comunicación es presentar los problemas que se nos han planteado a la hora de seleccionar unas variables para tratar de establecer una comparación entre diversos países del área latinoamericana, basada en el análisis de componentes principales y cluster a partir de dichas variables.

"Ocurre a menudo que el geógrafo se encuentra en la obligación de tratar una masa ingente de datos" (BEGUIN, H., 1979, p. 152). Es éste un hecho bastante frecuente en la investigación geográfica, complicado, además, por la distinta naturaleza de la información.

Para hacer manejable un conjunto de datos se hace necesario recurrir a técnicas estadísticas que ayuden a sintetizar la información en un doble sentido: reduciendo la masa de información y poniendo de manifiesto las relaciones entre las variables portadoras de la misma (BEGUIN, H., 1979).

Si tenemos en cuenta que la aplicación de métodos estadísticos está relacionada con la variación de alguna característica, la aplicación de técnicas estadísticas, concretamente del análisis multivariante, pretende explicar la variación experimentada por una o más variables en términos de la variación conjunta con otras, de la covariación (MALLO, F., 1984), respondiendo así al doble sentido a que hacía alusión BEGUIN anteriormente.

Sin embargo, uno de los primeros y principales problemas que se le plantean al investigador es precisamente el de la selección de las variables a medir sobre una serie de observaciones o individuos. "La elección de datos a introducir en el análisis es sin duda la etapa más importante" (CICERI, M.-F. et al., 1977, p. 20), puesto que los factores o componentes resultantes simplemente traducen la estructura de los datos originales.

Estos datos se toman en consideración porque, a priori, se piensa que son relevantes y que aportan información al análisis; no obstante, en muchas ocasiones, quedan relegados del mismo ciertos temas o ciertos aspectos que el investigador juzga como importantes o incluso como esenciales, simplemente

porque "pueden estar mal o nada representados en las estadísticas disponibles" (CICERI, M.F., et al., 1977, p. 21), lo cual puede originar importantes desequilibrios.

En definitiva, como señala BAILLY refiriéndose al medio urbano, "la ecología de la ciudad viene, de hecho, definida por la matriz de datos iniciales, antes que por el análisis factorial" (BAILLY, A.S., 1978, p. 154).

Para nuestro trabajo hemos utilizado un total de 50 variables referidas a 18 países del área de América Latina (Anexo 1).

Las variables de partida, los parámetros medidos sobre los individuos, - son dependientes, correlacionados, y no están sometidos a ninguna restricción; como corresponde a un análisis exploratorio, no hay ninguna hipótesis previa que considere a alguna variable más importante que las demás.

Pretendiendo abarcar un amplio espectro se han tomado variables que aportan información sobre aspectos demográficos , económicos (con datos sobre la estructura, dinámica y evolución de ambos) y socioeconómicos (educación y sanidad).

El límite viene dado, en principio, por la disponibilidad de los datos, ya que hemos tenido que centrarnos en aquellos que en las fuentes manejadas (1) estaban presentes en los 18 países, viéndonos obligados a prescindir de bastante información que, por lo menos a priori, se veía como importante (tal es el caso, entre otros muchos ejemplos que se podrían poner, de los datos referidos a "Gastos públicos en defensa, en % del P.I.B."). El segundo límite viene impuesto por la propia capacidad del ordenador utilizado (2).

Así es como, de 59 variables reunidas en un primer momento, seleccionamos las 50 reseñadas en el Anexo I y que han sido la base del análisis. (3)

2. Metodología

El análisis de componentes principales a que se ha sometido a las 50 variables se encuadra dentro del análisis multivariante, "conjunto de métodos cuyo fin esencial es poner de relieve las relaciones existentes entre individuos, entre los parámetros que los caracterizan yentre los individuos y los parámetros" (MALLO, F., 1984, p. 7).

El núcleo central del análisis de componentes principales lo constituye la consideración de la interdependencia entre las variables como un todo, a partir del cual la información suministrada es susceptible de ser sintetizada

en términos de otras nuevas variables. La diferencia esencial con el análisis factorial radica en que este modelo considera a toda variable compuesta de - dos partes, una específica de cada variable y otra común con otras variables y expresable en términos de factores comunes (BATISTA, J.M., 1984).

Las nuevas variables son las componentes principales, obtenidas de tal modo que "cada componente sucesivo explique la máxima varianza (posible) restante después de la extracción de las componentes precedentes. Las componentes principales son una transformación matemática y no exigen la adaptación a un modelo" (MATHER, P.M., 1981, p. 146).

Las componentes principales son ortogonales, estadísticamente independientes, y son combinaciones lineales de las variables originales. La trasnformación es puramente matemática y no estadística, de manera que no es necesaria ninguna hipótesis previa, como ya hemos visto. En este sentido, "la información contenida en los datos iniciales no disminuye ni aumenta en el curso de esta transformación, sino que simplemente se presenta bajo una forma nueva - que puede revelar las relaciones importantes difíciles de descubrir entre los datos originales" (CICERI,M.-F., et al.,1977, p. 15).

Esta técnica no solamente transforma las variables originales sino que - tiene por finalidad, además, reducir su dimensión al obtener "un número más reducido de variables (...) que por su mayor relevancia conceptual pueden, en posteriores aplicaciones, sustituir a las primitivas variables" (BATISTA, J.M., 1984, p. 25-26), eliminando de esta manera la información redundante.

La última etapa de nuestro trabajo ha consistido en un análisis cluster - aplicado a las componentes principales obtenidas, en sustitución de las va-- riables originales, tal como apunta BATISTA más arriba.

El análisis cluster tiene como "objetivo principal el agrupamiento "razonable" de los individuos objeto de estudio" (MALLO, F., 1984, p. 108) a partir de una serie de caracteres observados sobre cada individuo.

Como en el caso de las componentes principales, este análisis cluster es un método exploratorio en el que no existen hipótesis previas sobre los individuos ni sobre las clases o conglomerados, sino que "sólo se dispone de una colección de observaciones cuya pertenencia a la clase es desconocida. El - objetivo operacional consiste, por tanto, en descubrir una estructura de conglomerados que se ajuste a las observaciones" (MALLO, F., 1984, p. 168).

Así pues, lo que intentaremos hacer es una agrupación objetiva de los individuos a partir de los datos originales transformados en componentes.

3. Desarrollo de los análisis

3.1. Componentes principales

Aplicando un primer análisis de componentes principales a las 50 variables originales ya aludidas obtenemos otras tantas componentes que explican el 100 % de la varianza de las variables. Sin embargo (Tabla I), solamente las 17 primeras componentes tienen una varianza superior a 0,00, explicando el total de la varianza, es decir, sintetizando toda la información original. No obstante, solamente las primeras 11 componentes tienen autovalores iguales o superiores a 1,0, dándonos una varianza acumulada del 94,03 %, porcentaje más que suficiente de la varianza total y que nos permite prescindir de las componentes 12 a 17 (inclusives).

Tabla I

VARIANZA EXPLICADA POR CADA COMPONENTE
VARIANZA ACUMULADA EXPLICADA POR LAS K PRIMERAS C.P RETENIDAS

C.P.	VARIANZA	VAR.EXPLICADA	VAR.ACUMULADA
1	13.370	26.740	26.740
2	6.724	13.449	40.189
3	6.411	12.823	53.013
4	5.519	11.038	64.051
5	3.014	6.028	70.079
6	2.773	5.546	75.624
7	2.258	4.515	80.140
8	2.108	4.216	84.356
9	2.030	4.059	88.415
10	1.600	3.200	91.615
11	1.199	2.399	94.013
12	.850	1.699	95.712
13	.717	1.433	97.146
14	.542	1.083	98.229
15	.399	.799	99.028
16	.269	.539	99.567
17	.216	.433	100.000

Hemos pasado, de esta manera, de trabajar con 50 variables, a todas luces imposibles de manejar y visualizar conjuntamente, a retenr 17 componentes incorreladas y que, además, resumen la información inicial. Estamos, pues, en camino de conseguir uno de los objetivos del análisis de componentes principales, el de reducir la dimensionalidad al considerar únicamente la información proporcionada por un conjunto de variables no observables y menor que el original (BATISTA; J.M., 1984).

Con todo, considerar de forma conjunta un total de 17 componentes tampoco es tarea fácil, por lo que es necesario reducir su número aunque se haga, eso sí, a costa de perder información. Esta pérdida no resulta tan grave si pasamos a retener, como ya hemos apuntado, las 11 componentes con

autovalor superior a 1,0, ya que este es un punto de ruptura cómodo porque "la cantidad de varianza que explican estos componentes es superior a la - contenida en una variable al menos" (CICERI, M.-F. et al., 1977, p. 27).

No obstante, esta pérdida del 6 % de la información apenas si nos aporta ventaja alguna a la hora de tratarla. Se plantea, entonces, una nueva - reducción del número de componentes hasta hacerlo manejable. En este caso lo hemos fijado, en un principio, en seis, que explican dada una más del 5 % de la varianza y que en conjunto da una varianza acumulada del 75,62 %.

Es decir, perdemos casi el 25 % de la información original pero, en contrapartida, las componentes retenidas tienen unos porcentajes de dependencia relativamente altos al menos con dos variables, lo cual hace que su interpretación pueda tener algún significado (Tabla II), aunque esto también nos presentará importantes problemas, como veremos más adelante.

Tabla II

Porcentajes de dependencia

Variables	C.P. 1	C.P. 2	C.P. 3	C.P. 4	C.P. 5	C.P. 6
1	21,39			-11,75		- 6,17
2			12,58		8,20	
6	51,00			38,81		
11		-38,69		36,08		
12						2,36
16		-10,64		43,82	4,55	
21	-49,89	14.84			-10,91	
26	59,66			22,89		
31	-57,76			-27,05		
36	57,52				-11,30	
41	-39,60	-12,31				- 4,75
46	-49,61		12,54		19,80	

Lo contrario, es decir, la consideración de las 11 o de las 17 primeras componentes nos hubiera permitido trabajar con menos o incluso sin - pérdida de información pero nos hubiese palanteado serios problemas a la fora de definir las componentes, debido precisamente a sus débiles porcentajes de dependencia con las variables.De todos modos, como hemos adelantado, estos problemas se nos han presentado en la interpretación de los resultados aún cuando tuvimos en cuenta las saturaciones superiores a 0,30, tal como aconsejan ciertos autores (CICERI, M.-F. et al., 1977).

Ahora bien, esta misma razón que utilizamos para desechar las componentes 7 a 11 podemos aducirla también para prescindir de la 5 y de la 6, ya que con ninguna variable alcanzan porcentajes de dependencia del 20 %, -

hecho éste que también se da en la componente 3.

Como vemos en la anterior Tabla II, prescindiendo ya definitivamente de las componentes 5 y 6, hay una serie de variables que se repiten en las componentes principales: solamente las variables 2, 21 y 36 están en una componente; el resto depende de dos, manteniendo con ellas incluso el mismo sentido positivo o negativo. Además, los pares de componentes 1-2 y 2-4 tienen dos variables comunes y el par 1-4 tiene tres: ? puede hablarse, sobre todo en este último caso, de "redundancia", de "solapamiento" o de "complementariedad" de las componentes?.

No se acaban aquí nuestros problemas, puesto que, por otra parte, si representásemos gráficamente la localización de los individuos sobre el espacio de las componentes a partir de la Tabla III, veríamos cómo en las seis combinaciones de las 4 componentes retenidas hay 10 individuos (el 55 %) - que se sitúan en el cnetro del gráfico porque tienen valores de 0,00 en las coordenadas de cada par de componentes.

Tabla III

COORDENADAS DE INDIVIDUOS EN EL ESPACIO DE LAS C.P.

INDIVI	CP 1	CP 2	CP 3	CP 4	CP 5	CP 6
1	-7.00814	-3.1126	.97515	.42136	-.63618	1.71593
2	-1.14493	-.47968	1.30868	.40394	-.61401	-.21347
3	.00161	.00037	.0009	.00076	-.00244	-.00279
4	.00028	.00065	-.0001	-.0008	-.00221	-.00059
5	.00193	-.00025	.00085	.0017	.0013	.0004
6	1.51658	-5.21283	1.32881	4.62005	2.04205	-2.82108
7	1.11275	-.55138	.32404	.24327	.36151	-.1039
8	-.0013	.00068	.0013	-.00092	-.00081	.00102
9	.0022	.00002	.00002	.00225	-.00078	-.00005
10	-.0014	-.00225	.00035	.00078	.00049	.0008
11	-3.02677	.52168	-6.74534	-.41187	2.14913	-.6267
12	-1.68137	-.19143	.19323	-1.32854	.29639	-.02225
13	-.00104	-.00132	-.00118	.00147	-.00168	.00009
14	.00004	.00215	-.00163	-.00082	.00087	.00001
15	.00224	.00036	-.001	.00111	-.00309	-.00085
16	-3.4919	-1.75801	1.35567	-3.97413	1.37882	-.64728
17	-.51474	-1.79122	-1.1465	.69732	.42773	.11182
18	.00132	-.00066	-.00045	.00015	.00112	-.00151

Así pues, se reúnen una serie de circunstancias negativas para desarrollar la interpretación de los resultados: 1ª) la varianza explicada por las componentes retenidas es baja (4); 2ª) tenemos una serie de variables que se repiten en las componentes; 3ª) débiles porcentajes de dependencia - de las componentes con las variables, y 4ª) más de la mitad de los individuos tienen coordenadas 0,00 en las componentes.

Llegamos a concluir, a la vista de todo ello, que las componentes principales obtenidas de este análisis tienen un poder discriminante muy bajo, una capacidad muy baja para separar los individuos.

Este problema nos lleva a replantearnos la cuestión de la selección de variables. Por ello, y sin entrar siquiera en la definición de las componentes, hemos hecho una segunda parte en el trabajo, consistente en elimnar aquellas variables que menos relevancia tenían para el análisis, al ser errónea una de las consideraciones iniciales, la de la "la variación del fenómeno estudiado es explicada por el conjunto de variables retenidas" (BEGUIN, H. 1979,p. 156).

El hecho es que si solamente las 17 primeras componentes tienen un autovalor superior a 0,00 ello significa que las 33 restantes no tienen ninguna importancia en la explicación de la información original. Esto nos abre una importante vía para reducir el volumen de datos a tratar sin pérdida de información puesto que podremos desechar ("to discard") aquellas variables con las que mayor porcentaje de dependencia tengan las componentes de autovalor 0,00.

La justificación del "discarding" de variables está en que si la última componente es la que menor autovalor tiene y que menos varianza explica, la variable más correlacionada con ella ser, lógicamente, la menos relevante, pudiendo prescindir de ella, y así sucesivamente hasta quedar con un número de variables que en nuestro trabajo hemos establecido en 17, inferior en uno al número de individuos. (RACINE, J.B.; REYMOND, H., 1973).

Es decir, prescindimos de las "variables que a priori se creía que iban a proporcionar información sobre el fenómeno en estudio, resultando su información nula" (MALLO, F. 1984, p. 21). De esta manera quedan fuera algunas variables (p.e. la 2,3 y 4) que mantenían altas correlaciones con otras no desechadas (la 5), es decir, eran redundantes.

Efectuado este proceso, pasamos a trabajar con las 17 variables marcadas en el Anexo I con un asterisco, de las que obtenemos las componentes de la Tabla IV. En este caso, las 16 primeras componentes explican ya el 100 % de la varianza, aunque solamente las 6 primeras tienen un autovalor superior a 1,0.

De esta manera, con igual número de componentes retenidos que en el primer análisis, tenemos menos pérdida de información. Otra ventaja es que, si bien en ningún caso son altos, los porcentajes de dependencia de las componentes se establecen con un reducido número de variables, con lo cual su identificación puede resultar más fácil que en el caso anterior.

Tabla IV

VARIANZA EXPLICADA POR CADA COMPONENTE
VARIANZA ACUMULADA EXPLICADA POR LAS K PRIMERAS C.P. RETENIDAS

C.P.	VARIANZA	VAR.EXPLICADA	VAR.ACUMULADA
1	5.703	33.549	33.549
2	3.121	18.357	51.906
3	1.913	11.252	63.158
4	1.582	9.307	72.465
5	1.281	7.537	80.002
6	1.068	6.281	86.284
7	.748	4.400	90.684
8	.451	2.653	93.337
9	.406	2.391	95.727
10	.308	1.813	97.541
11	.158	.931	98.472
12	.131	.773	99.244
13	.097	.568	99.812
14	.020	.120	99.932
15	.007	.043	99.975
16	.004	.025	100.000
17	.000	.000	100.000

LA VARIANZA DE UNA COMPONENTE ES EL AUTOVALOR CORRESPONDIENTE.

No obstante, seguimos ante el problema, aunque atenuado, de la repetición de variables en las distintas componentes (Tabla V).

Tabla V

Porcentajes de dependencia

Variables	C.P. 1	C.P. 2	C.P. 3	C.P. 4	C.P. 5	C.P. 6
1	-52,43	10,28		5,60	12,94	
3			13,70	5,78		
4					-32,09	7,52
5	11,43			-11,39		
6			-32,35			47,42
9						56,57
11			-49,15			9,22
12	13,08					
13	12,32	42,67				
15			-11,12	7,64		
16					-41,05	

Por último, y a diferencia también del primer intento, ya no nos encontramos con ningún individuo con coordenadas 0,00 para todas las componentes, sino que los factors retenidos en este segundo análisis tienen mayor poder discriminatorio entre los individuos (Tabla VI).

De todos modos, la interpretación de los resultados tampoco va a ser tarea fácil puesto que en ningún caso los porcentajes de dependencia componentes/variables son altos. Por lo tanto haremos solamente un ligero comentario, sin que ello obste para que sean tratados más en extenso en otro contexto.

Tabla VI

COORDENADAS DE INDIVIDUOS EN EL ESPACIO DE LAS C.P.

INDIVI	CP 1	CP 2	CP 3	CP 4	CP 5	CP 6
1	-4.62858	.35849	-1.89335	-.34338	1.84626	-.08877
2	-.38717	.69033	-.33614	.09893	-.08277	.01155
3	-1.95767	-1.56524	2.1782	.44953	.71887	.65831
4	-.11355	.0033	-.04473	-.00073	-1.71475	-4.44588
5	.00481	.75334	-.613	-.5662	-.83585	-.07712
6	-.00276	-1.29066	1.04207	-1.54964	.47786	-2.34089
7	-.36308	-.22577	-.10685	.00208	-.16345	-.08261
8	.35252	.90833	.23396	.36082	.51149	.96043
9	-.4767	.0569	-.05846	.08132	.00257	-.27745
10	1.81965	.71937	-.69476	-.05576	.48962	-.85663
11	.05462	.00407	-.42476	1.46868	2.29482	-1.45576
12	.17626	.65277	-.2169	-.37856	.21098	-.14807
13	-.8407	-1.64252	.66704	-.18396	.07415	-.32861
14	-.23225	-.28355	.26153	-.00041	-.08854	-.00999
15	1.45734	.39175	-.07774	-.38621	-.3673	-.04997
16	-.17909	.11261	.00456	3.57171	-.53462	.62489
17	-.41933	.14437	.24332	-.33942	.14261	-.74042
18	-3.90378	-2.2859	.83332	1.02724	-.17556	1.40205

En la primera componente confluyen cuatro variables de distinta naturaleza. La correlación más alta se da con una variable demográfica (esperanza de vida) a la que se oponen otras de tipo económico (exportaciones, P.I.B.) y socioeconómico (escolarización 6-11 años), aunque con mucho menos peso en la definición de la componente, a la que hemos llamado, con las mayores reservas extensibles al resto de las denominaciones, "Grado de madurez socioeconómica".

La segunda componente solamente tiene correlaciones significativas con dos variables, las dos en sentido positivo, pero sólo una con peso suficiente para definirla: la escolarización de 6 a 11 años (var. 13), complementada por la esperanza de vida (var. 1), por lo que la denominamos "Escolarización básica".

La componente tercera viene definida principalmente por la alta correlación negativa que tiene con las variabñes 11 y 6 (consumo de energía y participación porcentual de la agricultura en el P.I.B., respectivamente) y con menos peso la variable 15 (escolarización de Tercer Grado), a las que se opone ligeramente la población activa en servicios. La denominamos "Estructura económica".

La cuarta componente el la que mayores problemas plantea puesto que los porcentajes de dependencia son realmente bajos con todas las variables; solamente es relativamente significativa la correlación negativa con la variable 5 (P.I.B.), mientras que en sentido positivo pueden tomarse en consideración, aunque con muy bajo significado, las correlaciones con las variables 15 y 3 (escolarización T.G. y población activa en servicios).

Por lo que se refiere a la quinta componente, está definida por las correlaciones negativas que mantiene con las variables 4 (evolución de la mortalidad infantil) y 16 (médicos por 1000 hab.), a la vez que se da una ligera dependencia positiva con la variable 1 (esperanza de vida). A esta componente la denominamos "Carencias sanitarias".

Por último, la sexta componente presenta los porcentajes de dependencia más altos de todas las retenidas, positivamente con las variables 9 y 6 (evolución de la participación porcentual de los servicios en el P.I.B. y participación porcentual de la agricultura en el P.I.B.), por lo que la hemos denominado "Terciarización".

3.2 Análisis cluster

Por último, y teniendo en cuenta que tanto la comparación como la diferenciación están en la base del conocimiento geográfico (RACINE, J.-B.; REYMOND, H., 1973), hemos intentado establecer una clasificación de los 18 países de América Latina objeto de estudio con un análisis cluster a partir de las 17 componentes principales obtenidas anteriormente.

La elección de las componentes como variables estriba en que aquéllas sintetizan la información de las variables originales, pudiendo sustituirlas (JOHNSTON; R.J., 1980).

La medida de similaridad aplicada ha sido la distancia euclídea, ya que "está ligada a conceptos de variación y de varianza a menudo evocados en materia de clasificación" (BEGUIN, H., 1979, p. 208), conceptos éstos intrínsecos a las componentes principales. Además esta medida permite trabajar con variables no correlacionadas como lo son, por propia definición, las componentes principales, llevando ventaja sobre otra de las medidas utilizadas frecuentemente, la distancia de Mahalanobis, ya que ésta tiene en cuenta las correlaciones entre las variables (BEGUIN, H., 1979; MALLO, F., 1984; MARTINEZ, E., 1984).

Los resultados concretos aparecen en la Tabla VII y en el Gráfico I y nos merecen, asimismo, un breve comentario.

Podemos apreciar la formación de tres grandes grupos cuya diferenciación entre sí es pequeña puesto que se unen a distancias muy próximas a las de la formación del propio grupo. Hay que señalar, además, que en general todos los individuos se unen a distancias altas: 12 lo hacen entre 5,90 y 6.05, mientras que sólo cuatro se unen por debajo de 5,80.

Tabla VII

Agrupación de los individuos en el cluster

Nivel	Grupo	Distancia
1	(8,14) = 8	5.7728
2	(10,16) = 10	5.7734
3	(5,15) = 5	5.84862
4	(3, 8) = 3	5.91538
5	(10,17) = 10	5.92115
6	(3, 7) = 3	5.93866
7	(5,10) = 5	5.954
8	(5, 6) = 5	5.96119
9	(1,11) = 1	5.96696
10	(3,18) = 3	5.97061
11	(4,13) = 4	5.97544
12	(3,12) = 3	5.98431
13	(5, 9) = 5	5.98959
14	(1, 4) = 1	5.99209
15	(1, 2) = 1	5.99855
16	(1, 3) = 1	6.00643
17	(1, 5) = 1	6.05727

Gráfico I

Dendograma

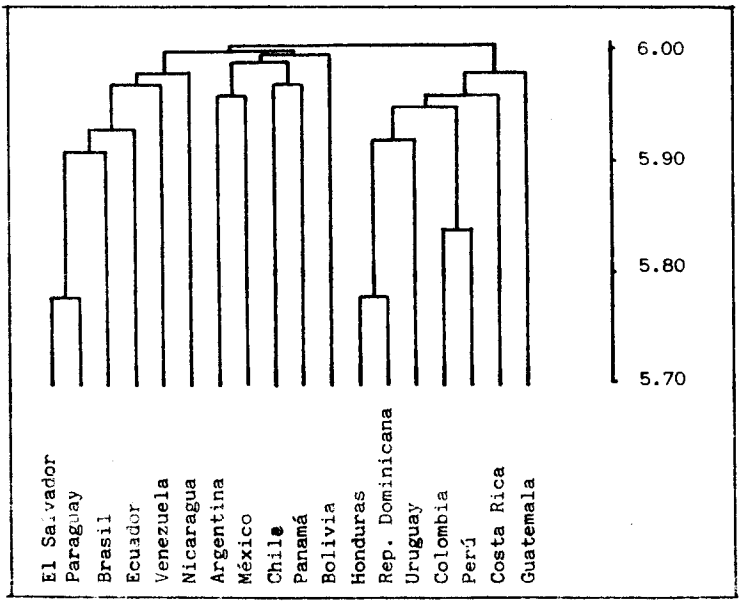


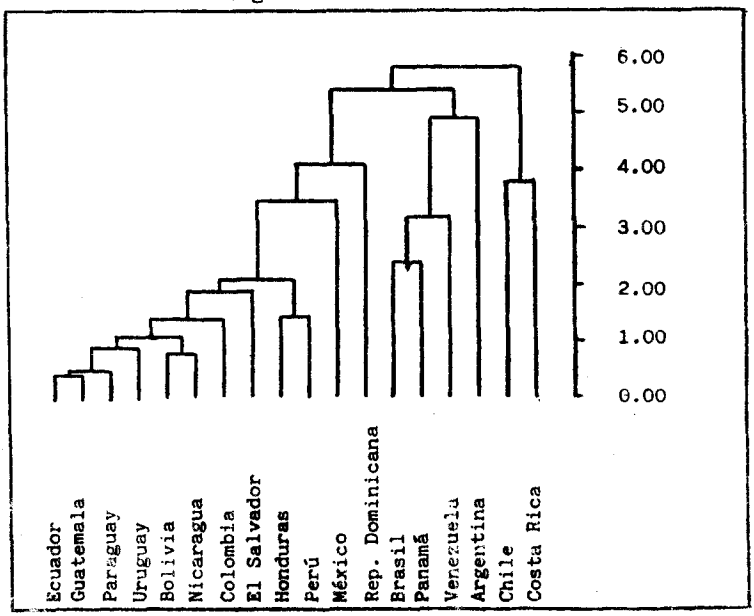
Tabla VIII

Agrupación de los individuos en el cluster

Nivel	Grupo			Distancia
1	(7, 9)	=	7	0.408566
2	(7,14)	=	7	0.506258
3	(2,12)	=	2	0.820273
4	(7,17)	=	7	0.896679
5	(2, 7)	=	2	1.11642
6	(2, 5)	=	2	1.37349
7	(10,15)	=	10	1.41725
8	(2, 8)	=	2	1.94426
9	(2,10)	=	2	2.11304
10	(3,13)	=	3	2.39969
11	(3,18)	=	3	3.19163
12	(2,11)	=	2	3.45948
13	(4, 6)	=	4	3.81424
14	(2,16)	=	2	4.12701
15	(1, 3)	=	1	4.91156
16	(1, 2)	=	1	5.36244
17	(1, 4)	=	1	5.7879

Gráfico II

Dendograma



El grupo más homogéneo es el formado por los individuos 1,11,4, 13 y 2, - que se unen todos ellos a distancias entre 5.966 y 5.998. De todas formas, a pesar de formar un grupo homogéneo, los individuos que lo componen (Argentina, México, Chile, Panamá y Bolivia) mantienen entre sí distancias relativamente considerables, ya que se unen muy cerca del límite superior del dendograma.

Los individuos 8,14,3,7,18 y 12 forman un grupo más heterogénero, en el - que destaca la "temprana" unión de El Salvador y Paraguay a 5.77, a los que - se van uniendo los demás (por orden, Brasil, Ecuador, Venezuela y Nicaragua, ya por encima de 5.90), como si de individuos espúreos se tratase.

El último grupo lo forman los restantes individuos (10,16,17,5,6 y 9) y - en él se pueden distinguir dos subgrupos formados, a su vez, por Honduras y - la República Dominicana, a los que se incorpora Uruguay, y por Colombia y Perú. Ambos subgrupos se unen más arriba y a ellos se les incorporan, también casi como espúreos, Costa Rica y Guatemala.

Así pues, como vemos, la validez de estas 17 componentes principales para establecer grupos entre los individuos es relativamente baja, por lo que debemos tomar solamente las 6 componentes con autovalor superior a 1,0, es decir, las más representativas.

Desgraciadamente para nosotros, por problemas derivados de los trabajos de ampliación del ordenador utilizado, no ha sido posible efectuar este último - paso hasta el momento de redactar este texto; ello nos hubiera permitido contrastar los resultados y establecer una clasificación definitiva; definitiva en la medida en que lo permiten la técnia de análisis y la información utilizada.

Notas

(1) Superada esta dificultad se presenta otra que ha tenido que ser soslayada en muchos caso; es la falta de coincidencia de las fechas a que se refieren algunos datos, sobre todo los demográficos.

(2) Hemos trabajado con un ordenador IBM, sistema 34, instalado en la E. U. de Estudios Empresariales de la Univ. de León. Su limitada capacidad para trabajos de investigación viene dada por su uso también en la administración del propio centro y en las tareas docentes, lo que lleva a su frecuente saturación.

(3) Los programas de ordenador del análisis de componentes principales y cluster forman parte del paquete AMBE (Análisis Multivariante en Basic Extended), desarrollado por Fernando MALLO FERNANDEZ, profesor del centro citado, a quien agradecemos su desvelo y preocupación porque los investigadores de - esta Universidad conozcamos estas técnicas. En la actualidad dicho paquete de programas está pendiente de publicación.

(4) Véanse, por ejemplo, los trabajos hechos con análisis factorial y de componentes principales de AZNAR; BATISTA; CABRER; FONSECA; JIMENEZ; MUÑOZ; - SANZ; SOLA (Vid. Bibliografía). Hay que hacer constar, no obstante, el recurso en estos trabajos a la rotación Varimax de los factores o componentes, que persigue, precisamente, obtener altas correlaciones con sólo unas pocas variables.

Fuentes

- Anuario El País. Ed. PRISA, Madrid. Años 1983, 1984 y 1985.
- Demographic Yearbook, 1982. United Nation, New York 1984.
- L'etat du monde. Edition 1983. Annuaire économique et géopolitique mondial. GEZE,F.; LACOSTE,Y.; VALLADAO,A.G.A. (Dirts.). Editions La Découverte, París 1983.
- El estado del mundo. 1984. Anuario económico y geopolítico mundial. GEZE,F.; LACOSTE, Y.; VALLADAO,A.G.A.; PAQUOT,Th. (Drts.). Eds. Akal. Madrid, 1984.
- EXTEBANK. Ed. Servicios de Estudios Económicos, Banco Exterior de España, - Madrid. Varios años; diversos números monográficos sobre dichos países.

Bibliografía

- AZNAR, A. (1974): "Infraestructura y regionalización de las provincias españolas: una aplicación del análisis factorial". Revista Española de Economía,nº 2, p. 137-166.
- BAILLY, A.S. (1978): La organización urbana. Teoría y modelos. Inst. de Estudios de Administración Local, Madrid, 278 p.
- BATISTA, J.Mª (1984): "Componentes principales y análisis factorial (Exploratorio y confirmatorio)", en SANCHEZ CARRION,J.J. (Editor): Introducción a las técnicas de análisis multivariable aplicado a las ciencias sociales, - Centro de Investigaciones Sociológicas, Madrid, p. 23-74.
- BEGUIN, H. (1979): Méthodes d'analyse géographique quantitative, Librairies Techniques, París, 284 p.
- CABRER,B.; PIQUERAS,J. (1980): "Tipificación de la población activa de España: 1955-1975. Un ensayo de aplicación del análisis de componentes principales". Estudios Geográficos, XLI, 159, p. 171-192.
- CICERI, M.F. et al. (1977): Introduction à l'analyse de l'espace. Masson, París, 173 p.
- FONSECA, Mª.L.; ABREU,D. (1984):"Permanencia e mudança das diferenciações - territoriais em Portugal no periodo 1950-1980". en III Coloquio Ibérico de Geografía. Barcelona, p. 563-575.
- FONSECA, Mª.L.; REIS, D. (1980): Crescimento e diferenciação das áreas suburbanas de Lisboa e do Porto. Estudos para o planeamiento REgional e Urbano nº 13, Centro de Estudos Geograficos, Univ. de Lisboa, 92 p.
- JIMENEZ, B.C. (1985): "La diferenciación sociodemográfica en los distritos municipales de Madrid". Aportación española. XXV Congreso Geográfico Internacional. París 1984. Real Sociedad Geográfica, Madrid, p. 173-187.
- JOHNSTON, R.J. (1980): Multivariate statistical analysis in Geography. Longman Group Limited, London, 280 p.

- MALLO, F. (1984): Análisis estadístico de datos multivariantes. Tomo I - Generalidades. I.C.E. de la Univ. de León. Documentación del Curso "Análisis estadístico de datos multivariantes", 232 p. (policopiado).

- MARTINEZ, E. (1984): "Aspectos teóricos del Análisis de Cluster y Aplicación a la caracterización del electorado potencial de un partido", SANCHEZ CARRION J.J. (Ed.), op. cit. en BATISTA,J.Mª (1984); p. 165-205.

- MATHER, P.M. (1981): "Factor analysis", en Quantitative Geography, ed. por N. WRIGLEY y R.J. BENNETT, Routledge and Kegal Paul LTD, London, Boston & Henley, p. 144-163.

- MUÑOZ, J. (1980): "Ensayo de clasificación sintética de los climas de la España peninsular y Baleares". Estudios Geográficos, XLI, 160, Madrid, p. - - 267-302.

- RACINE,J.B.; REYMOND,H. (1973): L'analyse quantitative en géographie. Presses Universitaires de France, Vendôme (France), 316 p.

- SANZ, E. (1981): "La ordenación del territorio y el sistema de ciudades. Un caso de aplicación de técnicas multivariantes a la definición del sistema - urbano", Ciudad y Territorio, 1, p. 63-89.

- SOLA-MORALES, M. (1970): "FActorialización de características de un área - suburbana". Revista de Geografía, nº 2, Univ. Barcelona, p. 159-186.

Anexo I

Individuos:
1.- Argentina
2.- Bolivia
3.- Brasil
4.- Chile
5.- Colombia
6.- Costa Rica
7.- Ecuador
8.- Es Salvador
9.- Guatemala
10.- Honduras
11.- México
12.- Nicaragua
13.- Panamá
14.- Paraguay
15.- Perú
16.- República Dominicana
17.- Uruguay
18.- Venezuela

Variables:
1.- Densidad en 1982 (hab./km^2).
2.- Natalidad (0/00) (fechas variables de 1975 a 1981)
3.- Mortalidad (0/00) "
4.- Mortalidad infantil (0/00) "
*(1) 5.- Esperanza de vida, en años "
6.- Población de 0-14 años (1984) (%)
7.- Población de 15-64 años (") (%)
8.- Población activa agraria en 1980 (%)
*(2) 9.- " " industria " "
*(3) 10.- " " servicios " "
11.- Evolución de la natalidad (1960= 100)
12.- " mortalidad "
*(4) 13.- " mortalidad infantil "
14.- " población potencialmente activa (15-64) "
15.- " población activa agraria "
16.- " " " industria "
17.- " " " servicios "
18.- P.N.B. "per cápita", 1982 ($)
*(5) 19.- P.I.B. de 1982 (millones de $)
20.- Crecimiento anual del P.I.B., 1982

21.- Tasa de inflación, 1983
22.- Tasa media de inflación 1970-1981
*(6) 23.- Participación agricultura en el P.I.B., 1982 (%)
*(7) 24.- " industria " " "
25.- " servicios " " "
26.- Consumo privado (% del P.I.B.)
27.- " público "
*(8) 28.- Inversión bruta "
29.- Ahorro bruto
30.- Balanza de re cursos
31.- Evolución participación agricultura en P.I.B. (1960=100)
32.- " " industria " "
*(9) 33.- " " servicios " "
*(10) 34.- " consumo privado (1960= 100)
35.- " " público "
36.- " inversión bruta "
37.- " ahorro bruto
38.- Producción de energía 1982 (millones TEC)
*(11) 39.- Consumo de energía " "
*(12) 40.- Exportaciones 1983 (millones $)
41.- Importaciones " "
42.- Deuda exterior " (mil millones $)
*(13) 43.- Escolarización de 6-11 años (%)
*(14) 44.- " 12-17 "
*(15) 45.- " Tercer Grado
46.- Tasa de analfabetismo (%)
47.- Gastos públicos en educación (% del P.I.B.)
48.- Aparatos de TV por 1.000 habitantes
*(16) 49.- Médicos por 1.000 habitantes
*(17) 50.- Calorías "per cápita" (% de las necesidades diarias)

Nota: la numeración de las variables con asterisco corresponde a la utilizada en el segundo análisis de componentes principales.