LA APLICACION Y ALGUNOS PROBLEMAS METODOLOGICOS DE LAS TECNICAS MULTIVARIANTES EN ECOLOGIA URBANA

Mª Jesús González González

(Universidad de León)

El objetivo que nos proponemos en este trabajo es dar a conocer algunas técnicas de análisis multivariante en ecología urbana (análisis de componentes principales, coordenadas principales y cluster). Este tipo de técnicas ha sido muy difundido en los países de habla anglosajona, a partir de la teoría de áreas sociales expuesta por SHEVKY y BELL (1955). El aporte esencial con respecto a otras teorias es que utiliza medidas cuantitativas para hacer posible una demostración empírica.

1. EL ANALISIS DE COMPONENTES PRINCIPALES

El análisis factorial junto con la familia asociada de las técnicas estadísticas multivariantes que han sido suficientemente probadas y usadas por varios geógrafos anglosajones (Berry, Openshaw, Herbert, Jhonston, etc.) ha llegado a ser uno de los métodos más usados en investigación social de todas las clases y es el preferido para identificar la diferenciación social (características socio-económicas y morfológicas) en las ciudades y para describir su expresión espacial. Estos estudios conocidos como ecología factorial, han tenido una gran aplicación en los estudios geográficos del espacio urbano y formaron las bases para una generalización acerca de la estructura socio-espacial urbana (KNOX, P., 1982, pp.78-80).

Este procedimiento ha sido muy utilizado por los planificadores y profesionales anglosajones en orden a ser capaces de proponer una política capaz de mejorar las áreas con mayores privaciones. El estudio de los indicadores sociales y económicos del espacio urbano son importantes de cara a formular una política espacial con la cual mejorar o rectificar los problemas, pero de hecho muy pocas veces tenidos en cuenta en la planificación de la ciudad.

En los últimos años ha aumentado la complejidad intrínseca de los

fenómenos urbanos y con ello los esfuerzos científicos y la aparición de nuevas técnicas para su mejor comprensión y tratamiento de las múltiples variables que pueden influir en los procesos urbanos. El estudio simultáneo de varios factores inter-relacionados ha sido posible mediante la utilización de ordenadores que permitan procesar y elaborar gran cantidad de datos. Estas técnicas nos permiten resolver cuestiones como la relación que hay entre los fenómenos en varias localizaciones o situaciones y si hay espacios diferentes en las condiciones del fenómeno presentado allí, permitiéndonos describir los procesos de cambio, interrelaciones e interdependencias que existen en la ciudad.

El objetivo del análisis factorial es reducir una matriz de datos a una matriz de factores mucho menor, simplificando o resumiendo la información urbanística original, con lo que se pone de manifiesto una estructura más simple subyacente. Los factores que se obtienen explican la mayor cantidad posible de varianza existente en la matriz de datos, nos quedamos con todos los factores que explican más varianza que cualquier otro indicador simple. El análisis factorial y de componentes principales han sido muy utilizados en los estudios de clasificación de la ciudad, investigaciones sobre la vivienda, análisis del área social y en una variedad de estudios de desarrollo económico (HARSTSHORN, T. A, 1980, pp. 469-470).

Vamos a considerar a continuación las características esenciales del análisis de componentes principales como la técnica estadística multivariante que presenta las mayores ventajas en cuanto a interpretación de los resultados en el estudio de las áreas urbanas y su diferenciación en términos de características sociales y económicas.

El análisis de componentes es una de las técnicas de la familia del análisis factorial frecuentemente empleado por los geógrafos. Mientras la diferencia con el análisis factorial más general es filosóficamente y metodológicamente, la interpretación es la misma. La diferencia que existe es que en el análisis de componentes se desprecia la unicidad (que indica la parte de la varianza que es "exclusiva" de la variable) y en este caso existen p componentes para explicar el total de la varianza de las variables (1). La utilización de este análisis viene justificada por la eliminación de la información resultante contenida en las variables,

transformación del conjunto inicial en uno de componentes y simplificación del análisis.

El fin principal es analizar la interdependencia estructural de un conjunto de variables y condensar lo esencial de la información por una serie de variables interdependientes, observadas directamente sobre un conjunto de individuos, en un número más restringido de variables fundamentales independientes. Estas nuevas variables (componentes) que son combinaciones lineales de los originales poseen las siguientes características: 1º Reducción dimensional (mediante la eliminación de variables redundantes y las que aportan información nula). 2º Ortogonalidad, es decir que sean estadísticamente independientes, (Cov. f_1 f_2 = 0) lo que permite la actividad de las influencias. 3º Deben explicar la mayor proporción posible de la variabilidad total, de forma que las variables compuestas resultantes tengan en conjunto varianza máxima. Este problema se resuelve imponiendo la restricción de normalización de los coeficientes de las variables en la transformación lineal ortogonal.

El hecho de que se exija varianza máxima se debe a que en cualquier estudio exploratorio el número de variables bajo consideración es demasiado grande para manejar y dado que el interés reside en las desviaciones de las variables, una posible forma de reducir el número de variables a tratar, es desechar las combinaciones lineales con varianza pequeña y estudiar las que la tienen grande.

Establecida la matriz de datos, donde cada elemento representa las contribuciones de cada variable a las observaciones, el análisis se puede esquematizar en los siguientes pasos:

1º. Transformación de los datos de la matriz original, con el fin de normalizar los valores de la distribución. Esto significa que se sustituye cada valor original por una nueva medida que se obtiene:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

x_{ij} = medida de la abundancia de la variable, para todas las n observaciones.

s_j = desviación típica correspondiente a la variable j para su media $\bar{x}_j$.

Se consideran los valores originales estandarizados. Esta transformación tiene por objeto colocar el origen de medidas en el valor medio de cada atributo y la unidad o unidades de medida originales sustituirlos por una nueva medida, la desviación típica, no afectando en absoluto el resultado final.

2ª. Obtención de la matriz de correlación entre variables, que contiene los coeficientes de correlación de todos los posibles pares de variables presentes en la unidad de observación, y consta por lo tanto de m filas y m columnas. Como medida de la ligazón entre las variables se utiliza el coeficiente de correlación de PEARSON, que está definido a partir de la varianza del par de variables y de su covarianza.

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

s_{jk} = covarianza de la variable j y k.
s_j = desviación tipo de la variable j.
s_k = desviación tipo de la variable k.

La dependencia total conduce a una correlación 1. Cuando j y k son independientes r es igual a 0. Una correlación de - 1 indica la dependencia perfecta pero inversa de las dos variables.

Si las variables se expresan estandarizadas el coeficiente de correlación entre los valores observados de las variables sobre un número determinado de muestras está relacionado con la representación vectorial de aquellas en el hiperespacio de tantas dimensiones como observaciones.

$$\cos \phi_{jk} = r_{jk} \quad (j, k = 1, 2, 3.... m)$$

Donde ϕ_{jk} es el ángulo de separación entre los vectores que representan a las variables j y k y r_{jk} es el coeficiente de correlación entre ellas. La relación entre los ángulos formados por los vectores y el coeficiente de correlación es inversa. Cuanto más pequeño es el ángulo más grande es la correlación y viceversa. Las nubes de puntos así formadas definirán las nuevas componentes dimensionales.

3ª. Cálculo de los valores propios o autovalores de la matriz de correlación, que expresan los valores de las direcciones de máxima variabi-

lidad de la nube de dispersión de los puntos que representan las variables, de forma que el primer componente deberá estar situado en la dirección que absorba la máxima proporción de la varianza total, el segundo componente se colocará de tal forma que absorba la máxima proporción de varianza remanente, lo cual significa que debe ser perpendicular al primer eje, e iguales condiciones deben reunir los restantes ejes hasta que el total de la varianza sea absorbida.

Los autovalores son las soluciones de la ecuación que cumple la condición de igualar a 0 la matriz que se obtiene al restar un valor determinado de los elementos de la diagonal principal de la matriz de correlación.

$$|R - \lambda I| = 0$$

Donde R es la matriz de correlación e I la matriz unidad.

4º. Cálculo de los vectores propios o autovectores. Para cada autovalor habrá un vector asociado a él, conocido como autovector y que satisface la ecuación siguiente:

$$R \cdot V = \lambda \cdot V$$

5º. Cálculo de los coeficientes o porcentajes de dependencia. Expresan la contribución de cada una de las variables en la formación de los nuevos componentes. Se obtienen según la expresión:

$$a_{ki} = x_{ki} \sqrt{\frac{\lambda_k}{x_{ki}^2}}$$

a_{ki} = Porcentaje de dependencia de la variable i en la componente k.
λ_k = Autovalor correspondiente al componente k.
x_{ki} = Autovector ligado al autovalor λ_k.

6º. Cálculo de las coordenadas de las observaciones. Las coordenadas se obtienen a partir de los valores estandarizados de las variables y de sus porcentajes de dependencia, según la función algebráica siguiente:

$$F_k = a\,k_1\,z_1 + a\,k_2\,z_2 + \ldots\ldots + a_{km}\,z_m$$

$$k = 1, 2 \ldots\ldots m$$

F_k = Representa las coordenadas de las observaciones con respecto al componente k.

a_{km} = Son los porcentajes de dependencia que indican la importancia relativa de cada variable en la componente considerada.

z_m = Son los valores estandarizados de las variables en la observación.

El método de componentes es una de las técnicas estadísticas que sirve de ayuda para el tratamiento de datos en ecología urbana de manera que se pueden poner de manifiesto claramente sus interrelaciones y exhibir su estructura.

2.ANALISIS DE COORDENADAS PRINCIPALES

Se encuadra dentro del análisis de escalogramas multidimensionales (métrico), cuyo objetivo consiste en, a partir de una matriz de diferencias o similaridades obtener una representación sintética e interpretable de las relaciones entre las variables representándolas en un espacio de pocas dimensiones de forma que las distancias en este subespacio conserven lo mejor posible, en algún sentido, las distancias o disimilaridades entre los datos originales. La calidad de representación es elevada si el porcentaje de dispersión total recogido por los componentes es alto (MALLO FERNANDEZ, F., 1985).

El proceso esquemático de la obtención de coordenadas principales es:

1º. Cálculo de la matriz A.

a) Para distancias: $a_{ij} = -1/2\ \delta_{ij}^2$

$a_{ii} = 0$

b) Para similaridades: $A = S$

2º. Transformación de la matriz A en una matriz.

B = H A H (producto interno centrado)

3º. Diagonalización de la matriz B (El principio de dualidad nos permite obtener las coordenadas diagonalizando tan sólo la matriz B). Se halla así los k autovalores $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ positivos de B con autovectores correpondientes $w = (w_1 \ldots w_k)$.

4º. Las coordenadas requeridas de los puntos x^i son $(w_{i1} \ldots w_{ik})$ $i = 1 \ldots n$, filas de w. Las coordenadas principales para una representación euclídea en dimensión h son las h primeras columnas de w.

Esta técnica nos permite clasificar o definir grupos de variables fuertemente correlacionadas entre si, visualizando la clasificación a través de una representación euclídea con la propiedad de que las variables estarán tanto más próximas cuanto más correlacionadas estén. Esto puede aprovecharse para eliminar alguna variable como fase previa a un análisis factorial.

El objetivo de este análisis es muy similar al de componentes principales, pero este destaca como método de representación de datos por el hecho de admitir cualquier tipo de variable incluyendo las dicotómicas (basadas en ausencia (-) o presencia (+) de caracteres cualitativos), basta para ello calcular una matriz de similaridades apropiada (CUADRAS, 1981, pp. 295-299 y 307-309).

La misma conclusión sobre la conglomeración de variables se puede obtener mediante un análisis cluster, usando el método "media aritmética no ponderada" sobre la matriz de distancias (euclídea) obtenida por la transformación estandar de la matriz de correlación.

3. EL ANALISIS CLUSTER

En orden a lograr una clasificación multivariada de las subáreas del censo de población se han utilizado procedimientos de agrupación como el análisis cluster, resultando una tipología con variaciones máximas y mínimas. En líneas generales consiste en formar a partir de un conjunto de observaciones, sucesivas particiones de tal forma que los elementos de cada participación sean lo suficientemente homogéneos entre si y distantes de los demás para justificar su inclusión en el cluster.

A partir de una matriz inicial, previamente estandarizada se calcula una nueva matriz de distancias. Se forma una matriz simétrica con las distancias calculadas entre las observaciones, así se formarán los amalgamientos, en función de las variables tratadas, buscando la distancia menor (las dos áreas cuya función en el espacio multidimensional están más próximos) (2). Las distancias más utilizadas son la euclídea y la de Mahalanobis.

La distancia euclídea sobre datos brutos puede ser muy insatisfactoria puesto que es poco sensible a los cambios de escala de las variables. Es invariante frente a transformaciones ortogonales de las variables. Ni siquiera cuando todas las variables están univocamente determinadas, excepto para cambios de escala, esta distancia puede preservar el ordenamiento de distancias. A causa de esto las variables se estandarizan frecuentemente, sin embargo se debe tener presente que ésta para las variables estandarizadas puede preservar distancias relativas.

Las propiedades de la distancia de Mahalanobis son: 1º. Es invariante por transformaciones lineales no singulares de las variables. En particular es invariante por cambios de escala. 2º. Esta distancia tiene en cuenta la interdependencia de las variables. Por tanto, considera esta medida, que es menor la distancia entre individuos que están en la dirección de la elipse formada por los autovalores y autovectores de la matriz inversa que la distancia de aquellos otros que en principio no lo estén. (MALLO FERNANDEZ, F., 1984, p.124).

Este análisis nos permite la formación de una serie de tipologías de áreas con características socio-económicas y ecológicas semejantes. Estas subáreas ecológicas de la ciudad caracterizadas por su homogeneidad interna y heterogeneidad entre ellas, es lo que nos va a permitir descubrir la segregación espacial de la ciudad. A través de las coordenadas de cada cluster en el espacio multidimensional nos permite detectar los factores que influyen más decisivamente en la formación de un espacio segregado, observando las características ecológicas de cada subárea formada y por medio de la escala de distancias se detectan aquellas áreas de segregación específica que debido a sus características se unen a las demás en distancias muy elevadas (son áreas no incluídas en el cluster).

4. LAS UNIDADES DE OBSERVACION Y CONDICIONES DE LOS DATOS

En este tipo de análisis hay que tener en cuenta para su interpretación las unidades de observación, pues el criterio de delimitación de estas, así como la utilización de un número más reducido por agrupación de otras más pequeñas, pueden producir diferencias en los resultados obtenidos, por lo que debe especificarse el nivel de escala en el que se realiza el trabajo (OPENSHAW, J. and TAYLOR, P. J., 1981).

Hay que señalar que los datos de los que podemos disponer para un estudio de áreas sociales, se recogen de las unidades censales que responden a necesidades administrativas y no a las de los investigadores, además el material recogido varía a lo largo del tiempo (en cuanto a datos y a veces unidades) con lo que se hace difícil una investigación comparativa (3). Los resultados y la metodología están, por tanto, sujetos a decisiones externas respecto a la naturaleza y calidad de los datos.

Las unidades que más se suelen utilizar son las secciones censales ya que ésta es la más desagregada de la cual se obtienen datos referidos al conjunto del territorio, pues cuanto más pequeñas sean mayor homogeneidad existe debido a la proximidad. Estas unidades ofrecen a nuestro juicio, las características idóneas, por su relativa homogeneidad interna y heterogeneidad entre ellas para iniciar un análisis ecológico de áreas, aunque algunas de ellas sobreexcedan la extensión media superficial, ya que la división se ha hecho en función del número de habitantes, englobando las superficies periféricas sin construir.

Las áreas de estudio deben ser lo más homogéneas posibles no en cuanto a que sean iguales, ya que la característica dominante de alguna de ellas es la heterogeneidad, sino a la probabilidad de que un individuo elegido arbitrariamente tenga una característica determinada es similar en todas las partes del área.

Las variables que se utilizan para establecer una diferenciación espacial urbana son: demográficos, socio-económicos y de vivienda. Con lo que los resultados empíricos dependen de la matriz de datos original.

Los valores originales se deben estandarizar, conveniente para nuestro propósito, ya que no afectan en absoluto al resultado final, corrigiéndose así la presencia de valores muy heterogéneos de las variables. Esto simplifica los mecanismos de análisis, fundamentalmente cuando la contribución de cada variable depende de su escala de medida, así como de la escala de las otras variables.

El problema que presenta la autocorrelación espacial (que existe cuando el valor de una variable no es independiente del valor de observaciones adyacentes) es una situación paradójica en términos del uso de los métodos estadísticos basados sobre el modelo lineal. Algunos datos no se ajustan al requisito del modelo lineal general cuyas observaciones son independientes de todas las otras. La resolución de esta cuestión depende del problema abordado. (JHONSTON, R. J., 1980, p. 259) (CLIFF, A.D. and ORD, J. K., 1981). Conviene, por tanto, hacer un test de linealidad para ver si las variables son linealmente independientes, sobre todo cuando se utilizan índices.

En situaciones donde la intercorrelación entre variables es baja y sugiere que su distribución es más diferente que similar hay que tener presente el peligro de la sobreinterpretación, así como el que varias variables definan un sistema cerrado de un fenómeno particular (4).

El contenido empírico es necesario también que coincida con el contenido teórico ya que son bien conocidos los peligros de la falacia ecológica al atribuir a datos de caracter general conclusiones que corresponden a casos particulares y lo mismo establecer conclusiones globales a partir de datos individuales.

CONCLUSIONES

La aplicación de estas técnicas ha producido resultados relativamente consistentes y parece claro que la mayoría de las variaciones concretas de las características de las subcomunidades urbanas pueden ser interpretadas en términos de 3 ó 4 categorías básicas que se refieren a la diferenciación en el status socio-económico y en la composición familiar. Sin embargo, la demostración de la invarianza de los factores no se

ve completada con la explicitación de las estructuras factoriales. Una vez establecida la validez empírica de los principales ejes de diferenciación, hay que elaborar la explicación de su significado y examinar las relaciones que tienen con otras facetas del comportamiento humano y de la estructura social, ya que la estadística no debe estar desprovista de teorías y conceptos para explicar los problemas de conflicto y procesos decisivos que determinan la organización urbana.

Los resultados de estas técnicas multivariantes no sólo dependen de la naturaleza de los datos utilizados y del método empleado, sino también de las inclinaciones teóricas de los investigadores que explican e interpretan el significado de las interrelaciones entre las diversas variables.

NOTAS

(1) Para un desarrollo de esta técnica y su aplicación véase JOHNSTON R. J. (1980) y para un desarrollo matemático MALLO FERNANDEZ, F. (1984).

(2) Para una aplicación de esta técnica véase CAMPO MARTIN, A. (1983).

(3) Para una discusión de los datos del censo en Gran Bretaña pero que se puede aplicar en general véase EVANS, S. I. (1981) y JOHNSTON R. J. (1976).

(4) Un planteamiento crítico de la ecología factorial puede verse en GIGGS, J. A. and MATHER, P. M. (1975).

BIBLIOGRAFIA

CAMPO MARTIN, A. (1983): "Una aplicación de ecología factorial al estudio de pautas espaciales de segregación social en el municipio de Madrid", Ciudad y Territorio 57-58, pp. 139-148.

CLIFF, A. D. and ORD, J. K. (1981): "Spatial and Temporal Analysis: Autocorrelation in Space and Time" in WRIGLEY, N. and BENNETT, R. J.: Quantitative Geography, London, Routledge and Kegan, pp. 104-122.

CUADRAS AVELLANA, C. M. (1981): Análisis Multivariante, Barcelona, Eurïbar.

EVANS, I. J. (1981): "Census Date Handling" in WRIGLEY, N. and BENNETT, T. J.: Quantitative Geography, London, Routledge and Kegan, pp.46-59.

GIGGS, J. A. and MATHER, P. M. (1975): "Factorial Ecology and Factor Invariance: an investigation", Economic Geography vol. 51, pp. 336-382.

HARTSHORN, T. A. (1980): Interpreting the City, New York, John Wiley and Sons.

JOHNSTON, R. J. (1976): "Residential Area Characteristics: Research Methods for Identifying Urban Sub-areas-Social Area Analysis and Factorial Ecology" in HERBERT, D. T and JOHNSTON, R. J.: Spatial

Processes and Form vol. 1, New York, Whiley and Sons.

JOHNSTON, R. J. (1980): Multivariate Statistical Analysis in Geography, London, Longman.

KNOX, P. (1982): Urban Social Geography, London, Longman.

MALLO FERNANDEZ, F. (1984): Análisis estadístico de datos multivariantes, León, Universidad de León.

OPENSHAW, S. and TAYLOR, P. J. (1981): "The Modificable Areal Unit Problem" in WRIGLEY, N. and BENNETT R. J.: Quantitative Geography, London, Routledge and Kegan, pp. 60-69.

SANCHEZ CARRION, J. J. (1984): Introducción a las técnicas de analisis multivariante aplicadas a las ciencias sociales, Madrid, Centro de Investigaciones Sociológicas.

SANCHO-ROYO, F. y GONZALEZ BERNALDEZ, F. (1972): "Estructura subyacente de datos urbanísticos en Sevilla", Ciudad y Territorio 3, pp. 6-13.

SHEVKY, E. y BELL, W. (1974): "Análisis del área Social" en THEODORSON, G. A.: Estudios de ecología humana vol. 1, Barcelona, Labor, pp. 337-417.

TIMMS, D. (1976): El mosaico urbano, Madrid, IEAL.

TRYON, R. C. (1965): Identification Social Area by Cluster Analysis, Berkeley, University California Press.