Indicador de calidad de los estratos para el Área Metropolitana de Medellín

Introducción. I. La cuantificación de variables cualitativas. II. Análisis de componentes principales. III. Metodología y aplicación. IV. Análisis de Cluster y tipología de los grupos. Conclusiones. Referencias.

Introducción

En la construcción de un indicador de calidad de los estratos, como un resumen de un conjunto de características de la vivienda y de su entorno, deberían emplearse técnicas estadísticas que permitan transmitir en forma óptima la información contenida en el conjunto de variables seleccionadas para tal fin. La optimalidad en este caso consiste en que el indicador debería tener máxima información del conjunto de variables seleccionadas.

Cuando las características seleccionadas son de tipo cuantitativo, el Análisis de Componentes Principales es un procedimiento estadístico adecuado para construir el indicador. Este se genera como la combinación lineal de las características (o transformaciones de ellas) que es capaz de explicar la mayor parte de la variación total de las variables originales, en otras palabras, que es capaz de conservar máxima información de ellas.

Sin embargo, en nuestro caso, las variables observadas para la discriminación de los estratos son variables de tipo cualitativo, es decir variables medidas en escala ordinal o nominal, y esta clase de medición no permite la utilización directa del Análisis de Componentes Principales. Una solución a este problema es la transformación de variables cualitativas a variables cuantitativas, lo que significa valorar de alguna manera las categorías de cada una de ellas. En algunos estudios, esta valoración o cuantificación ha sido realizada por expertos, los cuales han asignado un valor (por ejemplo, un puntaje de 0 a 100) a cada una de las clases o niveles de las variables. Sin embargo, este proceso presenta al menos dos dificultades: en primer lugar, este tipo de valoración procede de un juicio subjetivo y podría cambiar de experto a experto; en segundo lugar, el experto generalmente valora las categorías de una variable sin tener en cuenta su relación con las categorías de las otras variables de interés, es decir, para cada variable cualitativa, el experto realiza una valoración unidimensional, perdiendo información sobre la relación multivariante del conjunto de variables.

Estas dificultades pueden ser resueltas empleando una técnica de análisis de datos denominada «Cuantificación óptima» (Young, 1981) la cual asigna valores numéricos a las categorías de las variables de forma tal que se maximice la relación entre las observaciones y el modelo de Componentes Principales, respetando el carácter de medición de los datos. Un procedimiento denominado PRINQUAL (Análisis de Componentes Principales Cualitativas, Kuhfeld, Sarle y Young, 1983)) implementa dicha metodología en el paquete estadístico SAS (Statistical Analysis System).

La filosofía del procedimiento es simple: cuantificar las categorías de las variables de tal manera que se maximicen las correlaciones entre todas las variables de interés. El resultado de este proceso en el estudio es muy importante y se traduce en que viviendas de baja calidad tienden a tener valores bajos en las características medidas y, por tanto, un valor también bajo en el indicador final de calidad del estrato. Contrariamente, las viviendas de buena calidad tenderán a obtener valores altos.

El plan de este documento es el siguiente: la sección 1 presenta la definición, formulación matemática y criterios de cuantificación y el procedimiento de cuantificación óptima y mínimos cuadrados alternantes; la sección 2 muestra el método de análisis de componentes principales ; hace un breve repaso del modelo de componentes principales tradicionales; la sección 3 presenta la metodología propuesta para la construcción del indicador y su aplicación al formulario del DNP. Finalmente, la sección 4 presenta la determinación de los estratos usando Análisis de Cluster.

I. La cuantificación de variables cualitativas

A. Definición de cuantificación

Por cuantificación entenderemos la transformación de una o varias variables categóricas en variables numéricas. La principal consecuencia de cuantificar variables cualitativas es la de permitir el uso de la técnicas estadísticas usuales tales como, por ejemplo, el Análisis de Componentes Principales, la Regresión Múltiple, el Análisis Discriminante, el Análisis de Factores. Durante mucho tiempo el uso de las técnicas de cuantificación estuvo ligado a esta conveniencia. Sin embargo, hoy en día se considera como un método fundamental de la estadística, pues es una manera de procesar variables de clases diferentes (numéricas y categóricas) colocándolas todas en la misma condición. Por ejemplo, suponga que tenemos un primer conjunto de n variables numéricas X_1, X_2,...,X_n, y un segundo conjunto de variables cualitativas Y_1, Y_2,...,Y_m, y que queremos hacer un análisis descriptivo de datos para todas las n+m variables a través de un método similar al de componentes principales. Existen cuatro posibilidades:

- Hacer un análisis de componentes principales con X_1, X_2,...,X_n y usar Y_1, Y_2,...,Y_m como variables adicionales representando las categorías de cada Y_k por el promedio de los individuos que pertenecen a ella. Entonces tenemos una representación de Y_k en el espacio de los individuos.

• Realizar un análisis de correspondencia múltiple de las Y_k y emplear las X_j como variables adicionales calculando el coeficiente de correlación de las X_j con las componentes principales. La representación de las X_j está en el espacio de las variables.

• Dividir en categorías las variables numéricas y realizar un análisis de correspondencia múltiple con las m+p variables cualitativas.

• Cuantificar las Y_k y hacer un análisis de componentes principales con las m+p variables cuantitativas.

En esta última posibilidad es en la que estamos interesados. La tercera posibilidad, parece diferente pero también es una técnica de cuantificación.

Realmente muchos métodos clásicos que tratan con variables categóricas pueden ser consideradas como métodos de cuantificación. Por ejemplo, el análisis de varianza o covarianza realiza la cuantificación de variables nominales denominadas «factores de variabilidad» cuando se estiman sus efectos sobre la variable dependiente (para el modelo de no-interacción).

B. Cuantificación y tipo de variables

Cuando una variable cualitativa Y es puramente nominal una cuantificación es la transformación de Y en una variable numérica discreta donde asignamos el mismo valor numérico a todos los individuos que pertenezcan a la i-ésima categoría de Y. Si la variable Y es ordinal, se recomienda usar solamente cuantificaciones que respeten el orden de las categorías y los valores asignados a las m categorías ordenadas serían tales que $a_1 <= a_2 <= ... <= a_m$. Nishisato, (1980) considera una situación más general en la que se permite un orden parcial de las categorías.

La cuantificación bajo restricciones de orden conduce a una teoría más sofisticada que la de la cuantificación sin restricciones, la cual usa

conos convexos en lugar de subespacios vectoriales (Barlow et al, 1972, Tenenhaus, 1981) y cálculos más complicados. Dejando a un lado las dificultades introducidas por las restricciones, es necesario considerar cuando se deben respetar. Supongamos, un problema de predicción donde una variable explicativa es ordinal y la variable que se va a predecir es numérica. Entonces la cuantificación con restricciones de orden postula la existencia de una relación monótona. ¿Deberíamos introducir tal restricción a priori sin haber estudiado la relación? Puede ser más interesante realizar el análisis sin imponer las restricciones y ver si la cuantificación obtenida respeta el orden de las categorías. Si no lo hace, será una prueba de que la relación no es monótona, dado que no existen errores en el muestreo. Ahora bien, las restricciones deberían ser usadas si se tienen fuertes razones para creer en su existencia. Por el contrario, si la variable dependiente es ordinal, debemos respetar su naturaleza, como en la situación donde tenemos que describir las relaciones entre varias variables ordinales.

En la mayoría de los casos la cuantificación asigna un sólo número a cada categoría. Sin embargo, la diferencia entre el proceso y su nivel de medida puede dar cabida al uso de más de un valor. Por ejemplo, un fenómeno puede ser continuo (la longitud de onda para la percepción del color) y la medición discreta (el color). Por tanto una cuantificación más general implica que una categoría puede ser representada por intervalo de valores.

Para mediciones ordinales asociadas a procesos continuos existe además una restricción de orden para los intervalos (Young, De Leeuw y Takane, 1979). Es importante observar que en este caso se busca la cuantificación de las observaciones en vez de las categorías.

C. Formulación matemática de la cuantificación

Suponga que Y es una variable cualitativa, con m categorías y E el conjunto de sus categorías. Si Q es el universo usual, Y es una función de Q sobre E.

Una cuantificación de Y está definida como una **a** función de E sobre R. Si introducimos las siguientes m variables indicadoras Pj de las categorías, j=1,2,..,m:

$Pj(w) = 1$ si $y(w) = j$

$= 0$ en otro caso,

donde w es un elemento de Q, obtenemos un resultado elemental pero fundamental: la variable cuantificada **a**oY («o» es el símbolo de composición de funciones) no es más que la combinación lineal de las variables indicadoras definidas por los valores a_j:

$$\mathbf{a}oY = \sum_{1}^{m} ajPj$$

Si no existen restricciones sobre los valores a_j, es decir se trata de variables puramente nominales, el conjunto de variables numéricas que constituyen una cuantificación de Y es un subconjunto cerrado de dimensión m del espacio vectorial generado por las Pj.

Si Y es una variable ordinal con el orden natural sobre sus categorías, una cuantificación de Y debe verificar que $a_1 <= a_2 <= ... <= a_m$. Este conjunto de restricciones puede ser escrito como

$a_1 = b_1 - b_0$

$a_2 = b_1 + b_2 - b_0$

....

$a_m = b_1 + b_2 + ... + b_m - b_0$

donde los b_j son números reales no negativos. Entonces la variable cuantificada **a**oY es igual a:

$$\sum_{1}^{m} ajPj = \sum_{1}^{m} (b1 + b2 + \ldots + bj - bo)Pj$$

$$= \sum_{0}^{m} bjP*j \quad \text{con } b_j>0$$

donde $P^*_j = \sum_{i<=j} pi$ y P^*_0=-1.

Los P^*_j son las «variables indicadoras del orden» en el siguiente sentido:

$$P^*(w) = 0 \text{ si } Y(w)>j$$
$$1 \text{ si } Y(W)<=j$$

El conjunto de todas las posibles cuantificaciones de Y con las restricciones de orden es el cono poliédrico convexo C generado por las variables P^*_j,

$$C= \{y^* \mid y^* = \sum_{0}^{m} bjP*j \quad , bj>=0\}$$

Si la variable Y ha sido observada sobre n individuos y es puramente nominal, entonces Y puede ser representada como una matriz con n filas y m columnas de las variables indicadoras. Una variable numérica Y^{**} obtenida de la cuantificación de Y se puede expresar como $Y^{**} = Xa^*$ donde $a^*=(a_1,\ldots,a_m)'$ es el vector de los valores de las categorías.

El conjunto de todas las variables cuantificadas es W, el subespacio de R^n de dimensión m definido por $W=\{Y^{**} \mid Y^{**} = Xa^*, a^* \text{ en } R^m\}$.

Por ejemplo, para una variable ordinal Y, por ejemplo con 3 categorías y para cinco individuos, tenemos,

$$\begin{bmatrix} a1 \\ a2 \\ a3 \\ a1 \\ a2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a1 \\ a2 \\ a3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} bo \\ b1 \\ b2 \\ b3 \end{bmatrix} = \begin{bmatrix} b1-bo \\ b1+b2-bo \\ b1+b2+b3-bo \\ b1-bo \\ b1+b2-bo \end{bmatrix}$$

$$C = \{Y^{**} \mid Y^{**} = X^{*}b^{*}, b_j >= 0 \}$$

Frecuentemente las variables tienen medias cero: si 1^{*} representa una variable con todos sus elementos iguales a 1, el conjunto de todas las posible Y^{**} se reduce a W interceptado con 1^{*-}, donde 1^{*-} es el subespacio vectorial ortogonal para 1^{*}.

Para variables nominales la equivalencia entre una cuantificación y una combinación lineal de variables indicadoras muestra que el estudio de las relaciones entre un conjunto de variables cuantificadas se reduce al análisis canónico de ellas lo que no es más que el estudio de relaciones lineales entre conjuntos de variables numéricas (que toman solamente valores 0 o 1).

D. Cuantificación óptima

A pesar de que en la cuantificación de variables cualitativas debemos respetar la naturaleza de las variables, el número de posibles cuantificaciones es infinito. La cuantificación tiene sentido solamente si tenemos un objetivo preciso, el cual generalmente consiste en la maximización de algún criterio de ajuste. Por ejemplo, si estamos trabajando solamente con dos variables nominales, parece natural que las variables cuantificadas deberían estar maximalmente correlacionadas de forma que permita la mejor predicción de una por medio de la otra al menos en el sentido de los mínimos cuadrados.

De la misma forma, si tenemos que predecir una variable (cualitativa o no) usando varias variables que también pueden ser cualitativas o no, existe un criterio natural de cuantificación: la maximización del cuadrado del coeficiente de correlación entre la variable dependiente (posiblemente cuantificada) y una combinación lineal de las (posiblemente cuantificadas) variables explicativas. Pero si tenemos que cuantificar simultáneamente más de dos variables nominales sin una variable dependiente externa, no existe un único criterio y habrá muchas cuantificaciones «óptimas», como lo veremos a continuación.

E. Cuantificación simultánea de varias variables cualitativas

Para el caso de dos variables cualitativas la solución formal está dada por el análisis canónico de los dos conjuntos de variables indicadoras X_1 y X_2: Las variables cuantificadas son las variables canónicas y los valores óptimos están dados por los vectores propios de los productos de los dos arreglos de frecuencias condicionales.

Para el caso de p variables nominales, la cuantificación simultánea tiene tantas soluciones como criterios, al contrario del caso p=2 donde se puede mostrar que todas los criterios son equivalentes. Esto se debe al hecho de que no existe una medida simple de correlación entre más de dos variables. Sin embargo, existen diferentes formas de cuantificar p variables, las cuales son relativamente fáciles de calcular (Saporta, 1983, Van de Geer, 1993). Una de las más importantes, busca una cuantificación de cada una de las variables de forma tal que obtengamos una representación óptima del conjunto de individuos sobre un subespacio de dimensión fija. El problema consiste en buscar una cuantificación de las variables de tal manera que la suma de las varianzas de las primeras k componentes principales sea maximizada. Otra forma busca la cuantificación de las variables de forma que se minimice el determinante de la matriz de covarianzas de las variables cuantificadas.

En lo que sigue utilizaremos la teoría de la cuantificación junto con el procedimiento de mínimos cuadrados alternantes y la técnica de la

Componentes Principales para la obtención de las variables cuantificadas.

F. Cuantificación optima y mínimos cuadrados alternantes

Para mejorar la forma de cuantificación, Young (1981) propone una técnica de análisis de datos denominada 'Cuantificación óptima', método, que junto con el procedimiento de 'Mínimos Cuadrados Alternantes' asigna valores cuantitativos a las categorías de las variables de manera que se maximicen las correlaciones entre ellas. En particular, Young, Takane y de Leeuw (1978) desarrollaron un procedimiento denominado PRINCIPALS que realiza el análisis de componentes principales sobre todo tipo de variables, incluyendo mezcla de variables cuantitativas y cualitativas. Más tarde Kuhfeld, Sarle y Young (1983) construyeron el procedimiento PRINQUAL (Componentes principales cualitativas) el cual es una mejora del PRINCIPALS y ha sido empleado en este estudio. PRINQUAL se encuentra implementado en el paquete estadístico SAS.

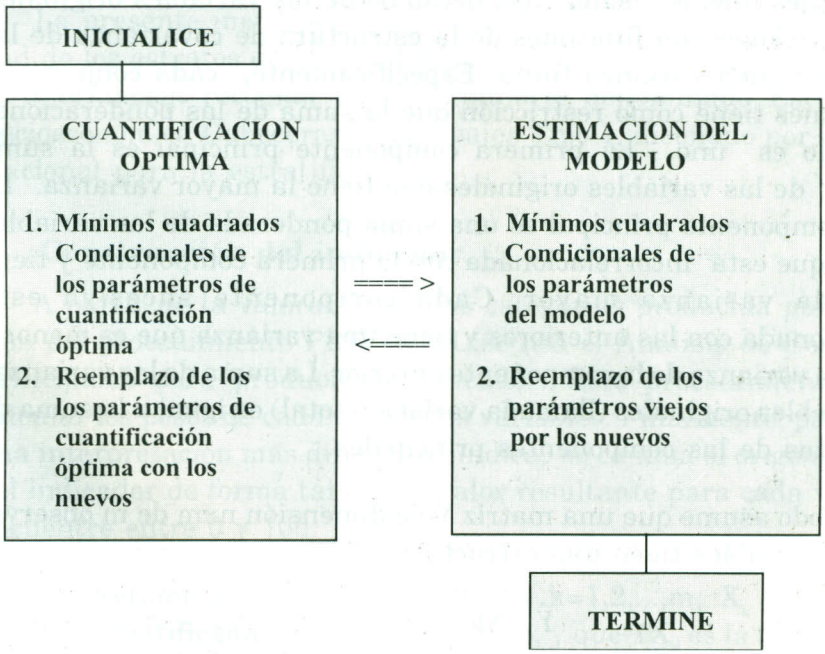
El procedimiento PRINQUAL usa un algoritmo basado en los principios de Mínimos Cuadrados Alternantes (ALS) y Cuantificación Óptima (OS) para obtener transformaciones no lineales de las variables cualitativas de modo que se maximice el ajuste de los datos al modelo de componentes principales lineal. El principio OS considera las observaciones como categóricas y representa cada categoría por medio de un parámetro. Este parámetro esta sujeto a las restricciones implicadas por las características de medición de la variable (por ejemplo, restricciones de orden para variables ordinales).

El principio ALS divide todos los parámetros en dos subconjuntos mutuamente excluyentes y exhaustivos: el primero consta de los parámetros del modelo y el segundo de los parámetros de los datos, denominados parámetros de cuantificación óptima. A su vez, cada subconjunto puede constar de varios subconjuntos que son mutuamente excluyentes y exhaustivos. El proceso de optimización se realiza encontrando las estimaciones mínimo cuadráticas de los parámetros en un subconjunto

suponiendo que los parámetros en todos los otros subconjuntos son constantes. Estas estimaciones son denominadas estimaciones mínimos cuadráticas condicionales, debido a que la naturaleza mínimo cuadrática es condicional sobre los valores de los parámetros en los otros subconjuntos. Una vez se han obtenido las estimaciones mínimo cuadráticas condicionales se reemplazan las estimaciones viejas de estos parámetros por las nuevas. Luego se pasa a otro subconjunto y se obtienen sus estimaciones mínimo cuadráticas condicionales. Alternativamente se obtienen las estimaciones en el subconjunto de parámetros del modelo, y seguidamente en los de los datos, hasta obtener convergencia. El cuadro 1 muestra el proceso ALS-OS.

Cuadro 1

Mínimos cuadrados alternantes y cuantificación óptima



La teoría sobre Mínimos Cuadrados Alternantes se encuentra en Wold & Lyttkens (1969). Young (1981) discute los aspectos sobre Cuantificación Optima y teoría de medición.

II. Análisis de componentes principales

El Análisis de componentes principales es un método multivariado que permite reducir un sistema p-dimensional a un sistema de bajas dimensiones (1 o 2, generalmente) por medio de combinaciones lineales de las variables cuantitativas originales. Una discusión más completa sobre este tema se encuentra en textos de análisis estadístico multivariado tales como Mardia, Kent y Bibby (1979), Johnson y Wichern (1984), Morrison (1976), Levard, Morineau y Warwick, (1984).

Las componentes principales son variables nuevas generadas como combinaciones lineales (sumas ponderadas) de las variables originales. Las ponderaciones son funciones de la estructura de covarianza de las variables y tienen varianza finita. Específicamente, cada conjunto de ponderaciones tiene como restricción que la suma de las ponderaciones al cuadrado es uno. La primera componente principal es la suma ponderada de las variables originales que tiene la mayor varianza. La segunda componente principal es una suma ponderada de las variables originales que está incorrelacionada con la primera componente y tiene la segunda varianza mayor. Cada componente sucesiva está incorrelacionada con las anteriores y tiene una varianza que es menor o igual que la varianza de la componente anterior. La suma de las varianzas de las variables originales (llamada variación total) es igual a la suma de las varianzas de las componentes principales.

El método asume que una matriz Y de dimensión nxm de m observaciones y n variables tiene una estructura

$$\hat{Y} = XF'$$

donde X es una matriz de mxr que contiene los valores de las r primeras componentes principales, y F es una matriz de nxr con las

ponderaciones de las n variables sobre las r componentes. Generalmente X y F son tales que X'X/m=I y F'F=D, donde D es diagonal y Z tiene sus columnas estandarizadas. El procedimiento de Hotellig (1933) encuentra X y F tales que:

$$\theta = \mathrm{tr}(Y - \hat{Y})'(Y - \hat{Y})$$

sea minimizada para un número predeterminado de componentes.

III. Metodología y aplicación

A continuación presentamos el proceso empleado en la construcción de un indicador de la calidad de los estratos.

A. Base de datos

La presente metodología para la elaboración del indicador de calidad de los estratos del Área Metropolitana de Medellín, está basada en la información recogida en el formulario denominado Estratificación Socioeconómica Cabeceras Municipales Tipo 1, diseñado por Planeación Nacional para la estratificación.

B. Construcción del indicador

A partir de la valoración de las categorías producida por la aplicación del procedimiento PRINQUAL se usa el Análisis de Componentes Principales para producir el indicador. Este procedimiento permite calcular los pesos de cada una de las variables. Finalmente, para obtener una interpretación más directa del índice, se cambia el origen y la escala del indicador de forma tal que el valor resultante para cada vivienda se encuentre entre 0 y 100.

Concretamente, suponga que para k=1,2,...,m, X_k es la k-ésima variable identificada como discriminante, y que TX_k es la transformación obtenida usando el procedimiento PRINQUAL. El Cuadro 2 resume el procedimiento.

Cuadro 2
Metodología para la construcción del indicador de calidad de los estratos

Entrada		Proceso	Salida
1.	Xk, k=1,...,m	Use PRINQUAL sobre todas las variables	TXk, k=1,...,m cuantificación
2.	TX1,...TXni	Análisis de Componentes Principales sobre las variables transformadas	Primera Componente Principal, CP1
3.	CP1	Cambio de origen y escala	Indic. de Calidad tipificado

C. Aplicación

Tal como se enuncia en las secciones anteriores, la metodología descrita se aplica a la Encuesta de Estratificación Socioeconómica diseñada por el Departamento Nacional de Planeación. De dicha encuesta se tomó una muestra de 3657 observaciones que representan igual número de lados de manzana para las comunas que están en los seis estratos socioeconómicos, de acuerdo con la estratificación actualmente existente.

Las variables seleccionadas proporcionan información sobre las características físicas de la vivienda y de su entorno; éstas se describen a continuación.

1. Descripción de variables

Vías de acceso: Hace referencia a las características de la calle o vía del lado de la manzana.

Focos de contaminación: Indaga sobre la existencia al lado de la manzana, o al frente, de focos de contaminación como: aguas negras a la vista, botaderos de basura, terminales de buses, talleres, cantinas, etc.

Anden: Se refiere a las características predominantes del anden en el lado de la manzana.

Antejardín: Indaga sobre la existencia de antejardines en el lado de la manzana y sobre sus características.

Garajes: Investiga sobre la existencia y características de los garajes en el lado de la manzana.

Material de las fachadas: Sondea las características predominantes del material de las fachadas en el lado de la manzana.

Material de la puerta principal: Averigua por las características predominantes del material de la puerta principal en el lado de la manzana.

Zona: Hace referencia a la zona a la cual pertenece el lado de la manzana.

En la tabla 1 se presentan las características de cada una de las variables, sus respectivas categorías y la cuantificación realizada a cada una de ellas.

2. Valoración de las categorías

En la tabla 1 se presentan los resultados de la valoración de las categorías de las variables identificadas por el DNP como discriminantes del estrato socioeconómico, llevada a cabo mediante la aplicación del procedimiento PRINQUAL.

En la columna denominada "Cuantificación" se encuentra la valoración de cada una de las categorías de cada variable. Para la primera variable (vías de acceso), por ejemplo, a la categoría 1 se le asigna el valor

Tabla 1

VARIABLES	DESCRIPCION DE LA CATEGORIA	CATEG.	CUANTIFIC	DISTANCIA
VIAS DE ACCESO (VIAS)	Sendero o camino	1	0.261	
	Peatonal	2	2.033	1.772
	Vehicular en recebo (balasto o gravilla) o en tierra	3 ó 4	3.041	1.008
	Vehicular en cemento, asfalto o adoquín	5	4.997	1.956
FOCOS DE CONTAMINACION (CONTAM)	Aguas negras a la vista, botaderos de basura, matadero, plaza de mercado o de ferias, talleres fábricas, terminales de buses, canchas de tejo, cantinas, billares, bares, etc.	0	0	
	Ninguna de las anteriores	1	1	1
ANDEN (ANDEN)	Sin andén	1	0.847	
	Con andén sin zona verde	2	2.16	1.313
	Con andén con zona verde	3	2.918	0.758
ANTEJARDIN (ANTEJAR)	Sin antejardín	1	0.962	
	Con antejardín pequeño	2	2.123	1.161
	Con antejardín mediano	3	2.947	0.824
	Con antejardín grande	4	3.214	0.267
GARAJES (GARAJE)	Sin garaje ni parqueadero	1	0.946	
	Con garaje cubierto usado para otros fines	2	3.101	2.155
	Con parqueadero o zona de parqueo	3	3.291	0.19
	Con garaje adicionado a la vivienda	4	3.647	0.356
	Con garaje sencillo que hace parte del diseño original de la vivienda	5	4.952	1.305
	Con garaje doble o en sótano	6	6.324	1.372

VARIABLES	DESCRIPCION DE LA CATEGORIA	CATEG.	CUANTIFIC	DISTANCIA
MATERIAL DE FACHADAS (FACHAD)	En guadua, caña, esterilla, tabla o desechos	1	-2.494	
	SIN CUBRIR (adobe, bahareque, tapia pisada,	2	2.059	4.553
	plac prefabricada, bloque o ladrillo común)			
	En revoque (pañete o repello) SIN PINTURA	3	2.858	0.799
	En revoque (pañete o repello) CON PINTURA	4	3.973	1.115
	Con enchapes, en ladrillo pulido o			
	en madera fina	5	5.109	1.136
MATERIAL DE LA PUERTA PPAL (PUERTAP)	Tabla, guadua, esterilla, zinc o tela	1	-0.886	
	Madera pulida, lámina metálica, armazón	2	2.003	2.889
	de hierro trabajado o labrado o aluminio			
	Madera fina tallada o completamente en vidrio	3	2.928	0.925
ZONA (NZONA)	Zona=1	1	-7.075	
	Zona=55	2	2.083	9.158
	Zona=44	3	2.083	0
	Zona=46	4	4.26	2.177
	Zona=36	5	4.995	0.735
	Zona=25, 26,30,33,34,49	6	5.85	0.855
	Zona=16,17,19,20,21,	7	7.258	1.408
	Zona=14	8	8.335	1.077
	Zona=9,10,11,12,13	9	8.806	0.471

de 0.261; a la categoría 2 un valor de 2.033; a las categorías 3 y 4 el valor de 3.041, para estas dos categorías el procedimiento arroja la misma cuantificación[1], y a la categoría 5 se le asigna un valor de 4.997. La columna "Distancia" es un indicador de que tan diferente es la calidad

1 Cuando el procedimiento asigna una misma valoración a dos categorías consecutivas significa que no hay separación entre estos dos niveles de la variable.

del estrato para lados de manzana en dos categorías distintas (esta distancia es calculada como la diferencia entre la cuantificación de una categoría y la de la anterior).

La cuantificación de las categorías responde a la asignación que haría un experto pero con las ventajas discutidas en la introducción. Se puede observar que los menores valores de la variable están asociados a una menor calidad del estrato y los mayores valores están asociados a una mayor calidad del estrato

3. Componentes principales y obtención del indicador de calidad del estrato

A continuación se presenta la primera componente principal obtenida de la matriz de correlación de las variables cuantificadas, la cual se interpreta como indicador de la calidad del estrato

CALESTR=0.349*(TVIAS-MD(TVIAS))/SD(TVIAS)+0.068*(TCONTAM-MD(TCONTAM))/SD(TCONTAM)+0.421*(TANDEN-MD(TANDEN))/SD(TANDEN)+0.354*(TANTEJAR-MD(TANTEJAR))/SD(ANTEJAR)+0.405*(TGARAJE-MD(TGARAJE))SD(TGARAGE)+0.422*(TFACHADA-MD(TFACHADA))/SD(TFACHADA)+0.163*(TPUERTA-MD(TPUERTA))/SD(TPUERTA)+0.446(TNZONA-MD(TNZONA))/SD(TNZONA)

Donde: MD(*) es la media de la variable *

SD(*) es la desviación estándar de la variable *

T(*) es la transformación de la variable * y

CALESTR es la variable que indica la calidad del estrato

La proporción de variación total explicada por la primera componente principal es 49.9%.

En la tabla 2 se presenta el análisis de componentes principales, los valores propios de la matriz de correlación y las ponderaciones sobre la primera componente principal.

Tabla 2

Eigenvalues of the Correlation Matrix

	Eigenval	Difference	Proportion	Cumulative
PRIN1	3.99269	2.91496	0.499086	0.49909
PRIN2	1.07773	0.16952	0.134716	0.6338
PRIN3	0.90821	0.17247	0.113526	0.74733
PRIN4	0.73574	0.26932	0.091967	0.83929
PRIN5	0.46642	0.15631	0.058302	0.8976
PRIN6	0.3101	0.02104	0.038763	0.93636
PRIN7	0.28907	0.069	0.036133	0.97249
PRIN8	0.22006		0.027508	1

Ponderación sobre la primera componente Principal

TVIAS	0.3498
TCONTAM	0.0682
TANDEN	0.4219
TANTEJAR	0.3547
TGARAJE	0.4051
TFACHAD	0.4229
TPUERTA	0.1631
TNZONA	0.4465

A través de un cambio de origen y escala se tipificó el indicador de calidad del estrato. de manera que el valor obtenido para cada lado de manzana estuviera en un rango de cero a cien, esto con el fin de hacer más fácil la interpretación del indicador. Se procede de la siguiente manera: se estandarizan los valores inicialmente obtenidos en la cuantificación, y a estos se les sustraen los valores mínimos para cada variable y se dividen por la sumatoria de los valores máximos.

D. Puntajes

En la tabla 3 se presentan los resultados de ésta tipificación y los puntajes finales para las categorías de las variables.

Tabla 3

Puntajes finales para las categorías de las variables DNP

VARIABLES	1	2	3	4	5	6	7	8	9
VIAS DE ACCESO	0	3.39	5.32	5.32	9.06				
FOCOS DE CONT.	0	1.59							
ANDEN	0	4.77	7.52						
ANTEJARDIN	0	3.34	5.72	6.49					
GARAJES	0	3.18	3.46	3.99	5.92	7.95			
MAT. FACHADAS	0	13.3	15.64	18.9	22.22				
MAT. PUERTA	0	17.93	23.65						
ZONA	0	12.38	12.38	15.32	16.32	17.47	19.38	20.83	21.47

IV Análisis de Cluster y tipología de los grupos

A. Análisis de Cluster

Con base en los puntajes obtenidos para cada variable y teniendo en cuenta el indicador de calidad del estrato, se empleó el método de clasificación empleando análisis de cluster para dividir la muestra en seis grupos (representando los 6 estratos). El procedimiento empleado es una variante del método **K-MEANS** (Mac Queen, 1967).[2] Una vez aplicada la técnica se obtuvo una clasificación de las observaciones en seis estratos.[3] En la tabla 4 se presentan las estadísticas descriptivas de cada grupo.

---

2 Se encuentra implementado en el paquete estadístico SAS bajo el nombre de **PROC FASTCLUS.** Este procedimiento usa un método denominado **Ordenamiento del Centroide Más Cercano** (Anderberg, 1973). En esta técnica se selecciona un conjunto de puntos llamados **semillas de grupos.** Estos puntos son una aproximación inicial a las medias de los grupos. A continuación, cada observación se asigna a la media más cercana para formar grupos temporales, luego las medias iniciales son reemplazadas por las medias de los grupos temporales y el proceso se repite hasta que no ocurran cambios en los grupos.

3. La existencia de seis grupos parece ser confirmada por el criterio del Pseudo t^2 usando usando los procedimientos de cluster jerárquico de la **mediana** y de Wad. También se confirma usando el criterio CCC bajo losprocedimientos del **centroide** y de **enlace completo.** Sin embargo, otra posibilidad que se desprende de los análisis anteriores es la existencia de cuatro grupos.

Tabla 4

Método de clasificación por Cluster

N : 368

Variable Label	Mean	Std Dev	Min	Max
NTVIAS	3.232	0.740	0	5.323
NTCONTAM	1.526	0.315	0	1.591
NTANDEN	0	0	0	0
NTANTEJA	0.072	0.488	0	3.348
NTGARAJE	0	0	0	0
NTFACHAD	13.35	0.396	0	18.901
NTPUERTP	17.934	0	0	17.934
NTNZONA	12.565	0.816	0	16.322
CALESTR	48.686	1.234	0	52.549

CALESTR: Variable que indica calidad del estrato.

Observac. en el rango (0,53] pertenece estrato uno.

N : 372

Variable Label	Mean	Std Dev	Min	Max
NTVIAS	6.501	2.780	3.392	9.068
NTCONTAM	1.399	0.519	0	1.591
NTANDEN	1.932	2.594	0	7.527
NTANTEJA	0.234	0.854	0	3.348
NTGARAJE	0	0.165	0	3.189
NTFACHAD	14.20	1.898	13.308	18.908
NTPUERTP	17.93	0	17.934	17.932
NTNZONA	15.29	23.30	12.385	17.479
CALESTR	57.51	2.724	52.888	62.728

CALESTR: Variable que indica calidad del estrato.

Observac en el rango (53,63] pertenece estrato dos.

N : 635

Variable Label	Mean	Std Dev	Min	Max
NTVIAS	7.975	2.203	3.3928	9.0685
NTCONTAM	1.248	0.655	0	1.5917
NTANDEN	4.794	1.430	0	7.527
NTANTEJA	0.537	1.230	0	3.348
NTGARAJE	0.489	1.213	0	3.997
NTFACHAD	17.58	2.201	13.308	18.901
NTPUERTP	17.934	0	17.934	17.934
NTNZONA	16.996	1.145	12.385	20.839
CALESTR	67.561	2.438	61.920	71.346

CALESTR: Variable que indica calidad del estrato.

Observac en el rango (63,71] pertenece estrato tres.

N : 1124

Variable Label	Mean	Std Dev	Min	Max
NTVIAS	9.061	0.157	5.323	9.068
NTCONTAM	1.484	0.399	0	1.591
NTANDEN	5.815	1.419	0	7.527
NTANTEJA	1.568	1.689	0	5.725
NTGARAJE	1.919	1.874	0	5.928
NTFACHAD	18.892	0.321	13.308	22.22
NTPUERTP	17.934	0.000	17.934	17.93
NTNZONA	18.491	1.307	16.322	21.47
CALESTR	75.1684	2.320	71.34	80.336

CALESTR: Variable que indica calidad del estrato.

Observac. en el rango (71,80] pertenece estrato cuatro.

N : 1050

Variable Label	Mean	Std Dev	Mini	Maxm
NTVIAS	9.068	0	9.068	9.068
NTCONTAM	1.561	0.217	0	1.591
NTANDEN	7.309	0.744	4.770	7.5279
NTANTEJA	4.306	1.773	0	6.495
NTGARAJE	5.378	1.032	3.189	7.959
NTFACHAD	19.796	1.473	18.901	22.220
NTPUERTP	17.934	0	17.934	17.934
NTNZONA	20.478	1.306	17.478	21.476
CALESTR	85.832	3.034	79.807	91.473

CALESTR: Variable que indica calidad del estrato.
Observac. en el rango (80,91] pertenece estrato cinco.

N : 106

Variable Label	Mean	Std Dev	Min	Max
NTVIAS	9.068	0	9.0685	9.0685
NTCONTAM	1.591	0	1.5917	1.5917
NTANDEN	7.241	0.8449	4.7706	7.5279
NTANTEJA	4.94	1.2482	3.3483	6.4958
NTGARAJE	6.627	1.367	3.468	7.950
NTFACHAD	21.061	1.589	18.901	22.226
NTPUERTP	23.497	0.954	17.934	23.659
NTNZONA	21.108	0.576	19.383	21.476
CALESTR	95.142	2.680	89.970	100

CALESTR: Variable que indica calidad del estrato.
Observac en el rango (91,100] pertenece estrato seis.

B. Tipología de los grupos según modelo

Por medio de las tablas anteriores se caracterizó cada grupo como se indica en la tabla 5.

Tabla 5

Tipología de los grupos

ESTRATO UNO. VARIABLES	DESCRIPCION DE LA CATEGORIA
VIAS DE ACCESO (VIAS)	Sendero o camino
	Peatonal
	Vehicular en recebo (balasto o gravilla) o en tierra
FOCOS DE CONTAMINACION (CONTAM)	Aguas negras a la vista, botaderos de basura, matadero, plaza de mercado o de ferias, talleres fábricas, terminales de buses, canchas de tejo, cantinas, billares, bares, etc.
	Ninguna de las anteriores
ANDEN (ANDEN)	Sin andén
ANTEJARDIN (ANTEJAR)	Sin antejardín
	Con antejardín pequeño
GARAJES (GARAJE)	Sin garaje ni parqueadero
MATERIAL DE LAS FACHADAS (FACHAD)	En guadua, caña, esterilla, tabla o desechos
	SIN CUBRIR (adobe, bahareque, tapia pisada, placa prefabricada, bloque o ladrillo común)
	En revoque (pañete o repello) SIN PINTURA
	En revoque (pañete o repello) CON PINTURA
MATERIAL DE LA PUERTA PRINCIPAL (PUERTP)	Tabla, guadua, esterilla, zinc o tela
	Madera pulida, lámina metálica, armazón de hierro trabajado o labrado o aluminio
ZONA (NZONA)	De la uno a la cinco

ESTRATO DOS. VARIABLES	DESCRIPCION DE LA CATEGORIA
VIAS DE ACCESO (VIAS)	Peatonal
	Vehicular en recebo (balasto o gravilla) o en tierra
	Vehicular en cemento, asfalto o adoquín
FOCOS DE CONTAMINACION (CONTAM)	Aguas negras a la vista, botaderos de basura, matadero, plaza de mercado o de ferias, talleres fábricas, terminales de buses, canchas de tejo, cantinas, billares, bares, etc.
	Ninguna de las anteriores
ANDEN (ANDEN)	Sin andén
	Con andén sin zona verde
	Con andén con zona verde
ANTEJARDIN (ANTEJAR)	Sin antejardín
	Con antejardín pequeño
GARAJES (GARAJE)	Sin garaje ni parqueadero
	Con garaje cubierto usado para otros fines
MATERIAL DE LAS FACHADAS (FACHAD)	SIN CUBRIR (adobe, bahareque, tapia pisada, placa prefabricada, bloque o ladrillo común)
	En revoque (pañete o repello) SIN PINTURA
	En revoque (pañete o repello) CON PINTURA
MATERIAL DE LA PUERTA PRINCIPAL (PUERTP)	Madera pulida, lámina metálica, armazón de hierro trabajado o labrado o aluminio
ZONA (NZONA)	De la dos a la seis

ESTRATO TRES. VARIABLES	DESCRIPCION DE LA CATEGORIA
VIAS DE ACCESO (VIAS)	Peatonal
	Vehicular en recebo (balasto o gravill) o en tierra
	Vehicular en cemento, asfalto o adoquín
FOCOS DE CONTAMINACION (CONTAM)	Aguas negras a la vista, botaderos de basura, Matadero, plaza de mercado o de ferias, talleres fábricas, terminales de buses, canchas de tejo, cantinas, billares, bares, etc.
	Ninguna de las anteriores
ANDEN (ANDEN)	Sin andén
	Con andén sin zona verde
	Con andén con zona verde
ANTEJARDIN (ANTEJAR)	Sin antejardín
	Con antejardín pequeño
GARAJES (GARAJE)	Sin garaje ni parqueadero
	Con garaje cubierto usado para otros fines
	Con parqueadero o zona de parqueo
	Con garaje adicionado a la vivienda
MATERIAL DE LAS FACHADAS (FACHAD)	SIN CUBRIR (adobe, bahareque, tapia pisada, placa prefabricada, bloque o ladrillo común)
	En revoque (pañete o repello) SIN PINTURA
	En revoque (pañete o repello) CON PINTURA
MATERIAL DE LA PUERTA PRINCIPAL (PUERTP)	Madera pulida, lámina metálica, armazón de hierro trabajado o labrado o aluminio
ZONA (NZONA)	De la tres a la ocho

ESTRATO CUATRO. VARIABLES	DESCRIPCION DE LA CATEGORIA
VIAS DE ACCESO (VIAS)	Vehicular en recebo (balasto o gravilla) o en tierra
	Vehicular en cemento, asfalto o adoquín
FOCOS DE CONTAMINACION (CONTAM)	Aguas negras a la vista, botaderos de basura, Matadero, plaza de mercado o de ferias, talleres Fábricas, t. De buses, canchas de tejo, cantinas, billares, bares
	Ninguna de las anteriores
ANDEN (ANDEN)	Sin andén
	Con andén sin zona verde
	Con andén con zona verde
ANTEJARDIN (ANTEJAR)	Sin antejardín
	Con antejardín pequeño
	Con antejardín mediano
GARAJES (GARAJE)	Sin garaje ni parqueadero
	Con garaje cubierto usado para otros fines
	Con parqueadero o zona de parqueo
	Con garaje adicionado a la vivienda
	Con garaje sencillo original de la vivienda
MATERIAL DE LAS FACHADAS (FACHAD)	SIN CUBRIR (adobe, bahareque, tapia pisada, placa prefabricada, bloque o ladrillo común)
	En revoque (pañete o repello) SIN PINTURA
	En revoque (pañete o repello) CON PINTURA
	Con enchapes, en ladrillo pulido o en madera fina
MATERIAL DE LA PUERTA PRINCIPAL (PUERTP)	Madera pulida, lámina metálica, armazón de hierro trabajado o labrado o aluminio
ZONA (NZONA)	De la cinco a la nueve

ESTRATO CINCO. VARIABLES	DESCRIPCION DE LA CATEGORIA
VIAS DE ACCESO (VIAS)	Vehicular en cemento, asfalto o adoquín
FOCOS DE CONTAMINACION (CONTAM)	Aguas negras a la vista, botaderos de basura, matadero, plaza de mercado o de ferias, talleres fábricas, terminales de buses, canchas de tejo, cantinas, billares, bares, etc.
	Ninguna delas anteriores
ANDEN (ANDEN)	Con andén sin zona verde
	Con andén con zona verde
ANTEJARDIN (ANTEJAR)	Sin antejardín
	Con antejardín pequeño
	Con antejardín mediano
	Con antejardín grande
GARAJES (GARAJE)	Con garaje cubierto usado para otros fines
	Con parqueadero o zona de parqueo
	Con garaje adicionado a la vivienda
	Con garaje sencillo que hace parte del diseño
	Con garaje doble o en sótano
MATERIAL DE LAS FACHADAS (FACHAD)	En revoque (pañete o repello) CON PINTURA
	Con enchapes, en ladrillo pulido o en madera fina
MATERIAL DE LA PUERTA PRINCIPAL (PUERTP)	Madera pulida, lámina metálica, armazón de hierro trabajado o labrado o aluminio
ZONA (NZONA)	De la seis a la nueve

ESTRATO SEIS. VARIABLES	DESCRIPCION DE LA CATEGORIA
VIAS DE ACCESO (VIAS)	Vehicular en cemento, asfalto o adoquín
FOCOS DE CONTAMINACION (CONTAM)	Ninguna de las anteriores
ANDEN (ANDEN)	Con andén sin zona verde
	Con andén con zona verde
ANTEJARDIN (ANTEJAR)	Con antejardín pequeño
	Con antejardín mediano
	Con antejardín grande
GARAJES (GARAJE)	Con parqueadero o zona de parqueo
	Con garaje adicionado a la vivienda
	Con garaje sencillo que hace parte del diseño original de la vivienda
	Con garaje doble o en sótano
MATERIAL DE LAS FACHADAS (FACHAD)	En revoque (pañete o repello) CON PINTURA
	Con enchapes, en ladrillo pulido o en madera fina
MATERIAL DE LA PUERTA PRINCIPAL (PUERTP)	Madera pulida, lámina metálica, armazón de hierro trabajado o labrado o aluminio
	Madera fina tallada o completamente en vidrio
ZONA (NZONA)	De la siete a la nueve

Conclusiones

La metodología presentada permite desarrollar un procedimiento "objetivo" para clasificar viviendas según sus características físicas, de entorno y de consumo.

A diferencia de los procedimientos del Departamento Nacional de Planeación -DNP- y Empresas Públicas Municipales -EPM-, la nueva metodología considera sólo las características individuales de la vivienda para su clasificación, minimizando así los errores de clasificación (hacia arriba o hacia abajo).

Referencias

ANDERBERG, M.R (1973), *Cluster Analysis for Applications*, New York , Academic Press, INC.

DE BOOR, C. (1978), *A Practical guide to Splines*, New York, Springer Verlag.

FISHER, R. (1938), *Statistical Methods for Research Workers*, 10ma ed., Edinburgh, Oliver and Press.

GORSUCH, R.L. (1983), *Factor Analysis*, 2da ed., Hillsdale New Jersey, Lawrence Erlbaum Associates, Inc.

HOTELLING, H. (1933), «Analysis of Complex Statistical Variables into Principal Components", *Journal of Educational Psychology*, 24, 498-520.

JOHNSON, R. y WICHERN (1988), *Applied Multivariate Statistical Methods*, 2da edición, Prentice Hall.

KRUSKAL, J.B. y SHEPARD, R.N. (1974), «A Nonmetric Variety of Linear Factor Analysis,» *Psychometrika*, 38, 123-157.

KUHFELD, W.F., SARLE, W.S. y YOUNG, F.W. (1985). *Methods for Generating Model Estimates in the PRINQUAL Macro*, SAS Users Group International Conference Proceedings: Sugi 10, Cary, NC:SAS Institute, 962-971.

LEVARD L., MORINEAU, A. y WARWICK, K.M. (1984), *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices*, New York, John Wiley & Sons.

MAC QUEEN, J.B (1967). " Some Methods for Clasification and Analysis of Multivariate Observations", *Procedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

MARCOTORCHINO, J.M. Proth Eds., Elsevier Science Publishers B.V., North-Holland.

MARDIA, K.V., KENT J.T. y BIBBY, J.M (1979), *Multivariate Analysis*, London, Academic Press.

MORRISON, D.F. (1976), *Multivariate Statistical Methods*, 2da, Ed. New York, MacGraw-Hill

SAPORTA, G. (1983), «Multidimensional data analysis and quantification of categorical variables,» en *New Trends in Data Analysis and Applications*, J. Janssen, J.F.

SARLE, W.S. (1984), en YOUNG et al (1985).

SAS/STAT User Guide (1990), Volume 2, Versión 6, Cuarta edición.

TENENHAUS, M. y VACHETTE, J.L. (1977), «PRINQUAL:Un Programme d'Analyse en Composantes Principales D'un Ensamble de Variables Nominales ou numeriques,» *Les Cahiers de Recherche* #68, CESA, Jouy-en-Josas, France.

VAN DE GEER, J.P. (1993), *Multivariate Analysis Categorical Data: Theory*, London, Sage Publications Inc.

WINSBERG, S. y RAMSAY, J.O. (1983), «Monotone Spline Transformations for Dimension reduction,» *Psychometrika*, 48, 575-595.

WOLD, H. y LITKENS, E, (1969), «Nonlinear Iterative Partial Least Squares (NIPALS) Estimation Procedures,» *Bulletin ISI*, 43, 29-47.

YOUNG, F.W. (1975), «Methods for Describing Ordinal Data with Cardinal Models,» *Journal of Mathematical Psychology*, 12, 416-436.

---------- (1981), «Quantitative Analysis of Qualitative Data,» *Psychometrika*, 46, 357-388.

YOUNG, F.W., TAKANE, Y. y de LEEUW, J. (1978), «The Principal Components of Mixed Measurement Level Multivariate Data: An Alternanting Least Squares Method with Optimal Scaling Features,» *Psychometrika*, 43, 279-281.

---------- (1985), «PROC PRINQUAL- Preliminary Specifications,» *Manuscrito no publicado, The University of North Carolina Psychometric Laboratory*, Chapel Hill NC.