Análisis multivariado aplicando componentes principales al caso de los desplazados

Ángel León González[1], Humberto Llinás Solano[2], Jorge Tilano[3]

Resumen

La complejidad de los fenómenos de las ciencias en general hace que los investigadores se vean obligados a enfrentarse a problemas donde intervienen múltiples variables y grandes volúmenes de datos que requieren conceptos avanzados y herramientas para su tratamiento e interpretación integral. Por esta causa, desde hace mucho tiempo se han desarrollado las técnicas multivariadas, pero sólo con la evolución de los computadores y diversos paquetes de software *que procesan amplios conjuntos de datos ha llegado a ser notoria la potencia de la estadística multivariada. Se ha tomado el problema del desplazamiento de personas del campo a la ciudad para consolidar el concepto y la aplicación de la técnica de componentes principales (ACP). En este artículo se presenta esta técnica dada su utilidad como paso previo a todos los análisis multivariados que se requiera aplicar.*

Para el estudio del ACP primero se desarrollan los conceptos fundamentales del álgebra matricial, para luego simular una situación problemática escogida como una forma de llevar a la práctica el marco teórico referente a la técnica objeto del artículo. Por otro lado, el desarrollo de la simulación mediante esta técnica conlleva el uso de otros conceptos relativos al ACP, los cuales se explican e interpretan a partir del análisis de los resultados obtenidos en los diferentes procesos.

En el análisis de los resultados se ha concluido que, a pesar de que el gobierno está tomando medidas para mejorar el bienestar de los desplazados, que es la situación escogida, todavía faltan mayores esfuerzos que conlleven a una solución integral de esta problemática.

Palabras claves: Componentes principales, autovalores, autovectores, matriz de componentes, comunalidad

Fecha de recepción: 5 de febrero de 2008
Fecha de aceptación: 25 de abril de 2008

[1] PhD. Gestión Industrial. Grupo de Investigaciones Productividad y Competitividad, Universidad del Norte. Profesor asociado de la Universidad del Norte. *agonzale@uninorte.edu.co*

[2] Dr. rer. nat. Estadística. Grupo de Investigaciones en Estadística e Investigación Operativa (GEIO), Universidad del Norte. Profesor asociado de la Universidad del Norte. *hllinas@uninorte.edu.co*

[3] Esp. Estadística. Grupo de Investigaciones en Estadística e Investigación Operativa (GEIO), Universidad del Norte. Profesor catedrático de la Universidad del Norte. *jtilano@uninorte.edu.co*

Dirección: Angel León González, Departamento de Ingeniería Industrial, Bloque B, segundo piso, Universidad del Norte Km 5, antigua vía a Puerto Colombia.

INGENIERÍA & DESARROLLO

Número 23
Enero-Junio, 2008
ISSN: 0122-3461

Abstract

In general, the complexity of phenomena in sciences makes researchers feel obligated to face problems where multiple variables and big volumes of information are presented. Those problems require advanced in order to decipher the concepts and tools for its treatment. For this reason, multivariate techniques were developed long ago, but only the computer evolution and several software packages have caused the power of the multivariate statistics to become important. The problem of the displacement of people from the countryside to cities is a reason for consolidating the concepts and the principal component application technique (PCA). This article explains PCA technique while prefacing the applyed multivariate analysis.

In order to study ACP, one first needs the fundamentals of matrix algebra concepts. When developed and then applied in a specific simulation, there is a way to carry and practice the theory related to the technique treated in this article. On the other hand, the simulation development using this technique needs to use other concepts associated with PCA which are explained and interpreted from the analysis of results obtained in the different processes.

The analysis of this article points to one conclusion. In spite of the government taking a role for the displaced persons well-being, there is an absence of major efforts that leads to these problematic solutions.

Keywords: Principal components, eigenvalue, eigenvector, matriz components, comunality.

INTRODUCCIÓN

La humanidad en su evolución necesita conocer los fenómenos que están a su alrededor porque éstos afectan su desarrollo dentro de todos los ámbitos (fenómenos de tipo social, económico, tecnológico, físico, etc.). Este conocimiento se logra mediante la construcción de modelos que puedan reproducir y explicar dichos fenómenos. Por tal motivo, es necesario que los profesionales, directivos e investigadores en las distintas áreas del saber estén familiarizados con las herramientas necesarias para la construcción y adecuación de modelos. Una de las herramientas más importantes para llevar a cabo este objetivo es la estadística, y en particular, muy a menudo, la estadística multivariada.

Según Peña [1] y Dallas [2] existen diversas definiciones acerca de las técnicas de análisis de datos multivariados, pero los dos coinciden en a conceptualizarla como "una herramienta que tiene como objetivo principal resumir grandes cantidades de datos por medio de pocos parámetros (simplificación), además busca encontrar relaciones entre:

Variables de respuesta
Unidades experimentales
Variables de respuesta y unidades experimentales" [1]

Según Peña [1], la mayoría de problemas que requieren la aplicación de la estadística exigen el tratamiento de muchos factores o variables y que por esto las técnicas del análisis de datos multivariados constituyen una herramienta poderosa para la toma de decisiones en las diferentes disciplinas, pues dan respuesta a necesidades palpables y plenamente identificables. Según Pérez [3], se puede observar que cuando existen muchas variables es posible que parte importante de la información sea redundante, en cuyo caso es necesario eliminar el exceso y dejar sólo variables que tengan representatividad dentro del conjunto. Esto se consigue con la aplicación de las técnicas multivariantes de reducción de la dimensión: análisis de componentes principales, factorial, correspondencias, escalamiento óptimo, homogeneidades, análisis conjunto.

Las técnicas multivariadas más utilizadas en el análisis de datos son: análisis de componentes principales; análisis factorial; análisis de clasificación entre los que se encuentran: discriminante, regresión logística y *clúster*; análisis multivariado de la varianza, y análisis de variables canónicas.

Con este artículo se desean integrar conocimientos teóricos y prácticos a través de la comprensión de las componentes principales, como una de las técnicas estadísticas que permiten estudiar la información que se dispone antes de entrar en el uso de los otros métodos que abordan el análisis de datos multivariados.

Por ser tan amplio el tema, este artículo sólo trata del análisis de componentes principales debido a su importancia dentro del desarrollo de las diversas técnicas de análisis de datos multivariados.

1. COMPONENTES PRINCIPALES

Siguiendo a autores como Peña [1] y Bramardi[4], el análisis de componentes principales (ACP) es una técnica estadística propuesta a principios del siglo XX por Hotelling (1933) quien se basó en los trabajos de Karl Pearson (1901) y en las investigaciones sobre ajustes ortogonales por míni-

[1] Dallas, E. Johnson. Métodos multivariados aplicados al análisis de datos. Thomson editores S.A. México. 2000.

mos cuadrados. Interpretando la definición de diversos autores, se puede decir que el ACP es una técnica estadística de análisis multivariado que permite seccionar la información contenida en un conjunto de p variables de interés en m nuevas variables independientes. Cada una explica una parte específica de la información y mediante combinación lineal de las variables originales otorgan la posibilidad de resumir la información, total en pocas componentes que reducen la dimensión del problema.

La mayor aplicación del ACP está centrada en la de reducción de la dimensión del espacio de los datos, en hacer descripciones sintéticas y en simplificar el problema que se estudia.

Para Peña [1], el ACP tiene una utilidad doble; por un lado, permite hacer representaciones de los datos originales en un espacio de dimensión pequeña y, por el otro, transformar las variables originales correladas en nuevas variables incorreladas que puedan ser interpretadas.

El ACP también se emplea con frecuencia cuando se desea dividir las unidades experimentales en subgrupos de acuerdo con la similaridad de los mismos. Igualmente, es útil para transformar un conjunto de variables respuesta correlacionadas en un conjunto de componentes no correlacionados, bajo el criterio de máxima variabilidad acumulada y, por tanto, de mínima pérdida de información.

Otra aplicación es el cribado, el cual permite el seguimiento sobre los componentes principales obtenidos para comprobar hipótesis establecidas en un estudio de análisis de datos multivariados y para identificar datos atípicos en el conjunto de datos.

De igual manera, García y Gil [5] afirman que el ACP es un criterio fundamental para hacer conjeturas sobre el numero de factores que se deben determinar en el análisis factorial y para probar si, en realidad, un grupo de variables $p > 2$ cae dentro de un espacio de dos o tres dimensiones que permita ser observado dentro del análisis de *clúster*.

Pérez [2] anota que el análisis de componentes principales es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por las componentes obtenidas. Éste siempre debe hacerse cuando se quiera obtener modelos en los que sea necesario el uso de las variables originales como explicativas para tratar con algunos problemas presentes, como la independencia.

Según Gil [6], en el análisis discriminante cuando se tienen menos observaciones que variables y es difícil encontrar nuevas observaciones, el ACP es útil para determinar un menor número de variables que resuma la máxima variabilidad de las originales y con las cuales se pueda construir la matriz de varianza-covarianza, de tal forma que sea invertible y permita elaborar una regla de discriminación necesaria para clasificar nuevas observaciones.

Finalmente, el ACP se usa como base para determinar si ocurre multicolinealidad entre variables predictoras en el análisis de regresión múltiple. Entendiéndose como multicolinealidad cuando en dos o más variables existe redundancia; esto es, la información de una o más variables ya está explicada en otra(s) variable(s) (véase por ejemplo, Peña [1], Dallas [2]).

2. NOTACIONES Y SÍMBOLOS

Siguiendo la simbología común de diversos autores, a continuación se presentan conceptos básicos del álgebra de matrices que son necesarios en el ACP.

Matriz de variable respuesta

La base para la utilización del ACP es la estructura de correlación (interdependencia) entre las variables cuantitativas definidas en una población, en donde cada individuo queda definido en términos de las mismas. La matriz de variable respuesta de doble entrada X está compuesta por filas que representan las unidades experimentales I_r, r=1,2,….,n y las columnas, por las variables X_j, j= 1,2,…., p, como se muestra a continuación:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_r \\ \vdots \\ I_n \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{r1} & x_{r2} & \cdots & x_{rp} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times p} \end{array}$$

x_{rj} : Valores de la r-ésima unidad experimental en la j-ésima variable respuesta

p **:** Cantidad de variables

n **:** Individuos o unidades experimentales sobre la cual se están midiendo las variables

Vectores de datos

Con el fin de tener un lenguaje común en los procesos de ACP, en adelante, los vectores siempre serán columnas *a* o *X*, etc., y la transpuesta de un vector cualquiera, por ejemplo *a*, se simboliza por a'.

Vectores de medias y matrices de varianza covarianza

La media de un vector X de variables aleatorias se denota por μ, definido por:

$$\mu = E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

La matriz de covarianza de X se denota por $\sum$, donde:

$$\sum = Cov(X) = E\left[(X-\mu)(X-\mu)' \right] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Con

$\sigma_{jj} = Var(X_j) = E[X_j - \mu_j]^2$, para $j = 1, 2, ..., p$, y

$\sigma_{ij} = Cov(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$, para $i \neq j = 1, 2, ..., p$,

Correlación y matriz de correlación

El coeficiente de correlación entre X_i y X_j se denota por ρ_{ij} : $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$

La matriz de correlación para un vector aleatorio X se denota por □ :

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

Matrices ortogonales y unitarias

Dentro del álgebra de matrices las rotaciones de un espacio vectorial son transformaciones lineales del espacio vectorial sobre sí mismo y están asociadas con matrices cuadradas, unitarias y ortogonales. Una matriz de éstas, Q, tiene tantas filas y columnas como sea la dimensión del espacio. Sus columnas son vectores unitarios (es decir, de longitud igual a la unidad) y tiene la particularidad de que al ser multiplicada por su transpuesta produce la matriz unidad. En otras palabras, $Q^{-1} = Q'$. En cambio, las traslaciones no son transformaciones lineales pero tienen la propiedad de no modificar la variabilidad de la nube de puntos. Es decir, las varianzas y covarianzas en la nube son las mismas antes y después de una traslación. Lo expuesto anteriormente, junto con algunas propiedades de la matriz de varianzas covarianzas Σ, constituye las bases sobre las cuales descansa la técnica de componentes principales.

3. PLANTEAMIENTO Y SOLUCIÓN DEL PROBLEMA DE LOS COMPONENTES PRINCIPALES

El ACP es una técnica descriptiva; sin embargo, no niega la posibilidad de que también pueda ser utilizado con fines de inferencia. Por otra parte, las aplicaciones del ACP son numerosas y entre ellas se pueden citar la clasificación de individuos, la comparación de poblaciones, la estratificación multivariada, entre otras. En el ACP se maneja un número p ($p \geq 2$) de variables numéricas. Si cada variable se representa sobre un eje, se necesitaría un sistema de coordenadas rectangulares con p ejes perpendiculares entre sí para ubicar las coordenadas de los puntos y poderlos dibujar. Cuando $p \geq 4$, para el ser humano es imposible hacer la representación gráfica. En estos casos el ACP permite buscar un nuevo sistema de coordenadas con origen en el centro de gravedad de la nube de puntos, de tal manera que el primer eje del nuevo sistema F_1 recoja la mayor cantidad posible de variación; el segundo eje F_2, la mayor cantidad posible entre la

variación restante; el tercer eje F_3 la mayor cantidad posible entre la variación que queda después de las dos anteriores y así sucesivamente. Las Figuras 1 y 2 permiten ver la representación gráfica de dos componentes.

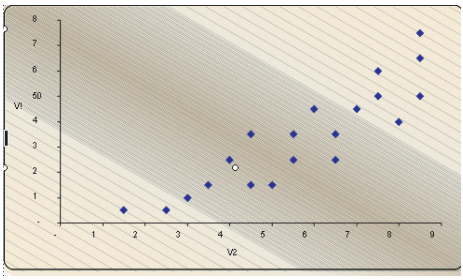

Figura 1.
Diagrama de dispersión divariado
Fuente: Elaboración propia

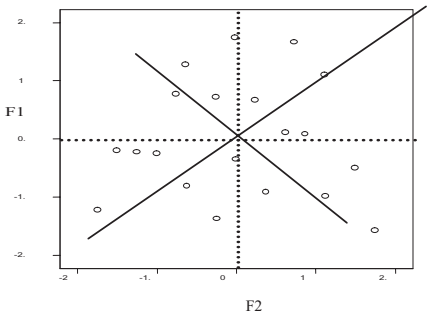

Figura 2.
Diagrama de dispersión rotado

Observando las figuras anteriores se puede concluir que el sistema de coordenadas de la derecha se logra después de dos movimientos de la nube de puntos: el primer movimiento corresponde a una traslación que permite situar el nuevo origen en el centro de gravedad de la nube. El segundo movimiento que se hace sobre la nube centrada es una rotación, usando el centro de gravedad como punto pivote. La rotación permite ubicar los ejes en dirección horizontal y vertical como se observa en la Figura 2. Esto indica que se desea encontrar un nuevo sistema de coordenadas que represente lo mejor posible los datos sin causar distorsiones, cuya forma de problema es equivalente a encontrar las nuevas variables del espacio reducido con una mínima pérdida de la información, y también a buscar un elipsoide de concentración que permita encerrar los datos originales.

Cuando ya se ha definido el problema es factible abordarlo. Según Peña [1], páginas 73-74, la matriz de varianza covarianza Σ es definida positiva, es decir, la forma cuadrática asociada a ella tiene todas sus raíces positivas. Lo anterior hace que esta matriz tenga p valores propios reales y diferentes, lo cual garantiza que sea diagonalizable. En términos matemáticos significa que existe una matriz A ortogonal, tal que $\Sigma = ADA^{-1}$ donde D es la matriz diagonal formada por los valores propios de Σ, denotados por $\lambda_1, \lambda_2,..., \lambda_p$. Es posible reordenar de acuerdo con su magnitud los valores propios de Σ de tal manera que $\lambda_1 > \lambda_2 >...> \lambda_p$. Esto simplemente se traduce en un reordenamiento de las columnas de la matriz A de manera que la primera sea el vector propio o componente asociado

con λ_1, la segunda sea un vector propio asociado con λ_2 y así sucesivamente. En particular, dichas columnas pueden estar formadas por vectores propios normalizados, es decir, perpendiculares entre sí y de longitud igual a la unidad. De esta manera se construye una matriz que produce la rotación deseada ya que, como puede probarse, el primer vector propio $a_1 = (a_{11}, a_{12},... a_{1p})'$ apunta en la dirección de máxima variabilidad de la nube centrada. Esta dirección se llama primera dirección principal. El segundo vector propio $a_2 = (a_{21}, a_{22},... a_{1p})'$ apunta en la siguiente dirección de máxima variabilidad de la nube centrada, llamada segunda dirección principal y así sucesivamente.

Una vez resuelto el problema de la rotación, bastará multiplicar la variable centrada $X_c = X - \mu = (X^c_1, X^c_2,..., X^c_p)$ por la matriz de rotación A para obtener la nueva variable, $Y = (Y_1, Y_2,... Y_p)$ llamada variable de componentes principales. Cada componente Y_i del vector aleatorio Y se llama una componente principal. Evidentemente se cumple que $Y_j = a_{j1}X^c_1 + a_{j2}X^c_2,..., + a_{jp}X^c_p$, es decir, cada componente principal es una combinación lineal de las variables originales centradas. Para hacer el análisis de los autovalores se necesita desarrollar los conceptos y las propiedades que se verifican. La traza de $\sum$, por ser la suma de las varianzas de las variables originales Y_i recibe el nombre de varianza total, resulta claro que $traza\ (\sum) = traza\ (ADA^{-1}) = \sum_{i=1}^{p} \lambda_i$. Se puede probar además que $V(Y_i) = \lambda_i$ para $i = 1,2..., p$ y que $Cov\ (Y_i, Y_j) = 0$, con $i \neq j$. Esto implica varios aspectos, a saber: La varianza total es igual a la suma de los valores propios de λ_i e igual a la suma de las varianzas de las componentes principales. Es decir, la varianza total es la misma con las variables originales que con las variables transformadas Y_i.

Las componentes principales son variables aleatorias no correlacionadas entre sí, obtenidas mediante la transformación lineal del vector de las variables originales centradas por la matriz de autovectores. Esto es: $Y_j = A_j X_c = a_{j1}X^c_1, + a_{j2}X^c_2 + ... + a_{jp}X^c_p$, para $j = 1,2..., p$.

Resulta claro que $E(Y_j) = 0$ para $j = 1,2..., p$. Si todas las variables originales X_i son normales, entonces todas las componentes principales son normales. Como puede deducirse de lo anterior, la varianza total se descompone en un número finito de partes disjuntas λ_j de tamaños cada vez menores, lo que en la práctica proporciona un mecanismo para estudiar la posibilidad de reducir la dimensionalidad de representación de las p varia-

bles originales a m. En efecto, si despreciamos las últimas $p - m$ componentes principales, las primeras m tendrán una tasa de representatividad igual a $\frac{\lambda_1 + \lambda_2 + ... + \lambda_m}{VT} 100\%$ de la varianza total VT de las variables originales. Muchas veces este porcentaje es bastante alto con un pequeño valor de m lo que se traduce en una alta representatividad en un espacio de pocas dimensiones.

En la práctica resulta importante el caso $m = 2$ ya que si, se obtuviera una tasa de representatividad alta, se habría logrado describir el problema sobre un plano con una pequeña pérdida de información. Por supuesto que si la reducción a un espacio de dos dimensiones conlleva una alta pérdida de representatividad no se habrá logrado un éxito y las técnicas que aquí se propondrán para visualización de individuos y variables no serán muy buenas.

La ecuación $Y = AX_c$ implica $X_c = A^{-1}Y = A^tY$ lo que permite obtener las variables centradas originales como combinaciones lineales de las componentes principales. Esto en particular va a permitir representar gráficamente las variables originales centradas dentro del espacio de componentes principales, llamado espacio factorial, como puntos cuyas coordenadas son los coeficientes de X_i en la combinación lineal correspondiente.

Teniendo en cuenta que sólo las componentes principales iniciales llevan la mayor parte de la representatividad se podrá reducir el espacio factorial a dos o tres dimensiones, lo que lleva a una representación de las variables originales como vectores sobre un plano (plano factorial) o sobre un espacio tridimensional. La representación sobre el plano factorial Y_1Y_2 es particularmente útil pues permite visualizar relaciones de correlación entre las variables originales y de éstas con los ejes factoriales, lo que rápidamente da una idea de cómo y en cuánto contribuye cada variable a la conformación de los primeros componentes y qué tan fuertes son las dependencias entre las diferentes variables y los componentes. La ausencia de correlación se traduce en vectores que tienden a formar ángulos rectos. Esto sugiere que la correlación entre dos variables se mida a través del coseno del ángulo que ellas forman. Igualmente es factible realizar una represtación de individuos, es decir, una proyección de la nube de individuos sobre el plano factorial Y_1Y_2, el cual reúne la mayor representatividad de VT. Las correlaciones entre las variables originales y los factores se conocen comúnmente como cargas factoriales, dadas en una matriz de carga C de orden pxm. Los elementos de la matriz C están dados por:

$$c_{ij} = \frac{\sqrt{\lambda_j}\, a_{ji}}{\sqrt{V(X_i)}}, \text{ para } i = 1, 2,..., p \text{ y } j = 1, 2,..., m$$

Criterios para determinación del número de componentes principales

Es importante saber el mecanismo para determinar el número de componentes principales (CP) que recojan la mayor variabilidad de las variables originales estandarizadas. Hay varios criterios para la selección de CP, los dos más extendidos son el criterio SCREEN y el de los porcentajes acumulados de varianza.

- **Criterio 1**: Según Dallas [3], en este criterio se utiliza la gráfica *screen* de los eigenvalores, la cual se construye tomando como eje X el número de eigenvalor y en el eje Y los valores propios, como se muestra en la figura siguiente.

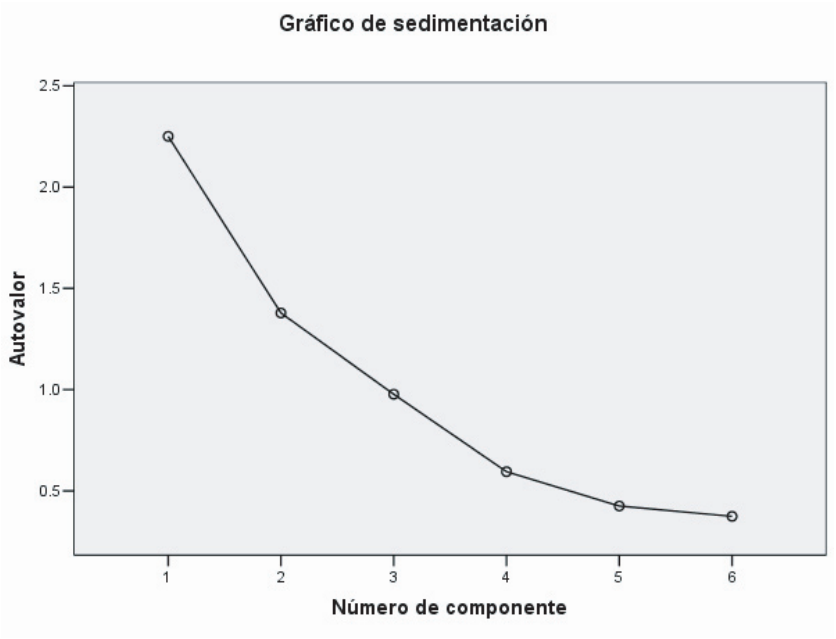

Figura 3. Gráfica *screen* para determinar número de componentes

- **Criterio 2.** Otro criterio, quizás más natural, consiste en retener tantos factores como sean necesarios para lograr un alto porcentaje de explicación de la varianza total. Para ello se usan los porcentajes acumulados de los valores propios con base en la varianza total del problema, junto con un criterio personal acerca de qué se considera un buen porcentaje de explicación. Los diversos investigadores sugieren que para datos tipo de laboratorio puede ser fácil explicar más del 95% de la variabilidad total con sólo dos o tres componentes principales y, que para datos de tipos de personas, negocios, estudios de mercados, etc., puede ser entre el 70% y el 75% de la variación total.

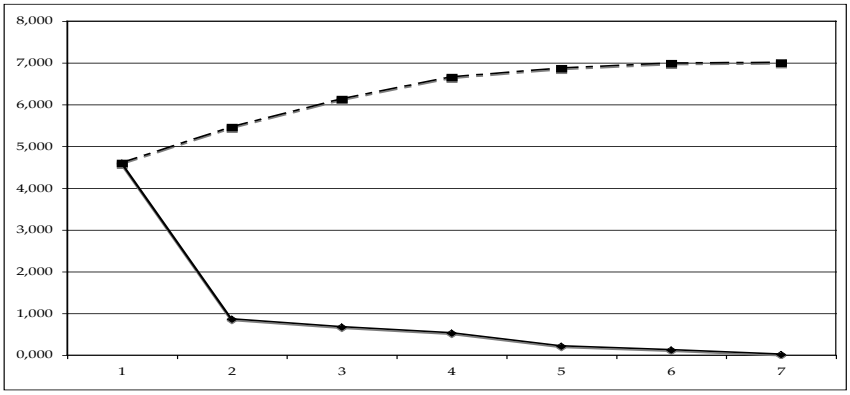

Figura 4. Porcentaje de varianza acumulado
Fuente: Elaboración propia con datos del caso presentado más adelante

ACP normado

Hay varios criterios. Tal vez los dos más extendidos son el criterio de Kaiser, según el cual se deben retener tantos factores como valores propios de la matriz $\sum$ estén por encima del promedio $^{VT}/_{p}$ y los diagramas de CATTELL [7]. Siendo VT : variabilidad total y p el número de variables originales. Todo lo mencionado anteriormente tiene un sentido geométrico y matemático muy claro pero en la práctica tiene un problema de interpretación. Por un lado, no tiene significado la combinación de variables cuyas naturalezas son diferente; por ejemplo las variables que representan la edad, ingreso, peso, etc., para la creación de un único factor. Por otra parte, el peso de cada variable original, traducido fundamentalmente en variabilidad, puede ser muy diferente para cada variable. Una variable muy dispersa puede contribuir enormemente a la varianza total mientras que una variable más homogénea contribuye menos. Esto finalmente determina

la participación de cada variable en la conformación de un factor. Las inquietudes anteriores tienen una solución: Realizar ACP con variables originales estandarizadas. Esto resuelve los dos problemas: De una parte, los valores de las variables estandarizadas son adimensionales, son simplemente números sin unidades en los cuales expresan las mediciones. De otra parte, la estandarización lleva todas las escalas de medida a una escala común de media 0 y varianza 1, con lo cual se elimina el problema de medición y variabilidad diferente de las variables originales. El ACP realizado con variables originales estandarizadas se llama ACP normado. Más adelante, en el ejemplo se puede ver que el ACP normado equivale al ACP corriente pero partiendo de la matriz de correlaciones R en vez de la matriz de varianzas covarianzas $\sum$.

Resulta claro que el ACP normado debe ser la técnica a seguir en cualquier caso, a menos que se quieran explorar algunas otras posibilidades de tipo teórico o que se tengan variables muy similares tanto en su naturaleza como en su escala de medida.

ACP a partir de una muestra

Finalmente, la matriz $\sum$ por ser desconocida no puede ser usada directamente en los cálculos. En la práctica, se usa la matriz de varianzas-covarianzas S, estimada a partir de una muestra observada de n individuos. Esta matriz constituye una estimación de $\sum$ y, por tanto, los resultados obtenidos con ella constituyen estimaciones de los correspondientes valores poblacionales. Se debe saber, sin embargo, que será necesaria una muestra aleatoria cuyo tamaño n sea mayor que el número p de variables consideradas.

De lo dicho anteriormente se obtienen algunas conclusiones:

- El ACP es una técnica que transforma ciertas variables correlacionadas en otras incorrelacionadas, de media cero, que pueden escribirse como combinaciones lineales de las primeras y que se llaman componentes principales, las cuales pueden ordenarse por la magnitud de su varianza, la cual está dada por un valor propio de la matriz.

- Las primeras y componentes principales bastan para describir en alto porcentaje la variabilidad total de las variables originales.

- Con frecuencia y vale 2 o 3, siendo el primero de ellos el caso más deseable.

- Cuando el porcentaje de variabilidad explicado por dos componentes principales es alto (70%) se puede realizar una representación gráfica de las variables originales y de los individuos de la muestra (mapas perceptuales) que muestran algunas relaciones de correlación o semejanza entre ellos.

4. UN CASO DE APLICACIÓN DEL ANÁLISIS DE COMPONENTES PRINCIPALES

Descripción del problema

Según la ley 387 de 1997, "Es desplazado toda persona que se ha visto forzada a migrar dentro del territorio nacional abandonando su localidad de residencia o actividades económicas habituales, porque su vida, su integridad física, su seguridad o libertad personales han sido vulneradas o se encuentran directamente amenazadas, con ocasión de cualquiera de las siguientes situaciones: Conflicto armado interno, disturbios y tensiones interiores, violencia generalizada, violaciones masivas de los Derechos Humanos, infracciones al Derecho Internacional Humanitario u otras circunstancias emanadas de las situaciones anteriores que puedan alterar o alteren drásticamente el orden público"[2].

Nuestro problema está basado en la lectura de la situación de los desplazados en un municipio de Colombia, donde se concentran el mayor porcentaje de estas personas que huyen de la violencia y el temor que generan las fuerzas oscuras en los campos del país.

Mediante entrevistas a expertos, a los mismos desplazados y la observación directa, se ha podido determinar problemas de diferente índole, tales como: la ubicación desordenada de los desplazados que han incomodado hasta llegar a roces con el personal que habita en los diferentes barrios, hacinamiento, inseguridad, y otros problemas de orden público.

Para analizar más profundamente esta problemática, los investigadores han recopilado información de fuentes (como, por ejemplo, el ministerio de Protección Social, el Sistema de Información de Hogares Desplazados por Violencia en Colombia – SISDES; el boletín sobre "Niños desplazados" editado por Codhes el 25 de octubre de 1997, entre otros) relativa a la población desplazada, con el propósito de contribuir desde la academia a ver técnicamente el problema con la ayuda del análisis de componentes principales.

[2] http://www.derechoshumanos.gov.co/modules.php?name=informacion&file=article&sid=120

Los datos de la Tabla 1 corresponden a la investigación exploratoria y estimaciones realizadas por los autores con el fin de encontrar los niveles de incidencia de los factores que conforman el problema de los desplazados en la comunidad.

Lo anterior se consigue mediante ACP, con lo cual se obtienen resultados útiles para ver más claro la gravedad del problema (véase resultados finales en esta sección). Para este estudio se han definido las variables que a continuación se nombran en los 25 lugares donde se ubican los desplazados: HPM: Horas promedio de movilidad diaria; NPM: Número promedio de desplazados por mes; NHS: Número de horas semanales que los centros de alimentación están en funcionamiento; ATR: Área total de recreación de uso común (en metros cuadrados); NBC: Número de centros del lugar de posible concentración; CCD: Cantidad de camas disponibles; NTC: Número total de cuartos; HHM: Horas-hombre mensual requeridas para atenderlos.

Tabla 1. Datos originales

Lugares	HPM	NPM	NHS	ATR	NBC	CCD	NTC
1	4	5	3	1,13	2	7	5
2	5	2,67	50	1,5	2	6	6
3	15,6	22,87	50	2,1	2	14	12
4	8	1,97	178	2,1	2	8	8
5	6,2	2,01	40	2,8	4	26	24
6	17,6	7,89	170	2,10	3	20	20
7	24,9	4	50	2,01	4	37	35
8	45,33	160,57	170	2,06	20	49	49
9	40,53	51,69	50	4,8	8	78	76
10	32,82	41,84	170	1,6	7	48	46
11	96,23	245,89	170	3	7	166	131
12	57,73	383,24	170	2,6	5	37	36
13	97,57	210,01	170	1,8	12	121	121
14	55,60	209,07	170	3,8	8	67	65
15	112,78	985	172	2,5	8	167	180
16	150,2	243,38	170	4,1	16	186	201
17	125,43	138,99	170	3	10	193	193
18	179,67	897	175	4,6	25	238	236
19	109,33	420	173	4,1	14	116	116
20	97,98	678,63	170	2,1	13	303	209
21	100,34	278,98	169	3,2	16	132	132
22	265,82	688,55	168	4,7	60	364	364
23	812,18	715,43	170	4,3	60	243	241
24	385,9	1569,68	168	3,7	20	541	454
25	89	379	167	3,1	10	293	195

Estos datos fueron procesados con SPSS y *Statgraphics* y se obtuvieron los resultados que aparecen a continuación, para sacar algunas conclusiones que sirven para consolidar el estudio sobre el ACP.

Estadísticos descriptivos

Tabla 2. Estadísticos descriptivos

Varaible	Media μ	Desviación típica □	Coeficiente de variación
HPM	117,4296	169,38660	144%
NPM	333,7344	392,80835	118%
NHS	139,3200	57,36486	41%
ATR	2,9120	1,08948	37%
NBC	13,5200	15,34090	113%
CCD	138,4000	134,15041	97%
NTC	126,2000	116,71725	92%

En la Tabla 2 se muestran la media, la desviación y el coeficiente de variación para cada una de las variables (análisis univariante). Estos valores permiten estimar la variable centrada tipificada Z (compárese con la Tabla 7). El objetivo de esta tipificación es homogenizar las unidades de medidas, buscando que todas pesen por igual en el análisis como se dijo anteriormente.

Matriz de correlaciones y prueba de independencia

Tabla 3. Matriz de correlaciones

		HPM	NPM	NHS	ATR	NBC	CCD	NTC
Correlación	HPM	1,000	0,614	0,335	0,520	0,823	0,613	0,667
	NPM	0,614	1,000	0,457	0,456	0,506	0,854	0,862
	NHS	0,335	0,457	1,000	0,306	0,359	0,458	0,479
	ATR	0,520	0,456	0,306	1,000	0,606	0,527	0,605
	NBC	0,823	0,506	0,359	0,606	1,000	0,585	0,674
	CCD	0,613	0,854	0,458	0,527	0,585	1,000	0,978
	NTC	0,667	0,862	0,479	0,605	0,674	0,978	1,000

Determinante = 0,000469886

El tener determinante bajo y coeficiente de correlaciones relativamente altas entre las variables originales es un buen indicador para utilizar la técnica de componentes principales que ayuda a resumir las variables en pocas dimensiones cuando se hace este tipo de análisis. Esto se debe a que las correlaciones altas implican dependencia lineal entre las variables, dando lugar a que se puedan explicar con un número menor de variables llamadas componentes principales Y_i. Todo lo anterior, y suponiendo normalidad de los datos, se puede corroborar con la prueba de independencia que se muestra en la siguiente tabla (p-valor=0 es menor que 0,05 y KMO es próximo a 1):

Tabla 4. KMO y prueba de Bartlet

Medida de adecuación muestral de Kaiser-Meyer-Olkin		,775
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	159,646
	Grados de libertad	21
	p-valor	,000

Comunalidades

En el análisis de componentes principales, las comunalidades son los elementos diagonales de la matriz analizada (la matriz de correlaciones o covarianza); en el análisis factorial, son las sumas de los cuadrados de las saturaciones para cada variable, utilizando todas las otras variables como predictores (ver Tabla 5). Esto indica, en el caso de la matriz de correlaciones, que la comunalidad es el porcentaje de varianza explicado por los componentes principales de la variable original determinada. Por ejemplo, para HPM, se observa que los tres componentes elegidos explican aproximadamente el 79,87% de la variabilidad; sin embargo, el porcentaje correspondiente al componente uno es de 83% frente al 16% y 0.3% en los componentes 2 y 3 respectivamente. Igual comportamiento se observa en las demás variables, excepto con NHS (número de horas semanales de los centros de alimentación en funcionamiento). Teniendo en cuenta estos porcentajes, podemos afirmar que todas las variables pueden resumirse con un solo componente principal, como se puede reconfirmar con los resultados mostrados en la Tabla 6. Para el caso en estudio, se podría pensar que el gobierno si está cumpliendo en parte con las necesidades de los desplazados, si se tiene en cuenta la alta correlación de las variables con el componente principal.

Tabla 5. Correlaciones entre variables y componentes, comunalidades y porcentaje de varianza explicada

	c_1	c_2	c_3	$\sum_{j=1}^{3} c^2{}_{ij}$	$c_{i1}^2 / \sum_{j=1}^{3} c_{ij}^2$	$c_{i2}^2 / \sum_{j=1}^{3} c_{ij}^2$	$c_{i3}^2 / \sum_{j=1}^{3} c_{ij}^2$
HPM	0,816630257	-0,35941546	0,05188733	0,79875674	0,834903719	0,161725674	0,003370607
NPM	0,857274396	0,29328938	-0,27385001	0,89593188	0,820284901	0,096010269	0,08370483
NHS	0,567356169	0,50330097	0,65030901	0,9981067	0,322503619	0,253792372	0,423704009
ATR	0,704094314	-0,34567018	0,15784723	0,64015243	0,774423064	0,186655349	0,038921586
NBC	0,808354752	-0,43939489	0,18301901	0,88000124	0,742541463	0,219395002	0,038063535
CCD	0,907653937	0,23540294	-0,26851144	0,9513486	0,86596613	0,05824841	0,075785459
NTC	0,948948613	0,14904507	-0,20225875	0,9636265	0,934494297	0,023052948	0,042452755

Autovalores y varianza explicada

Los autovalores se relacionan con la varianza explicada y permiten determinar el número de componentes principales adecuado (ver Tabla 6). En el caso de valores tipificados, el número de componentes principales está dado por aquellos autovalores mayores que uno. En este caso, solamente habría un solo componente principal correspondiente a lambda ($\lambda_1 = 4{,}597$); sin embargo, para el caso se trabajará con tres componentes. Los valores de porcentaje de la varianza se estiman a partir del cociente entre λ_j y la traza de la matriz de R.

Tabla 6. Valores de λ y Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	λ_j	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4,597	65,675	65,675	4,597	65,675	65,675
2	,859	12,267	77,942	,859	12,267	77,942
3	,672	9,600	87,542	,672	9,600	87,542
4	,526	7,519	95,061			
5	,211	3,016	98,077			
6	,121	1,736	99,813			
7	,013	,187	100,000			

Técnica de extracción: Análisis de Componentes principales.

La Figura 5 permite ver con mayor claridad los datos estimados en la tabla anterior.

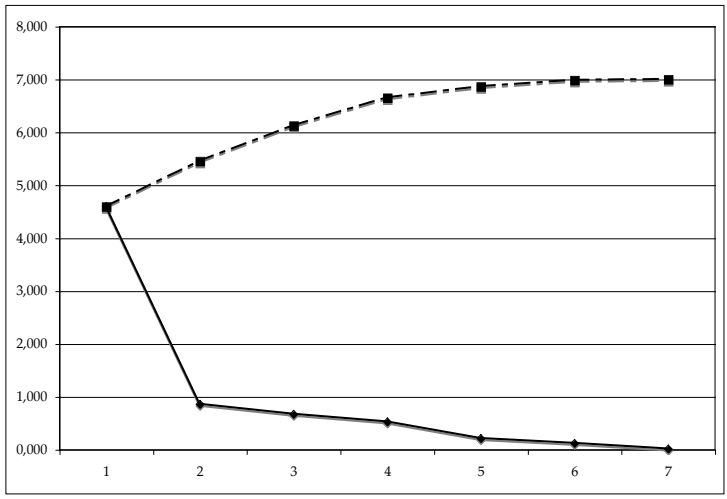

Figura 5. Gráfico de sedimentación individuales (serie 1) y acumulados (serie 2)

En la Tabla 7 se muestra la matriz de componentes C y la matriz de coeficientes de puntuaciones que permiten estimar los componentes principales.

Tabla 7. Matriz de componentes C y autovectores: $a_j = c_j / \sqrt{\lambda_j}$

	c_1	c_2	c_3	a_1	a_2	a_3
HPM	0,816630257	-0,35941546	0,05188733	0,38086158	-0,38793105	0,06329544
NPM	0,857274396	0,29328938	-0,27385001	0,39981727	0,31655861	-0,33405951
NHS	0,567356169	0,50330097	0,65030901	0,26460466	0,54323226	0,79328794
ATR	0,704094314	-0,34567018	0,15784723	0,32837686	-0,37309524	0,192552
NBC	0,808354752	-0,43939489	0,18301901	0,37700205	-0,47425596	0,22325813
CCD	0,907653937	0,23540294	-0,26851144	0,4233134	0,25407953	-0,32754718
NTC	0,948948613	0,14904507	-0,20225875	0,44257247	0,16087012	-0,24672797

Las Tablas 8, 9 y 10 muestran los procesos para determinar las coordenadas de los sitios en nuevo sistema de componentes principales mediante los datos estimados anteriormente y haciendo uso de las siguientes fórmulas o relaciones.

$$Z_{rj} = \frac{X_{rj} - \mu_j}{\sigma_j} \; ; \; Y_{rj_j} = Z_r^{'} a_j \; ; \; a_j = c_j / \sqrt{\lambda_j}$$

Donde $r = 1,2,..., n$; $j = 1,2,..., p$.

Tabla 8. Estimación de Z'_r

r	Z_{r1}	Z_{r2}	Z_{r3}	Z_{r4}	Z_{r5}	Z_{r6}	Z_{r7}
1	-0,66964918	-0,83688242	-2,37636787	-1,63564844	-0,75093359	-0,97949754	-1,03840691
2	-0,66374553	-0,84281406	-1,5570509	-1,29603569	-0,75093359	-0,98695186	-1,0298392
3	-0,60116679	-0,79138949	-1,5570509	-0,74531231	-0,75093359	-0,92731731	-0,97843291
4	-0,64603456	-0,8445961	0,67428044	-0,74531231	-0,75093359	-0,97204322	-1,01270377
5	-0,65666114	-0,84449427	-1,73137366	-0,1028017	-0,62056317	-0,83786548	-0,87562035
6	-0,58935948	-0,82952514	0,53482223	-0,74531231	-0,68574838	-0,88259139	-0,9098912
7	-0,5462628	-0,83942819	-1,5570509	-0,82792081	-0,62056317	-0,75586797	-0,7813755
8	-0,42565114	-0,44083686	0,53482223	-0,7820272	0,42240014	-0,66641614	-0,6614275
9	-0,45398868	-0,71802038	-1,5570509	1,7329429	-0,35982234	-0,45024088	-0,43009923
10	-0,49950586	-0,74309622	0,53482223	-1,20424846	-0,42500755	-0,67387046	-0,68713065
11	-0,12515512	-0,2236317	0,53482223	0,08077276	-0,42500755	0,20573921	0,04112503
12	-0,35244582	0,12602991	0,53482223	-0,28637616	-0,55537796	-0,75586797	-0,77280778
13	-0,11724422	-0,31497396	0,53482223	-1,020674	-0,09908151	-0,12970515	-0,04455211
14	-0,3650206	-0,31736698	0,53482223	0,8150706	-0,35982234	-0,53223839	-0,52434408
15	-0,02744963	1,65797291	0,56968678	-0,37816339	-0,35982234	0,21319353	0,460943
16	0,19346512	-0,23002159	0,53482223	1,09043229	0,16165931	0,35482559	0,64086499
17	0,0472316	-0,4957746	0,53482223	0,08077276	-0,22945193	0,40700583	0,57232328
18	0,36744582	1,43394508	0,62198361	1,54936844	0,74832618	0,74245019	0,94073497
19	-0,04781724	0,21961244	0,58711906	1,09043229	0,0312889	-0,16697675	-0,08739068
20	-0,11482372	0,87802513	0,53482223	-0,74531231	-0,03389631	1,22698094	0,7094067
21	-0,1008911	-0,13939215	0,51738995	0,26434722	0,16165931	-0,04770764	0,04969274
22	0,87604567	0,90327917	0,49995768	1,64115567	3,02980843	1,68169441	2,03740234
23	4,1015664	0,97170949	0,53482223	1,27400675	3,02980843	0,77972179	0,98357354
24	1,58495651	3,14643414	0,49995768	0,72328337	0,42240014	3,00110891	2,80849658
25	-0,16783854	0,11523584	0,4825254	0,17255999	-0,22945193	1,15243775	0,58945871

La Tabla 8 se puede representar la matriz normalizada de los datos originales consignados en la Tabla 1 que se utiliza para el cálculo de las coordenadas de los sitios en el nuevo sistema de ejes (componentes principales). El cálculo de las coordenadas se hace a través de la matriz de autovectores que se presenta a continuación.

Tabla 9. Autovectores a_j

	a_1	a_2	a_3
HPM	0,38086158	-0,38793105	0,06329544
NPM	0,39981727	0,31655861	-0,33405951
NHS	0,26460466	0,54323226	0,79328794
ATR	0,32837686	-0,37309524	0,192552
NBC	0,37700205	-0,47425596	0,22325813
CCD	0,4233134	0,25407953	-0,32754718
NTC	0,44257247	0,16087012	-0,24672797

En la Tabla 10 se presentan los valores de los componentes principales en el nuevo espacio generado por los tres componentes principales. Básicamente, se muestra que los datos se han reducido de un espacio de siete dimensiones a tres, lo que facilita la interpretación del problema.

Tabla 10. Valores de los componentes principales Y_j

r	Y_1	Y_2	Y_3
1	-2,91285905	-0,74559587	-1,5535249
2	-2,58402976	-0,43190798	-0,83549457
3	-2,31079557	-0,62195594	-0,77488609
4	-1,79283664	0,57386072	1,03324182
5	-2,05578636	-0,97421572	-0,80079009
6	-1,69418382	0,48924022	0,88105068
7	-2,12727927	-0,61420765	-0,84694064
8	-0,86923829	0,13182349	0,86979075
9	-0,81952521	-1,55651146	-0,51712258
10	-1,49086838	0,61817926	0,70438182
11	-0,02397398	0,49860804	0,33418315
12	-0,9077381	0,52101785	0,6189791
13	-0,47621251	0,62398372	0,35689028
14	-0,449762	-0,02136298	0,88749709
15	0,83758796	1,28502532	-0,44038125
16	0,97608354	-0,14758679	0,48506961
17	0,32689152	0,38943295	0,28267996
18	2,39937152	0,05628299	0,02775755
19	0,48545604	-0,07114595	0,68256914
20	1,02467317	1,33304386	-0,60431606
21	0,19229562	0,09665388	0,54097679
22	4,12183338	-1,07648216	0,08922985
23	4,41811932	-2,54888088	0,78294081
24	4,90406699	1,39691427	-1,99653083
25	0,82870989	0,79578681	-0,20725139

Tabla 11. Asignación de sitios por componentes

De acuerdo con el mayor valor absoluto de las coordenadas se presenta a continuación una clasificación de los sitios por componentes:

- **Componente 1**: 1-7, 10, 12, 16, 18 , 22-25
- **Componente 2**: 9, 11, 13, 15, 17, 20
- **Componente 3:** 8, 14, 19, 21

La Tabla 10 muestra los valores de coordenadas de los sitios en los componentes principales que fueron estimados. En dicha tabla aparecen en negrilla las coordenadas absolutas mayores de cada sitio asociadas con cada componente. Al asignar cada sitio a un componente siguiendo este criterio (véase Tabla 11), nuestro el modelo será como se muestra a continuación:

$$Y_{r1} = 0{,}380871Z_{r1} + 0{,}399827Z_{r2} + 0{,}264611Z_{r3} + 0{,}328385Z_{r4} + 0{,}377011Z_{r5} + 0{,}423324Z_{r6} + 0{,}442583_{r7}$$

Siendo r= 1,....., 25. Las otras dos columnas se estiman de igual forma utilizando como coeficientes los valores de las columnas dos y tres de la matriz de autovectores.

5. CONCLUSIONES

La aplicación del análisis de datos multivariado, específicamente de la técnica de componentes principales, al caso de los desplazados de un municipio de Colombia permite sacar las siguientes conclusiones:

Las siete variables originales estudiadas en el caso de los desplazados quedan resumidas en tres índices (componentes principales), que están explicando el 87,542% de la variabilidad total (ver Tabla 7).

Según la matriz de autovectores, el primer componente principal asocia las variables NPM, CCD y NTC, explicando un 65,675% de la variabilidad total, equivalente al 75% del total explicado por los tres componentes (ver Tablas 7 y 10).

Al segundo componente principal le corresponden las variables HPM, ATR y NBC, explicando el 12,267% de la variabilidad total, equivalente al 14% del total explicado por los tres componentes, ver tablas 7 y 10.

Al tercer componente principal le corresponde la variable NHS, explicando el 9,6% de la variabilidad total, equivalente al 11% del total explicado por los tres componentes (ver Tablas 7 y 10).

Analizando las variables por componentes se pueden evidenciar los siguientes aspectos (recordar que por ser un caso, las conclusiones sólo hacen referencia a los datos seleccionados):

Según los resultados de la Tabla 10 de autovectores, el componente 1 al asociar con más peso las variables NPM, CCD y NTC se puede definir como el índice que mide la atención que el gobierno local ofrece a los desplazados. Sin embargo, creemos que no son suficientes las acciones realizadas por el gobierno local y nacional dado que para la movilidad de los desplazados hay poca disponibilidad de camas y cuartos. Hay que anotar que sólo en algunos sitios se percibe que el gobierno cumple a cabalidad el compromiso adquirido con la sociedad. Por ejemplo, los sitios 22, 23 y 24.

En el componente 2, las horas promedio de movilidad diaria (HPM), que al compararlo con las área total de recreación de uso común (ATR) y números centros de lugar de posible concentración (NBC), explica en gran parte el interés del gobierno por concentrar a los desplazados en lugares determinados proporcionándoles áreas de recreación común. Esto se puede evidenciar en los sitios 9, 22 y 23, donde las áreas recreacionales y/o de concentración responden a la movilidad de los desplazados más adecuadamente que los demás sitios.

El componente 3 se puede definir como el índice de atención alimenticia de los desplazados. El número de horas semanales en que los centros de alimentación están funcionando, parece no ser suficiente para atender al personal que se presenta en los diferentes sitios.

Referencias

[1] D. Peña. *Análisis de datos multivariados*. Madrid: Mac Graw Hill, 2002, pp. 133-158.
[2] E. J. Dallas. *Métodos multivariados aplicados al análisis de datos*. México: Thomson, 2000, pp. 93-396.
[3] C. Pérez. *Técnicas de análisis multivariante de datos. Aplicaciones con SPSS*. Madrid; Pearson, 2004, pp. 121- 154.
[4] S. J. Bramardi. *Estrategias para el análisis de datos en la caracterización de recursos fitogenéticos*. Tesis doctoral, Universidad Politécnica de Valencia, Valecia, 2000, p 47-52.
[5] García Jiménez y J. Gil Flores. *Análisis factorial. Cuadernos de estadística*, Valencia, España: La Muralla, 2001.
[6] J. Gil Flores. *Análisis discriminante. Cuadernos de estadística*, Valencia, España: La Muralla, 2000.
[7] R. Cattell. The screen test for the number of factors. *Multivariate Behavioral Research*, 1, pág. 245-276, 1966.