Propuesta de Función PCA en R: Una aplicación para la detección de focos de pobreza en la Gran Caracas

Autor: Econ. Zeus Grafe Pérez

Palabras clave: Análisis de Componentes Principales, Pobreza, Caracas

La capital de Venezuela se caracteriza, entre otras cosas, por su distribución geoespacial y político-territorial. Compuesta de 22 parroquias en su municipio Libertador y de cuatro municipios compartidos con el Estado Miranda, la pobreza como factor poblacional se distribuye a lo largo de la ciudad con concentraciones distintas en las diferentes parroquias.

Por otra parte, en el lenguaje R existen funciones y librerías que permiten el desarrollo de estudios para el hallazgo de situaciones cotidianas. Para el uso de los datos en la toma de decisiones, se requiere llevar grandes volúmenes no sólo de datos, sino de estadísticos y cálculos a cifras relevantes que permitan entender la situación a analizar. La función prcomp() de la librería stats de R contiene elementos que facilitan el análisis de componentes principales ("ACP"), sin embargo, no incluye en su salida inmediata cifras en formato sencillo o estadísticos de interés para el análisis, mientras que la función PCA() de la librería FactoMineR si bien posee una gama más completa de datos para el análisis, requiere un nivel superior de conocimiento estadístico para su entendimiento en completitud.

Por ello, el presente trabajo consiste en el desarrollo de una función en R que permita el ACP de forma más sencilla y directa, tal que los resultados que muestre la salida del código sean, además de afables, completos en cuanto a requerimientos de información para el análisis situacional. Para ello, se da una aplicación práctica a través de la detección de focos de pobreza en las principales parroquias de Caracas y los municipios Baruta, Chacao, El Hatillo y Sucre del Estado Miranda, lo que permite comparar la pobreza por región y las causas que así lo determinan.

Función ACP

El problema principal de la función prcomp() es que no incluye información de interés para el investigador, como la variabilidad total o la contribución de las variables al factor, lo que permite conocer las tendencias y diferencias entre las clases a estudiar. Si bien la librería FactoMineR a través de la función PCA() contiene mayor información y elementos que permiten un PCA más versátil y una mejor interpretación de los resultados, contiene también elementos que requieren un dominio avanzado a nivel teórico para su entendimiento, lo que complica que los usuarios intermedios de R y estadísticos puedan ver directamente los aportes entre las variables.

Por ello, la función toma una forma simplificadora donde, primerísimamente, se toma las correlaciones entre las variables, adicionando los valores p de las mismas para conocer su significancia en el análisis. El problema con esta significancia es que no existe una función que permita hacer ese cálculo automáticamente, sino por parejas de variables, por lo tanto, a través de un bucle for, se rellena el espacio de una matriz vacía con las mismas etiquetas de las variables originales tal que se calcule los valores p de las correlaciones, lo cual resulta de interés para el investigador.

Por otra parte, de la función prcomp() se extraen los resultados a mostrar, modificando el cálculo de algunos de ellos para mostrar otros estadísticos relevantes. Por ejemplo, para mostrar la variabilidad total y la varianza explicada por cada autovalor, se crea un data frame que calcule el porcentaje de la varianza y su acumulado, creando las nuevas columnas con formato tidyverse y a través de las fórmula: eigenvalues / sum(eigenvalues) y cumsum(eigenvalues / sum(eigenvalues). Con esto, se obtiene la variabilidad total y el aporte del factor a la variabilidad total en el componente principal.

Posteriormente, a través de una lista, se seleccionan las correlaciones, valores p de las correlaciones, la variabilidad total, los elementos arrojados en la función prcomp() y la contribución al factor, que no es otra cosa que la "rotation" del ACP elevado al cuadrado. Por último se le asigna a

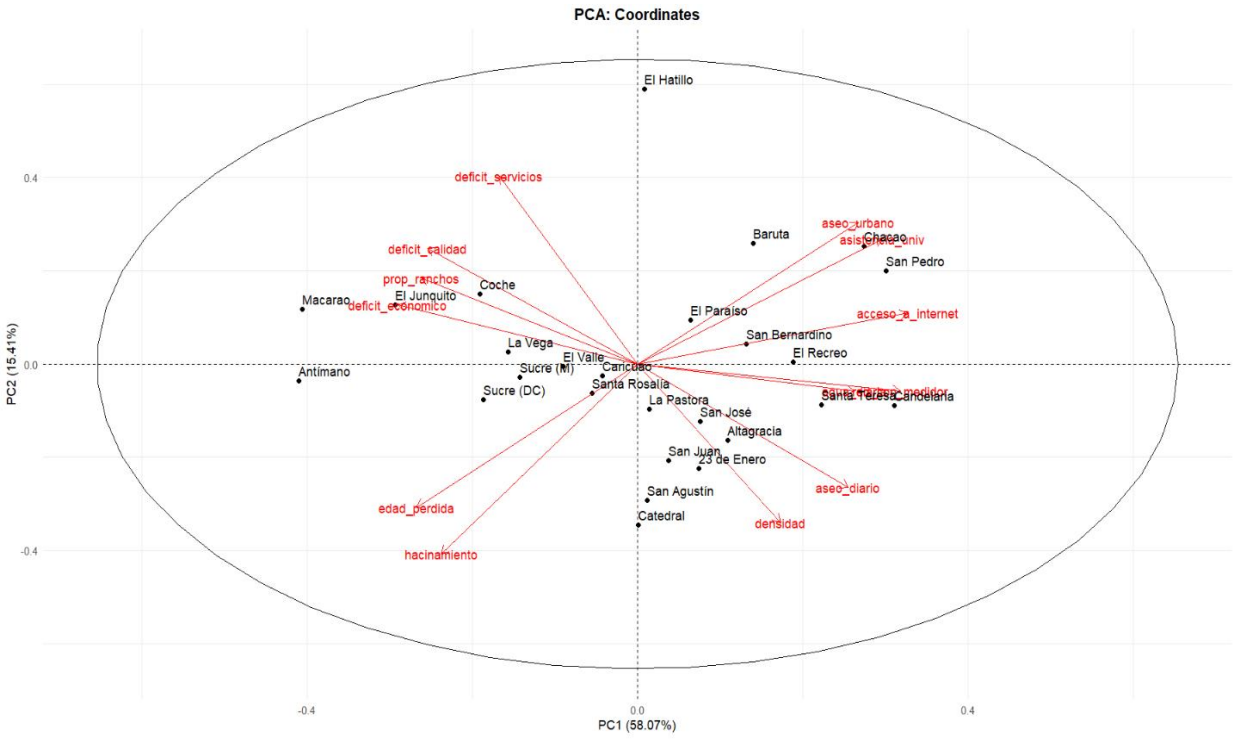
esta función una clase llamada PCA. Dado el gran volumen de información que trae esta función creada, se configura una función print() para esta nueva función, la cual muestra, primeramente, la variabilidad total para que el investigador conozca hasta qué componente principal vale la pena tomar, posteriormente toma las contribuciones al factor de los primeros cinco componentes principales para no sobrecargar el output de la función y separando cada tabla de la anterior.

De lo expuesto, para los datos de pobreza en la Gran Caracas, se obtiene la siguiente salida como resultado:

```
Principal Component Analysis (PCA)

Eigenvalues:
         Eigenvalues Percent. of Variance Cumulative Variance
Factor 1      2.7476                0.2114               0.2114
Factor 2      1.4153                0.1089               0.3202
Factor 3      1.1012                0.0847               0.4049
Factor 4      0.8395                0.0646               0.4695
Factor 5      0.6839                0.0526               0.5221

Contributions to the Factor:

                          PC1         PC2           PC3          PC4         PC5
acceso_a_internet 0.11854786 0.013132988 0.0005221231 1.100426e-05 0.014511442
red_con_medidor   0.11216503 0.003708001 0.0232585275 2.723000e-02 0.001450784
asistencia_univ   0.09722593 0.079242305 0.0001033299 4.030035e-03 0.037316097
deficit_economico 0.09277786 0.017659008 0.0011790110 4.213368e-02 0.033870052
aseo_urbano       0.07918709 0.102583223 0.0059468811 2.670146e-02 0.001737225
```

Con ello, se tiene información de cuál es la variabilidad total de cada autovalor, y que, para el estudio de la pobreza en la Gran Caracas, los primeros cinco componentes principales contienen el 52.21% de la variabilidad total acumulada. Asimismo, con base en la contribución al factor, se determina que el primer componente se concentra en los servicios tecnológicos, mientras que el segundo componente en el servicio sanitario y educativo.

Empleando un gráfico de tipo Biplot, se puede obtener las coordenadas de cada parroquia de la Gran Caracas que clasifica a cada parroquia de acuerdo a su calidad de vida y eficiencia en sus servicios, para lo cual, el estudio designó a cuatro tipos de parroquias, aquellas con buena y mala calidad de vida y aquellas con buena y mala calidad en sus servicios. Se muestra que los municipios Mirandinos poseen mayor calidad de vida y eficiencia en sus servicios que las parroquias del municipio Libertador, siendo El Paraíso, San Bernardino, El Recreo y San Pedro las parroquias en Caracas con mejor calidad de vida, mientras que Macarao, Coche, El Junquito y La Vega aquellas con peor calidad de vida y deficiencias en los servicios.



Comparativa: Nueva función PCA vs prcomp vs FactoMineR

La función prcomp() contiene, en su salida vía print() los datos de las desviaciones estándares de los componentes principales junto a sus coeficientes de coordenadas de acuerdo a su componente, mientras que en su salida vía summary() posee únicamente las desviaciones de cada componente principal y su aporte a la variabilidad total. La problemática radica en que, primerísimamente, los cálculos hechos por la función summary() respecto al aporte con la variabilidad está errado, ya que la variabilidad total se mide en varianza, que es el aproximador estadístico al autovalor, y la función utiliza la desviación estándar para su cálculo. Además, para el investigador es de interés conocer el aporte de cada componente principal a las variables y viceversa, lo que requeriría de cálculos manuales en R para su hallazgo.

Por otra parte, la función PCA() de FactoMineR contiene en su versión summary() los aportes de las varianzas a la variabilidad total junto a los aportes de las variables a los individuos y viceversa, además de los coeficientes de coordenadas. Esta versión es más completa para un PCA, sin embargo, requiere de conocimientos avanzados en estadística y sobre el trasfondo matemático de este método para comprender e interpretar adecuadamente estos resultados, lo cual hace que entusiastas y aprendices de estadística y data science obtengan resultados que entienden a medias.

La principal ventaja de la nueva función PCA() es que permite tanto a los aficionados en ciencia de datos como a aquellos investigadores que deseen conocer a las variables e individuos desde una perspectiva más correlacional puedan encontrar resultados directos que les permitan conocer con una sola función los aportes a la variabilidad, a las variables y a los individuos de cada componente principal. Ahora bien, ¿Qué ocurre si se desearía ver los coeficientes de coordenadas? Para ello, se configura en la función print() la alternativa de ver o no los coeficientes, siendo su estándar no verlos para comprimir la información a fin de no saturar al investigador de datos que en primera vista no requiere.

Por último, cabría señalar, ¿Qué aporte a nivel visual esperaría la nueva función PCA()? La función contiene por defecto un set gráfico hecho con ggplot2 que permite evaluar el gráfico biplot desde el punto de vista tanto de las variables como del individuo en simultáneo, tal que el análisis de datos por la vía del tabulado esté acompañado por la vista gráfica. Para ello se crea por defecto un gráfico con autoplot() con circunferencia hecha por stat_ellipse() y la respectiva introducción de los datos dentro del gráfico. Tanto la vista gráfica como la vista tabulada vienen incluidas en la función print() de la respectiva clase correspondiente a la nueva función PCA().