Universidad de Costa Rica - Sede de Occidente
Revista Pensamiento Actual - Vol. 15 - No. 24, 2015
ISSN impreso: 1409-0112 / ISSN electrónico: 2215-3586 ● *73 - 81*

Evaluación del estado socioeconómico del cantón de San Ramón: una aplicación del Método HJ-Biplot

Evaluation of the socioeconomic status of the canton of San Ramon: an application of the method HJ-Biplot

Carlomagno Araya Alpízar[1]

Recibido 6-10-2014 / Aprobado: 16-06-2015

Resumen

A partir de los datos del Censo de Población de Costa Rica de 2011, el presente trabajo es una evaluación estadística del estado socioeconómico del cantón de San Ramón (Alajuela), con la utilización el método HJ-Biplot. El lugar tiene una población de 80566 habitantes con una densidad de 79 habitante/km². El análisis HJ-Biplot permitió identificar dos grupos distritos en función de su comportamiento en las variables de estudio. Las variables que presentaron mayor nivel de correlación entre ellas son: la población de habitantes, número de hogares y número de estudiantes de los distritos.

Palabras claves: San Ramón; población; HJ-Biplot.

Abstract

Based on the data from the 2011 Costa Rican Population Census, this research study is a statistical evaluation of the socioeconomic status of the canton of San Ramon in the province of Alajuela, using HJ-Biplot. This canton has a population of 80566 inhabitants with a density of 79 inhabitants / km². The HJ-Biplot analysis identified two district groups based on their behavior of the study variables. The variables that show a higher level of correlation among them are: population, number of households and number of students in the districts.

Keywords: San Ramón; population; Biplot.

1. Introducción

San Ramón es el cantón 202 de la República de Costa Rica y el número dos de la provincia de Alajuela. Al considerar de los resultados del Censo de Población y Vivienda del año 2011, se observa que la población es de 80566 habitantes y que la densidad de la población del cantón es de 79 habitantes/km².

Por otra parte, los Métodos Biplot se usan para representar conjuntos de datos multivariados contenidos en una matriz de datos compuesta por p variables y n individuos, sin hacer supuestos sobre modelos subyacentes ni distribuciones poblacionales. De igual manera que un diagrama de dispersión muestra la distribución conjunta de dos variables (X, Y), un Biplot representa tres o más variables en un espacio de dimensión reducida (Gabriel y Odoroff, 1990). El prefijo "bi" se refiere a la superposición, en la misma representación de individuos y variables. Las representaciones de las variables son normalmente vectores, y coinciden con las direcciones donde se muestra mejor el cambio individual de cada variable.

El desarrollo del trabajo se realiza en tres secciones: en la primera, se establecen los principios básicos del método Biplot; en la segunda sección se presenta un análisis univariado y multivariado del estado socioeconómico del cantón, con el uso de los métodos Biplot mediante el programa MULTBIPLOT (Vicente, 2014). Esta técnica permite ubicar los distritos del cantón según un conjunto de variables de interés, y en la última sección se presentan las conclusiones de la evaluación socioeconómica del cantón de San Ramón.

(1) Doctor en Estadística de la Universidad de Salamanca. Magister en Estadística de la Universidad de Costa Rica. Profesor en la Universidad de Costa Rica, Sede de Occidente, San Ramón, Costa Rica. Correo electrónico: carlomagnocr@gmail.com

2. El Método Biplot

El objetivo del método Biplot es realizar una representación plana de una matriz $\mathbf{X}_{np}$ compuesta por ***n*** individuos (filas) y las ***p*** variables medidas sobre los individuos (columnas), mediante marcadores (vectores) $g_1,\cdots,g_n$ para las filas y $h_1,\cdots,h_n$ para las columnas, de forma que el producto interno $\mathbf{g}_i^T\mathbf{h}_j$ represente al elemento x_{ij} de la matriz de partida, tan bien como sea posible (Gabriel, 1971). Si se consideran los marcadores $g_1,\cdots,g_n$ como filas de una matriz **G** y los marcadores $h_1,\cdots,h_n$ como filas de una matriz **H**, entonces se puede escribir:

$$\mathbf{X} \cong \mathbf{G}\mathbf{H}^T$$

donde el símbolo $\cong$ significa que la matriz **X** se puede aproximar con el producto de la derecha. Se considera como ejemplo una matriz **X** con 3 variables y 4 individuos,

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ \mathbf{x}_{41} & x_{42} & x_{43} \end{bmatrix} \cong \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \\ g_{31} & g_{32} \\ \mathbf{g}_{41} & \mathbf{g}_{42} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{11} & h_{12} & h_{13} \\ \mathbf{h}_{21} & h_{22} & h_{23} \end{bmatrix}$$

Cada elemento de la matriz de partida se puede expresar como un producto de una fila de **G** por una columna **H** de . A manera de ejemplo $\mathbf{x}_{41} = \mathbf{g}_{41}\mathbf{h}_{11} + \mathbf{g}_{42}\mathbf{h}_{21}$. Asimismo, estos elementos de la matriz de partida pueden expresarse como un producto de una fila de **G** por una columna de $\mathbf{H}^T$. En general, se parte de una descomposición en valores singulares de la matriz $\mathbf{X}_{np}$ de rango **p**,

$$\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$$

Donde:	**U**	Es una matriz de dimensión (np) cuyos vectores columna son ortonormales y corresponden a los vectores propios de $\mathbf{X}\mathbf{X}^T$.
	V	Es una matriz ortogonal de dimensión (pp) cuyos vectores columna son los vectores propios de $\mathbf{X}^T\mathbf{X}$.
	D	Es una matriz diagonal de dimensión (pp) que contiene los valores singulares de **X**, ordenados de mayor a menor. Los valores singulares coinciden con los valores propios de $\mathbf{X}^T\mathbf{X}$.

Al aproximar una matriz de $\mathbf{X}_{np}$, de rango r, con una matriz de rango menor, $\mathbf{X}_q$, se está "perdiendo información", ya que la representación Biplot es aproximada. Una forma de medir esta pérdida es a través de la Calidad de Representación de los puntos fila y columna, cuanto más cercano esté a 100, mayor cantidad de información está siendo recogida por la representación Biplot.

Por otro lado, para que la representación Biplot sea útil, es necesario imponer una métrica de forma que la descomposición y el Biplot resultantes sean únicos. La elección de distintas métricas hará posible que la representación tenga diferentes propiedades, por lo cual la elección de métricas distintas puede manifestar diversos aspectos relevantes de los datos. En este sentido, entre los

métodos Biplot, se encuentran el GH-Biplot, JK-Biplot y HJ-Biplot.

El **GH-Biplot**, es una representación simultánea de individuos y variables, donde las variables tienen máxima calidad de representación. Los productos escalares de las columnas de **X**, coinciden con los productos escalares de los marcadores **H**.

$$\begin{aligned}\mathbf{X^TX} &= \left(\mathbf{GH^T}\right)^{\mathbf{T}}\left(\mathbf{GH^T}\right)\\ &= \mathbf{HG^TGH^T}\\ &= \mathbf{HU^TUH^T}\\ &= \mathbf{HH^T}\end{aligned}$$

A este Biplot se llama CMP-Biplot (Column Metric Preserving) ya que preserva la métrica euclídea usual entre las columnas de **X** donde se obtiene una alta calidad de representación para estas. La aproximación de los productos escalares (variancias y covariancias), en la dimensión reducida es óptima en el sentido de los mínimos cuadrados. La longitud al cuadrado de los vectores h_j aproxima la varianza de la variable **x**, por tanto, la longitud aproxima la desviación estándar. Además, el coseno del ángulo que forman dos marcadores columna aproxima la correlación entre las variables. Entonces, si los vectores son casi perpendiculares, el coseno del ángulo es próximo a cero y, por tanto, las variables son independientes, si el ángulo es cercano a cero estas presentan una correlación positiva alta y si el ángulo es próximo a 180 grados, la correlación es negativa y alta. Cuando se habla de ángulo se refiere al que forman los vectores en las direcciones crecientes de ambas variables (Villardón, 1992).

El **JK-Biplot**, es una representación simultánea de individuos y variables, donde los individuos tienen máxima calidad de representación, razón por la cual se conoce RMP-Biplot (Row Metric Preserving). A este Biplot Gabriel lo denominó JK-Biplot porque utilizó **J=UD** para denotar la matriz de marcadores fila y **K=V** para la matriz de marcadores columna., quien tiene las siguientes propiedades: los productos escalares de las filas **X** coinciden con los de los marcadores fila, los marcadores filas coinciden con las coordenadas de los individuos para las componentes principales, la similitud entre las columnas se aproxima utilizando como métrica la inversa de la matriz de dispersión entre los individuos y proporciona la mejor calidad de representación para filas.

El **HJ-Biplot** es una representación gráfica multivariada de las líneas de una matriz $\mathbf{X}_{np}$ mediante los marcadores $j_i,\cdots,j_n$ para las filas y $h_i,\cdots,h_n$ para las columnas, elegidos de tal forma que puedan superponerse en el mismo sistema de referencia con máxima calidad de representación (Galindo, 1985; Galindo y Cuadras, 1986). Se parte de la descomposición en valores singulares de la matriz $\mathbf{X}_{np}$**:**

$$\mathbf{X = UDV^T}$$

donde **J=UD** y **H=DV**

Esta elección de marcadores es equivalente a introducir en el espacio de las filas la métrica asociada a la inversa de la matriz de covariancias entre las variables y en el espacio de las columnas la métrica asociada a la inversa de la matriz de dispersión de las mismas unidades de medida de las variables.

Los elementos de la matriz están centrados por filas y columnas, razón por lo cual la métrica introducida en el espacio de las filas es equivalente a la inversa de la matriz de covarianzas entre variables, mientras que en el espacio de las columnas la métrica es equivalente a la inversa de la matriz de dispersión entre unidades de estudio (o sea individuos, objetos, instituciones, etc.). Dado que en el HJ-Biplot se puede hacer una representación simultánea de filas y columnas se le denomina también RCMP-Biplot (Row Column Metric Preserving).

El HJ- Biplot permite interpretar las posiciones de las filas, de las columnas y las relaciones fila-

columna a través de los factores (ejes), como en el caso del Análisis Factorial de Correspondencias (Benzecri, 1973; Greenacre, 1984) lo cual genera, además, la ventaja de que un análisis Biplot puede aplicarse sobre cualquier tipo de datos. Esta fue la principal razón que llevó a utilizarlo como herramienta de análisis multivariado en este trabajo del Estado del Cantón de San Ramón.

2.1 Bondad de la representación.

Cuando se realiza una aproximación multidimensional en un subespacio de menor dimensión que el de partida, se pierde información. Por lo tanto, se hace necesario valorar cuál es la cantidad de información que se consigue explicar, es decir, la bondad de ajuste de la representación en un subespacio de máxima inercia. Para ello se utiliza la absorción de inercia, de manera que cuanto más se acerque al valor 1 (o a 100%) más fiable será la representación.

Es evidente que cuanto mayor sea la dimensión del espacio de partida (es decir, de la matriz de datos) más difícil será tener absorciones de inercia próximas a 1 en el primer plano factorial, aunque las características más relevantes se manifiestan. Sin embargo, una tasa de absorción de inercia alta es necesaria para la fiabilidad al interpretar la posición de los puntos en los gráficos factoriales, pero no es suficiente.

La interpretación de las proyecciones de la nube de puntos en el subespacio de máxima inercia resultante de un análisis HJ-Biplot, se hace mediante la el uso de una serie de índices de contribuciones propuestos por Galindo en 1985.

3. Análisis de Resultados

La población total de San Ramón es de 80566 habitantes, de los cuales el 51% son mujeres según el Censo de Población de 2011. El distrito con mayor población es San Juan (11695 habitantes), además en el distrito central de San Ramón es donde se presenta la mayor diferencia entre la población según el sexo, ya que el 54% son mujeres. El distrito de Zapotal es el más pequeño respecto al tamaño de la población, con 391 habitantes de los cuales el 54% son mujeres.

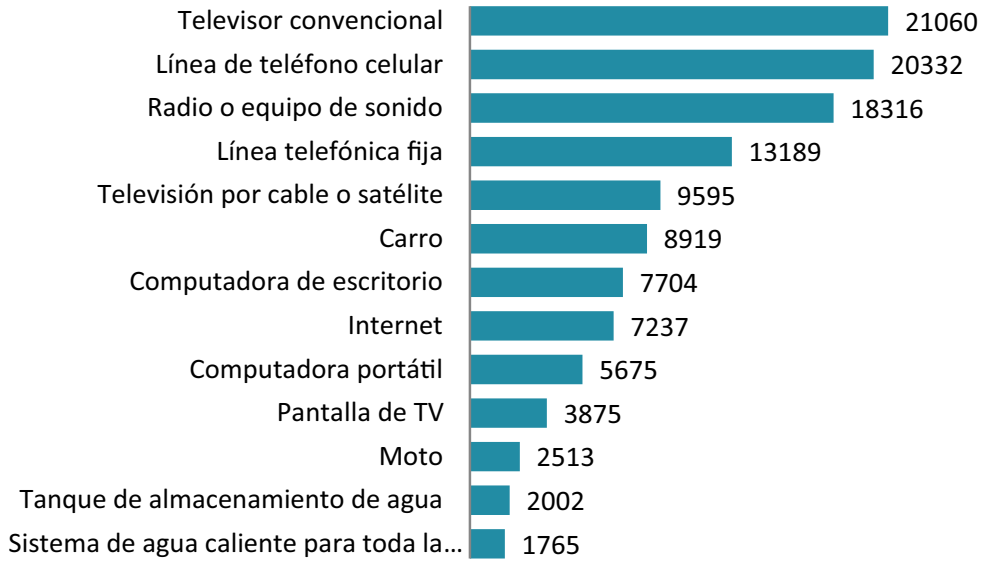
El cantón de San Ramón presenta la una tasa bruta de natalidad de 14,4 por mil y la tasa de fecundidad general es de 134,1 por mil. Por otra parte la tasa bruta de mortalidad del cantón es 3,2 por mil y la tasa de mortalidad infantil es de 4,5 por mil.

En cuanto las viviendas, en el área urbana del cantón existen un total de 13679 de las cuales 93% se encuentran ocupadas. En tanto, en la zona rural el 80% están en condición de ocupación de un total de 13191 viviendas.

Respecto al estado de las viviendas ocupadas en el cantón se tiene que un 7% de las viviendas se encuentra en mal estado, el 26% en un estado regular y un 67% están en buen estado. En este sentido, el 85% de las viviendas tienen acceso al abastecimiento de agua potable.

El equipamiento de las viviendas del cantón de San Ramón se puede observar en el Gráfico 1 se destaca que un 90% viviendas tienen al menos un televisor convencional y un 17% poseen una pantalla de televisión. Además, el 31% de las viviendas cuenta con servicio de internet, y en un 24% algún miembro posee computadora portátil.

Gráfico 1. San Ramón: Viviendas individuales ocupadas según equipamiento, 2011.



Fuente: Censo Nacional de Población y Vivienda de Costa Rica, Instituto Nacional de Estadística y Censos, 2011.

Por otra parte, existe un promedio de 3,4 personas por vivienda. El 73,7% son casas propias, en tanto, el 17,7% son alquiladas. El 7,6% de los hogares posee sistema de agua caliente, el 8,6% tiene tanque de almacenamiento de agua, el 38,3% posee carro y 10,8% motocicleta.

En el cantón hay 31182 personas que pertenecen a la población económicamente activa, de estos 97% se encuentran ocupados. Ahora bien, la caracterización de la población ocupada según el sector en el cual se desempeñan y el sexo, muestra que en el sector público en el cantón se desempeñan 5738 personas, de las cuales 52% son hombres, mientras que en el sector privado el 71% son hombres de un total de 24528 empleados.

El 17% de la población labora en el sector primario, mientras en el sector secundario se emplea el 18% de los ramonenses y en el sector terciario se emplea el 65%. Se tiene una tasa neta de participación de 51%, una tasa de ocupación de 49%, el desempleo es del 3%, mientras la población económicamente activa es de 49% y la relación de dependencia económica es de 2%.

Respecto de la caracterización educativa del cantón hay 83 escuelas diurnas. Para el año 2012 la matrícula de I y II ciclo fue de 8110 estudiantes y en la matrícula en Educación Preescolar fue de 1365 estudiantes. En cuanto a los colegios diurnos y nocturnos en el cantón, en total son 15, y reportaron una matrícula de 7205 estudiantes en el 2012. El 26% la población de 5 años asisten a un centro educativo público y el 6% a un centro de educación privado. En el nivel parauniversitario existen 567 hombres y 666 mujeres; en el nivel de instrucción universitaria en el cantón hay matriculados 5375 hombres y 6655 mujeres. La alfabetización en el cantón alcanza a 66677 personas, de estas 51% son mujeres.

A modo de resumen de la información anterior para el 2011, del Informe Estado de la

Nación se desprende, respecto del nivel educativo de la población ramonense, que el 4% no posee ningún tipo de educación formal, el 15% posee algún grado de Educación Primaria pero esta no fue completada, el 27% poseen Educación Primaria completa, el 18% de los pobladores posee Educación Secundaria incompleta, el 12% posee secundaria completa y el 23% de la población posee un nivel de Educación Superior.

A continuación, se presenta el nombre de la variable y la simbología que se presenta en los gráficos Biplots.

Variables	Simbología
Población de habitantes	pob
Número de hogares	hog
Número de estudiantes	est
Número de personas pensionados	pen
Porcentaje de personas ocupadas	ocu
Índice de desarrollo social	ids

Fuente: Elaboración propia.

El **índice** aborda condiciones esenciales para el desarrollo social en las diferentes dimensiones como se indica a continuación (MIDEPLAN, 2013).

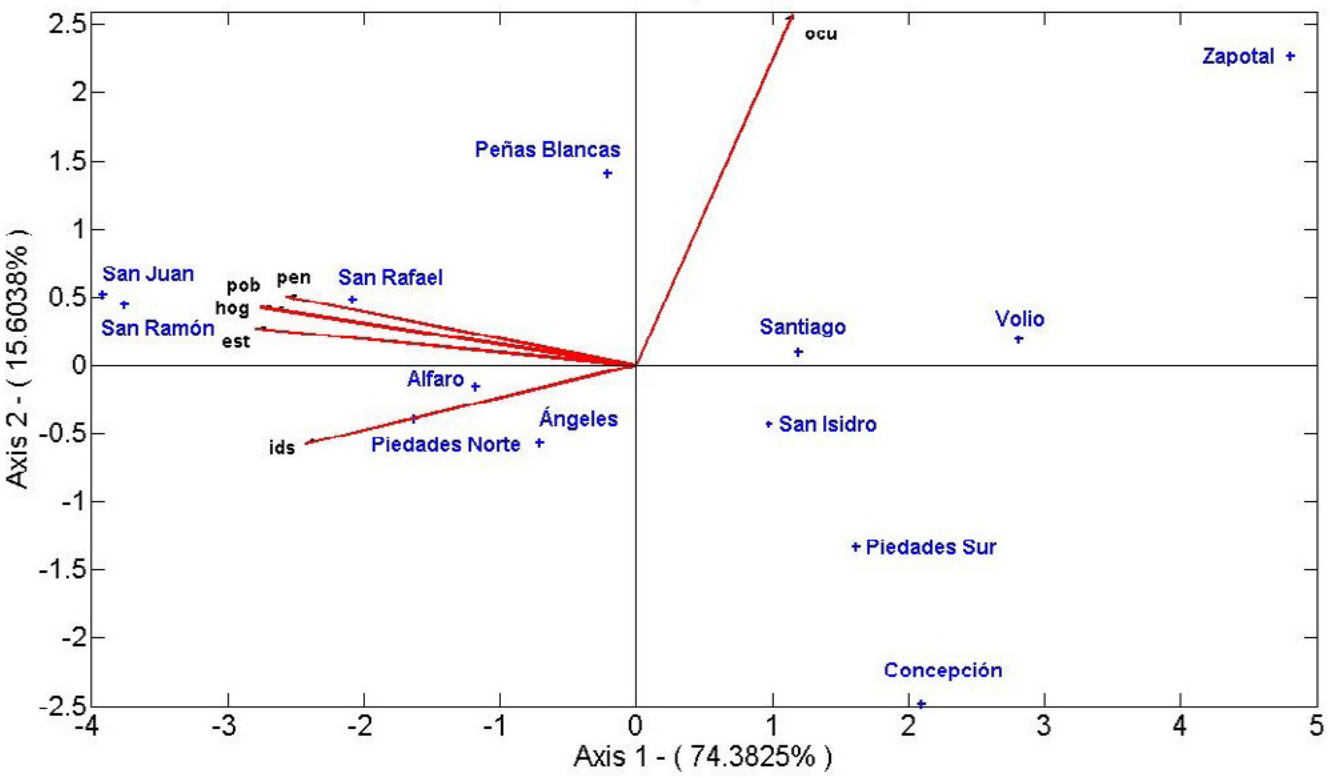
- ***Dimensión de Educación.*** Infraestructura educativa, programas educativos especiales, escuelas unidocentes y reprobación escolar.
- ***Dimensión de Salud.*** Bajo peso en niños(as), calidad del agua potable residencial, nacimientos en madres adolescentes solteras.
- ***Dimensión de Economía.*** Viviendas con acceso a internet, consumo residencial de electricidad).
- ***Dimensión de participación electoral.***

Para el análisis de los datos utilizando los métodos Biplot, se pueden utilizar distintos tipos de transformaciones en los datos (centrado, estandarización). Para este estudio, se tomó la decisión de estandarizar por variables (columnas) debido a las diferentes unidades de medida para poder compararlas y permitir un mejor descubrimiento de patrones en los datos.

La bondad en el ajuste global para el primer plano factorial (ejes 1 y 2) es del 90%. Todas las variables están bien representadas, ya que las calidades de representación acumuladas para los dos primeros ejes son superiores a 800. Las calidades se evalúan en una escala de 0 a 1000 puntos. Esta contribución permite saber las variables que tienen relación con los ejes factoriales, responsables de la colocación de los distritos sobre las proyecciones en cada uno de los ejes. Respecto de las calidades de las representaciones de las filas (CR_{filas}) se obtienen altas calidades para todos los distritos del cantón.

En la representación Biplot (**Gráfico 2**) se observa que el distrito central de San Ramón, San Juan y San Rafael, están ubicada más cerca de las variables número de personas pensionados (pen), población de habitantes (pob), número de hogares (hog) y número de estudiantes (est), esto implica que los distritos están caracterizados por mayores valores en las variables mencionadas. En tanto, Piedades Sur y Concepción poseen menor cantidad de habitantes, pensionados, estudiantes y hogares.

Gráfico 2. Representación Biplot de los distritos y variables en los planos factoriales 1 y 2.



Fuente: Elaboración propia.

Por otro lado, los distritos de Alfaro y Piedades Norte se encuentran en la dirección del índice de Desarrollo Social (ids); se indica que tienen altos valores, aunque San Juan, San Ramón y San Rafael también tienen altos niveles de desarrollo relativos. El distrito de Zapotal se encuentra en la dirección contraria del índice indicando que es el distrito con el menor desarrollo.

El distrito de Peñas Blancas tiene el mayor porcentaje de personas ocupadas, en tanto, Concepción el porcentaje más pequeño, esto se visualiza, al observar que la ubicación del distrito en la figura Biplot, es la más distinta con respecto al final del vector que representa la variable porcentaje de población ocupada.

Al interpretar la longitud de los vectores que representan las variables que proporciona este HJ-Biplot, se observa el porcentaje de población ocupada, es la que tiene mayor variabilidad entre los distritos.

Como se señaló con anterioridad, el ángulo de los vectores que representan las variables indica el grado de correlación. En este sentido, las variables que presentan mayor asociación entre ellas, son la población de habitantes, número de hogares y número de estudiantes de los distritos.

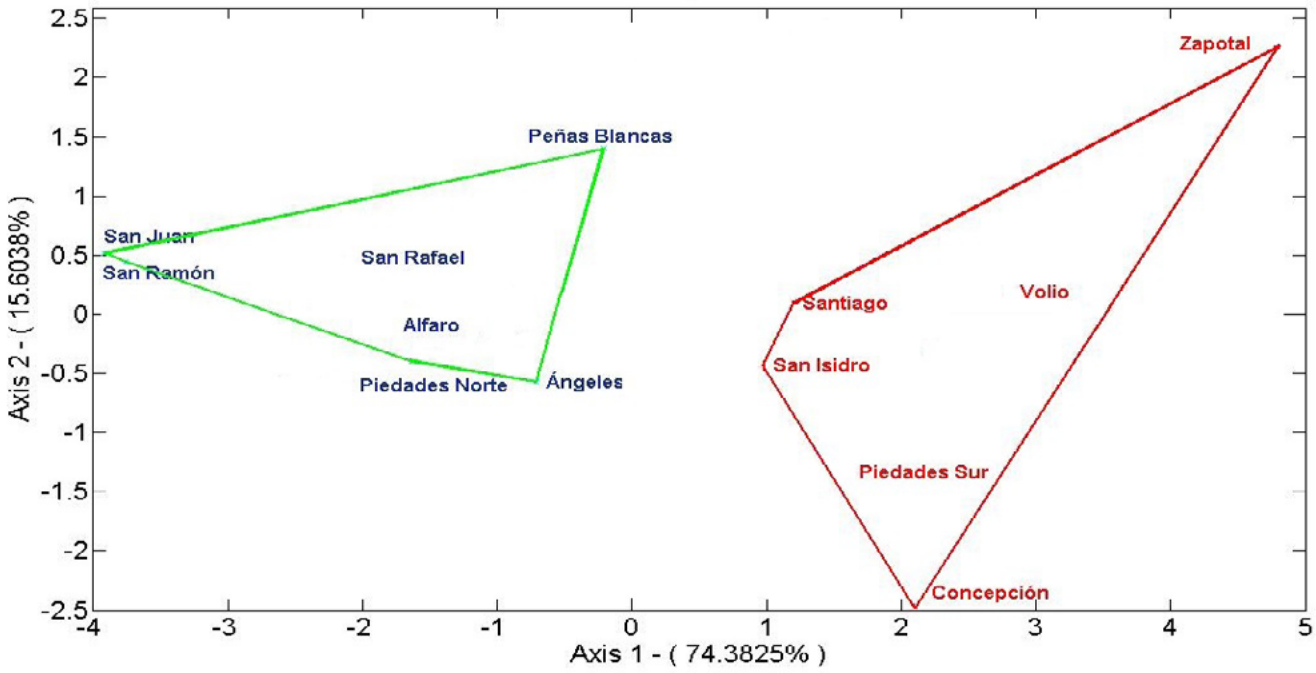
Agrupamiento de los distritos

El análisis de grupos (en inglés llamados cluster) es un método estadístico multivariado de clasificación que trata, a partir de una tabla de datos (distritos y variables), de situarlos en grupos homogéneos o conglomerados, de manera que

los distritos que pueden considerarse similares sean asignados a un mismo grupo. A través de las coordenadas Biplot se han calculado los conglomerados (método K-medias, distancia euclídea). Se observa en el Gráfico 3, la agrupación de los distritos en función de su comportamiento en las variables de estudio en el plano 1-2. Este método estadístico da como resultado dos conglomerados:

- **Grupo 1.** San Juan, San Ramón, San Rafael, Alfaro, Piedades Norte, Ángeles y Peñas blancas.
- **Grupo 2.** Concepción, Piedades Sur, San Isidro, Santiago, Volio y Zapotal.

Gráfico 3. Representación Biplot de la agrupación de los distritos según las variables de estudio en los planos factoriales 1 y 2.



Fuente: Elaboración propia.

Finalmente, una vez determinado los 2 grupos de distritos, se debe interpretar el perfil de cada uno de ellos. En todas las variables (excepto el porcentaje de ocupación de la población) existen diferencias estadísticas significativas entre los promedios aritméticos de los dos grupos a un nivel de confianza del 95%.

4. Conclusiones

En este estudio se ha analizado la caracterización socioeconómica de los distritos de cantón de San Ramón (Alajuela) en función de un conjunto de variables. A partir de la representación Biplot es posible disponer de información aproximada para evaluar visualmente de qué manera las variables explican la situación de los distritos.

A través de las coordenadas Biplot y utilizando el método K-medias, se encontraron dos conglomerados distritos en función de su comportamiento en las variables de estudio. En todas las variables (excepto el porcentaje de ocupación de la población) existen diferencias estadísticas significativas entre los promedios aritméticos de los dos grupos a un nivel de confianza del 95%.

Las variables que presentaron mayor nivel de correlación entre ellas son: la población de habitantes, número de hogares y número de estudiantes de los distritos.

Es posible afirmar que en este trabajo se presenta un método alternativo para analizar un conjunto de variables dentro de un espacio multidimensional. Por otra parte, el análisis univariado de datos indica que en el cantón de San Ramón el distrito de San Juan es el que tiene mayor cantidad de habitantes. Del total de viviendas el 67% están en buen estado.

La mayor parte de la población tiene agua potable y servicio de electricidad. El porcentaje de personas desempleadas es pequeño, lo cual da como resultado que el sector privado es el principal empleador. En cuanto, al sector educativo aproximadamente el 35% de las personas tienen un nivel de educación de secundaria completa y estudios universitarios.

Bibliografía

Benzécri, J. P. (1973). *L'Analyse des Donnðees,* Tome 2: L'Analyse des Correspondences Dunod, 519 - 521.

Gabriel, K. R. (1971). *The Biplot Graphic Display of Matrices with Application to Principal Component Analysis.* Biometrika, 58, 453-467.

Galindo, M. P. (1985*). Contribuciones a la Representación Simultánea de datos Multidimensionales*. Tesis doctoral. Universidad de Salamanca.

Galindo, M. P. (1986). Una Alternativa de Representación Simultánea: HJ-Biplot. Questió, 10, 13-23.

Galindo, M. P., Cuadras, C. (1986*). Una extensión del método Biplot y su relación con otras técnicas.* Publicaciones de Bioestadística y Biomatemática 17, Universidad de Barcelona, España.

Greenacre, J. (1984). *Theory and Application of Correspondence Analysis. London: Academic Press, Inc.*

Instituto Nacional de Estadística y Censos (2012). *X Censo de Población y VI de Vivienda: Características de las Viviendas/Instituto Nacional de Estadística y Censos San José, Costa Rica.*

MIDEPLAN (Ministerio de Planificación Nacional y Política Económica) (2013). *Costa Rica:* Índice de Desarrollo Social.

Vicente, J. L. (1992). Una alternativa a las técnicas factoriales basada en una generalización de los métodos Biplot. Tesis Doctoral, Universidad de Salamanca, España.

_____ *(2014). MULTBIPLOT: A package for Multivariate Analysis using Biplots*. Departamento de Estadística. Universidad de Salamanca. http://biplot.usal.es/ClassicalBiplot/index.html