J. health med. sci.,
6(1):45-50, 2020.

Análisis Clúster para Big Data: una aplicación con variables demográficas en provincias del Ecuador

Cluster analysis for big data: an application with demographic variables in provinces of Ecuador

Jaramillo-Feijoo Leyda Elizabeth[1], Galindo-Villardon María Purificación[2], Real-Cotto Jhony Joe[3]

JARAMILLO, L.; GAINDO, M. & REAL, J. Análisis clúster para big data: una aplicación con variables demográficas en provincias del ecuador. *J. health med. sci., 6(1)*:45-50, 2020.

RESUMEN: Los métodos de clasificación permiten explorar y analizar grandes conjuntos de datos visualmente, lo cual es de gran utilidad para tomar decisiones rápidas. El objetivo fue comparar dos métodos de análisis de clúster para big data en variables demográficas de las provincias del Ecuador. Se hizo uso de un estudio observacional de tipo comparativo mediante la representación simultanea del HJ-Biplot y el método Two Step (clúster bietápico), a través del software MultBiplot y SPSS. Los datos corresponden a variables demográficas de interés sociosanitarias tasa de mortalidad general, tasa de mortalidad infantil, tasa de natalidad, densidad poblacional, porcentaje urbano y esperanza de vida, medidas en las provincias del Ecuador. Se utilizaron datos provenientes del Instituto de Estadísticas y Censos INEC. Se analizó la asociación entre variables y se identificaron clústeres de las provincias del Ecuador según estas variables demográficas. Según la representación simultánea del HJ-Biplot se identificaron 3 clústeres, el clúster 1 son provincias con mayor densidad poblacional y tasas de mortalidad general, pero valores bajos de tasas de natalidad, el clúster 2 agrupa provincias con mayor esperanza de vida y tasas de mortalidad infantil pero bajos valores de tasa de natalidad y el clúster 3 están las provincias con valores altos de tasas de natalidad y valores bajos de densidad poblacional, esperanza de vida, tasas de mortalidad general y mortalidad infantil, distintos resultados se obtuvieron con el método Two Step. Se pudo concluir que estos métodos son de utilidad para explorar las similitudes entre las provincias según variables demográficas.

PALABRAS CLAVE: clúster, demográficas, HJ-Biplot, método two step.

INTRODUCCIÓN

Según cifras del censo del año 2010 por el Instituto Nacional de Estadísticas y Censos (INEC), Ecuador registra una población de 14 millones de habitantes, de los cuales el 66% es población urbana. El crecimiento de la población se ha visto afectada por la reducción de la tasa bruta de natalidad de 32,4 a 11,4 nacimientos por 1000 habitantes entre 1981 y 2010 (Lucio et al., 2011), la disminución de la tasa de mortalidad de 6,7 muertes por 1000 habitantes en 1981 a 4,3 en 2008 y la tasa de mortalidad infantil en 2009 fue de 20 por 1000 nacidos vivos.

Las técnicas de minería de datos son herramientas que tienen como propósito descubrir conocimiento, siendo el agrupamiento o clúster uno de estos métodos, (Shirkhorshidi et al., 2014).

El análisis de clúster es un método que permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori pero que pueden ser útiles una vez que se han encontrado, (Vicente, 2007) es una técnica no supervisada que se utiliza para clasificar grandes conjuntos de datos en grupos correlativos, tiene aplicación en muchos campos, se incluyen otras técnicas para el análisis de big data, tales como: el aprendizaje automático, el reconocimiento de patrones, entre otros. La agrupación consiste en clasificar objetos similares en grupos distintos, es decir la partición de un conjunto de datos en subconjuntos, los mismos que tienen características similares.

La generación de indicadores socio demográficos, epidemiológicos y de producción,

[1] Ingeniera en Estadística e Informática. Departamento Gestión de la Información y Productividad SOLCA, Guayaquil, Ecuador.
[2] Vicerrectora de ordenación académica y profesorado en la Universidad de Salamanca, España.
[3] Docente de la Universidad de Guayaquil. Departamento Gestión de la Información y Productividad SOLCA, Guayaquil, Ecuador.

permiten conocer la situación de salud de la población, facilitando la comparación y análisis de los avances en la salud individual y colectiva (OPS, 2020). Resulta interesante e importante identificar clústeres de estos indicadores epidemiológicos medidos en las áreas geográficas, determinar hallazgos y analizar patrones, siendo una herramienta que contribuye a determinar los puntos críticos y falencias a ser superadas; así como, un aporte significativo e importante en la epidemiología y salud pública. (Gonzalez, 2015). Desde el punto de vista de la información sanitaria, es beneficioso determinar si hay similitudes entre los países de cada región y de distintas regiones (Verhasselt & Mansourian, 1991).

En términos del tipo de datos que se utiliza, se pueden considerar para la aplicación de clúster, el jerárquico que está limitado a conjuntos de datos pequeños, K-Means está restringido a valores continuos y Two Step que permite crear modelos de clúster basado tanto en variables continuas como categórica y el número de clúster se determina automáticamente, se muestra la aplicación del método Two Step, y del HJ-Biplot, dando a conocer y destacando sus ventajas (Şchiopu, 2010) (Bacher et al., 2004).

Este tipo de estudio no se ha realizado en el país y más aún con la aplicación de diversos métodos para evaluar variables sociodemográficas de impacto sanitario, por lo que se tuvo como objetivo comparar dos métodos de análisis de clúster para big data en variables demográficas de las provincias del Ecuador.

MATERIAL Y MÉTODO

Estudio de tipo observacional y comparativo (Santos, 2015) donde se realizó un análisis de clúster con variables demográficas de interés sociosanitarias, medidas en provincias del Ecuador de la población femenina. Los datos fueron tomados de los registros administrativos del año 2017 del INEC. Las variables evaluadas fueron: tasa de natalidad, tasa de mortalidad, tasa de mortalidad infantil, esperanza de vida, densidad poblacional, población urbana. Se utilizaron métodos multivariantes y de clasificación, tales como: HJ-Biplot y Two step.

HJ-Biplot

El HJ-Biplot es una representación gráfica multivariante de una matriz Xnxp mediante los marcadores j_1…. jn para sus filas y $h_1 \ldots h_p$ para sus columnas,(Gabriel, 1971) (Galindo, 1986) elegidos de forma que ambos marcadores puedan ser superpuestos en un mismo sistema de referencia con máxima calidad de representación. Se utilizó el software MultBiplot (Vicente, 2010) para realizar el gráfico donde se obtiene la representación simultanea de las provincias, los indicadores epidemiológicos y la identificación de clúster.

$X = UDV^T \qquad J = U\,D$

$H = V\,D$

Two Step (cluster bietápico)

El origen del método "TwoStep" fue el algoritmo BIRCH, (Zhang et al., 1996) que fue desarrollado por Chiu et al. (2001) para realizar análisis de clúster, maneja variables de tipo continuas y categóricas. Consiste en dos fases: primero se realiza un proceso de pre-clusterización para todo el conjunto de registros agrupando éstos en muchos pequeños subclústeres y posteriormente se agrupan estos subclústeres mediante un algoritmo de agrupamiento jerárquico hasta obtener el número deseado de clúster.

Siguiendo esta metodología, como el número de elementos a procesar es mucho menor que el número total original de registros y dado que requiere un análisis para todos ellos, este algoritmo es muy eficiente desde el punto de vista de coste operacional. El método Two-Step permite dos tipos de medidas en función del tipo de variables que se tiene en la matriz, las mismas que son:

- Distancia Euclídea
- Distancia Máxima Verosimilitud

Además, maneja dos criterios de agrupamiento, el AKAIKE y el SCHWARTZ:

AKAIKE: $AIC = -2 * lnL(\theta) + 2K$

SCHWARTZ: BIC $= -2 * lnL(\theta) + (\ln(n) * K)$

El modelo con el valor BIC más bajo se considera el mejor en explicar los datos del análisis con el mínimo número de parámetros. Se utilizó el software SPSS versión 20 (Bacher et al., 2004) para realizar este análisis.

RESULTADOS

La Tabla I, muestra la calidad de representación de las provincias del Ecuador sobre el HJ-Biplot. Una vez preparada la matriz de datos, con las provincias como filas y las variables demográficas como columnas; siendo valores relativos, para el análisis HJ-Biplot se estandarizó por columnas y se realizó una descomposición de valores singulares, considerando una dimensión de dos componentes.

La Figura 1, resume la representación simultánea de la matriz de datos a través del HJ-Biplot y la técnica del clúster jerárquico con coordenadas biplot y distancia euclídea. El análisis HJ-Biplot permite hacer un ordenamiento de las provincias al proyectar los marcadores filas sobre las variables demográficas, a continuación, se detallan los clústeres formados:

- Clúster1: Guayas y Pichincha.
- Clúster2: Bolívar, Cotopaxi, Carchi, Azuay, Tungurahua, Chimborazo, Loja, Santa Elena, El Oro, Los Ríos, Manabí, Cañar, Imbabura.
- Clúster 3: Esmeraldas, Santo Domingo de los Tsáchilas, Orellana, Pastaza, Zamora Chinchipe, Sucumbíos, Morona Santiago, Napo.

Adicionalmente, se aplicó otra técnica de clasificación, el análisis de clúster bietápico con el software SPSS, dicha técnica combina variables continuas y categóricas, se incorporó la variable categórica tipo de región, que agrupa las provincias en 4 regiones que son: costa, sierra, oriente e insular, esta última fue excluida por corresponder a una provincia con poca población. Estos resultados son mostrados en la Figura 2.

Tabla I. Calidad de representación de las provincias del Ecuador sobre el HJ-Biplot.

Calidad de Representación de las Provincias			
	Provincias	Axis 1	Axis2
1	Azuay	443	603
2	Bolivar	204	771
3	Cañar	39	110
4	Carchi	445	691
5	Cotopaxi	3	593
6	Chimborazo	299	432
7	Imbabura	643	654
8	Loja	242	270
9	Pichincha	568	605
10	Tungurahua	527	696
11	Santo Domingo de los Tsáchilas	155	881
12	El Oro	400	748
13	Esmeraldas	562	809
14	Guayas	366	872
15	Los Ríos	9	333
16	Manabí	11	252
17	Santa Elena	130	221
18	Morona Santiago	764	826
19	Napo	495	655
20	Pastaza	981	982
21	Zamora Chinchipe	610	615
22	Sucumbíos	738	796
23	Orellana	847	859

DISCUSIÓN

La Tabla I mostró un 67,7% como varianza explicada del análisis, lo cual pone de manifiesto que todas las variables se encuentran bien representadas en el eje 1 y 2, aunque la tasa de natalidad está mejor representada en el eje 1; por otro lado, las provincias de Pastaza, Zamora Chinchipe, Sucumbíos, Orellana están bien representadas en el eje 1 y las demás provincias en el eje 2.

En la Figura 1, se pueden destacar tres clústeres, donde se caracterizó al clúster 1 con las provincias de alta concentración y desarrollo; el clúster 2, agrupó las provincias medianamente desarrolladas, mientras que el clúster 3 son las provincias de poco desarrollo, evidenciando una agrupación de forma más homogénea en los clústeres con las variables demográficas del estudio.

La clasificación con la técnica del HJ-Biplot, describe al clúster 1 por tener valores altos de densidad poblacional, población urbana, mortalidad general, mortalidad infantil y esperanza de vida. En contraste, se observan bajas tasas de natalidad, este grupo es considerado por tener provincias con alta concentración y desarrollo; el clúster 2, presenta valores altos de esperanza de vida, tasas de mortalidad general y mortalidad

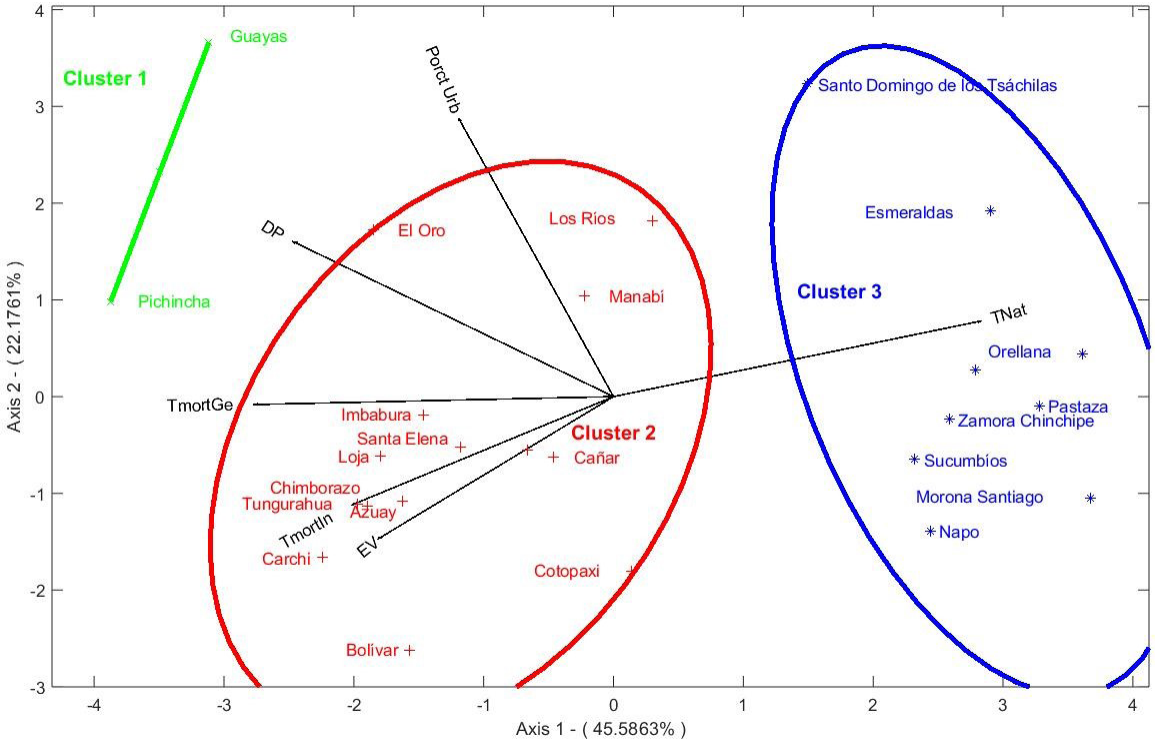

Fig. 1. Clúster y representación HJ-Biplot de provincias del Ecuador según variables demográficas.

Agrupaciones

Importancia de entrada (predictor)
1,0 0,8 0,6 0,4 0,2 0,0

Clúster	Etiqueta	Descripción	Tamaño	Entradas						
1			47,8% (11)	region Sierra (100,0%)	tmortG 4,26	tmortI 9,68	Tnat 16,47	porcurb 0,51	denspob 94,39	EV 77,03
2			26,1% (6)	region Oriente (100,0%)	tmortG 2,95	tmortI 8,20	Tnat 21,83	porcurb 0,42	denspob 6,98	EV 75,60
3			26,1% (6)	region Costa (100,0%)	tmortG 4,08	tmortI 8,53	Tnat 17,98	porcurb 0,66	denspob 106,27	EV 75,47

Fig. 2. Clúster Bietápico de regiones del Ecuador y variables demográficas.

infantil. En contraste tiene valores bajos de tasas de natalidad, densidad poblacional y población urbana; sin embargo, las provincias de El Oro, Los Ríos y Manabí muestran valores altos de población urbana. Este clúster se lo identifica como provincias medianamente desarrolladas; finalmente el clúster 3, se caracterizan por tener valores altos de tasa de natalidad. En contraste tienen valores bajos de esperanza de vida, densidad poblacional, población urbana, tasa de mortalidad infantil. Estas provincias se las identifica como de poco desarrollo.

Resultados similares se puede observar en el estudio realizado por Lucila Blanco y Carlos Mujica, donde se propuso una metodología aplicando la técnica de escalamiento multidimensional (MDS), para definir una configuración de países latinoamericanos con respecto a características sociodemográficas y económicas. Los resultados sugieren que las técnicas aplicadas pueden ofrecer una visión global del comportamiento sociodemográfico y económico de los países latinoamericanos, apoyados de análisis multivariante (Blanco & Mujica, 1996).

Según los resultados mostrados en la Figura 2, se identificaron tres clústeres donde la clasificación fue según la variable tipo de región, así se tiene que el clúster 1 representa el 47,8% y corresponde a la región sierra, el clúster 2 representa el 26,1% y corresponde a la región del Oriente y el clúster 3 representa el 26,1% y corresponde a la región costa.

En el clúster 1 se encuentran las 11 provincias que corresponde a la región Sierra, las mismas que son: Bolívar, Cotopaxi, Carchi, Azuay, Tungurahua, Chimborazo, Loja, Cañar, Imbabura, Santo Domingo de los Tsáchilas y Pichincha; en el clúster 2, están 6 provincias de la región Oriente que son: Orellana, Pastaza, Zamora Chinchipe, Sucumbíos, Morona Santiago, Napo; y en el clúster 3 están 6 provincias que corresponden a la región Costa: Guayas, Santa Elena, Los Ríos, Manabí, El Oro y Esmeraldas.

Con la técnica de Two Step, se obtuvo que la mayoría de los valores promedio del clúster 1 y 3 son similares, se diferencia porque el clúster 1 presenta mayor valor promedio de la tasa de mortalidad general, tasa de mortalidad infantil y de esperanza de vida; en cambio el clúster 3, presenta mayor promedio en la población urbana y densidad poblacional. El clúster 2, tiene mayor valor promedio de la tasa de natalidad y promedios bajos de densidad poblacional; dicho método, evidenció tres clústeres que son similares a la variable región, siendo una estructura distinta al método antes descrito.

Los resultados obtenidos según las técnicas de clúster reflejan la existencia de tres grandes estructuras sociodemográficas en el Ecuador; siendo una estructura, con la mayor concentración y desarrollo, la segunda estructura de mediano desarrollo y finalmente una tercera estructura de poco desarrollo.

Esta investigación, provee información sobre técnicas de clasificación aplicados a variables demográficas e indicadores básicos de salud medidos en las provincias del Ecuador; la identificación de clúster en provincias del Ecuador con características similares dentro del clúster y diferentes entre los mismos, ayudan a comprender la importancia de las técnicas de clasificación que permiten combinar variables socio-demográficas y las interacciones con las áreas geográficas (Mueller et al., 2019).

Limitaciones

Se han representado diferentes métodos para demostrar su utilidad y la aplicación en variables sociodemográficas, sería importante evaluar otras variables que complemente a este estudio a fin de brindar una información adecuada que refleje grupos prioritarios en la parte sanitaria, y por ende sea de apoyo a la toma de decisiones.

CONCLUSIONES

El HJ-Biplot, como técnica de representación gráfica multivariante, ha demostrado ser de mejor utilidad, ya que permite conocer las relaciones entre las variables, realizando un adecuado ordenamiento de los individuos y clasificación de los clústeres, siendo un aporte al análisis multivariante, reflejando información de posibles estructuras de conglomerados de las variables sociodemográficas medidas en las provincias del Ecuador.

Por otro lado, el método Two Step identificó clústeres en función de las variables estudiadas distintos al método HJ-Biplot, obteniéndose de este último una agrupación de forma más homogénea en los clústeres con las variables demográficas del estudio; con lo cual basados en estos análisis, se obtuvieron tres estructuras de desarrollo a nivel nacional.

JARAMILLO, L.; GAINDO, M. & REAL, J. Cluster analysis for big data: an application with demographic variables in provinces of Ecuador. *J. health med. sci., 6(1):*45-50, 2020.

ABSTRACT: The classification methods allow to explore and analyze big data sets visually, which is very useful for making quick decisions. This work aimed to compare of two methods of cluster analysis for big data in demographic variables of the provinces of Ecuador. An observational study of comparative type was carried out through the simultaneous representation of the HJ/Biplot and the Two Step method (two-stage cluster), through the MultBiplot and SPSS software. The data correspond to demographic variables of socio-health interest, general mortality rate, infant mortality rate, birth rate, population density, urban percentage and life expectancy, measured in the provinces of Ecuador. Data from Statistics and Census Institute were used. The association between variables was analyzed and clusters of the provinces of Ecuador were identified according to these demographic variables. According to the simultaneous representation of the HJ-Biplot, 3 clusters were identified, cluster 1 are provinces

with higher population density and general mortality rates, but low birth rates values, cluster 2 are provinces with higher life expectancy and mortality rates infantile but low birth rate values and cluster 3 are the provinces with high birth rates values and low population density, life expectancy, general mortality and infant mortality rates, different results were obtained with the Two Step method. It was concluded that these methods are useful for exploring the similarities between provinces according to demographic variables.

KEY WORDS: cluster, demographic, HJ-Biplot, two step method.

REFERENCIAS BIBLIOGRÁFICAS

Blanco, L. & Mujica, C. Representación de variables sobre una configuración de objetos obtenida a través de un escalamiento multidimensional. Rev. Venez. Anál. Coyunt., 4(2):223-36, 1998.

Bacher, J.; Wenzig, K. & Vogler, M. SPSS TwoStep Cluster-a first evaluation. SSOAR Social Sciences Open Access Repository, 2004.

Gabriel, K. R. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58(3):453–67, 1971.

Gonzalez, M. V. Modelo extendidos para el análisis espacial en epidemiología del cáncer. Universidad Nacional de Córdoba. Trabajo de Tesis para optar al Título de Magister en Estadística Aplicada, 2015.

Lucio, R.; Villacrés, N. & Henríquez, R. Sistema de salud de Ecuador. Salud Pública Méx., 53(2):s177–s87, 2011. Disponible en: https://www.redalyc.org/pdf/106/10619779013.pdf

Mueller, E.; Sandoval, J. S. O.; Mudigonda, S. & Elliott, M. 2019. A Cluster-Based Machine Learning Ensemble Approach for Geospatial Data: Estimation of Health Insurance Status in Missouri. ISPRS Int. J. Geo-Inf., 8(1):13, 2019. Disponible en: https://doi.org/10.3390/ijgi8010013

Santos, C. Two-step Cluster" en SPSS y técnicas relacionadas. Universidad de Salamanca. Máster en Análisis Avanzado de Datos Multivariantes. Trabajo de Fin de Máster. 2015.

Şchiopu, D. Applying TwoStep cluster analysis for identifying bank customers' profile. Buletinul, 62(3): 66–75, 2010.

Shirkhorshidi, A. S.; Aghabozorgi, S.; Wah, T. Y. & Herawan, T. Big data clustering: a review. International Conference on Computational Science and Its Applications. ICCSA 2014. Springer, pp. 707–20, 2014.

Organización Panamericana de la Salud (OPS). Situación de la salud, 2020. Disponible en: https://www.paho.org/ecu/index.php?option=com_content&view=article&id=25:situacion-salud&Itemid=135

Verhasselt, Y. & Mansourian, B. Método para la clasificación de los países de acuerdo con sus indicadores de salud, 1991. Disponible en: https://iris.paho.org/handle/10665.2/16636

Vicente, J. MULTBIPLOT: A package for Multivariate Analysis using Biplots. Departamento de Estadistica. Universidad de Salamanca, 2010.

Vicente, J. Introducción al análisis de clúster. Departamento de Estadística. Universidad de Salamanca. 22pp., 2007.

Galindo, M. Una alternativa de representacion simultánea: HJ-Biplot. Qüestiió: quaderns d'estadística i investigació operative, 10(1):13–23, 1986.

Zhang, T.; Ramakrishnan, R. & Livny, M. BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Record, 25(2):103–14, 1996.

Direccion para Correspondencia:
Leyda Jaramillo Feijoo
Departamento Gestión de la Información y Productividad SOLCA- Guayaquil
Av. Pedro Menéndez Gilbert y Atahualpa, parroquia Tarqui.
Guayaquil
ECUADOR
Teléfono: (593) 3718300

Email:
leydaj14@hotmail.com; ljaramillo@solca.med.ec

Recibido: 20-12-2019
Aceptado: 30-01-2020