Middle-East Journal of Scientific Research 27 (1): 06-11, 2019
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2019.06.11

K-Means Clustering Application: A Monetary Poverty Analysis of Colombian Regions

Víctor Daniel Gil Vera and Isabel Cristina Puerta Lópera

Universidad Católica Luis Amigó, Transversal 51A # 67B 90. Medellín, Colombia

Abstract: Poverty in Colombia is one of the issues that most affects the majority of the population. The lack of employment opportunities makes it impossible to have a decent quality of life, which in turn contributes to the violence and crime increase. The objective of this paper is to implement the clustering algorithm "k-means" for the grouping of Colombian departments that recorded the largest concentration of monetary poverty over the last ten years. We used the statistical software SPSS 25 and the data recorded by the National Administrative Statistics Department (DANE), corresponding to the monetary index of poverty from the year 2008 until the year 2017 in 24 Colombian departments. For the data analysis was used the k means cluster function of the software. This paper concludes that the departments that are located at the ends of the country are those that have the highest concentration of monetary poverty, a situation that is mainly due to the limited number of economic activities that develop. Strategies should be adopted that will permit the diversification of economic activities in these departments, as the development of agro-industrial and engineering projects, that allow to employ the population occupationally inactive.

Key words: Clustering · Colombia · K-means algorithm · Machine Learning · Poverty

INTRODUCTION

At present, large amounts of data are collected every day from satellite images, biomedical, security, marketing, web search, geo-spatial or other automated devices. The knowledge of these large volumes of information beyond human capabilities [1]. The grouping or clustering, is one of the methods of data mining more important for knowledge discovery in multidimensional data [2]. The clustering or grouping are methods used to identify groups of similar objects in a set of multivariate data. The different methods of clustering are: partitioning, hierarchical clustering, diffuse, based on density and in models [3].

The object of clustering is to identify patterns or groups of similar objects in a set of interest data [4]. In the literature, it is known as "pattern recognition" or " automatic learning not monitored". Not monitored because it is guided by a priori ideas that establish that type of variables or samples belong to the groups. Learning because the algorithms of machine learn how to cluster [5]. Here are some of the clustering applications found in the literature review:

- In the health area research has been conducted to classify patients with cancer in subgroups according to their gene expression profile. This allows you to identify the molecular profile of patients with good or bad prognosis, as well as for the understanding of the disease
- In marketing for the market segmentation through the identification of subgroups of clients with similar profile that may be receptive to a particular form of advertising
- In territorial planning to identify groups of houses, according to their type, value and location.

Corresponding Author: Víctor Daniel Gil Vera, Universidad Católica Luis Amigó Transversal 51A # 67B 90. Medellin, Colombia.

The objective of this paper is to implement the clustering algorithm "k-means" for the clustering of Colombian departments that recorded the largest concentration of monetary poverty over the last ten years. We used the statistical software IBM SPSS 25 and the data recorded by the colombian National Administrative Statistics Department (DANE), corresponding to the monetary poverty index from the year 2008 until the year 2017 in 24 departments of Colombia. For the data analysis, the distances are computed using simple Euclidean distance.

This paper concludes that the departments located in coastal areas, are the ones who recorded the highest rates of poverty, this may be because its main economic activity is tourism, which remains constant throughout the year, only in seasons. The departments located in the center of the country, are registered by lower rates of monetary poverty. The diversification of economic activities and the development of agro-industrial and engineering projects can contribute to solve this problem.

K-Means Algorithm: K means is the learning algorithm does not monitored, more commonly used for partitioning a given data set in a set of k groups, where k represents the number of prespecified groups by the analyst [6]. This algorithm classifies objects in multiple groups so that objects within the same cluster are as similar as possible while objects of different groups are the most diverse posible [7]. In the k-means clustering, each group is represented by its center which corresponds to the average of points assigned to the group [8]. The basic idea behind the k-means clustering, is to define groups so that the total variation within the cluster is minimized. There are several k-means algorithms available. The standard algorithm is the Hartigan-Wong algorithm [9], that defines the total variation within the group as the sum of the squared distances. The Euclidean distances between elements and the centroid for:

$$W(C_k) = \sum_{x_i \in C_k}^{n} (x_i - \mu_k)^2 \tag{1}$$

where:

x_i Corresponds to a data point belonging to the group C_k

μ_k This is the mean value of the points assigned to the group C_k.

Each observation (x_i) is assigned to a given group so that the sum of squares (SS), distance from the observation of their assigned cluster centers μ_k is minimal. The total variation within the cluster is calculated as:

$$tot.withinss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k}^{n} (x_i - \mu_k)^2 \tag{2}$$

First, it specifies the number of groups (k) that will be generated in the final solution. The algorithm starts by selecting at random *k* objects in the data set to serve as initial centers for the clusters. The selected objects are also known as centroids. Then, each one of the remaining objects is assigned to your nearest centroid, which defines the closest using the Euclidean distance between the object and the average in the cluster. This step is called the step of assigning cluster.

$$D_{euc}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

where X and Y are two vectors with lenght n. After the passage of assignment, the algorithm calculates the new average value for each group. The term update of the cluster centroid is used to design this step. Now that the centers have been recalculated, each observation is checked again to see if he could be closer to a different group. All objects are re-mapped using the cluster media up to date. The allocation of cluster and the upgrade steps of the centroid are repeated iteratively until the cluster assignments stop change, until that is achieved convergence. That is to say, the conglomerates formed in the current iteration are the same as those obtained in the previous iteration. In summary, the K-means algorithm includes the following steps:

- Specify the number of clusters (K) that will be created
- Randomly select k objects in the data set as initial media centers or of the group
- Assign each observation to your nearest centroid, based on the Euclidean distance between the object and the centroid
- For each of the k clusters, update the centroid of the cluster by calculating the new average values of all the data points in the cluster. The centoide of a group *K - th* is a vector of length p that contains the averages of all variables for observations in the kth; p is the number of variables.

Table 1: Monetary poverty

Department	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Antioquia	38,3	35,1	31,3	29,3	26,8	24,2	24,3	23,7	21,9	21,3
Atlantico	48,0	47,9	43,9	37,8	33,9	32,4	28,6	25,7	25,0	24,3
Bogota D.C.	19,7	18,3	15,4	13,1	11,6	10,2	10,1	10,4	11,6	12,4
Bolivar	58,3	57,1	49,4	43,7	44,2	41,8	39,9	39,3	41,0	38,2
Boyaca	58,0	48,0	47,1	39,9	35,6	39,3	38,2	35,4	32,0	28,7
Caldas	42,8	41,7	39,6	36,6	35,4	32,2	29,2	27,9	27,6	26,7
Caqueta	47,7	51,6	44,3	40,8	42,1	42,4	39,0	41,3	35,8	35,1
Cauca	66,4	66,1	64,7	62,0	62,1	58,4	54,2	51,6	50,7	48,7
Cesar	63,2	58,6	53,6	47,2	46,8	44,8	40,9	42,3	41,9	40,7
Choco	73,1	68,3	64,9	64,0	68,0	63,1	65,9	62,8	59,8	58,7
Cordoba	62,0	61,8	63,6	61,5	60,2	51,8	46,3	46,6	44,8	45,8
Cundinamarca	30,8	26,2	25,4	21,3	23,3	18,9	16,9	17,0	17,3	14,7
Huila	58,5	57,5	53,3	48,2	45,4	47,3	43,9	44,3	45,9	35,7
La Guajira	69,9	66,7	64,6	57,4	58,4	55,8	53,0	53,3	52,5	52,6
Magdalena	64,5	58,3	58,0	57,5	52,3	50,5	48,1	44,8	50,0	48,5
Meta	32,2	36,0	32,4	30,0	29,5	27,1	23,3	21,8	24,5	25,1
Nariño	56,1	55,1	56,4	50,6	50,8	47,6	42,9	40,0	45,7	40,2
Norte de Santander	50,7	47,5	43,1	40,6	40,4	39,4	39,9	40,0	40,4	40,0
Quindio	43,3	49,9	43,4	40,2	38,9	35,6	31,7	31,7	30,3	26,4
Risaralda	35,1	32,3	33,3	27,0	28,4	28,8	23,7	22,3	19,6	16,3
Santander	30,6	27,2	21,6	21,8	20,8	19,5	19,6	17,9	18,0	18,9
Sucre	66,6	66,2	63,7	53,0	51,5	47,3	43,9	44,7	46,7	41,6
Tolima	46,4	48,4	45,1	43,1	42,3	34,8	32,5	32,9	31,4	29,1
Valle del Cauca	33,4	33,3	30,7	30,0	26,9	27,2	22,7	21,5	22,6	21,1
Total	42,0	40,3	37,2	34,1	32,7	30,6	28,5	27,8	28,0	26,9

Source: [1]

- To minimize the total within the sum of the square. That is to say, repeat the previous two steps until the cluster assignments no longer change or reaches the maximum number of iterations.

Monetary Poverty in Colombia: Poverty in Colombia is one of the major social problems affecting the majority of the population. A large number of Colombians are considered poor. Unemployment, violence, low educational level and the lack of opportunities, are the main factors that exacerbate this problem in the country [10]. Table 1, presents the percentage (%) of monetary poverty in Colombia since the year 2008 until 2017, which measures the percentage of the population with incomes below the minimum monthly income defined as necessary to meet the basic needs:

Table 2, presents the geographic, demographic and economic characterization of the colombian departments:

The main colombian regions are: Caribe andina, Pacific, Orinoquía, Amazonic and insular. The Caribe region is composed by Guajira, Magdalena, Atlantico, Bolivar, Sucre, Cordoba and Cesar departments. The Pacific region by Choco, Valle del Cauca and Nariño departments. The Andina region by Cauca, Valle del Cauca, Choco, Antioquia, Risaralda, Caldas, Quindio, Tolima, Huila, Cundinamarca, Boyaca, Santander, North Santander, Cesar, Arauca, Casanare, Caqueta and Putumayo departments. The Amazonic región by Putumayo, Caqueta, Amazonas, Vaupes and Guainia departments. Finally, the insular region by San Andrés, Providence and Santa Catalina islands and the small islands; Alicia, Quitasueño, Serrana, Serranilla, Roncador and Albuquerque.

Methodology: In this study used the statistical software R Cran 3.4.3 and the data recorded by the National Administrative Department of Statistics (DANE), corresponding to the monetary poverty registered in the 32 departments of Colombia from the year 2008 until the year 2017. The data correspond to continuous variables. The role of R, scale () so that the algorithm does not depend on a variable unit arbitrary. Was used the "factoextra" package and the *kmeans()* function:

kmeans (x, centers, iter.max = 10, nstart = 1)

Table 2: Colombian departments characterization

Department	N° of Township	Population (Habitant)	Extention (km²)	Export Per Capita ($)	Import Per Capita($)	% Participation in National PIB
Antioquia	125	6.534.857	63.612	663.6	1.012.5	16,01
Atlantico	23	2.489.514	3.386	547.6	986.3	4,25
Bogota D.C.	1	7.980.001	1.775	306.8	2.753.7	12,05
Bolivar	46	2.121.956	25.978	643.0	1.235.4	5,1
Boyaca	123	1.278.107	23.189	218.2	76.0	3,01
Caldas	27	989.934	7.888	692.3	346.4	2,52
Caqueta	16	483.846	88.965	1.2	0.8	1,48
Cauca	42	1.391.836	29.308	207.7	237.7	2,49
Cesar	25	1.041.204	22.905	2.687.3	215.3	2,86
Choco	30	505.016	46.530	42.4	$3.4	1,43
Cordoba	30	1.736.170	23.980	23.980	26.2	1,49
Cundinamarca	116	2.721.368	24.210	505.1	1.498.9	4,33
Huila	37	1.168.869	19.890	387.1	15.8	1,79
La Guajira	15	985.452	20.848	1.521.2	489.9	2,01
Magdalena	30	1.272.442	23.188	471.2	155.3	2,3
Meta	29	979.710	85.635	635.7	40.9	2,01
Nariño	64	1.765.906	33.268	64.4	111.4	3,63
N. Santander	40	1.367.708	21.648	87.7	63.3	2,72
Quindio	12	568.506	1.845	461.7	141.5	1,8
Risaralda	14	957.254	4.140	553.6	496.7	1,53
Santander	87	2.071.016	30.537	371.7	276.3	7,50
Sucre	26	859.913	10.670	286.7	13.3	2,72
Tolima	47	1.412.220	23.562	126.4	58.6	4,17
V. del Cauca	42	4.660.741	22.195	406.4	799.9	10,8
Total	1047	47343546	659152	23.980	3568,7	100 %

Source: [1]

where:

- x: Numeric array, framework of numeric data or a numeric vector
- centers: possible values are the number of clusters (k) or a set of initial cluster centers. If you select a number, you choose a random set of rows (different) in x as initial centers
- iter.max: maximum number of iterations. The default value is 10
- nstart: number of Random boot partitions when centers is a number.

RESULTS AND DISCUSSION

The departments were classified in four clusters in each of the years considered (2008-2017) (Table 3), a total of 10 iterations were developed (Table 4):

Table 5, presents the classification of the departments in each cluster and the distance of the average value.

Table 3: Clustering classification

	Cluster			
Year	1	2	3	4
2008	32,20	73,10	19,70	58,30
2009	36,00	68,30	18,30	57,10
2010	32,40	64,90	15,40	49,40
2011	30,00	64,00	13,10	43,70
2012	29,50	68,00	11,60	44,20
2013	27,10	63,10	10,20	41,80
2014	23,30	65,90	10,10	39,90
2015	21,80	62,80	10,40	39,30
2016	24,50	59,80	11,60	41,00
2017	25,10	58,70	12,40	38,20

Source: author elaboration

Table 4: Change in clusters centers

	Cluster			
Iteration	1	2	3	4
1	9,416	23,464	13,142	5,486
2	1,177	3,352	3,286	,610
3	,147	,479	,821	,068
4	,018	,068	,205	,008
5	,002	,010	,051	,001
6	,000	,001	,013	9,290E-5
7	3,592E-5	,000	,003	1,032E-5
8	4,490E-6	2,849E-5	,001	1,147E-6
9	5,613E-7	4,070E-6	,000	1,274E-7
10	7,016E-8	5,815E-7	5,013E-5	1,416E-8

Source: author elaboration

In the cluster 1, Antioquia, Atlantico, Caldas, Meta, Quindio, Risaralda and Valle del Cauca departments were grouped. In the cluster 2, Cauca, Choco, Cordoba, Guajira, Magdalena and Sucre were grouped. In the cluster 3, Bogota D.C, Cundinamarca and Santander were

Table 5: Cluster of belonging

Case	Department	Cluster	Distance
1	Antioquia	1	11,160
2	Atlántico	1	15,933
3	Bogotá D.C.	3	17,523
4	Bolívar	4	6,172
5	Boyacá	4	15,848
6	Caldas	1	10,302
7	Caquetá	4	10,302
8	Cauca	2	6,764
9	Cesar	4	13,707
10	Chocó	2	27,375
11	Córdoba	2	11,977
12	Cundinamarca	3	9,251
13	Huila	4	13,538
14	La Guajira	2	6,457
15	Magdalena	2	13,553
16	Meta	1	10,762
17	Nariño	4	15,942
18	Norte de Santander	4	11,443
19	Quindío	1	21,009
20	Risaralda	1	14,287
21	Santander	3	9,604
22	Sucre	2	17,924
23	Tolima	4	19,180
24	Valle del Cauca	1	12,972

Source: author elaboration

Table 6: Final clusters centers

	Cluster			
	1	2	3	4
2008	39,01	67,08	27,03	54,86
2009	39,46	64,57	23,90	52,97
2010	36,37	63,25	20,80	49,04
2011	32,99	59,23	18,73	44,26
2012	31,40	58,75	18,57	43,45
2013	29,64	54,48	16,20	42,18
2014	26,21	51,90	15,53	39,65
2015	24,94	50,63	15,10	39,44
2016	24,50	50,75	15,63	39,26
2017	23,03	49,32	15,33	35,96

Source: author elaboration

grouped. Finally, in the cluster 4, Bolivar, Boyaca, Caqueta, Cesar, Huila, Nariño, North Santander and Tolima were grouped.

The maximum distance between clusters was the distance of the cluster 2 and cluster 3 (121,464). The minimun distance was the distance between the cluster 2 and cluster 4 (41,002). The average of the distance between clusters was 67,954.

CONCLUSIONS

The departments that are located at the ends of the country are those that have the highest concentration of monetary poverty, a situation that is mainly due to the limited number of economic activities that develop. Strategies should be adopted that will permit the diversification of economic activities in these departments. The development of agro-industrial and engineering projects, photovoltaic solar projects, eolic projects, the tourism reinforcement, the creation of technical programs for youngs and adults, can contribuite to poverty reduction in these regions. The national goverment must make significant investments in this populations.

REFERENCES

1. DANE, 2017. Monetary and Multidimensional Poverty in Colombia. Colombia. National Administrative Statistics Department, Bogotá D.C., Retrieved March 10, 2018 from: https:// www.dane.gov.co/ index.php/estadisticas-por-tema/ pobreza-y-condiciones-de-vida/ pobreza-y-desigualdad/pobreza-monetaria-y-multidimensional-en-colombia-2017#pobreza-monetaria-por-departamentos-2017.
2. Qian, Y., X. Liang, Q. Wang, J. Liang, B. Liu, A. Skowron and C. Dang, 2018. Local rough set: A solution to rough data analysis in big data, Int. J. Approx. Reason., 97: 38-63.
3. Gan, G. and M.K.P. Ng, 2017. K-Means Clustering With Outlier Removal, Pattern Recognit. Lett., 90: 8-14.
4. Ismkhan, H., 2018. I-k-means-+: An iterative clustering algorithm based on an enhanced version of the k-means, Pattern Recognit., 79: 402-413.
5. Nerurkar, P., A. Shirke, M. Chandane and S. Bhirud, 2018. Empirical Analysis of Data Clustering Algorithms, Procedia Comput. Sci., 125: 770-779.
6. da, L., F. Costa, F.N. Silva and C.H. Comin, 2018. A pattern recognition approach to transistor array parameter variance," Phys. A Stat. Mech. its Appl., 499: 176-185.
7. MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in Proceedings of the fifth Berkeley symposium on mathematical Statistics and Probability, 1(14): 281-297.

8. Li, Y. and H. Wu, 2012. A Clustering Method Based on K-Means Algorithm, Phys. Procedia, 25: 1104-1109.
9. Luz López García, M., R. García-Ródenas and A. González Gómez, 2015. K-means algorithms for functional data, Neurocomputing, 151(1): 231-245.
10. Hartigan, J.A. and M.A. Wong, 1979. Algorithm AS 136: A k-means clustering algorithm, J.R. Stat. Soc. Ser. C (Applied Stat., 28(1): 100-108.
11. Heath, J. and H. Binswanger, 1996. Natural resource degradation effects of poverty and population growth are largely policy-induced: the case of Colombia, Environ. Dev. Econ., 1(1): 65-84.