
Available online at www.jpit.az14 (1)
2023

Comparative analysis of k-means and fuzzy c-means algorithms on demographic data using the PCA method

Eltun Y. Ahmadov

Institute of Information Technology, B. Vahabzade str., 9A, AZ1141 Baku, Azerbaijan

eltunehmedov95@gmail.com

 <https://orcid.org/0000-0003-2710-3248>

ARTICLE INFO

<http://doi.org/10.25045/jpit.v14.i1.03>

Article history:

Received 6 October 2022

Received in revised form 9 December 2022

Accepted 23 December 2022

Keywords:

Data mining

Demography

Clustering

k-means

Fuzzy c-means

PCA

ABSTRACT

The concept of demography, which includes the processes such as birth, death, natural increase, improvement of employment and standard of living of the population, migration, etc., occupies a unique place in the global processes of the modern era. In this regard, this article uses clustering algorithms, which are estimated as a demographic data mining technology. For the analysis of demographic data, experiments are performed using k-means and fuzzy c-means clustering algorithms in the Python programming language. The experiment uses PCA method to reduce the size and get more effective results. Silhouette, Calinski-Harabasz and Davies-Bouldin indices, and algorithm execution time indicators are used to evaluate the quality of the algorithm. The result of the experiment shows the possibility of achieving an effective result through the k-means and fuzzy c-means clustering algorithms by applying the PCA method in the demographic data analysis.

1. Introduction

Data mining technology in the field of ICT is considered to be a very beneficial and important tool, since the issue of processing large volumes of data is a rather complex process and has accelerated the interest in the application of data mining. Data mining refers to discovering useful knowledge and patterns in large datasets to make certain decisions about future actions. Data mining technology emerged as a tool to extract previously unknown patterns and trends (useful knowledge) from raw data. Clustering, one of the data mining methods, is considered to be one of the most popular methods for grouping datasets. Data clustering is the process of identifying natural groupings or clusters within multidimensional data based on some dimension metrics (e.g., the Euclidean metric)

(Jain, Duin & Mao, 2000). The qualitative clustering method provides clusters with high intra-cluster similarity and low inter-cluster similarity. By applying the PCA method proposed in this study, k-means and fuzzy c-means algorithms are comparatively analyzed according to clustering efficiency.

2. The concept of demography

Demography as a multidisciplinary field of study that explores the regularities of events and processes occurring in the structure, location, migration and dynamics of the population based on the social, economic, cultural, medical-biological, geographical, and other factors (Alguliyev & Yusifov, 2021). The concept of demography, which includes the processes such as birth, death, natural increase, improvement of employment and standard of living of the

population, migration, etc., occupies a unique place in the global processes of the modern era. Demographic data such as population size, distribution and dynamics are considered important factors in assessing the needs of a country in terms of education and health facilities, physical infrastructure, employment and overall economic development.

Demography is a field of science that explores and describes population in general. More specifically, demography uses census data, surveys, and statistical models to analyze population size, migration, and structure. Demography not only studies the existing population but also works together with the factors affecting the population change. Moreover, demographers view how a population develops, changes, and reproduces over a period of time. Birth, death and migration rates are the key aspects of demographic analyses. Age and gender are important features determining these aspects. Some examples of demographics may include:

1. Age of death: Death in the early years of life and death at the age of 93 have very different consequences for a person. However, both affect the calculation of the total life expectancy of the population and its age structure.

2. Mother's age at the first birth: It makes a big difference whether a woman is 22 or 36 when she gives birth to her first child. Her chances of having more children throughout her lifetime are higher in the first case rather than in the second. In this regard, populations with a lower average age of mothers when she gives birth to her first child tend to have higher fertility rates and larger family sizes.

3. Gender (sex) distribution among newborn children: If there is a disparity in the gender distribution among newborn children for one or more generations, that is, if there is a clearly defined difference in the number of newborn boys and girls, this will affect their lives. For example, it will lead to the lack of partners to start a family, which may affect the birth rate (Ahmadov, 2021).

3. k-means algorithm

The k-means algorithm is a simple, numerical, iterative method and one of the

unsupervised clustering algorithms of machine learning. This algorithm has advantages and disadvantages, but it is perhaps one of the most popular algorithms due to its fast performance when applied to large data. The k -means algorithm is a partition-based clustering method that attempts to identify a user-specified number of k clusters. These clusters are represented by their centers (means). The algorithm consists of two separate steps: the first step is the random selection of k centers, where the value of k is predetermined. The next step is to assign each point to the nearest center. Distance metrics are used to calculate the distance between each point and the cluster centers. When all the data are assigned to certain clusters, the centroid of each cluster is recalculated.

The k -means algorithm is an iterative method consisting of dividing a set of n number of objects into k number of clusters so that the objects within the cluster should be similar to each other and different from the objects in other clusters. Let $N = \{x_1, \dots, x_n\}$ be a set of n number of objects to be grouped by the similarity criterion, where $x_i \in \mathbb{R}^d$ $i = 1, \dots, n$ and $d \geq 1$ are the number of dimensions. Additionally, let $k \geq 2$ be an integer.

The basic working principle of the k -means algorithm is to minimize the objective function (SSE).

$$SSE = \sum_{j=1}^k \sum_{x_i \in c_j} \|x_i - \mu_j\|^2,$$

$$\text{here } \mu_j = \frac{1}{n_j} \sum_{x_i \in c_j} x_i$$

c_j denotes the cluster, n_j - the number of objects in cluster c_j , μ_j - the cluster center, and x_i - the point.

The working principle of the k means algorithm is to always try to approach the local minimum. The particular local minimum obtained depends on the starting cluster centers. The k -means algorithm updates the cluster centers until a local minimum is reached. Distance and center calculations are performed until the k -means algorithm converges, and iterations are performed several times. Let the positive integer t is defined as the number of k -means iterations. The exact value of t varies depending on the starting cluster centers even within the same dataset. Thus, the

computational complexity of the algorithm is $O(n * k * t)$, where n is the total number of data, k is the number of clusters, and t is the number of iterations. For a multidimensional dataset, the computational complexity is $O(n * k * t * d)$, where d is the number of attributes (dimensions) in the dataset.

The complexity of correct determination of the number of clusters is one of the main disadvantages of the k -means algorithm. During clustering, factors such as noise in the data are not taken into account when selecting similar groups. Therefore, when applying the k -means algorithm to data with noise points, several difficulties may arise. The biggest drawback of the k -means algorithm is the selection of starting centers. Regarding these shortcomings, various scientific studies have been conducted and new modified versions of the k -means algorithm have been proposed (Oti et al, 2021).

Although the k -means algorithm has several disadvantages, it also has many advantages. One of the most widely used clustering techniques, k -means is considered to be one of the main solutions for clustering very large datasets. The k -means algorithm's ease of use, computational efficiency, and low memory storage have made it very popular even compared to other clustering methods. Correspondingly, the k -means clustering method has the advantage of enabling an unknown number of clusters to be searched in a dataset.

4. Fuzzy c-means algorithm

In hard clustering, each element in the dataset refers to only one cluster. Assume that the dataset is divided into k number of clusters, and a set of variables m_{i1}, m_{i2}, m_{ik} is defined, which represent the membership degree of the element i to the cluster k . In hard clustering algorithms, one of these variables will be 1 and the rest will be 0.

This indicates that each element is classified into only one cluster. In fuzzy or soft clustering, the elements may refer to more than one cluster, and each element is assigned a degree of membership to the clusters it refers to. In fuzzy clustering, membership degrees are assigned to all clusters. m_{ik} can be in the range of 0 and 1 in

this case. Membership degrees indicate the power of association between that element and a particular cluster.

The fuzzy c-means (FCM) algorithm is one of the most widely used fuzzy clustering algorithms. This algorithm was proposed by Dunn J. in 1973 and improved by Bezdek J. in 1981 (Grover, 2014). It is one of the most popular fuzzy clustering methods with the approach that elements have membership degrees with cluster centers to be updated iteratively (Chattopadhyay, Pratihar, Sarkar, 2011). Objects on the boundaries between several classes may not fully refer to one of the classes, but instead are given degrees of membership in the range of 0 and 1, indicating their partial membership (Suganya & Shanti, 2012). The computational complexity of the fuzzy c-means algorithm is $O(n * d * k^2 * t)$, where n is the total number of data, d is the number of attributes, k is the number of clusters, and t is the number of iterations. Fuzzy c-means is widely used in astronomy, chemistry, geology, and medical diagnosis (Yong, Chongxun & Pan, 2004).

Steps for fuzzy c-means clustering are as follows:

Specify the number of clusters c , where c must be in the range $2 \leq c \leq n$, and set a value for the parameter m in the range $1.25 \leq m \leq 2$. $U^{(0)}$ the membership degree matrix. Each step in this algorithm will be denoted as r , where $r = 0, 1, 2 \dots$

Calculate c number of center vectors $\{v_{ij}\}$ for each step.

$$v_{ij} = \frac{\sum_{k=1}^n (\mu_{ik})^m x_{kj}}{\sum_{k=1}^n (\mu_{ij})^m} \quad (1)$$

1. Calculate the distance matrix $D_{[c,n]}$.

$$D_{ij} = (\sum_{j=1}^m (x_{kj} - v_{ij})^2)^{1/2} \quad (2)$$

2. Update the membership matrix $U^{(R)}$ for the r th step.

$$u_{ij}^{r-1} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ij}^r}{d_{jk}^r} \right)^{2/m-1}} \quad (3)$$

If $\|U^{(k+1)} - U^{(k)}\| < \delta$, the algorithm stops, otherwise it goes back to step 2, iteratively updating the cluster centers as well as the

membership values for the elements (Ghosh & Dubey, 2013).

5. PCA method

Principal Component Analysis or PCA is one of the finest methods used for efficient analysis of multi-dimensional data. It was first developed by Pearson in 1901 and improved independently by Hotelling in 1933. Like many other multivariate methods, the PCA method was also not widely accepted or used until the advent of computers, but is now integrated almost to all statistical software packages and widely applied. PCA is the general name for a method that uses mathematical principles to convert a number of correlated variables into a smaller number of variables called principal components (Mishra et al., 2017).

PCA is used to reduce the attributes of a large dataset by retaining most of the information. The application of the PCA method speeds up the algorithms execution and facilitates data visualization.

6. Experiments

This section presents the results of the experiment for conducting a comparative analysis of k -means and fuzzy c -means clustering algorithms on demographic data of different sizes. Python programming language (Python3.8) is used to perform the experiments. Experiments are conducted on five different sizes of data. Experiments were conducted on demographic datasets "Voter Registration", "Economic Community of Central African States Statistics, 2013", "Income Inequality", "Educational Attainment" and "CBSA to Zip Crosswalk 2012" (dataworld.io). Table 1 illustrates the characteristics of the dataset.

Table 1. Characteristics of the dataset

Name	Number of objects	Number of attributes
Voter Reg.	96531	8
Economic.	2809	21
Income	4789	9
Educational.	166663	8
CBSA	165509	6

During the experiment, the PCA method is used to make k -means and fuzzy c -means

clustering algorithms more effective. The PCA method reduces the data size, makes algorithms run faster, and at the same time, minimizes the data loss. The number of attributes in each dataset is reduced to 2 through the PCA method. Experiments are performed on all datasets to check the results of k -means and fuzzy c -means algorithms when the PCA method is applied and when the PCA method is not applied. The experiment results are comparatively analyzed using different evaluation indices.

Silhouette, Davies-Bouldin, Calinski-Harabasz indices and the algorithm execution time are used to evaluate the efficiency of the proposed method and used algorithms (Mamat et al, 2018; Wijaya et al, 2021; Wang & Xu, 2019). The Silhouette index takes a value in the range of $[-1,1]$, with -1 as the worst case and 1 as the best case for clustering. Moreover, the Calinski-Harabasz index provides good performance in its higher values. Lower value of the Davies-Bouldin index and the algorithm execution time are considered good for clustering.

As Table 2 shows, during experiments with the fuzzy c -means algorithm on the "Voter Registration" dataset, when the PCA method is applied, Silhouette, Davies-Bouldin and the algorithm execution time provide good performance in all cluster values, whereas, the Calinski-Harabasz evaluation index performs better only in 7 out of 9 indicators. When the PCA method is applied during the experiments using the k -means algorithm on the same dataset, Calinski-Harabasz and Davies-Bouldin indices show better performance in all cluster values, and Silhouette only in 7 out of 9 indicators. However, the algorithm execution time performs worse in all cluster values when the PCA method is applied.

During experiments with the fuzzy c -means algorithm on the dataset "Economic community of central African states statistics, 2013", when the PCA method is applied, Calinski-Harabasz, Davis-Buldin and the algorithm execution time perform high performance in all cluster values, and the Silhouette evaluation index performs higher performance only in 7 out of 9 indicators. Calinski-Harabasz, Davis-Buldin indices and the algorithm execution time perform well in all cluster values, and Silhouette only in 6 out of 9 indicators when the PCA method is applied

during experiments using the k-means algorithm.

During the experiments with the fuzzy c-means algorithm on the dataset “Income inequality”, when applying the PCA method, Silhouette, Calinski-Harabasz and Davies-Bouldin show better performance in all cluster values, and the algorithm execution time only in 6 out of 9 indicators. When applying the PCA method during experiments using the k-means algorithm on the same dataset, the Silhouette, Calinski-Harabasz, and Davies-Bouldin indices show better performance in all cluster values, and the algorithm execution time only in 7 out of 9 indicators.

During experiments with the fuzzy c-means algorithm on the dataset “Educational attainment”, when applying the PCA method, Silhouette, Calinski-Harabasz and Davies-Bouldin show better performance in all cluster values, and the algorithm execution time only in 6 out of 9 indicators. When the PCA method is applied during the experiments through the k-means algorithm, the Silhouette, Calinski-Harabasz and Davies-Bouldin indices show better performance in all cluster values, and the algorithm execution time only in 7 out of 9 indicators.

During experiments with the fuzzy c-means algorithm on the dataset “Cbsa to zip crosswalk 2012”, when applying the PCA method, Silhouette, Calinski-Harabasz and Davies-Bouldin show better performance in all cluster values, and the algorithm execution time only in 4 out of 9 indicators. When the PCA method is applied during the experiments through the k-means algorithm, the Silhouette, Calinski-Harabasz and Davies-Bouldin indices show

better performance in all cluster values, and the algorithm execution time only in 7 out of 9 indicators.

5 datasets are used during the experiment. Experiments are conducted on each dataset using the fuzzy c-means and k-means algorithms applying and not applying the PCA method. Experiments are carried out for values of the number of clusters (k) from 2 to 10. Both k-means and fuzzy c-means algorithms are evaluated using Silhouette, Calinski-Harabasz, Davies-Bouldin, and time indicators. Thus, a total of 45 results are obtained for each indicator in order to compare the effectiveness of the PCA method application when taking into account the 5 datasets and 9 clusters for each algorithm.

During the experiments through the fuzzy c-means algorithm, when the PCA method is applied, the Silhouette index shows better performance in 43 out of 45 values, and in 40 out of 45 values though the k-means algorithm.

During the experiments through the fuzzy c-means algorithm, when the PCA method is applied, the Calinski-Harabasz index shows better performance in 43 of the 45 values, and in all 45 values though the k-means algorithm.

Davies-Bouldin index shows better performance in all 45 values when the PCA method is applied during experiments using both fuzzy c-means and k-means algorithm.

During the experiments through the fuzzy c-means algorithm, when the PCA method is applied, the algorithm execution time shows better performance in 34 out of 45 values, and in 30 out of 45 values through the k-means algorithm.

Table 2. Results of fuzzy c-means and k-means algorithms

VOTER REGISTRATION (96531*8)									
Fuzzy c-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.7667	79403.25	0.3515	109.66	2	0.8323	114148.64	0.2362	106.40
3	0.6750	46020.36	0.9743	105.24	3	0.7508	68848.23	0.5588	105.04
4	0.6987	49811.10	0.7730	108.92	4	0.7674	53266.38	0.5405	107.11
5	0.6492	38897.30	0.9262	113.25	5	0.7296	41339.43	0.6626	108.97
6	0.6621	35959.77	0.8831	114.39	6	0.7124	33729.10	0.7228	113.94
7	0.6008	26876.41	1.2404	117.83	7	0.7081	29013.56	0.6990	113.48
8	0.6126	26308.67	1.1745	138.12	8	0.6960	25040.89	0.7789	128.32
9	0.6223	24242.28	1.1353	127.89	9	0.7101	103423.27	0.6321	122.46
10	0.6345	21948.36	1.0785	127.54	10	0.7105	94173.18	0.6322	120.31

k-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.7674	79429.75	0.3485	105.19	2	0.8325	114151.39	0.2357	105.47
3	0.7719	96490.95	0.3356	105.59	3	0.8372	198963.95	0.3094	105.86
4	0.7957	120112.08	0.5228	108.06	4	0.7566	184040.94	0.5115	127.67
5	0.7977	115325.77	0.5279	105.48	5	0.7598	212258.48	0.5173	110.51
6	0.7034	120246.09	0.6437	105.87	6	0.7769	269975.81	0.4786	110.19
7	0.7147	138177.51	0.6484	107.66	7	0.7810	291837.67	0.4845	109.25
8	0.7177	135204.34	0.6367	106.85	8	0.7814	327419.19	0.4515	117.66
9	0.6782	131273.29	0.6925	114.22	9	0.7509	319980.09	0.5131	123.38
10	0.6845	137153.20	0.7229	109.34	10	0.7508	339776.76	0.5133	133.31
ECONOMIC COMMUNITY OF CENTRAL AFRICAN STATES STATISTICS, 2013 (2809*21)									
Fuzzy c-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.9804	4612.86	0.5836	0.6844	2	0.9813	5254.93	0.5190	0.1845
3	0.9762	6163.90	0.6444	0.2650	3	0.9778	8045.76	0.5215	0.1904
4	0.9695	5321.80	0.8220	0.4741	4	0.9714	7447.87	0.6486	0.2963
5	0.9677	6123.24	0.8175	0.6330	5	0.9703	9746.42	0.6379	0.2897
6	0.9641	5701.88	0.9583	0.6460	6	0.9662	9837.91	0.6850	0.4575
7	0.9624	5501.16	0.9102	1.1435	7	0.9589	9535.98	0.6751	0.5635
8	0.9558	5342.97	0.9189	1.2202	8	0.9468	9021.36	0.6912	0.4480
9	0.9411	4160.10	0.9572	1.2228	9	0.9465	8369.48	0.6632	0.4594
10	0.9403	4539.19	1.0746	1.4222	10	0.9464	7885.83	0.7471	0.7569
k-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.9815	4730.41	0.5612	0.3660	2	0.9820	5379.03	0.4996	0.2938
3	0.9757	6095.51	0.6599	0.2942	3	0.9772	7975.66	0.5402	0.2832
4	0.9675	5213.86	0.7835	0.3332	4	0.9701	7279.92	0.6139	0.3161
5	0.9604	4343.46	0.9344	0.4338	5	0.9643	6297.72	0.6979	0.4253
6	0.9406	3597.18	0.9434	0.4268	6	0.9463	5284.55	0.7061	0.3802
7	0.9234	3014.27	0.9819	0.4885	7	0.9305	4438.73	0.7388	0.4458
8	0.8936	2591.05	0.9996	0.5270	8	0.8911	3816.90	0.7326	0.4672
9	0.8761	2278.69	1.1344	0.6228	9	0.8519	3344.73	0.7977	0.5322
10	0.8384	2026.85	1.1349	0.7094	10	0.6151	2973.13	0.8216	0.5680
INCOME INEQUALITY (4789*9)									
Fuzzy c-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.7680	5686.90	0.3734	0.5056	2	0.8853	10394.18	0.1651	0.5261
3	0.6581	3380.84	0.9785	0.5574	3	0.8363	5880.21	0.6065	0.5411
4	0.7043	2701.89	0.8925	0.6254	4	0.7149	4095.54	0.6840	0.5885
5	0.6581	2110.48	0.9641	0.7032	5	0.7095	3120.99	0.7166	0.8027
6	0.6407	1721.93	1.0556	1.0831	6	0.7065	2532.23	0.9731	0.7929
7	0.6268	1444.59	1.3136	1.0719	7	0.6838	2120.91	0.9501	1.0890
8	0.6354	1311.41	1.3080	1.4869	8	0.6814	1830.00	1.1108	1.1081
9	0.6108	1137.45	1.4586	1.7222	9	0.6763	1606.98	1.1324	1.2683
10	0.5991	1033.82	1.4870	1.7268	10	0.6658	1429.45	1.1172	1.4992
k-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.7684	5687.86	0.3703	0.5042	2	0.8855	10394.63	0.1644	0.5076
3	0.7686	3432.59	0.9895	0.5189	3	0.8535	5904.10	0.5852	0.5335
4	0.7465	3521.41	1.0914	0.6103	4	0.8416	76328.99	0.3244	0.5525
5	0.7103	6415.76	0.6743	0.5914	5	0.7240	94186.85	0.4703	0.5674
6	0.7203	12069.67	0.5616	0.6627	6	0.7282	99618.66	0.5186	0.5864
7	0.7013	11707.54	0.6462	0.7031	7	0.7200	108551.30	0.5713	0.6642
8	0.6829	11128.35	0.7519	0.7587	8	0.7039	110052.32	0.6227	0.6292
9	0.6879	11599.76	0.7202	0.8348	9	0.6986	111084.80	0.6562	0.7141
10	0.6561	13009.63	0.7447	0.8955	10	0.6954	111764.51	0.6831	0.7949
EDUCATIONAL ATTAINMENT (166663*8)									
Fuzzy c-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.4476	56004.39	0.9885	323.65	2	0.6304	121059.67	0.5588	319.58
3	0.3333	37056.26	1.3710	309.22	3	0.5664	86447.04	0.5846	311.25
4	0.2680	27424.67	1.7084	310.75	4	0.5254	65741.20	0.6003	310.90
5	0.2135	21423.24	2.1647	309.05	5	0.5175	52766.15	0.6249	309.06

6	0.2363	23287.98	1.4911	317.44	6	0.5043	43719.80	0.6562	308.88
7	0.2012	19659.62	1.7451	320.41	7	0.4829	37170.87	0.6872	312.30
8	0.2241	18692.38	1.3863	323.31	8	0.4749	32394.23	0.6887	310.91
9	0.1991	16559.45	1.5567	327.72	9	0.4722	28691.13	0.7114	312.58
10	0.2204	16721.53	1.3282	337.76	10	0.4608	25693.27	0.7420	315.29
k-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.4557	56315.47	0.9698	319.39	2	0.6311	121074.36	0.5594	316.35
3	0.4767	43593.98	1.1312	317.07	3	0.5710	86590.43	0.5914	307.84
4	0.3604	36609.96	1.1283	307.43	4	0.5257	65803.68	0.6134	305.01
5	0.3886	33304.33	1.0224	307.39	5	0.5313	169461.96	0.5477	307.39
6	0.3838	47042.58	0.9993	307.38	6	0.5242	151334.21	0.5629	306.54
7	0.3427	43114.64	1.0614	305.25	7	0.5107	133795.23	0.5893	302.12
8	0.4076	42394.48	0.9118	307.71	8	0.4903	118712.17	0.6301	300.40
9	0.3625	42032.53	0.8999	305.57	9	0.4826	107123.18	0.6109	299.99
10	0.3749	43647.65	0.8543	308.47	10	0.4871	178513.36	0.6056	324.95
CBSA TO ZIP CROSSWALK 2012 (165509*6)									
Fuzzy c-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.4730	191313.48	0.8654	308.31	2	0.5723	306148.14	0.6468	309.33
3	0.3405	131455.55	1.2219	312.17	3	0.4551	236448.42	0.8430	312.61
4	0.2718	112708.70	1.4316	311.68	4	0.3998	236149.42	0.9560	313.09
5	0.2760	97895.42	1.4007	312.74	5	0.4407	241893.00	0.7967	311.42
6	0.2958	95980.99	1.1999	323.76	6	0.4614	251099.19	0.7178	313.63
7	0.3144	95156.23	1.1771	324.69	7	0.4910	281029.57	0.6694	354.35
8	0.3281	95019.99	1.0442	333.69	8	0.4862	273841.77	0.7579	319.92
9	0.3432	95326.15	1.0481	320.99	9	0.4300	255141.78	0.8378	328.84
10	0.3372	89179.35	1.2186	339.74	10	0.4543	284408.76	0.7375	327.94
k-means									
PCA method not used					PCA method used				
Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)	Cluster Number	Silhouette	Calinski-Harabasz	Davies-Bouldin	Time (sec.)
2	0.4735	191358.71	0.8636	308.32	2	0.5725	306171.31	0.6460	307.01
3	0.3457	133002.31	1.1954	310.19	3	0.4683	236744.80	0.8033	308.36
4	0.2865	115761.40	1.4015	308.28	4	0.3974	236892.17	0.9572	307.38
5	0.3044	106494.23	1.2178	308.33	5	0.4399	244491.62	0.7588	304.21
6	0.3133	100427.22	1.1099	310.62	6	0.4629	255586.77	0.7273	304.63
7	0.3342	98201.90	1.0452	305.83	7	0.4920	283283.47	0.6784	306.46
8	0.3399	96650.46	1.0175	306.28	8	0.5032	294908.29	0.6930	306.59
9	0.3448	92222.11	1.0824	308.23	9	0.5097	299517.67	0.6446	303.25
10	0.3543	93525.29	1.0520	305.67	10	0.5224	312876.69	0.6076	304.18

7. Conclusion

This article examined the advantages and disadvantages of data mining, demography, k-means, and fuzzy c-means algorithms. Fuzzy c-means and k-means algorithms were used to perform the experiment of different volumes of demographic data. Through the PCA method, the efficiency of these algorithms increased, the quality of the clustering results improved and a comparative analysis were performed.

One of the main goals of the PCA method is to save time by reducing the number of attributes. In the dataset “Economic community of central African states statistics, 2013”, which included the largest attribute in the experimented data, when the PCA method was applied with both the fuzzy c-means and

k-means algorithm, it showed good performance at all values of the algorithm execution time, that is, less time was spent. An analogous result was not observed in other datasets.

During experiments, the PCA method showed better performance in 165 out of 180 values in all indicators with the fuzzy c-means algorithm, and in 160 out of 180 values with the k-means algorithm. It can be concluded from the results of the experiment that if all indicators are taken into account the PCA method is applied, the fuzzy c-means algorithm showed better performance than the k-means algorithm.

References

- Ahmadov, E. (2021). Intelligent analysis of demographic data using K-means clustering algorithm (in Azerbaijani). 2nd International Science and Engineering Conference With The Joint Organization By The Ministry Of Education Azerbaijan Republic. 369-371. https://beu.edu.az/root_panel/upload/files/beu.edu.az/documents/Engineering_Book_2021%20%281%29.pdf
- Alguliyev, R. M. and Yusifov, F. F. (2021). Architectural principles of creating a national e-demography system (in Azerbaijani). Information Society Problems, 1, 3-17. [10.25045/jpis.v12.i1.01](https://www.ijsr.net/archive/v9i5/SR20507222308.pdf)
- Chattopadhyay, S., Pratihari, D. K., Sarkar, S. (2011). A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms. Computing and Informatics, 30, 701-720. https://www.researchgate.net/publication/285788940_A_comparative_study_of_fuzzy_c-means_algorithm_and_entropy-based_fuzzy_clustering_algorithms
- Ghosh, S. & Dubey, S. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4, 35-39. <https://dx.doi.org/10.14569/IJACSA.2013.040406>
- Grover, N. (2014). A study of various fuzzy clustering algorithm. International Journal of Engineering Research, 3, 177-181. [10.17950/ijer/v3s3/310](https://www.researchgate.net/publication/285788940_A_comparative_study_of_fuzzy_c-means_algorithm_and_entropy-based_fuzzy_clustering_algorithms)
- Jain, A., Duin, R., Mao, J. (2000). Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 1, 4-37. [10.1109/34.824819](https://dx.doi.org/10.1109/34.824819)
- Mamat, R., Mohamed, F. S., Mohamed, M., A., Rawi, N., M., Awang, M. I. (2018). Silhouette index for determining optimal k-means clustering on images in different color models. International Journal of Engineering & Technology, 7, 105-109 <https://www.sciencepubco.com/index.php/ijet/article/view/11464>
- Mishra, S., Sarkar, U., Traphder, S., Datta, S., Swain, D. P., Saikhom, R., Panda, S., Laishram, M. (2017). Multivariate statistical data analysis-principal component analysis (pca). International Journal of Livestock Research, 7, 60-78. <https://www.semantic scholar.org/paper/Multivariate-Statistical-Data-Analysis-Principal-Mishra-Sarkar/3ad314f33dbdf486999f521ed3ba061006a2d2b2>
- Müllensiefen, D., Hennig, C., Howells H. (2017). Using clustering of rankings to explain brand preferences with personality and sociodemographic variables. Journal of Applied Statistics, 45, 1-21. [10.1080/02664763.2017.1339025](https://doi.org/10.1080/02664763.2017.1339025)
- Oti, E. U., Olusola, M. O., Eze, F. C., Enogwe, S. U. (2021). Comprehensive review of k-means clustering algorithms," International Journal of Advances in Scientific Research and Engineering, 7, 64-69. <https://dx.doi.org/10.31695/IJASRE.2021.34050>
- Sharma, R. D. (2020). Python tools for big data analytics. International Journal of Science and Research (IJSR), 9, 597-602. <https://www.ijsr.net/archive/v9i5/SR20507222308.pdf>
- Suganya, R. & Shanthi, R. (2012). Fuzzy c-means algorithm – a review. International Journal of Scientific and Research Publications, 2, 1-3. <https://www.ijsrp.org/research-paper-1112.php?rp=P11381>
- Wang, X. & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. IOP Conference Series: Material Science and Engineering, 569, 1-6. https://www.researchgate.net/publication/335081976_An_improved_index_for_clustering_validation_based_on_Silhouette_index_and_Calinski-Harabasz_index
- Wijaya, Y. A., Kurniady, D. A., Setyanto, E., Tarihoran, W. S., Rusmana, D., Rahim, R. (2021). Davies Bouldin index algorithm for optimizing clustering case studies mapping school facilities. TEM Journal, 10, 1099-1103. <https://docplayer.net/219118152-Davies-bouldin-index-algorithm-for-optimizing-clustering-case-studies-mapping-school-facilities.html>
- Yong, Y., Chongxun, Z., Pan, L. (2004). A novel fuzzy c-means clustering algorithm for image thresholding. Measurement Science Review, 4, 11-19. <https://www.semanticscholar.org/paper/A-Novel-Fuzzy-C-Means-Clustering-Algorithm-for-Yong-Chong-un/ebbf7e8b7a1ea133999561ab279e51b961d31>