Article 6

Comparison of Clustering Algorithms on Air Quality Substances in Peninsular Malaysia

Sitti Sufiah Atirah Rosly, Balkiah Moktar, Muhamad Hasbullah Mohd Razali
Faculty of Computer & Mathematical Sciences,
Universiti Teknologi MARA Perlis Branch, Malaysia

Abstract
Air quality is one of the most popular environmental problems in this globalization era. Air pollution is the poisonous air that comes from car emissions, smog, open burning, chemicals from factories and other particles and gases. This harmful air can give adverse effects to human health and the environment. In order to provide information which areas are better for the residents in Malaysia, cluster analysis is used to determine the areas that can be clustering together based on their air quality through several air quality substances. Monthly data from 37 monitoring stations in Peninsular Malaysia from the year 2013 to 2015 were used in this study. K-Means (KM) clustering algorithm, Expectation Maximization (EM) clustering algorithm and Density Based (DB) clustering algorithm have been chosen as the techniques to analyze the cluster analysis by utilizing the Waikato Environment for Knowledge Analysis (WEKA) tools. Results show that K-means clustering algorithm is the best method among other algorithms due to its simplicity and time taken to build the model. The output of K-means clustering algorithm shows that it can cluster the area into two clusters, namely as cluster 0 and cluster 1. Clusters 0 consist of 16 monitoring stations and cluster 1 consists of 36 monitoring stations in Peninsular Malaysia.

Keywords: *Air Quality, Clustering Algorithm, WEKA, k-mean, density based, expectation maximization*

Introduction

Nowadays, a severe environmental problem such as air pollution has attracted much attention in many developed countries. People who are exposed to unhealthy air quality can suffer from short-term throat, eye and nose irritation. Moreover, a citizen who has lung and heart disease, children and senior citizen will be at dangerous risks because of air pollution. When the air is polluted, the population that is exposed to this air will experience increasingly adverse health effects. Different areas have their own air pollutant index which some of these areas will experience same environmental problems

Cluster analysis divides data into groups that are meaningful and useful. In other words, cluster analysis is the main method of data description in assorted sectors like image analysis, data mining, pattern recognition, machine learning, and bioinformatics. Cluster analysis is also recognized as an important method for classifying data and discovering clusters of a dataset based on similarities in the same cluster and dissimilarities between different clusters.

The aim of this study was to determine the areas that can be cluster together based on several air pollution substances. Three clustering algorithm namely as Expectation Maximization (EM), Density Based (DB) and *k*-mean were applied on the training dataset.

Related Works

i. Air Quality Analysis

Air is one of the most important elements for the existence of life on this world. However, an environmental problem such as air pollution could give negative impacts on human health and environment. Atmospheric pollution affects all societies; no matter the level of socioeconomic development they have where finally a large impact on human health (Cortina–Januchs et al., 2015). Generally, the level of air pollutions at industry and urban areas are increasing because of pollutant emission from factories, residential areas, and transportation into the atmosphere. Bishoi et al. (2009) stated that air pollution is a well-known environmental issue linked with ur-ban areas around the world. Saddek et al. (2014) shows that Fuzzy Inference System (FIS) can be a useful method for health effect identification and development of early warning systems to curb the nature of risk disease according to the Air Quality Index (AQI). Recent study by Ehsanzadeh et al. (2015), proved that Classification and Regression Tree algorithm (CART) method can be used to make decision and solve problem of air quality management better. Their main objective in this study is to imitate hourly air quality index through CART method using air pollutants and meteorological variables. They obtained the data of measured variables that are based on hourly slots throughout the year of 2007 and 2008 at the Gholhak monitoring station.

ii. Clustering Algorithms

A study conducted by Ghosh and Dubey (2013) found that Fuzzy C-Means clustering produces close results to K-Means clustering but it still needs more calculation time than K-Means clustering. Similarly to another study conducted by Bora and Gupta (2014) reported that K-Mean performance is better than Fuzzy C-Mean performance regarding computational time. They also conclude that K-Means algorithm is suitable for exclusive clustering task; meanwhile, Fuzzy C-Means is suitable for overlapping clustering task.

Whilst, Cibeci and Yildiz (2015) conducted a study to compare the competence of K-Means and Fuzzy C-Means algorithms on synthetically created datasets consisting of differently shaped clusters scattering with regular and non-regular patterns in two-dimensional space. As a nutshell, they conclude that there is no any algorithm, which is the best for all cases. Therefore, in order to determine for a suitable algorithm, the datasets should be carefully analysed for shapes and scatter of clusters.

In addition, a study conducted by Saithan and Mekparyup (2012), used cluster analysis as an approach to classify numerous variables that are present in the air and to determine the pattern of ozone in Thailand. The data of air quality are from year the 2006 to 2010 and has been obtained from two monitoring stations, which are General Education Centre, Mueang District, Chonburi and Map Ta Phut Health Office, Mueang District, Rayong. The result of this study clarifies that four clusters from variables in the air which are air quality variables as cluster 1, pressure as cluster 2, wind speed, temperature and sun radiation as cluster 3 and last cluster are wind direction, relative humidity, and rain. They also demonstrate that there are three different clusters based on the time of the day.

Methodology

i. Data Collection

The data for this study obtained from Department of Environment collected from 37 monitoring sites in Peninsular Malaysia during 2013 to 2015 summarized in Figure 1. Five types of air quality substances namely carbon monoxide (*CO*), ozone (*O3*), sulphur dioxide (*SO2*), nitrogen dioxide (*NO2*) and suspended particulate matter of less than l0 microns in size (*PMl0*) which believed contributed to the pollutants were used in this study.

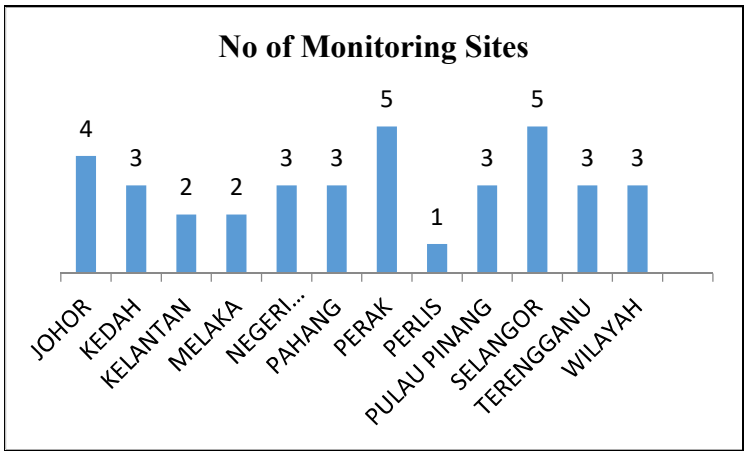

Figure 1: Total Monitoring Site

ii. Expectation Maximization (EM) Clustering Algorithm

Steps:

1. Initialization:
   Randomly select initial parameters of k distributions.
2. Iteration:
   E-step:
   i. Compute the $P(Ci|x)$ for all objects x by using the current parameters of the distributions.
   ii. Re-label all objects according to the computed probabilities.
   M-step:
   i. Re-estimate the parameters of the distribution to maximize the likelihood of the objects assuming their current labelling.
3. Stopping Criteria:
   At convergence-when the change in log-likelihood after each iteration becomes small.
4. Repeat:
   Repeat the step 2 until the stopping condition occurs.

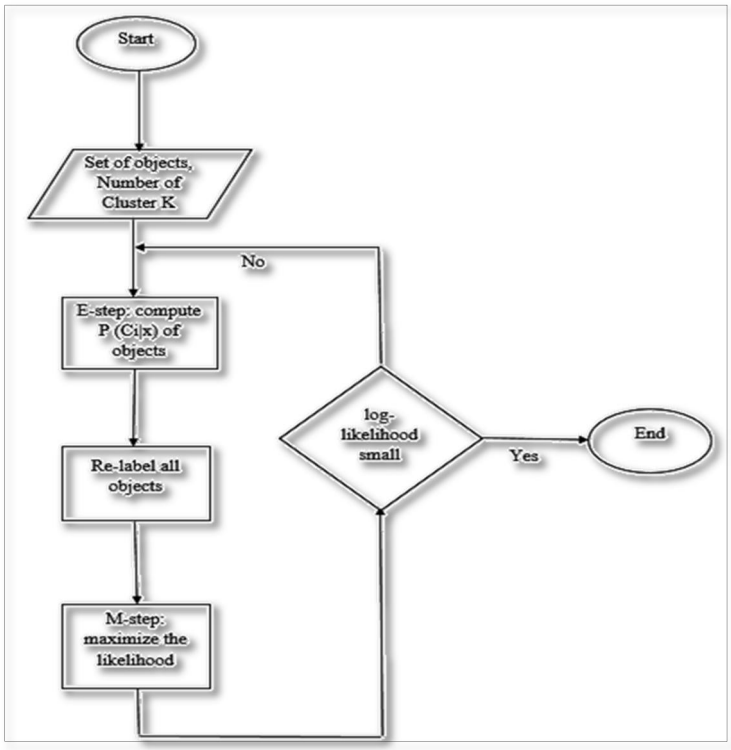

Figure 2: Flowchart for EM Algorithm Process

iii. Density Based (DB)Clustering Algorithm

In density-based clustering algorithms, dense areas of objects in the data space are considered as clusters, which are separated by low-density area (noise). Therefore, density-based is an impressive basic clustering algorithm for data streams (Karrar and Mutasim, 2016).

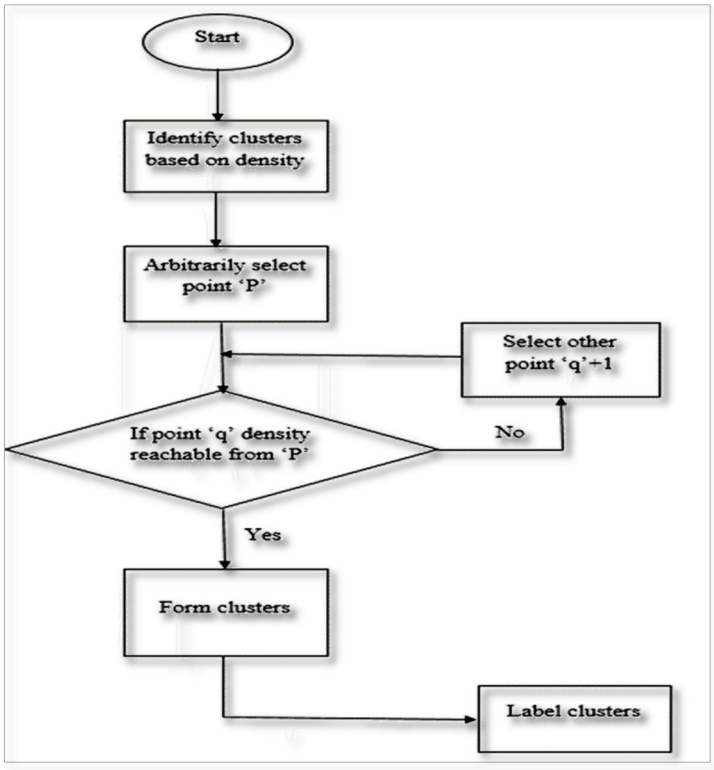

Figure 3: Flowchart for Density-Based Algorithm Process

iv. K-means (KM) Clustering Algorithm

Step 1: Decide the number of clusters *k*

Assume that there are two variables of dataset. Each variable will have n data points $x_i, i = 1,2,..., n$ that have to be divided into parts in *k* clusters. The number of clusters should fit the data. An inaccurate selection of the number of clusters will invalidate the entire process.

Step 2: Initialize the center of the clusters

Initialize the data points that have been clustered and calculate the mean of that cluster to choose *k* starting points, which are used as initial estimates of the cluster centroids. They are taken as the initial starting values $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ randomly.

Step 3: Attribute the closest cluster to each data point

Consider each point in the dataset and assign it to the cluster, which its centroid is nearest to it. Assign each point by calculating the distance between data point and initial cluster centroid by using square Euclidean distance measure. Based on square Euclidean distance, each data point is assigned to one of the clusters, which are based on minimum distance. For square Euclidean distance is calculated as:

$$d_{ik} = \sum_{j=1}^{n} (x_{ij} - x_{kj})^2, \quad (1)$$

Where d_{ik} is the distance between variables x_{ij} and x_{kj} and j is the number of variables which are $j = 1,2,...,n$

Step 4: Recalculate cluster centers by finding mean of data points belonging to the same cluster

When each point in dataset has been assigned to its cluster, recalculate the new positions of the cluster centroids. Assign new position of the cluster centroids based on minimum Euclidean distance and update them to the same cluster.

Step 5: Repeat steps 3 and 4 until all the data points are convergence

The steps of 3 and 4 needed to be repeated until the cluster centroids no longer move and not change.

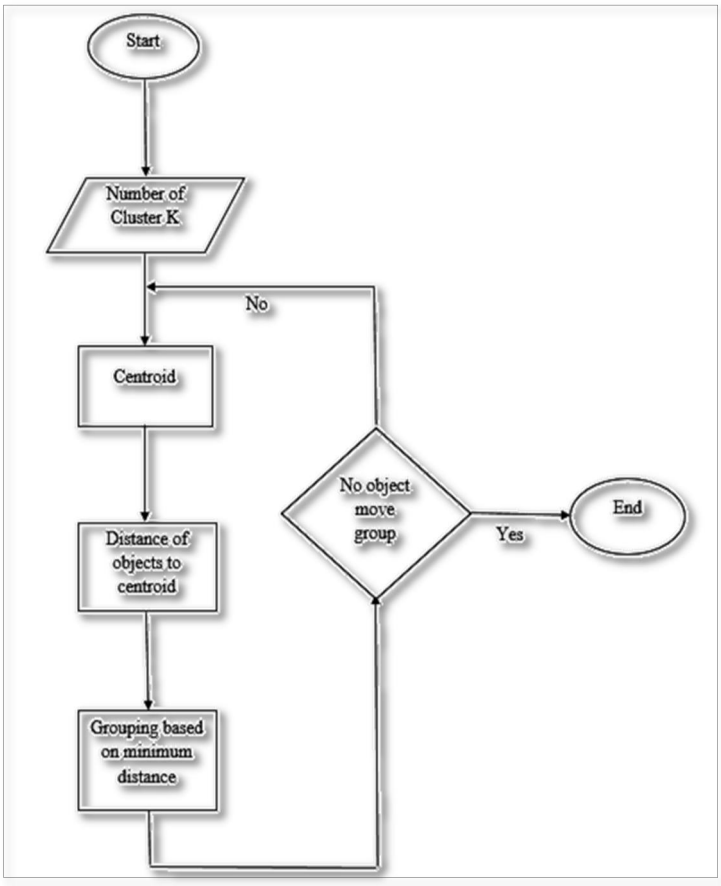

Figure 4: Flowchart for K-Means Algorithm Process

v. Knowledge Flow using WEKA

The data pre-processing and data mining was performed using the WEKA Data Mining tool. Figure 5 shows the data mining process to perform the cluster analysis. The process is started by loading the dataset using *CSV Loader*. The dataset was then split by default into 66% of training and 34% of test using *Train Test Split Maker*. Three cluster algorithm mentions previously were then applied. The performance for each algorithm were evaluate using *Clusterer Performance Evaluator* and the result will be view using *Text Viewer*

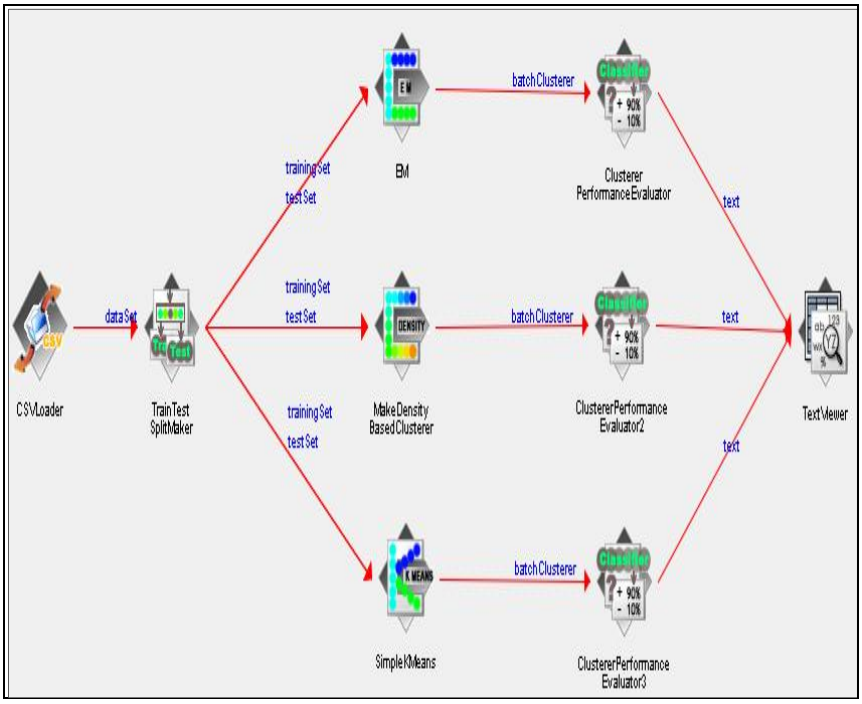

Figure5: Building Cluster using WEKA

Results

The summary of the analysis was presented in table 1. The output was summarized based on the model and evaluation of the test split. *K-means* and DB algorithm is much better than EM algorithm in time to build the model. Using log- likelihood value, DB Clusters have a negative value that does not show its perfection in results.

Every algorithm has their own importance and we use them on the behaviour of the data, but based on this study we found that *k*-means clustering algorithm is the simplest algorithm as compared to other algorithms due to its simplicity and shortest time to build the model. Figure 6 to 8 visualized the cluster formed by the algorithm. For the purposes of illustration, only two attributes were chosen, which is ozone (*O3*) versus nitrogen dioxide (*NO2*).

Table 1: Comparisons of algorithm

Algorithm	No of cluster	Log likelihood	Clustered instances					Time to build model
			0	1	2	3	4	
EM	5	2.2593	102 (23%)	90 (20%)	100 (22%)	114 (25%)	47 (10%)	12.29 second
DB	2	-0.1331	309 (68%)	144 (32%)				0.02 second
k-means	2		304 (67%)	149 (33%)				0.02 second

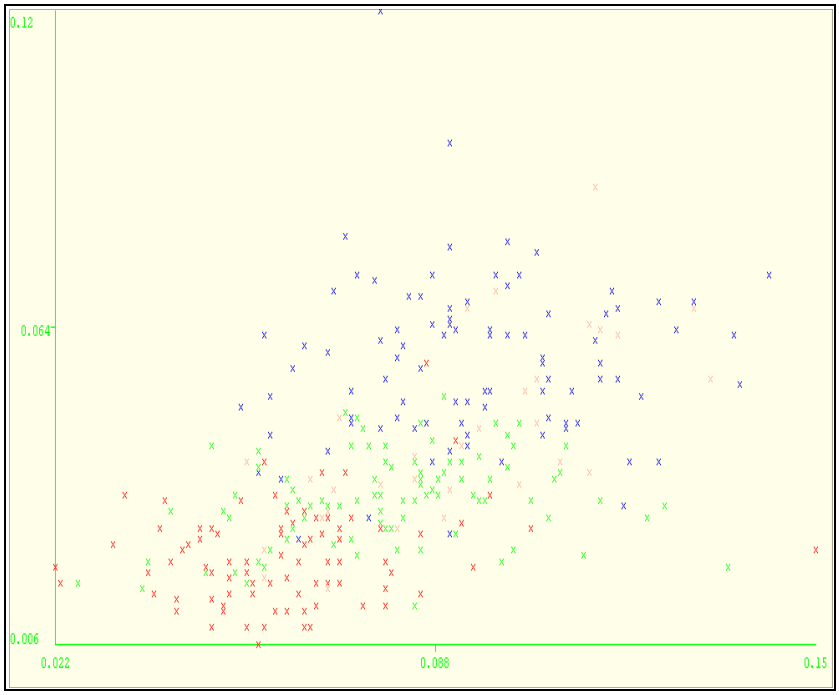

Figure 6. Selected Cluster Visualization using EM (O3 vs N02)

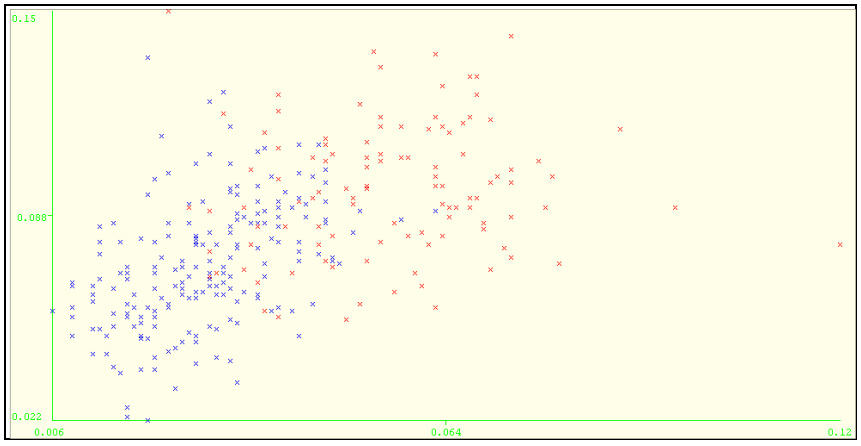

Figure 7:. Selected Cluster Visualization using DB (O3 vs NO2)

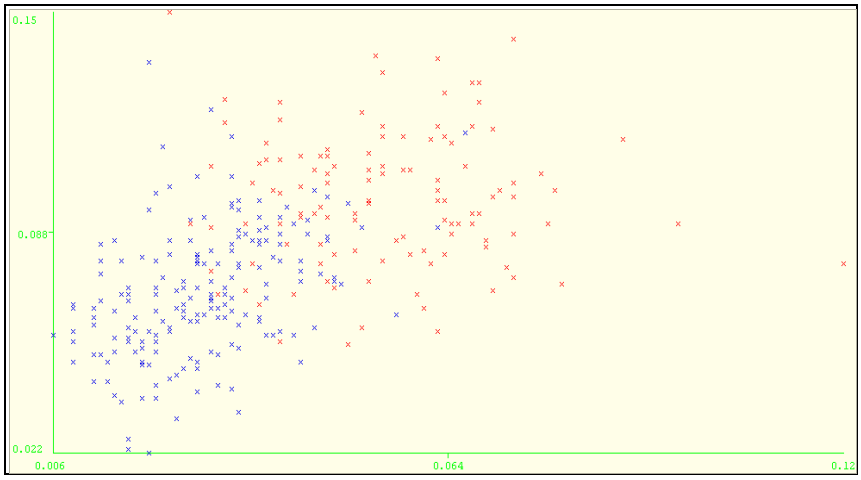

Figure 8: Selected Cluster Visualization using *k-mean* (O3 vs NO2)

Conclusion

As a conclusion, cluster analysis is able to cluster the area based on substances in air quality index. There are several algorithms for cluster analysis that can be applied to the dataset of air quality index. In this study, K-Means clustering algorithm, Expectation Maximization (EM) clustering algorithm and Density-based clustering algorithm have been applied in the dataset of air quality index using WEKA tools for cluster the area based on air quality's substances. The results of this study showed that the area that can be clustered together based on air quality through their fives substances. The output of K-means clustering algorithm shows that it can cluster the area into two clusters, which are cluster 0 and cluster 1. Cluster 0 has clustered 16 sites of the monitoring station and cluster 1 has clustered 36 sites of monitoring station in Peninsular Malaysia. From the WEKA's result, it can conclude that K-means clustering algorithm is the best method among other algorithms due to its simplicity and time took to build the model.

Acknowledgements

The authors would like to express our gratitude to Department of Environment for the data courtesy.

References

Bishoi, B., Prakash, A., & Jain, V. K. (2009).A comparative study of air quality index based on factor analysis and US-EPA methods for an urban environment. *Aerosol and Air Quality Research*, *9*(1), 1-17.

Saddek, B., Chahra, B., Wafa, B. C., &Souad, B. (2014). Air quality index and public health: modelling using Fuzzy Inference System. *American Journal of Environmental Engineering and Science,1*(4), 85-89.

Ehsanzadeh, A., Nejadkoorki, F., Talebi, A., &Bahrami, S. (2015). Simulating hourly air quality index using the classification and regression tree (CART).*International Conference on Architecture, Urbanism, Civil Engineering, Art, Environment. Future Horizons & Retrospect ICAUCAE 2015* (pp. 1-8). Tehran, Iran: Institute of Art and Architecture (SID).

Ghosh, S., & Dubey, K. S. (2013).Comparative analysis of K-Means and Fuzzy C-Means algorithms.*International Journal of Advanced Computer Science and Applications, 4*(4), 35-39.

Bora, D. J., & Gupta, A. K. (2014).A comparative study between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. *International Journal of Computer Trends and Technology (IJCTT), 10*(2), 108-113.

Cibeci, Z., &Yildiz, F. (2015).Comparison of K-Means and Fuzzy C-Means algorithms on different cluster structures.*Journal of Agricultural Informatic, 6*(3), 13-23.

Saithan, K., &Mekparyup, J. (2012). Clustering of air quality and meteorological variables associated with high ground ozone concentration in the industrial areas, at the east of Thailand. *International Journal of Pure and Applied Mathematics*, *81*(3), 505-515.

Karrar, A. E., &Mutasim, M. (2016). Comparing EM clustering algorithm with density based clustering algorithm using weka tool. *International Journal of Science and Research (IJSR), 5*(7), 1199-1201.