Journal of Theoretical and Applied Information Technology
30th April 2020. Vol.98. No 08

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENT CLUSTERING TECHNIQUES FOR URBAN AREA ANALYSIS BASED ON SATELLITE IMAGES

[1]A. TOBAL, [1]H. FAROUK, [2]S. MOKHTAR, and [2]H. ZIDAN

[1] Associate Prof., Electronics Research Institute, Computers and Systems, Cairo, Egypt

[2]Researcher. Electronics Research Institute, Computers and Systems, Cairo, Egypt

E-mail: {atobal, hesham, sahar, hbanna}@eri.sci.eg

ABSTRACT

Urban analysis provides the urban planners with the information about how to optimally utilize the land resources and the infrastructure. It gives solutions to many problems such as dense population, population growth, and limited resources. The Arab Republic of Egypt is one of the most densely populated countries in the region. In addition to the accumulation of the population in a specific area of the land, there is limited natural resources that directly affects the nature of life, agriculture and the population spreading. So, it is vital to study the changes in the geography of the land in order for the urban planners to set short- and long-term strategies to improve the quality of life and set a sustainable development plan. The Suez region was selected for such study. In this work, satellite images have been categorized into four segments: Desert, agriculture, residential and water using three clustering techniques to study the increment and decrement of urbanization, water resources, agricultural patch and dissertation. The three clustering techniques are; Fuzzy, Kohonen Neural Network and k-means. Each technique was applied on a low-quality resolution Google satellite images of Suez area across 16 years from 2001 to 2016. A comparison between each technique behavior on this image style and a ready-made program ArcGIS has been done. Astoundingly, the results show that the Fuzzy clustering is the best technique for such kind of images.

Keywords: *Clustering Techniques, Unsupervised Learning, Fuzzy Clustering, K-means Clustering, Kohonen Neural Network.*

1. INTRODUCTION

Due to the continuous successes of the space and earth explorations, satellites became more popular now. There are about 5000 operational satellites orbiting the earth according to United Nations Office for Outer Space Affairs (UNOOSA) [1]. The main purposes of the operational satellites are: Communications: 777 satellites. Earth observation: 710 satellites. Technology development/demonstration: 223 satellites. Navigation / Positioning: 137 satellites. Space science / observation: 85 satellites. Earth science: 25 satellites. Taking into consideration that there are many satellites can be multipurpose. Also due to the advances in the image processing and remote sensing technologies, satellite images are widespread used in different aspects of life. There are three main types of satellite images: Panchromatic (black and white), multispectral (RGB which stands for red, green, blue) and this is what have been used in the paper, and finally the hyperspectral images (images records hundreds of narrow band spectrum in order to cover the continuous spectrum of light instead of discrete bands).

Clustering is one of the data mining techniques which is mainly an unsupervised learning method [2]. The word 'clustering' means grouping similar things together.

Clustering methods are used to identify groups of similar objects in a multivariate data sets collected from fields such as marketing, bio-medical and geo-spatial. There are different types of clustering methods [3], including:

1- Partitioning methods
2- Hierarchical clustering
3- Fuzzy clustering
4- Density-based clustering
5- Model-based clustering

This work will apply three out of the five previous techniques in unsupervised manner. The

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

selected three techniques are: Partitioning methods, Fuzzy clustering, and Model-based clustering. The Hierarchical model will be excluded because this technique uses one of two to level methods, either agglomerative or divisive. The first is supervised and the second is similar to K-means which was applied in this study. The model-based clustering will be used as a testing model represented in "ArcGIS" ready-made software as a comparative system to this study results.

Using clustering in image segmentation is to classify the image into a finite number of clusters, where the number of clusters can be user defined or calculated using smart algorithm. In this process there is no training stages, but train themselves using the available data. Based on some criteria, the pixels are grouped together and form the cluster [3]. In this research, three types out of five are applied with their associated clustering technique which are tabulated in table 1.

Table 1: Methods and Techniques Used.

Method Used	Technique
Partitioning methods	K-means
Fuzzy clustering	Fuzzy
Model-based clustering	Kohonen Neural Network

Clustering techniques are applied to a set of satellite images for the special part in Egypt which is Suez Governorate. This part of Egypt has special nature because it includes different regions ranging from Agricultural Land, water, population, industry communities and desert. It is also a promising area for development [4]. Suez governorate is the castle of oil industry in Egypt. Many international oil companies are operating and have made massive discoveries. Thus, Suez has really become the oil capital of Egypt. Suez also has many tourism attraction sites. The governorate has great potentials that can enhance the capacities of the national economy, promote the competitiveness of the Egyptian production in the international market, attract Arab and foreign investments and create job opportunities. These are represented in the availability of natural resources, basic infrastructure and issuance of a special law for North West of Suez Gulf zone which gained the policy makers' concern to develop and market the zone internationally.

In this paper, the multispectral images were acquired from the satellite images via Google Earth [5] through 16 years, from year 2001 to 2016.

In this research, Clustering techniques are used to identify the nature of the land and the increase and decrease in the agricultural, residential, desert, and water areas in order to help decision makers in their strategic plans for this part of Egypt.

2. DATA MINING AND APPLICATIONS FOR SATELLITE IMAGES ANALYSIS AND CURRENT RESEARCH

Data Mining is to extract knowledge and discover patterns in large amounts of data using machine learning techniques. The two main machine learning techniques used in Data Mining is classification and clustering [6]. In classification training, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples which is called supervised learning. In clustering, groups of examples that belong together are sought. When there is no specified class, clustering is used to group items that seem to fall naturally together. The challenge is to find these clusters and assign the instances to them—and to be able to assign new instances to the generated clusters as well. This type of learning is called unsupervised learning [7].

The success of clustering is often measured subjectively in terms of how useful the result appears to be to a human user. It may be followed up by a second step of classification learning in which rules are learned that give an intelligible description of how new instances should be placed into the clusters. Clustering techniques apply when there is no class to be predicted but the instances are to be divided into natural groups. Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. Moreover, it is possible that there is no knowledge of how many groups to look for. This is what distinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict.

Clustering methods are based on measuring distances between records and between clusters. Records are assigned to clusters in a way that tends to minimize the distance between records belonging to the same cluster.

In this research, Clustering techniques are used to identify the nature of land and the decrease and increase in cultivation, building, desert, water areas in order to help decision makers in their strategic

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

plans for this part of Egypt. Among different clustering methods, K-means clustering, fuzzy clustering and Kohonen Neural Networks are used in this research [8].

In our research, the data which is extracted from the satellite image through 16 years from 2010 to 2016 is unclassified the given examples are unsupervised so clustering is the suitable learning to find concepts that describe different clusters in the map under examination. Among different clustering methods, K-means clustering and Kohonen Neural Networks are used in this research.

2.1 K-means Clustering

The classic clustering technique is called k-means. To begin the process, the number of clusters are being sought should be specified which is called the parameter k. Then K points are selected randomly as the Cluster centers. All objects in the input set are assigned to their nearest cluster center after calculating the ordinary Euclidean distance metric between each object and each cluster center. Next the centroid, or mean, of the objects in each cluster is calculated—this is the "means" part. These centroids are taken to be new center values for their corresponding clusters. Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain the same forever.

Once the iteration results in no changes in the cluster centers determination, each point is assigned to its closest cluster center, so the overall objective is to let the total squared distance from all points to their cluster centers as minimum as possible. But the minimum is a local one; there is no guarantee that it is the global minimum. The final clusters are quite sensitive to the initial cluster centers. Completely different groups are generated from small changes in the initial random choice. In fact, this is true of all practical clustering techniques: It is almost always infeasible to find globally optimal clusters.

The mathematical formula that K-means works with is:

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$\min arg \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x - \mu_i\|^2 = \min arg \sum_{i=1}^{k} |S_i| \, Var \, S_i \quad (1)$$

where μi is the mean of points in Si.

First, In K-means algorithm, a cluster center is determined based on the given data, then objects are assigned to clusters according to the distance calculated from each object to the cluster center, the object with the smallest distance from the cluster center is assigned to this cluster. The centers of clusters are recalculated and hence objects are rearranged iteratively. In each iteration, the quality of k-means is calculated and the process stops if there is no improvement in the quality. The K-means algorithm can be presented as follows:

1. The number of clusters k should be specified by the user
2. The variables with metric values are converted to have values between 0 and 1,
3. The centers of the k clusters are determined as follows:
   (a) The initial cluster center is taken from the values of the first instance in the data set.
   (b) Distances are determined from all instances to the centers of clusters that are calculated so far.
   (c) The instance values with the largest distance from the cluster centers is considered to be a new cluster center.
   (d) When the number of clusters defined by the user equals the number of clusters determined by the procedure, the process stops.
4. The distance between each instance in the data set and each center of a cluster is calculated as the squared Euclidean distance. The instance or the object is assigned to the cluster center with the minimal distance.
5. The average number of objects assigned to certain cluster is used as a measure to update the cluster centers.
6. If a maximum number of iterations done or there is no update in cluster centers rearrangement, the process stops. The user can define another threshold for the change that will stop the iterations.

2.2 Kohonen Neural Network

Kohonen neural networks are a type of neural network that perform unsupervised learning. It is also called a self-organizing map [9]. The self-organization map is created according to the training data. Similar objects are grouped together, the most similar objects are assigned to one class according to the learnt pattern in the map that is used to group similar objects. This process is done

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

automatically. The input layer includes the user defined input variables as input vector. The parameters in the output layer is adjusted after the learning process is completed and hence different patterns included in the input data are learnt and be ready for new input vectors to be clustered. This learning process is called unsupervised learning as there is no target variable is determined.

When new input vector is showed to the model after learning process, the output layer results in the neuron that represents the most similar cluster learnt.

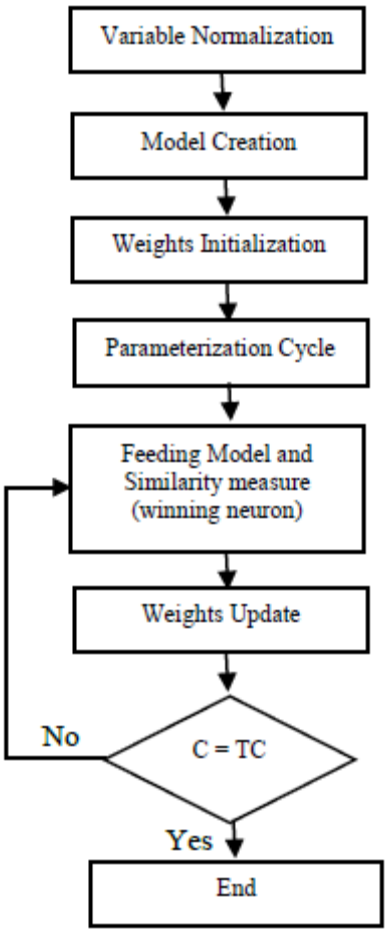


Figure 1: Kohonen Neural Network Operation Flowchart

2.3 Fuzzy Clustering

There are two types of Clustering: Overlapping clustering which is called soft clustering, and Exclusive Clustering which is called Hard clustering. Soft clustering uses fuzzy sets to cluster data in this technique each instance may be assigned to one or two clusters with different membership degrees. The membership value is associated with data presented to the model. Soft Clustering is considered to be more natural than hard clustering in some applications. In soft clustering technique, the instances on the borders among many clusters shouldn't be forced to belong to one cluster otherwise they are assigned to a membership degree between 0 and 1 which is called partial membership. In Hard Clustering, each instance is assigned to one cluster in exclusive way. Fuzzy C means (FCM) is one of the most popular soft clustering methods [10,11,12] while k-means is one of the most important methods of hard clustering. In fuzzy clustering, the membership is spread among all clusters so an individual object could be classified to more than one cluster, while in hard clustering techniques, if a certain instance belongs to a definite cluster then it could not be included in another cluster.

Some researches were published in the area of applying Datamining –specially clustering - techniques [13,14,15,16] on satellite images to discover information. Paper [17] author used K-means clustering technique to discover different clusters for sugarcane planting in Brazil using time series satellite images. It used K-means Dynamic Time wrapping time (DTW) series clustering to support agricultural monitoring as well as to automatically determine sugarcane fields' expansion using a clustering analysis with the Dynamic Time Warping distance function [17]. In [18] supervised classification is used to classify different satellite images using clustering ensemble based on different cluster techniques.

3. PROBLEM DESCRIPTION

Environmental changes, represented in climate changes, water poverty, desertification, farmland, and urbanization is a major concern for all countries worldwide. Egypt suffers from many of those challenges and the need for developing a forecasting model is a necessity not luxury anymore to be ready to solve the resulted effects and helps in applying a perfect planning across all vertical; urban communities; searching for new water resources; expanding farmland, etc.

Getting satellite images is a very efficient source for doing this forecasting with the merging of ICT as clustering techniques; to analyze the mobility of each segment of the area under study (the work is applied on four segments: Water, Desert, Agriculture and Residential Community.

Journal of Theoretical and Applied Information Technology
30th April 2020. Vol.98. No 08

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. THE PROPOSED MODEL

This work uses historical 16 satellite images for Suez Canal area across 16 years (2001 to 2016). Figure two shows three original images (years 2001, 2009, and 2016 selected) and the corresponding luminance version for clustering processing needs.

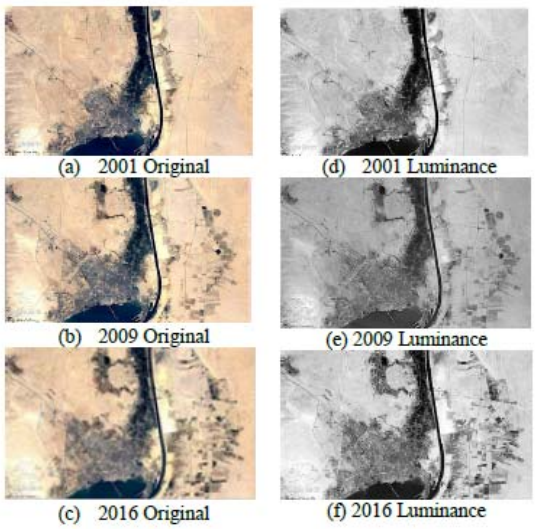


Figure 2: Original Satellite Images of years (a) 2001, (b)2009 and (c) 2016 for Suez Canal area and the corresponding luminance version for the same years (d) 2001, (e)2009, (f) 2016

Figure three represents the exact working areas in the study.

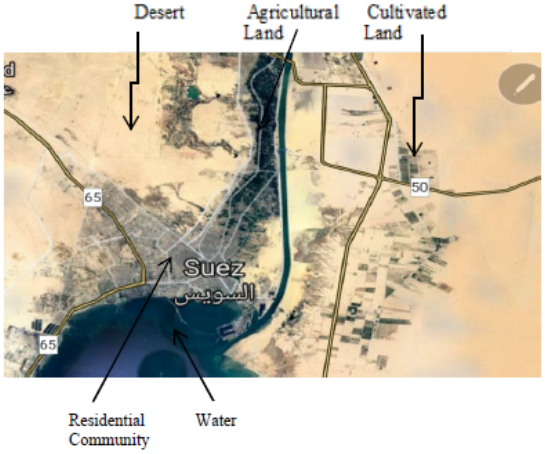


Figure 3: Live recent satellite Image for the selected working area out of Suez Canal region

This work selected three clustering techniques; K- means, Kohonen neural network and Fuzzy respectively.

A decision to use more than one technique to achieve many objectives:

- Achieve more accurate results
- Take a voting between the most accurate ones
- Give a recommendation about the most suitable technique for similar applications

The proposed algorithm passes through three steps:

Step 1: Applying clustering techniques.

Step 2: Analyzing the four segments resulted from each clustering technique.

Step 3: Comparison with a ready-made program "ArcGIS".

From Figure three, it is clear that the largest segment area represents the "Desert", Next segment large size represents the "Residential". The third segment, size-based represents the "Agriculture" and the last one is the "Water".

Step one: Applying clustering techniques

• Technique 1: K-means

K-means technique was applied on google satellite images every December of each year from year 2001 to 2016. This technique was applied in Matlab version 2018b. The RGB parameters of the used image sets were used as the input data to the algorithm and the k parameter used was 4 to resemble the four expected clusters: Desert, Green lands, Water, and Residential regions. The results are shown in figures 4, 5 which illustrates a complete separation among different clusters.

(a)
2016 Original Image Luminance Band

ISSN: www.jatit.org E-ISSN: 1817-3195

(b)
2016 Clustered by K-Means Technique

Figure 4: Original (a) and Clustered (b)images by K-Means for year 2016

• Technique 2: Kohonen Neural Network

Kohonen Neural Network/Self Organizing Map (SOM)/Unsupervised neural network [19] was applied on, the same data set used, google satellite images every December of each year from year 2001 to 2016. This technique was applied using Matlab version 2018b. The used SOM classifies four clusters, so uses four output nodes, randomly initialized, used 1000 e-books with 0.001 resolution, and 60% of the data sets were used to train the network while the rest were used for validation and testing.

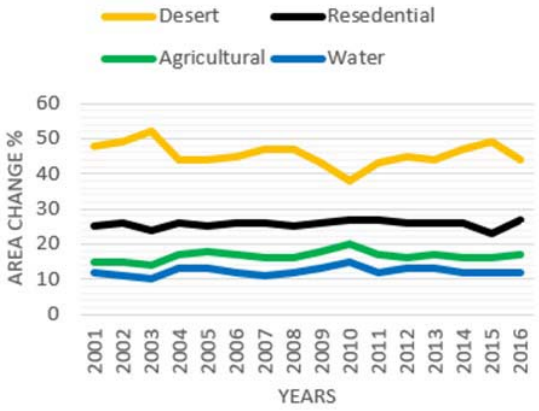


FIGURE 5. K-Means Clustering Results

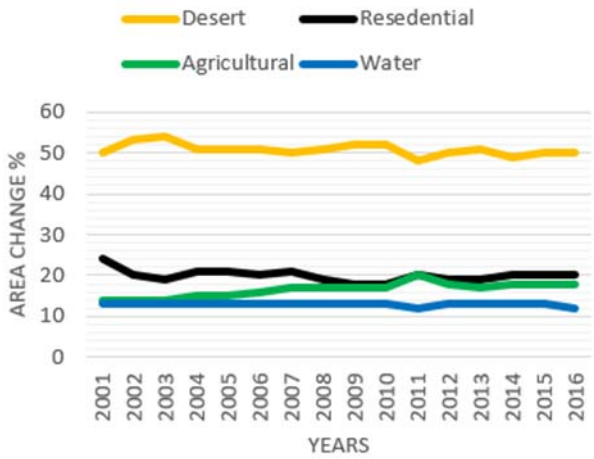
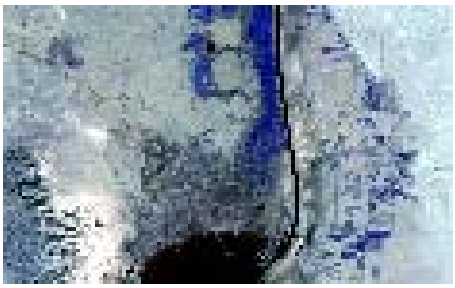


FIGURE 6. Kohonen Neural Network Results

• Technique 3: Fuzzy Clustering

Fuzzy Clustering technique was applied on the same used google satellite images for December of each year from year 2001 to 2016. This technique was applied in MATLAB version 2018b. Ten trim membership functions were used in the input layer, twenty rules were used and four output trim membership functions were applied to classify the four targets (categories) we are looking for.

(a)



(b)

Figure 7: Original (a) and Clustered (b)images by Fuzzy Clustering for year 2016

Journal of Theoretical and Applied Information Technology
30th April 2020. Vol.98. No 08

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

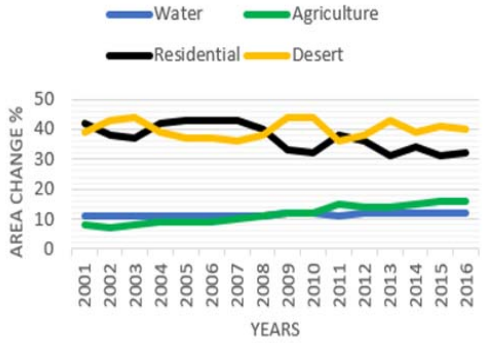


FIGURE 8. Fuzzy Clustering Results

Step two: Analyzing the four segments resulted from each clustering technique.

As a result from technique 1; K-means, it is clear that we have three segments increasing from year to year and one segment is decreasing which is the desert and this is a logical result due to the continuous development for this area in the field of land cultivation and establishing new residence communities. The development deteriorated little bit during the period of Egypt revolution (year 2011) so the agriculture area decreased. The development of agriculture increased rapidly after the year 2011 especially in the eastern side of the Suez Canal.

As a result from technique 2; Kohonen neural network, the four segments are smoothly constant with no dramatic changes in the four segments.

As a result from technique 3; Fuzzy clustering, two segments (water and agriculture) are almost constant (smooth changes) while the other two segments (desert and residential community) are changing alternatively.

The difference in percent for each segment is coming from the resolution of the image and the internal design of each algorithm.

Step three: Comparison with "ArcGIS"

In order to verify the results of the proposed techniques, the same objectives - to cluster the satellite images - was applied through a ready-made program "ArcGIS" [20]. The ArcGIS uses a supervised maximum likelihood clustering technique on which it is required to give the program a sample of the segments to be clustered. Four samples where provided for the program to be trained and cluster with. The results of two years 2001 and 2016 (first and last images in the provided data set) are presented in table two.

Table 2: Comparison between the used techniques and ArcGIS program for the year (a) 2001 (b) 2016

(a)

2001	Fuzzy	Kmeans	Kohonen	ARCGIS
Desert	58%	48%	50%	66%
Water	2%	12%	13%	2.7%
Agriculture	10%	15%	14%	4.7%
Residential	30%	25%	24%	26%

(b)

2016	Fuzzy	Kmeans	Kohonen	ARCGIS
Desert	53%	44%	50%	53.4%
Water	1%	12%	12%	2.4%
Agriculture	14%	17%	18%	7.1%
Residential	32%	27%	20%	36%

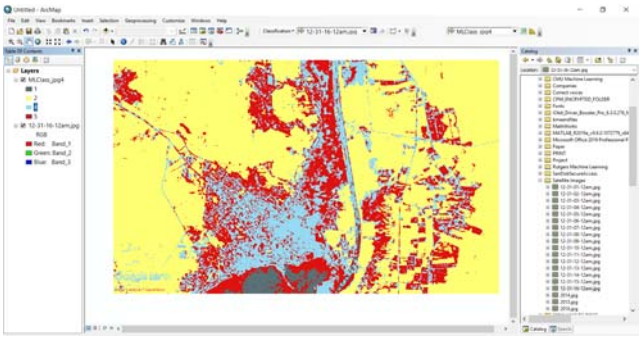


Figure 9. Arcgis Clustering Screenshot

5. RESULTS AND DISCUSSION

From the previous three techniques, it has been noticed that: our clustering techniques are rational clustering techniques and by this we mean that it could not be considered as absolute area calculations directly from the satellite images but it will give an indication about the relation of any segment with respect to the others.

In the first technique; K-means: Desert segment varying within 10% range specially going down as in 2010 due to the enlargement of "residential community" and ''land cultivation" especially in east bank of Suez Canal as detected from the images (Figure 4). The "Agricultural Land" seems to be in a steady state situation. The "residential community" is in reverse relation with the "Desert" which sounds logic. Finally, "Water" must vary in a very narrow range which is in the real case represents mainly, the Suez Canal and the north section from Suez golf which also sounds rational.

In the second technique; Kohonen Neural Network:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

All the segments behavior resulted from the Kohonen Neural network clustering technique is the same as segments behavior resulted from K-means clustering technique i.e. the change from going up and down for each segment for its lifetime from 2001 to 2016 is identical with the relevant segment behavior in K-means Clustering technique.

In the third technique Fuzzy Clustering, the same notice as the technique segments behavior is there. All the segments behavior resulted from the Fuzzy clustering technique is the same as segments behavior resulted from both K-means clustering technique and Kohonen Neural Network clustering technique i.e. the change from going up and down for each segment for its lifetime from 2001 to 2016 is identical with the relevant segment behavior in both K-means Clustering technique and Kohonen Neural Network clustering technique.

On the other hand, the resolution in the used satellite images data set (distance between pixels) represents on average 26.3 cm for each two neighbor pixels which means that it is not a high satellite image resolution to process more detailed accurate data compared by other specialized satellite systems.

As a global remark from the three applied techniques results and according to the comparison with ArcGIS the more accurate one regarding segments clustering behavior and also area more accurate calculation is the Fuzzy logic, so, the recommendation from this work for such similar applications and for similar type if data (satellite images, luminance band) is to use Fuzzy logic clustering.

6. CONCLUSIONS

Manual clustering of satellite images is extremely difficult especially if low quality resolution images are used. Also, it will be time consuming if there is a stream of satellite images. Machine learning is the suitable and efficient solution represented for the clustering techniques.

In this work, three clustering techniques were applied on examples of satellite images on the same location and same period of time (12:00 AM), annually based (December 1st of each year). Period of 16 years were studied (from 2001 to 2016), Suez governorate area, in Egypt. The clustering techniques were guided to have four segments representing: Desert, Agriculture land, Residential communities and Water.

The clustering techniques showed that three segments are increasing, and one segment is decreasing which is the desert and this is a logical result and identical with the real situation. Fuzzy clustering tracked the changes of the segments accurately so the recommendation for similar applications is to use directly the Fuzzy clustering technique. A forecasting image could be derived from the previous years' analysis and the same technique may be used for any other regions to provide better strategic planning for urbanization to protect available resources and secure our population needs in the future.

ACKNOWLEDGMENT

The teamwork would like to express gratitude and thankfulness for the Information Technology institute, ITI, GIS department specially engineer Ahmed Rashad, the head of the GIS department and his team for the great support and consultation in applying the comparison between this research work and ArcGIS software.

REFRENCES:

[1] http://www.unoosa.org/oosa/en/spaceobjectregister/index.html last seen at 30/04/2019

[2] King Ronald S., *Cluster Analysis and Data Mining: An Introduction,* Mercury Learning & Information, Dulles, VA 20166, USA, 2015.

[3] Rodriguez, Mayra Z. et al, "Clustering algorithms: A comparative approach," *PLOS ONE,* 14(1), 2019.

[4] https://www.sczone.eg/English/Pages/default.aspx
last seen 29/04/2019

[5] https://www.google.com/earth/
last seen at 30/04/2019

[6] Witten Ian H., Frank Eibe, Hall Mark A., *Data Mining Practical Machine Learning Tools and Techniques* (3rd ed.), Morgan Kaufmann, Burlington, MA 01803, USA, 2011.

[7] Borra S., Thanki R., Dey N, *Satellite Image Analysis: Clustering and Classification*, Springer Briefs in Applied Sciences and Technology, Singapore, 2019.

[8] Dhanachandra N. and Chanu Y. J., "A survey on Image Segmentation Methods using Clustering Techniques," *European Journal of Engineering Research and Science*, Vol. 2, No. 1, 2017.

[9] Törmä Markus, "Kohonen self-organizing feature map and its use in clustering," *Proceedings of the SPIE,* Vol. 2357, pp. 830-835, 1994.

[10] Sadaaki Miyamoto, Hidetomo Ichihashi, and Katsuhiro Honda., *Algorithms for Fuzzy*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Clustering: Methods in C-Means Clustering with Applications (1st ed.), Springer-Verlag Berlin Heidelberg, 2008.

[11] Suganya R. and Shanthi R.,"Fuzzy C- Means Algorithm- A Review," *International Journal of Scientific and Research Publications*, Vol. 2, 2012.

[12] Bora D. J. and Gupta A. K., "A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm," *International Journal of Computer Trends and Technology (IJCTT),* Vol. 10, pp. 108 – 113, 2014.

[13] Han Jiawei and Kamber Micheline, *Data Mining: Concepts and Techniques* (2nd ed.), Morgan Kaufmann, San Francisco, CA 94111, USA, 2006.

[14] Chitra A. K. and Maheswari D., "Comparative Study of Various Clustering Algorithms in Data Mining," *International Journal of Computer Science and Mobile Computing*, Vol. 6, pp. 109 – 115, 2017.

[15] Jassar Kamalpreet Kaur and Kanwalvir Singh Dhindsa., "Comparative Study and Performance Analysis of Clustering Algorithms." *International Journal of Computer Applications*, 2016.

[16] Tan Pang-Ning, Steinbach Michael, Karpatne Anuj, and Kumar Vipin, *Introduction to Data Mining* (2nd ed.), Pearson Publisher, 2018.

[17] R. R. do Valle Gonçalves, J. Zullo, L. A. S. Romani, B. F. do Amaral and E. P. M. Sousa, "Agricultural monitoring using clustering techniques on satellite image time series of low spatial resolution," *9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, Brugge, pp. 1-4, 2017.

[18] Radhika K. and Sourirajan Varadarajan, "Image classification of different resolution images using cluster ensemble techniques," *International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET),* 2017.

[19] Kohonen Teuvo, *Self-Organizing Maps* (3rd ed.), Springer Series in Information Sciences, 2001.

[20] ESRI white paper, "Architecting the ArcGIS Platform: Best Practices", 2018.