

Climate Signal Clustering Using Genetic Algorithm for Precipitation Forecasting: A Case Study of Southeast of Iran

Banafsheh Zahraie,
Assistant Professor, School of Civil Engineering, University of Tehran, Tehran, Iran,
bzahraie@ut.ac.ir

Abbas Roozbahani,
Graduate Student, School of Civil Engineering, University of Tehran, Tehran, Iran,
roozbahany@ut.ac.ir

Abstract

In this paper, an innovative method for clustering of climate signals is developed using Genetic Algorithm (GA). In this model, the relation of the signals with the variations of another climatic variable is considered in clustering algorithm. In the case study, the model is used for clustering the Sea Surface Temperature (SST) data in Omman Sea, Arabian Sea, and Indian Ocean considering the precipitation variations in Sistan-Balouchestan Province in Southeast of Iran. For this purpose, the precipitation data is classified to the three categories of below normal, normal, and above normal and the fitness function of the GA model is formulated to minimize the variance of the precipitation for each selected cluster. The results show that the model can be effectively used for prediction of low and high precipitation seasons in the study area using the SST variations in the defined clusters.

Keywords: *Climate Signals, Clustering, Genetic Algorithm, Sea Surface Temperature.*

Introduction

Considering climate signal variations is an important part of many hydrologic and water resources management related studies. Results of analysis of these variations have been used in many previous studies for long-term water resources planning. A number of clustering methods have been developed for different applications and among these methods, K-means, Map to Map, and Fuzzy C-means are the most popular, which have been used in different sciences and engineering fields.

GA-Clustering is a new and efficient method, which has been mostly used for clustering of economical and social data. Development of GA-Clustering goes back to early 90s. Alippi & Cucchiara (1992) and Bandyopadhyay & Maulik (2002) tried to develop the method and presented new approaches for GA-Clustering using real encoding of genomes and un-known number of clusters. Another efficient method was presented by Garai & Chaudhuri (2004) and Hwiei-GenLin et al. (2005) based on binary encoding with un-known number of clusters.

In this paper, a new method has been presented that in addition to it's innovation in using GA-Clustering for climate signals, the model's framework is improved in a

way that choosing the clusters for climate signals (such as SST) is not only based on temporal and spatial variations of these signals, but also variations of another climate variable like precipitation and relation between them, is also considered. In the following sections, results of this model are compared with K-means method. In the case study, the results of the GA-Clustering model are used to prepare a set of precipitation prediction guidelines. The model structure and results are presented in the following sections.

Methodology

GA-Clustering is a relatively new clustering method, which is used in this study for climate signal clustering. In the presented method in this paper, length of chromosomes are variable with respect to the number of selected zones for analyzing climate signals variations. For example, if number of zones and clusters are equal to 3, the length of chromosomes is 9. (See Fig.1).

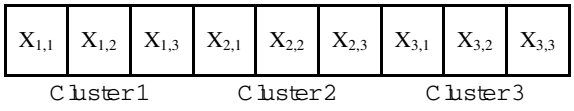


Fig 1. *Sample of chromosomes in the GA-Clustering model*

The centers of the clusters are considered as the genes and real encoding has been used. As it can be seen in Fig. 1, $X_{i,j}$ is the measured value of the climate signal for the center of the i^{th} cluster for the j^{th} geographical zone. In this study, the variable, which is being clustered, is Sea Surface Temperature (SST) in three different geographical zones, which are explained later in the paper.

In the first step, considering range of observed SST in each geographical zone, for each of genomes in each chromosome, a random number is generated, which showes center of each cluster. For example in Fig. 1:

$$SST_{\min,j} \le X_{i,j} \le SST_{\max,j} \qquad (1)$$

Where $SST_{max,j}$ and $SST_{min,j}$ are the maximum and minimum values of SST in j^{th} geographical zone, respectively. In the next step, using input data table, containing SST and precipitation time series, the clustering process starts. Each row of this table is related to a specific year of the historical input data as follows:

$(SST_{1,t},SST_{2,t},SST_{3,t},P_t)$

$SST_{j,t}$ is the value of the SST in j^{th} zone in year t. P_t is value of the secondary variables that in this study is considered to be precipitation in a specific season in year t.

Euclidean distance of SST from randomly selected centers of clusters for each year has been used for allocating SST data to different clusters as follows:

$$d_{i,t} = \sqrt{\sum_{j=1}^{3} \left(SST_{t,j} - SST_{i,j}^{\,c} \right)^2} \qquad (2)$$

Where $d_{i,t}$ is the Euclidean distance of observed SST in year t from selected center for the i^{th} cluster. $SST_{t,j}$ is the observed SST in year t in the j^{th} zone and $SST_{i,j}^{c}$ is the center of i^{th} cluster in j^{th} zone, generated randomly in GA-Clustering model. This value for each year is calculated and the minimum value of these distances indicates what years belong to each cluster.

It is important to mention that in the proposed method, the precipitation variation has not any effect on allocation of SST data to different clusters. In the next step, in order to estimate the fitness of each chromosome, precipitation variations have also been taken to account as follows:

$$\text{Min [Max } \{ Var(\underline{P}_i) / \text{ i= 1, 2, 3} \}] \qquad (3)$$

Where, (*Var*) is the variance of data and $\underline{P}_i$ is the set of precipitation values for the years in which SST values are allocated to cluster i. Equation (3) indicates that the best chromosomes are selected based on the minimum variance of observed precipitation for the years in each cluster.

Finally, in each generation, the chromosomes having the minimum fitness function are selected as the best chromosomes. If maximum number of clusters is equal to 3, the model's results will suggest the proper clustering from 1 to 3 clusters. The algorithm for GA-Clustering model is shown in Fig. 2.

K-means Clustering

K-means is one of the simplest unsupervised learning algorithms, which has been used in this study to evaluate the efficiency of the proposed GA-Clustering model. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define K centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early clustering is done. At this point, we need to re-calculate K new centroids as the centers of the clusters resulting from the previous step. After we have these K new centroids, a new binding has to be done between the same data set points and the nearest new centroid.

As a result of this loop we may notice that the K centroids change their location step by step until no more changes are done. Finally, this algorithm aims at minimizing an objective function, as follows:

$$J = \sum_{i} \sum_{t} d^{2}_{i,t} \qquad (4)$$

In this study, we have used this method to cluster the SST data and the results are compared with the GA-Clustering model.

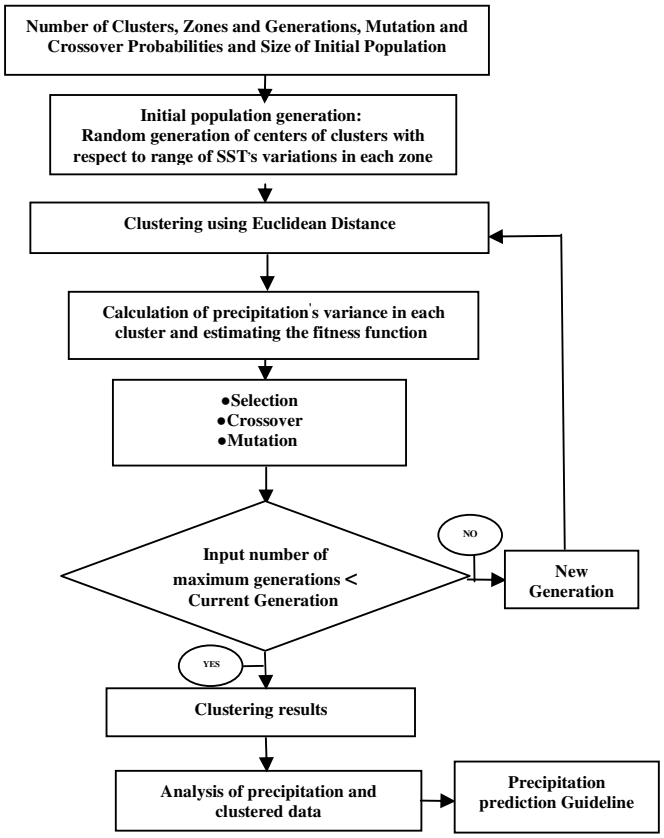


Fig 2. *GA-Clustering of SST data with respect to precipitation variations*

Case study

To analyze the efficiency of the model, case study of Sistan & Baluchestan Province in Southeast of Iran is selected. This part of Iran with area of 181,000 km^2 and population of 1.7 billion, has annual precipitation of 19,954 m^3. This Province has a dry climate condition and very limited precipitation mostly in the form of heavy Monsoon storms. Extreme fluctuations of surface water resources availability, ground water limitations and high speed winds are some of the main characteristics ot this area. It shows the importance of water resources planning and management in this Province.

There are 70 rain gauges in this province. 20 of these rain gauges having long term data are selected in this study to analyze the efficiency of the developed methodology. For studying the effects and relations between the precipitation and large scale climate signals, SST data in three zones, which are considered to be effective on the precipitation of the study area, are selected and position of this zones and their geographical information are shown in Fig. 3 and Table 1.

After studying SST maps in these zones and precipitation variables, the precipitation data in the months of January, February, and March in the years of

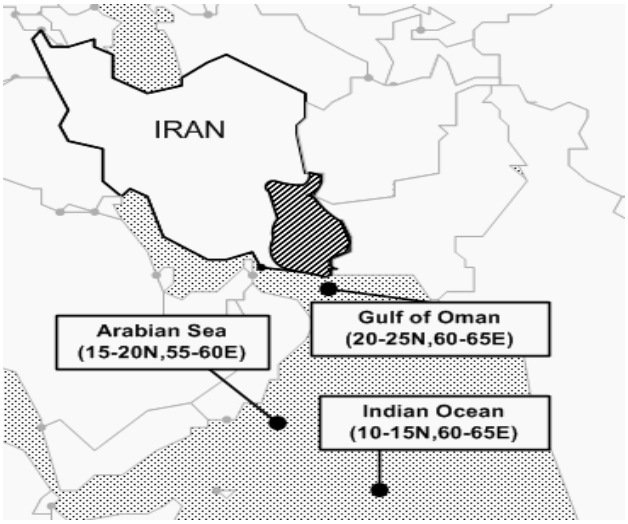


Fig 3. *Geographical zones for SST's data*

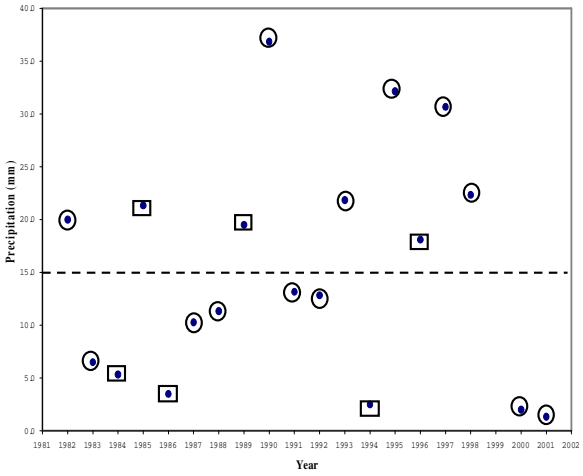
1983-2002 and the SST data in the months of August, September, and October in the years of 1982-2001 are used for clustering. It should be noted that these months are selected in such a way that SST data in the selected clusters can be used for predicting the precipitation variations in the months of January, February, and March (High Precipitation Season). This model easily can change the number of zones from 1 to 3. Therefore, the effects of SST on precipitation have been studied zone by zone and for all three geographical zones.

Table 1.*Gographical information of studied zones for SST data*

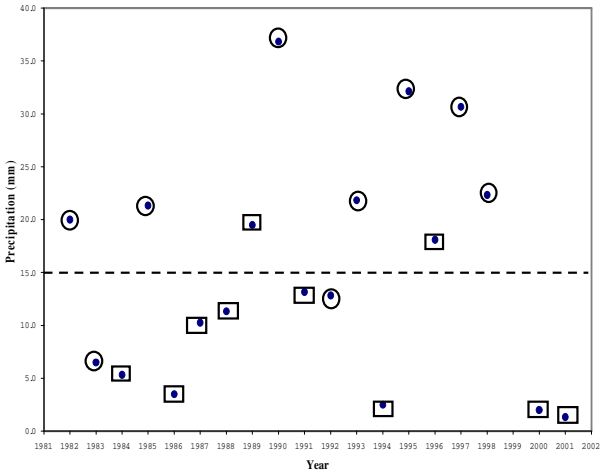
Geographical zone	Longitude	Latitude
Omman Sea	60E-65E	20N-25N
Arabian Sea	55E-60E	15N-20N
Indian Ocean	60E-65E	10N-15N

Clustering Results

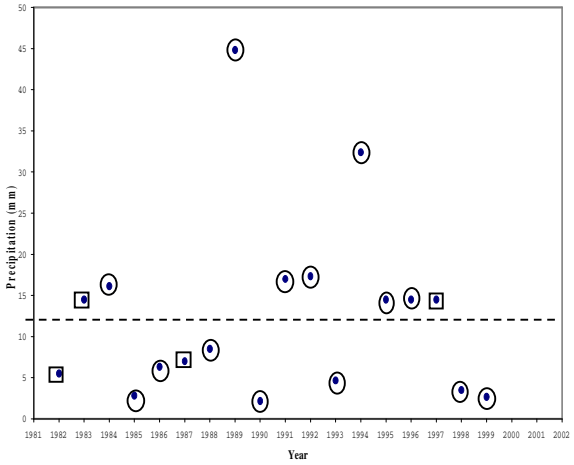
The precipitation data in Saravan, Kajdar, Ladiz, and Chabahar rain gauges with the SST in Omman Sea, Arabian Sea, and Indian Ocean are used as the model input data. In the presented model, the number of initial population and the best values of mutation and crossover probabilities are selected based on sensitivity analysis and are equal to 50, 0.06, 0.6, respectively. Fig. 4 shows the results of K-means and GA-Clustering methods for Saravan and Ladiz rain gauge and Omman Sea and Indian Sea SST data as an example. As it can be seen in this figure, incorporating precipitation variations in GA-Clustering model has significantly affected the ordinary K-means clustering results.



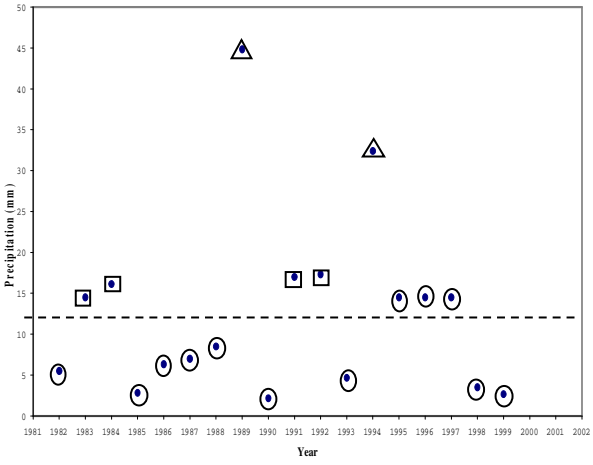
4-a: *Saravan rain gauge-Omman Sea (K-means)*



4-b : *Saravan rain gauge -Omman Sea (GA)*



4-c**:** *Ladiz rain gauge -Indian Ocean (K-means)*



4-d: *Ladiz rain gauge -Indian Ocean (GA)*

Fig 4. *Results of Clustering SST with and without respect to precipitation in GA and K-means methods (◯,□ ,△ show the selected clusters)*

When we consider the effects of the SST in Omman Sea on the precipitation of Saravan Station (Fig. 4-b), the GA-Clustering model select the two clusters that fist cluster consists of 9 years. 78 percent of the precipitation data in this cluster has been higher than normal and 80 percent of the remained years, in second cluster, have been lower than normal As it can be seen in Fig. (4-a), non of the two clusters selected by K-means method can be used for predicting precipitation level. Similar results have been also observed for Ladiz rain gauge as shown in Fig. (4-c, 4-d).

Precipitation Prediction

Based on the results of this study, low and high precipitation spells, with respect to SST variations in selected geographical zones can be predicted. Precipitation forecasting guidelines for selected rain gauges in this study are shown in Table 2.

Table 2. *Results of clustering the SST data with respect to precipitation variations in the selected rain gauges in Sistan- Balouchestan Province*

Rain Gauge	Predictor Geographical zone (SST)	Range of variations SST(C°)	Precipitation Prediction
Saravan	Omman Sea	$27.7 \leq SST_O \leq 28$	P > 15.3 mm
		$26.4 \leq SST_O \leq 27.7$	P < 15.3 mm
	Arabian Sea	$26.2 \leq SST_A \leq 26.3$	P > 15.3 mm
Kajdar	Omman Sea	$27.4 \leq SST_O \leq 27.7$	P < 30mm
Ladiz	Indian Ocean	$26.5 \leq SST_I \leq 26.8$	P >12.7 mm
		$26.9 \leq SST_I \leq 27.9$	P < 12.7 mm
Chahbahar	Omman Sea	$26.5 \leq SST_O \leq 26.8$	P > 19.2 mm
		$27.7 \leq SST_O \leq 28$	P < 19.2 mm
	Omman Sea,Arabian Sea,Indian Ocean	$26.2 \leq SST_O \leq 26.3$ $25.4 \leq SST_A \leq 26.5$ $26.8 \leq SST_I \leq 28$	P < 19.2 mm

SST_A : Sea Surface Temperature of Arabian Sea
SST_I: Sea Surface Temperature of Indian Ocean
SST_O : Sea Surface Temperature of Omman Sea

Using Table 2, we can predict the average precipitation in the months of January, February, March. For example, in Ladiz rain gauge, if estimated average of SST in Indian Ocean in the months of the August through October is between 26.5 and 26.8 C°, average precipitation in winter will be higher than 12.7 mm and if the SST value is between 26.9 and 27.9 C°, average precipitation in winter is lower than 12.7 mm.

Conclusion

In this study, a GA-Clustering method is presented, which has been developed for clustering the SST data which respect to precipitation variations. Comparison between the result of this method and ordinary K-means method, shows that the proposed algorithm has been able to significantly change the clustering results in a way, that it can be used for predicting the precipitation variations. Results of the case study of Southeast of Iran show that this model has been able to predict below and above normal winter precipitation with overall more than 80 percent accuracy. Studying SST variations in more locations in water bodies in South of Iran, can also help in further development of precipitation prediction guidelines.

References

Alippi, C., and Cucchiara, R.(1992), "Cluster partitioning in image analysis and classification: A genetic algorithm approach", Proc. CompEuro 1992 on Comput. System Software Eng., IEEE Computer Society Press, Silver Spring, MD, pp. 139-144.

Bandyopadhyay, S., and Maulik, U. (2002), "Genetic algorithm-based clustering technique", *Journal of Pattern Recognition Society*., Vol. 33, pp. 1455-1465.

Garai, G., and Chaudhuri, B.B. (2004), "A novel genetic algorithm for automatic clustering", *Journal of Pattern Recognition Letters*, Vol. 25, pp. 173-187.

GenLin H., F. Yang, and Y. Ta Kao (2005), "An Efficient GA-based Clustering Technique", *Journal of Science and Engineering*. 8(2), pp. 113-122.