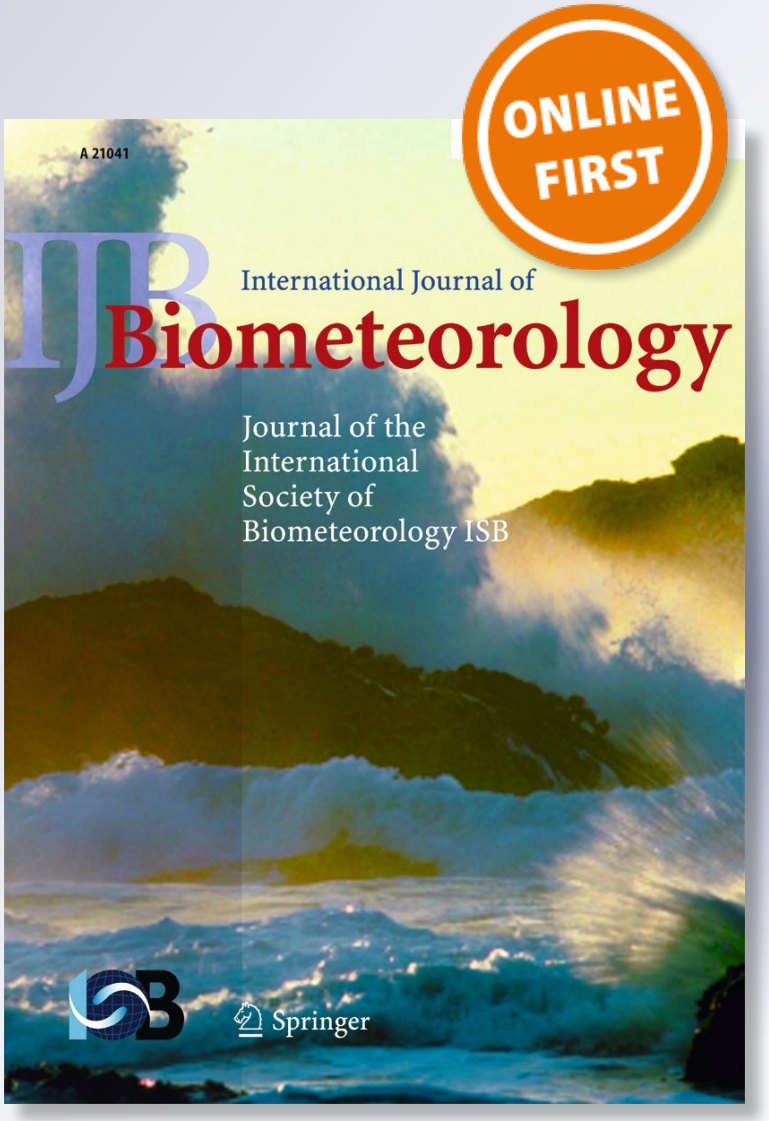
Climate data clustering effects on arid and semi-arid rainfed wheat yield: a comparison of artificial intelligence and K-means approaches

Nasrin Salehnia, Narges Salehnia, Hossein Ansari, Sohrab Kolsoumi & Mohammad Bannayan

International Journal of Biometeorology

ISSN 0020-7128

Int J Biometeorol
DOI 10.1007/s00484-019-01699-w



Springer

International Journal of Biometeorology
https://doi.org/10.1007/s00484-019-01699-w

ORIGINAL PAPER

Climate data clustering effects on arid and semi-arid rainfed wheat yield: a comparison of artificial intelligence and K-means approaches

Nasrin Salehnia[1,2] · Narges Salehnia[3] · Hossein Ansari[1] · Sohrab Kolsoumi[2] · Mohammad Bannayan[1]

Received: 28 September 2018 / Revised: 5 February 2019 / Accepted: 16 February 2019

Abstract

Clustering algorithms are critical data mining techniques used to analyze a wide range of data. This study compares the utility of ant colony optimization (ACO), genetic algorithm (GA), and K-means methods to cluster climatic variables affecting the yield of rainfed wheat in northeast Iran from 1984 to 2010 (27 years). These variables included sunshine hours, wind speed, relative humidity, precipitation, maximum temperature, minimum temperature, and the number of wet days. Seven climatic factors with higher correlations with detrended rainfed wheat yield were selected based on Pearson correlation coefficient significance (P value < 0.1). Three variables (i.e., sunshine hours, wind, and average relative humidity) were excluded for clustering. In the next step based on Pearson correlation (P value < 0.05) between the yield, and the seven climate attributes, fitness function, and silhouette index, only four attributes with higher correlation in its cluster were selected for reclustering. Four climate attributes had an extensive association with yield, so we used four-dimensional clustering to describe the common characteristics of low-, medium-, and high-yielding years, and this is the significance of this research that we have done four-dimensional clustering. The silhouette index showed that the best number of clusters for each station was equal to three clusters. At the last step, reclustering was done through the best-selected method. The results yielded that GA was the best method.

Keywords Fitness function · Attribute · Rainfed wheat · Silhouette index · Genetic algorithm · Ant colony

Introduction

Rainfed agriculture is the primary source of staple food production in many regions over the globe, and it supports the livelihood of a large number of farmers in the semi-arid regions of Iran (Bannayan et al. 2011). Wheat (*Triticum aestivum* L.) is the most important source of calories in the area, providing more than 40% of human required energy (Romero et al. 2013). Wheat production is the third largest cereal produced globally (FAO, Statistical Pocketbook 2015).

Meteorological conditions significantly affect wheat yield. Li et al. (2010) studied the correlation between wheat yield and climatic factors at five different spatial scales in China. They concluded that under the current climatic conditions, the relationships between wheat yield and each of precipitation and temperature became weaker and stronger, respectively, with an increase in spatial scale. Ahmed and Hassan (2011) studied temperature and solar radiation as determinant factors for spring wheat grain yield; results indicated that yield was directly proportional to solar radiation and temperature. Lobell and Field (2007) identified a 0.6 to 8.9% reduction in wheat yield per 1 °C rise in temperature. Luo et al. (2005) found that temperature increase had some impacts on wheat yield, but its effects were much smaller than that of rainfall. Wheeler et al. (2000), Bannayan et al. (2010), and Ahmed et al. (2016) reported similar results.

Iran is located in South Asia, and its climate is predominantly semi-arid and arid. Agricultural production (especially of rainfed crops) often fluctuates in Iran's water-limited ecosystems due to periodic meteorological droughts and erratic rainfall. Long-term trends of precipitation (Tabari and Talaee 2011b) and temperature (Tabari et al. 2011; Tabari and Talaee

✉ Hossein Ansari
Ansary@um.ac.ir; Ansariran@gmail.com

[1] Faculty of Agriculture, Ferdowsi University of Mashhad, P.O. Box 91775-1163, Mashhad, Iran

[2] AgriMetSoft, Roshd Center, Ferdowsi University of Mashhad, P.O. Box 9177-949207, Mashhad, Iran

[3] Department of Economics, Ferdowsi University of Mashhad, Mashhad, Iran

Published online: 22 May 2019

2011a) show changes that could further exacerbate productivity swings. Assessing the relationships between climatic factors and rainfed wheat production provide insights for policymakers to adopt appropriate adaptation measures to alleviate the likely adverse effects of climate change on future national food security.

Data mining is a critical data analysis technique used in a wide variety of science fields. One of the types of data mining methods is clustering analysis. A cluster is a collection of objects where objects are similar to each other, where objects in different clusters are dissimilar from each other, and where there is a good separation between clusters. Cluster analysis is appropriate for illustrating associations between climatic parameters and crop yield (Vardhan et al. 2014). In addition to clustering methods, a variety of artificial intelligence (AI) approaches have been developed for monitoring changes based on evolutionary algorithms, including genetic algorithm (GA), ant colony optimization (ACO), and artificial neural network optimization (ANN) approaches.

Several studies have applied ANN to agriculture. Alvarez (2009) explored the effects of soil and climate factors on wheat yield in the Argentine Pampas with artificial neural networks (Alvarez 2009). Abdullah et al. (2014) created a model hybrid of the ANN and GA to predict reference evapotranspiration in arid and semiarid regions, which improved the efficiency of predicted evapotranspiration. Kim and Ahn (2008) applied K-means clustering and GA. They suggested a new clustering algorithm GA K-means.

Similarly, Krishna and Narasimha Murty (1999) and Mualik and Bandyopadhyay (2000) proposed a novel clustering method. They pointed out that the performance of the proposed algorithm is better than a single method. Laszlo and Mukherjee (2007) developed a genetic algorithm that exchanges neighboring centers for k-means clustering. Kuo et al. (2005) presented a novel clustering method, ant K-means (AK) algorithm. Xu et al. (2010) used the K-means algorithm and ACO algorithm for clustering image. Ant-based clustering sorting was first introduced by Deneubourg et al. (1991) to explain the phenomena of corpse clustering and larval sorting in ants. Machnik (2006) used ACO, K-means and single link, and average link methods for clustering two collections of documents. The results revealed that the ACO method performed better compared with the other methods. Tsai et al. (2004) proposed an efficient clustering approach for large databases, i.e., ACO with different favor algorithm. This algorithm performed better than the fast self-organizing map K-means, and genetic K-means algorithm. Handl and Meyer (2002) have compared the performance of ACO with two classical algorithms (K-means and agglomerative average link), and their results indicated that ACO performed better. Shelokar et al. (2004) evaluated the ACO approach for clustering three different examples (such as flowers, the dataset contains chemical analysis and thyroid diseases). This approach is compared with other stochastic algorithms, i.e., GA, simulated annealing, and tabu search. The results showed that ACO and GA had close results.

This study employed three methods of data clustering techniques: K-means, GA, and ACO. Among the clustering techniques, K-means is one of the most popular and frequently used methods (Sung and Jin 2000; Laszlo and Mukherjee 2007; Kao et al. 2008; Niknam and Amiri 2010; Celebi et al. 2013). K-means is greedy, which means that every time it converges to a local minimum, it is expected to converge to a locally optimal solution only and not to the globally optimal solution. This problem has been nearly solved by stochastic methods (Kim and Ahn 2008). The second algorithm used was GA, an optimization technique based on the principles of heredity of living organisms. The third method, ACO (Dorigo et al. 1991), has not previously been used to cluster meteorological variables in the context of predicting wheat yield.

This study aims to (1) compare two AI methods (ACO and GA) with K-means using the fitness function metric, (2) select the best method for clustering the most useful attributes for determining detrended wheat yield, and (3) to analyze different achieving clusters and assess relationship between detrended wheat yield, effective climatic data, and drought years.

Materials and methods

Data and location

The study sites are Mashhad, Sabzevar, and Torbat Heydarieh, located in Khorasan-Razavi province, northeastern Iran. The physiographic characteristics of the study locations are presented in Table 1 and Fig. 1. According to the De Martonne aridity index (De Martonne 1926), Mashhad and Torbat Heydarieh have a semi-arid climate, while Sabzevar has an arid climate. The De Martonne aridity index is computed by the following equation:

$$I_{DM} = \frac{P}{T + 10} \tag{1}$$

where I_{DM} refers to De Martonne aridity index, P is the annual mean precipitation in mm, and T is the yearly mean air temperature in °C. According to the I_{DM} values, De Martonne classified the climate into six groups, namely, arid, semi-arid, dry sub-humid, humid, and very humid.

Meteorological data used include sunshine hours (h), wind (m/s), average relative humidity (%), total precipitation (mm), April precipitation (mm), average maximum temperature (°C), February and May average maximum temperature (°C), average minimum temperature (°C), and number of

Table 1 Physiographic details of study locations. The average temperature (T_{mean}) and total precipitation of three locations

Station	Lat (°N)	Long (°E)	Elevation (m)	T_{mean} (°C)	Total precipitation (mm)	Crop and climate data period
Torbat H	35° 16′	59° 13′	1450.8	14.2	279.4	1984–2010
Mashhad	36° 16′	59° 38′	999.2	14	257.5	1984–2010
Sabzevar	36° 12′	57° 43′	977.6	17.3	189.1	1984–2010

wet days (day) for the period of 1984–2010 (27 years). These data were collected from meteorological stations co-located at each study site, and homogenization of weather data was implemented by the national meteorological organization of Iran (www.weather.ir) before the release of such data to users. Historical wheat yield data were obtained from the Organization of Agricultural Khorasan Research Station. To control for the effects of technological improvements on wheat yields over time, we carried detrended of wheat yield using linear regression (Kettlewell et al. 1999).

Applied algorithms

We used the fitness function (FF) to evaluate the three clustering methods, with lowest minimized FF reached iteratively (Mirkin 2011; De Amorim and Mirkin 2012; De Amorim 2016). For all three methods, the equation for FFs was the same; however, their values and the times of iteration to access the minimized value may differ. FFs for ACO and GA were calculated using MATLAB (R2017b, version 9.1) codes, and K-means FFs were computed using SPSS (version 18.0).

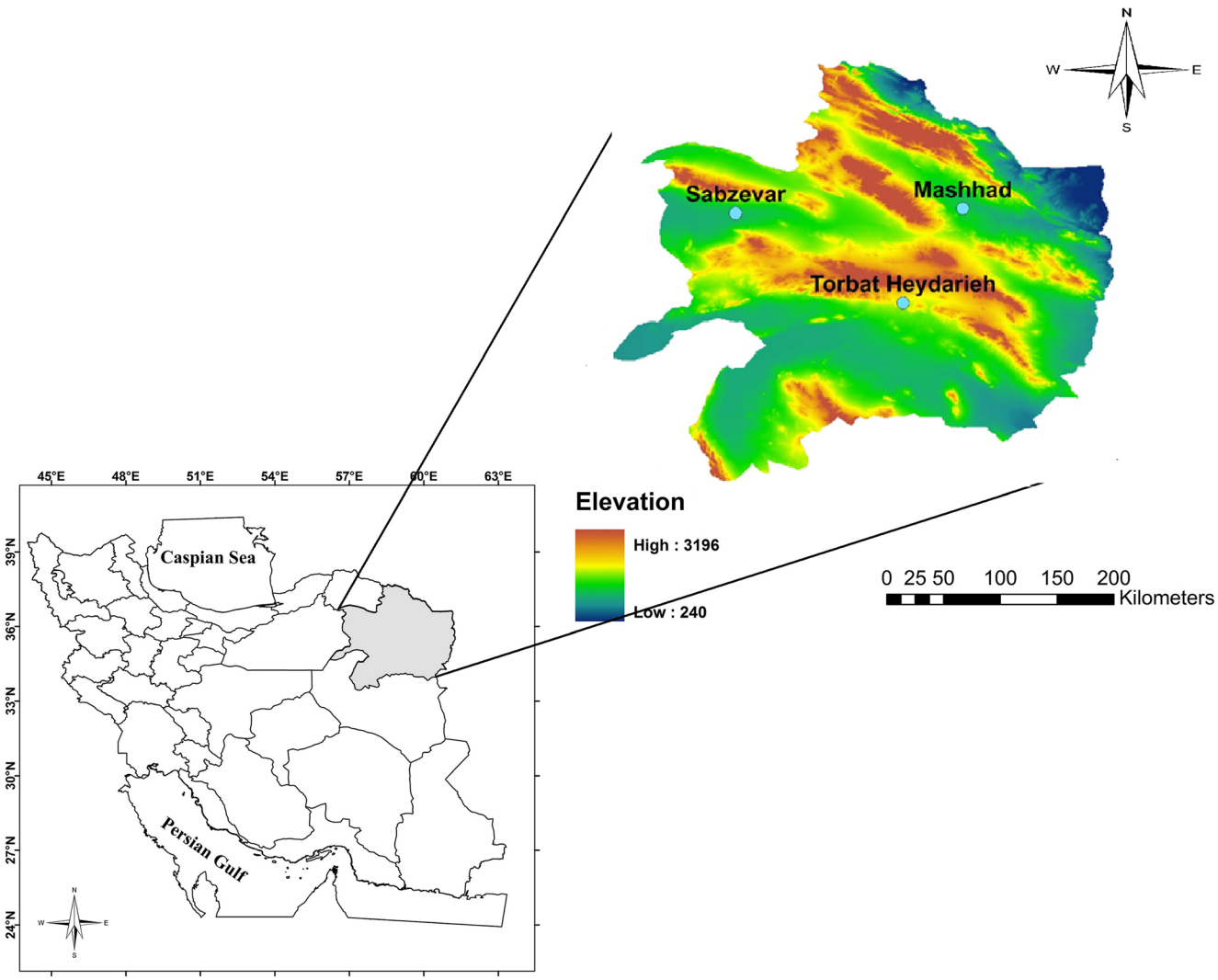


Fig. 1 Geographical study locations and synoptic weather stations

K-means algorithm

One of the most popular and commonly used methods for clustering is K-means (Niknam et al. 2011; Olgun et al. 2016). In this algorithm, the member k (number of clusters) was randomly selected among n elements as cluster centers. The n-k remaining member allocated to the closest cluster. After the allocation of all members, the cluster centers are recalculated and members reallocated until the cluster centers remain constant. These steps are repeated until the FF is minimized (Kim and Ahn 2008). The objective function (i.e., FF) is given by

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \sum_{v=1}^{n} \|X_{ijv} - C_{ijv}\|^2 \quad (2)$$

where K is the number of clusters, N is the number of members in the clusters, n is the number of the attribute, $\|X_{ijv} - C_{ijv}\|^2$ is a chosen distance measure between a data point X_{ijv}, and the cluster center C_{ijv} is an indicator of the distance of the n data points from their respective cluster centers.

Genetic algorithm

The GA is a stochastic search algorithm that begins with an initial random set of solutions, called chromosomes (Mualik and Bandyopadhyay 2000), which are composed of elements (genes). Genetic algorithms are characterized by attributes such as FF, encoding the input data, genetic operators, and population size. The most important operators used in the process include:

1. Selection: The selection operator selects a solution from the current population for the next population with probability proportional to its fitness value. It means this operator randomly selects a chromosome from the previous population according to the distribution. During selection, pairs of individuals are chosen from the population according to their fitness (Hertz and Kobler 2000). The most common way to select variables is the roulette wheel used in this study.

2. Crossover: This operator is used to combine the pairs of strings chosen (parents) to create new strings that potentially have a higher fitness than either of their parents. A probabilistic process exchanges information between two parent chromosomes for generating two child chromosomes. The middle parts of the two parents are then swapped to create two new offspring. We used a two-point method of the Crossover operator.

3. Mutation: Mutation process changes the structure of chromosomes typically by negating a randomly chosen bit. It randomly modifies the gene values at selected locations. Each chromosome undergoes mutation with a fixed probability. We used a uniform method of the Mutation operator (Krishna and Narasimha Murty 1999).

An objective chromosome consists of a set of classes' number for each year. The FF for GA was calculated using Eq. 2. Further explanation of the GA method is illustrated in Fig. 2, which shows one chromosome consisting of 27 genes (gen). Each gen represents the cluster's number of a year; in other words, each gene represents the class number, and we have considered three bits for each gene. Since we have 27 years (1987–2010), we have 27 gens. Because we run the GA with 2–6 clusters, every gen has three bits of zero and one, representing the number of clusters.

ACO algorithm

ACO was carried out for 27 years of actual data sets for seven meteorological parameters affecting wheat yield. In this algorithm, each ant builds a solution by walking from one point to another point. All steps to produce ACO clustering in this research were undertaken based on Shelokar et al. (2004). To solve an ACO clustering problem for achieving an optimal assignment of N objects to one of the K clusters, the sum of squared Euclidean distances between each object and the center of the belonging cluster was minimized. The pheromone matrix τ with the size of $N \times K$ (N samples and K clusters) at the beginning of the algorithm was initialized at some small value τ_0. The value of τ_{ij} at location (i, j) represents the pheromone concentration of sample i associated to the cluster j. After obtaining an answer (a population), a local search was run. When an agent intended to produce another solution, τ matrix should be updated. Assumed that N numbers of objects, divided into K cluster, to calculate FF to obtain optimal solution, Eqs. 3 and 4 are used:

$$MinF(w, m) = \sum_{j=1}^{K} \sum_{i=1}^{N} \sum_{v=1}^{n} w_{ij} \left\| x_{iv} - m_{jv} \right\|^2 \quad (3)$$

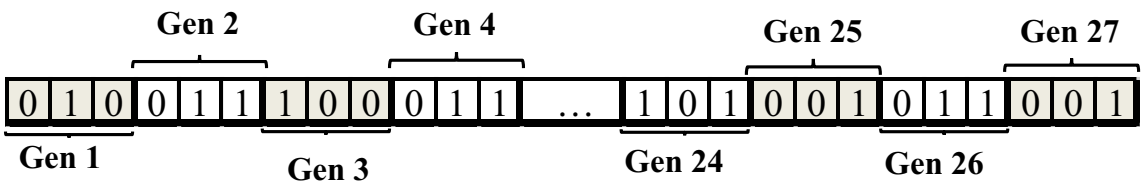


Fig. 2 One chromosome and 27 gens, according to the period of 1984–2010 (27 years)

$$m_{jv} = \frac{\sum_{i=1}^{N} w_{i,j} x_{iv}}{\sum_{i=1}^{N} w_{ij}}, j = 1, \ldots, K, \quad v = 1, \ldots, n \tag{4}$$

where m is a cluster center matrix of size $K \times n$, x_{iv} is a value of vth attribute of ith sample, m_{jv} an average of the vth attribute values. The amount of w refers to a weight matrix of size $N \times K$, w_{ij} is a weight for object x_i with cluster j, if object i is contained in cluster j, $w_{i,j} = 1$, otherwise $w_{i,j} = 0$. With respect to $i = 1 \ldots N, j = 1 \ldots K$. For more details, refer to Shelokar et al. (2004). In this method, 27 years are the objects with seven attributes (weather variables).

Select the effective variables

We examined ten variables to assess their effect on wheat yield: sunshine hours (h), wind (m/s), average relative humidity (%), maximum and minimum temperature (°C), total precipitation (mm), maximum temperature of April (°C), maximum temperature of May (°C), April precipitation (mm), and number of wet days (#) (Li et al. 2010; Chen et al. 2012; Eyshi Rezaie and Bannayan 2012; Jing-Song et al. 2012; Rahimi et al. 2014; Carvalho et al. 2015). In this present study, a wet day is defined as a day with the rainfall amount of at least 0.1 mm (Chou et al. 2012). To reduce the computational burden, we used a filtering process for the most critical variables and best methods of analysis as follows. We filtered the variable list based on correlation and significance level to seven variables, then implemented the three clustering methods. The best method based on FF and silhouette index (SI) was selected. After calculating correlations for each region, the correlation and significance level was used to limit the second analysis to the four most effective variables. In other words, reclustering was run with four variables using the best method.

Fitness validation

Many validity measures have been suggested for evaluating clustering results, such as Dunn method, Calinski–Harabasz (CH) method, sum-of-squares (SS) method, and silhouette index (Halkidi et al. 2001; Liu et al. 2010; Rendón et al. 2011). In this study, we used two methods for assessing the validation of cluster results, FF and SI. Silhouette analysis is not only used to choose an optimal value for the number of clusters in a problem, but also it can be a validation of consistency within clusters of data. The optimal outcome of the cluster analysis is to achieve the highest average silhouette value with the fewest clusters.

If x is an object in the cluster C_k and n_k is the number of objects in C_k, then the silhouette width of x is defined by the ratio of Eq. 5:

Table 2 Pearson correlation between wheat yield and minimum temperature (T_{min}, °C), maximum temperature (T_{max}, °C), precipitation (Precip, mm), precipitation April (Precip Apr, mm), sunshine hours (Nhour, h), wind (knot), maximum temperature April ($T_{max.}$ Apr), maximum temperature May ($T_{max.}$ May), relative humidity (Rhm, %), and wet day (#) across study locations

	Mashhad		Torbat H		Sabzevar	
	r	*P* value	*r*	*P* value	*r*	*P* value
$T_{min.}$	0.36	0.065***	0.31	0.099***	0.32	0.097***
$T_{max.}$	−0.33	0.092***	−0.34	0.082***	−0.32	0.097***
Nhour	−0.25	0.208	0.11	0.626	−0.12	0.551
Precip	0.59	0.001*	0.42	0.029**	0.35	0.073***
Wind	−0.15	0.452	0.21	0.293	0.01	0.96
T_{max} Apr	−0.51	0.008*	−0.53	0.004*	−0.39	0.044**
T_{max} May	−0.54	0.003*	−0.50	0.007*	−0.33	0.092***
Rhm	0.18	0.368	0.02	0.921	0.25	0.208
Precip Apr	0.33	0.091***	0.54	0.004*	0.49	0.009*
Wet day	0.51	0.006*	0.49	0.009*	0.46	0.016**

*Significant at 1% level, ** significant at 5% level, *** significant at 10% level

$$S(x) = \frac{b(x) - a(x)}{\max[b(x), a(x)]} \tag{5}$$

where $a(x)$ is the average distance between x and all other objects in C_k. The value of "k" refers to the number of clusters. It can be calculated using Eq. 6 as follows:

$$a(x) = \frac{1}{n_k - 1} \sum_{y \in C_k, y \neq x} d(x, y) \tag{6}$$

and $b(x)$ is the minimum of the average distances between x and the objects in the other clusters (Eq. 7),

$$b(x) = \min_{h=1,\ldots,K, h \neq k} \left[\frac{1}{n_h} \sum_{y \in C_k} d(x, y) \right] \tag{7}$$

Table 3 Value of fitness function (FF) and silhouette index (SI) in three methods over three locations with seven attributes

Method	Mashhad		Torbat H		Sabzevar	
	FF	SI	FF	SI	FF	SI
GA	32,740	0.37	29,516	0.54	21,238	0.46
ACO	32,740	0.37	29,516	0.54	21,392	0.44
K-mean	34,353	0.34	29,631	0.49	23,996	0.45

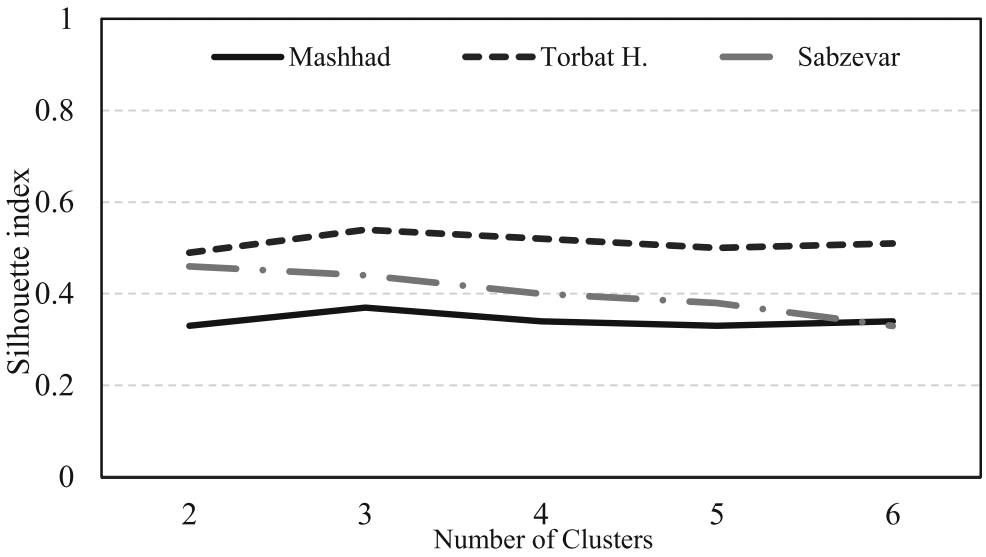


Fig. 3 The values of SI in different number of clusters through GA method over three stations

Finally, the global SI is defined by Eq. 8 as follows:

$$S = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{n_k}\sum_{x\in C_k} S(x)\right] \tag{8}$$

The SI measure has a range of [− 1, 1]. When the amount of S is close to − 1, the object is expected to be assigned to the wrong cluster; when S is close to 0, object is equally likely to be assigned to any of the two clusters; when S is close to + 1, object is considered to be clustered correctly (for more details, refer to Rousseeuw 1987).

Results and discussion

Variable selection

Pearson correlation between yield and each variable are listed in Table 2. The results showed that seven out of ten variables were both strongly correlated and significant about wheat yield ($r > 30$ and P value < 0.1). These were maximum and minimum temperatures, precipitation, wet day, April $T_{max.}$, May $T_{max.}$, and April precipitation. Our analysis showed that (Table 2) the highest correlation coefficient was associated with the precipitation in Mashhad ($r = 0.59$), and the lowest correlation coefficient was for wind in Sabzevar ($r = 0.01$). However, there was no significant correlation (P value > 0.1) between the detrended wheat yield and wind, Rhm. and Nhour variables across three locations, throughout the study (1984–2010).

Clusters analysis and select the best method

The results of the FF and SI calculations for three methods with seven selected variables are presented in Table 3. The smallest value for FF and the closest amount of SI to + 1 indicates the best method. In this case, the GA and ACO techniques were much better than the k-means method, with the best FF resulting from the GA method. SI showed that the clusters were best in Torbat H, with SI ≥ 0.49. Based on these results, we selected the GA method for the second analysis step.

For the exact number of clusters, we implemented the SI's value with differing cluster numbers through five runs of GA. According to the SI results with GA method in Fig. 3, the best number of clusters were obtained with three clusters. At Mashhad and Torbat, SI values (0.37 and 0.54, respectively) showed three clusters were best. Though two clusters (SI = 0.46) was slightly better than three (SI = 0.45) at Sabzevar station, we selected three clusters for all three stations for uniformity of analysis approach.

Assessing the results of the GA method

The average correlation values for three clusters under the GA method are provided in Table 4. Based on correlation and

Table 4 Average value of Pearson correlation between seven attributes and wheat yield (t/ha) through GA method across study locations

	Mashhad		Torbat H		Sabzevar	
	r	P value	r	P value	r	P value
T_{min}	0.41	0.033**	0.38	0.05	0.19	0.342
T_{max}	0.20	0.227	0.50	0.003**	0.30	0.208
Precip	0.40	0.038**	0.57	0.001*	0.48	0.011**
T_{max} Apr	0.42	0.029**	0.59	0.001*	0.24	0.227
T_{max} May	0.26	0.191	0.53	0.004	0.43	0.025**
Precip Apr	0.18	0.368	0.52	0.005	0.44	0.022**
Wet day	0.40	0.044**	0.60	0.001*	0.40	0.033**

*Significant at 1% level, ** significant at 5% level

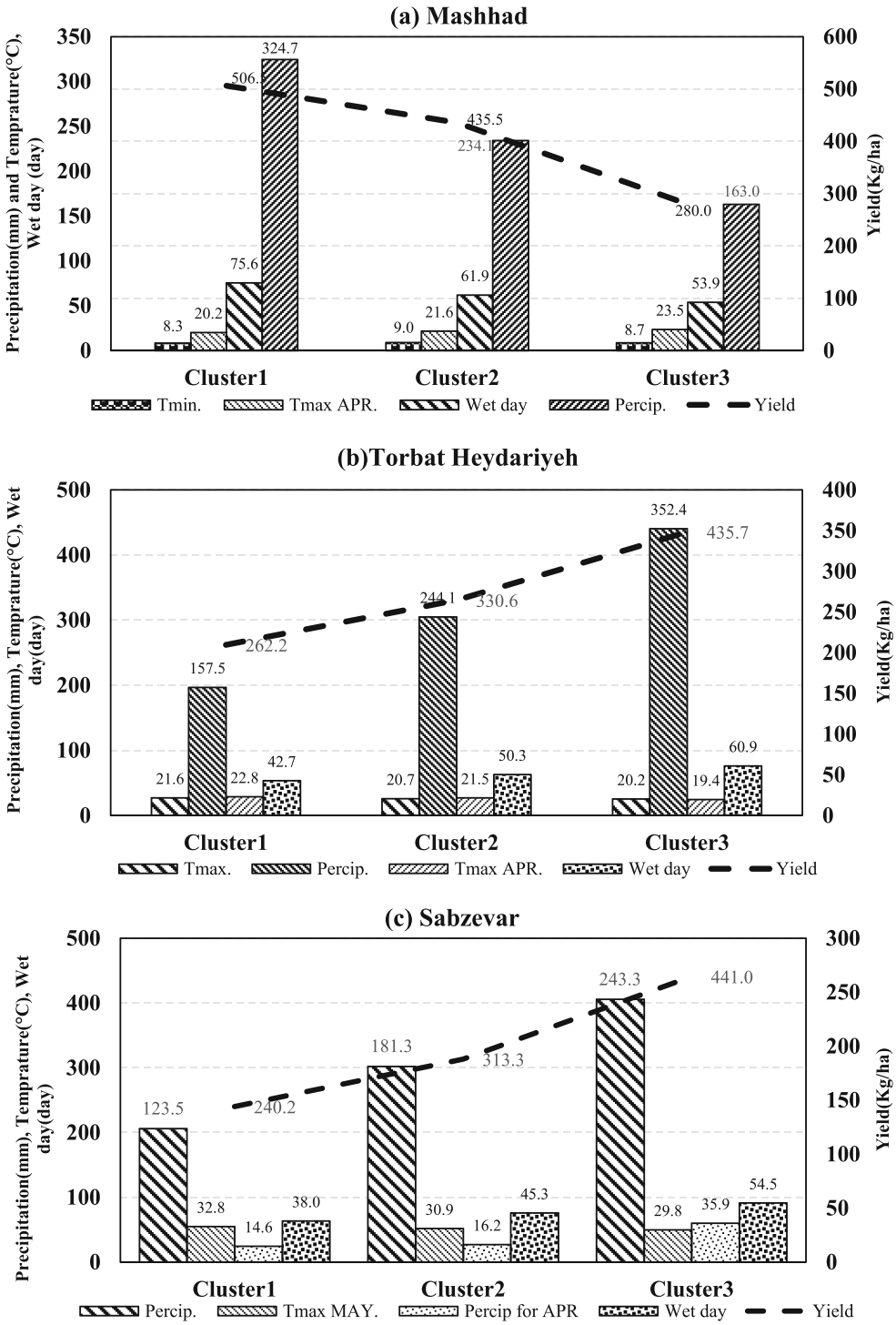


Fig. 4 Clusters of average rainfed wheat yield (kg ha^{-1}), precipitation (mm), T_{max} and T_{min} (°C), and the number of wet days (day) in **a** Mashhad, **b** Torbat Heydarieh, and **c** Sabzevar

significance values, three variables were omitted. The maximum temperature of May, April precipitation, and the maximum temperature had low correlations and were omitted from the remaining calculations in Mashhad. Minimum temperature, maximum temperature, and precipitation in April had the weakest *r* values for Torbat H. In Sabzevar, minimum temperature, maximum temperature, and maximum April temperature were eliminated from further analyses. Four variables were selected for the final studies using GA clustering (the highlighted values in Table 4).

Analyzing the result of clusters and wheat yield

In this study, three yield categories (high-yielding, mid-yielding, and low-yielding) were defined. The average values for each cluster are shown in Fig. 4. Based on GA clustering, the

Table 5 Different classes for the sites clustered by GA. Classes 1, 2, and 3 are shown in italics, bold, and bold italics, respectively

Mashhad	***1984***	*1985*	*1986*	**1987**	*1988*	**1989**	***1990***	*1991*	*1992*	***1993***	*1994*	***1995***	**1996**	**1997**	*1998*	***1999***	***2000***	***2001***	**2002**	**2003**	*2004*	**2005**	**2006**	**2007**	*2008*	***2009***	***2010***
Torbat	*1984*	**1985**	**1986**	**1987**	**1988**	**1989**	*1990*	***1991***	***1992***	***1993***	**1994**	***1995***	***1996***	***1997***	***1998***	***1999***	**2000**	*2001*	**2002**	**2003**	***2004***	**2005**	*2006*	**2007**	*2008*	**2009**	*2010*
Sabzevar	**1984**	*1985*	***1986***	**1987**	***1988***	**1989**	*1990*	***1991***	***1992***	***1993***	*1994*	*1995*	***1996***	**1997**	***1998***	**1999**	*2000*	*2001*	***2002***	***2003***	***2004***	**2005**	**2006**	***2007***	*2008*	**2009**	**2010**

Fig. 5 The variations of average values for precipitation, wet days, minimum temperature, and maximum temperatures of clusters in **a** Mashhad, **b** Torbat Heydarieh, and **c** Sabzevar

clusters 1, 3, and 3 were the high-yielding category, respectively, in Mashhad, Torbat H, and Sabzevar. Cluster 2 was mid-yielding in all locations. Low-yielding cluster in Mashhad, Torbat H, and Sabzevar were clusters 3, 1, and 1, respectively. The clusters with the highest yield had the highest precipitation, wet day, and the lowest T_{max}. Three clusters in each station are presented with different colors for 27 years (Table 5).

In Mashhad and Torbat H during the study period, the number of years with mid-yield were 11 and 12 years, respectively, with 9 years of relatively high yield. The selected region was severely affected by drought during the years 2000, 2001, and 2008, especially in Mashhad (Ministry of Jihad-e-Agriculture (Iran) 2009; USDA 2010; Rostami Khaleghi et al. 2014; Salehnia et al. 2017). Wheat yield clustering via GA shows this impact well (Table 5).

In Mashhad, cluster 1 was the high-yielding cluster (Fig. 4). In this cluster, precipitation and wet days were high, and T_{min} and T_{max} in April were low (Fig. 5a). Cluster 2 was mid-yielding and is shown in Fig. 5a, showing average values for precipitation and wet days, minimum temperature values higher than the other clusters, and maximum April temperatures in the middle relative to the other clusters (Fig. 5a).

The mean yield in Torbat H was low in cluster 1 (Fig. 4), with higher maximum temperature and maximum April temperature than the values of clusters 2 and 3 (Fig. 5b). The years included in cluster 1 had less precipitation and wet days relative to the other clusters. Cluster 2 has the mid-yields that cover the years where the maximum April temperature, maximum temperature, precipitation, and wet days are average (Fig. 5b). Cluster 3 includes high yield years with small maximum temperature and maximum April temperature values, and higher precipitation and wet days including the highest values than the other clusters (Fig. 5b).

Sabzevar shows cluster 1 of low yields with low precipitation amount, April precipitation and the number of wet days, and a high maximum temperature of May, relative to the other clusters (Fig. 4). Cluster 3 contains high yield years, with high amounts of precipitation, April precipitation, and number of wet days and lower maximum temperatures than the other clusters (Fig. 5c).

Mashhad and Torbat H both have the semi-arid climate, so it is not surprising that they share three attributes of their clusters in common (Fig. 5): precipitation, T_{max} April, and wet days. However, they differ in the fourth most important attribute, with Mashhad having T_{min} and Torbat H having T_{max}. This difference could be due to elevation. FF from the GA and ACO methods agreed for all locations. The K-means method

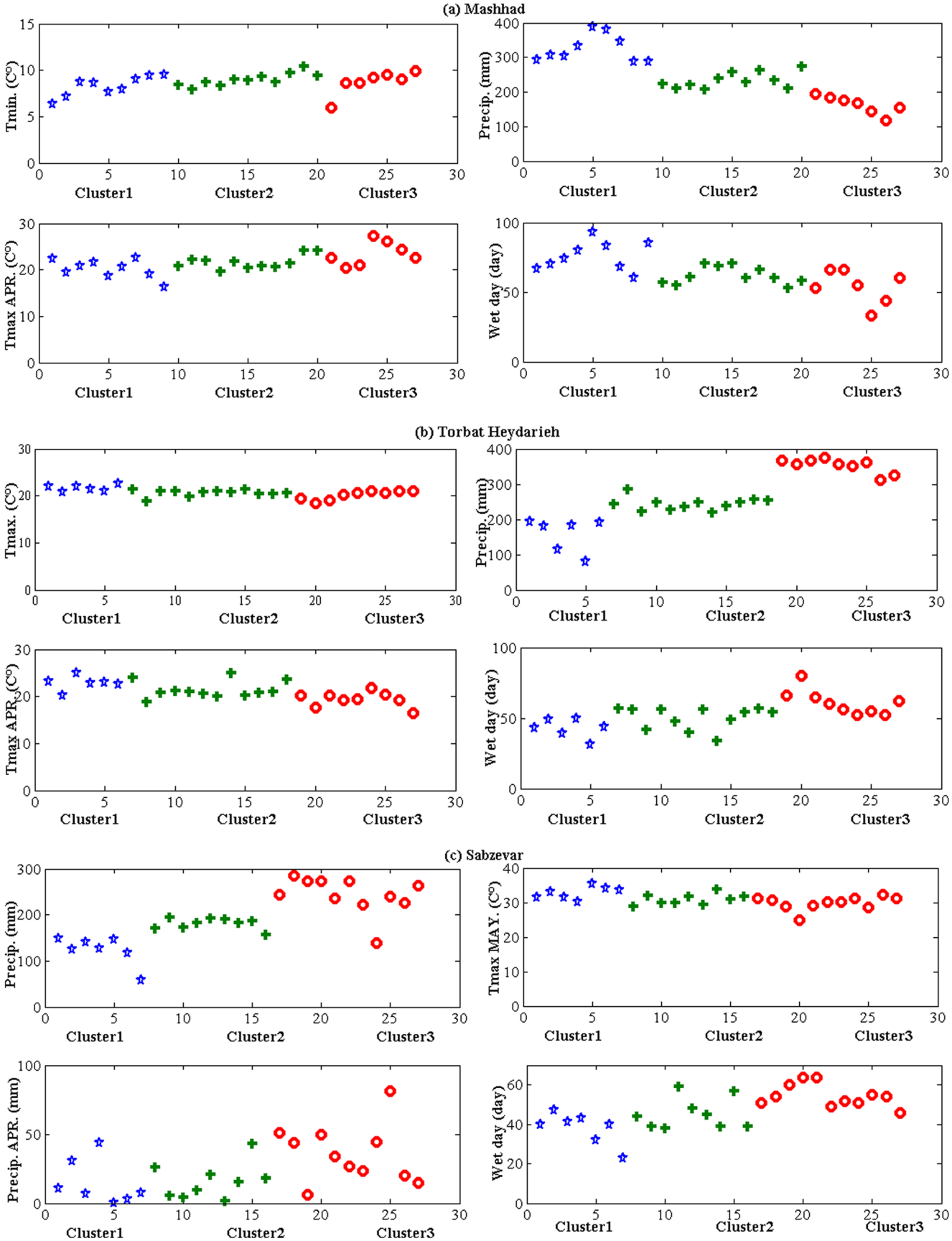


Table 6 The ranges of each variable and the related yield class, over the three locations

	T_{min}				Yield
		Precip	T_{max} Apr	Wet day	
Mashhad	$9 < T_{min} < 10$	285–398	16–22	60–93	High
	$8 < T_{min} < 10$	209–274	19–24	53–71	Mid
	$6 < T_{min} < 10$	119–193	20–27	33–66	Low
	T_{max} Apr				Yield
		Precip	T_{max}	Wet day	
Torbat H	T_{max} Apr > 16	313–376	18–21	52–80	High
	T_{max} Apr > 19	180–287	19–23	34–57	Mid
	T_{max} Apr > 23	81–116	21–22	31–39	Low
	T_{max} May				Yield
		Precip Apr	Precip	Wet day	
Sabzevar	$24 < T_{max}$ May < 32	7–81	223–286	46–64	High
	$28 < T_{max}$ May < 33	2–45	157–195	38–59	Mid
	$30 < T_{max}$ May < 35	0–44	57–150	23–51	Low

performed less well than GA and ACO, based on FF and SI values (Zhang et al. 2006; Kim and Ahn 2008). In this case, the AI methods were better at clustering accuracy of weather data.

The findings (Fig. 4 and Fig. 5) can allow us to describe the range of weather variables that result in different wheat yield categories, namely high, mid, and low yields, for Mashhad and Torbat H with a semi-arid climate, and Sabzevar with an arid environment. Table 6 shows the ranges of each variable. According to the results of the table, in Mashhad location, the domain changes of the T_{min} is 4 °C between 6 and 10, in different yield classes. In the Torbat H location, the amount of T_{max} April is greater than 16 °C in every yield class, whereas the number of wet days in Torbat H is greater than Mashhad station in every similar yield classes. Moreover, the amount of maximum total precipitation in Mashhad is more significant than Torbat H, in the high-yield category. In Sabzevar, the domain changes of T_{max} May is equal to 12 °C, and the total amount of precipitation over the study period in the high-yield class is less than the two other locations.

Conclusions

This study compared three (GA, ACO, and K-means) AI methods to cluster ten climatic attributes affecting wheat yield. Clustering based on K-mean was less reliable than the other applied methods. To determine the appropriate number of clusters, FF and silhouette index were implemented. Associations between ten climate attributes and wheat yield differed among the three study locations. Four climate attributes had an extensive association with yield, so we used four-dimensional clustering to describe the common characteristics of low-, medium-, and high-yielding years. The use of AI and its links with sciences like meteorology and agriculture opens new visions in related research to reduce errors and improve accuracy. In the future study, we aim to evaluate other variables that may affect wheat yield such as potential evapotranspiration and soil characteristics via clustering in different periods under several conditions and regions.

Acknowledgments We would like to thank K. Grace CRUMMER (Institute for Sustainable Food Systems, University of Florida, USA) for editing and improving the language of the manuscript.

Funding This study is supported by a grant from the Ferdowsi University of Mashhad, Iran.

References

Abdullah SS, Malek MA, Mustapha A, Aryanfar A (2014) Hybrid of artificial neural network-genetic algorithm for prediction of reference evapotranspiration (ET_0) in arid and semiarid regions. J Agric Sci:Published by Canadian Center of Science and Education 6(3): 191–200. https://doi.org/10.5539/jas.v6n3p191

Ahmed M, Hassan F (2011) Cumulative effect of temperature and solar radiation on wheat yield. Not Bot Horti Agrobo 39(2):146–152. https://doi.org/10.15835/nbha3925406

Ahmed M, Akram MN, Asimc M, Aslam M, Hassan F, Higgins S, Stöckle C, Hoogenboom G (2016) Calibration and validation of APSIM-wheat and CERES-wheat for spring wheat under rainfed conditions: models evaluation and application. Comput Electron Agric 123:384–401. https://doi.org/10.1016/j.compag.2016.03.015

Alvarez R (2009) Predicting average regional yield and production of wheat in the argentine pampas by an artificial neural network approach. Eur J Agron 30:70–77. https://doi.org/10.1016/j.eja.2008.07.005

Bannayan M, Sanjani S, Alizadeh A, Sadeghi Lotfabadi S, Mohamadian A (2010) Association between climate indices, aridity index, and rainfed crop yield in northeast of Iran. Field Crop Res 118:105–114. https://doi.org/10.1016/j.fcr.2010.04.011

Bannayan M, Lakzian A, Gorbanzadeh N, Roshani A (2011) Variability of growing season indices in northeast of Iran. Theor Appl Climatol 105:485–494. https://doi.org/10.1007/s00704-011-0404-1

Carvalho M, Serralheiro R, Corte-Real J, Valverde P (2015) Implications of climate variability and future trends on wheat production and crop technology adaptations in southern regions of Portugal. Water Utility Journal 9:13–18 http://hdl.handle.net/10174/14564

Celebi M, Kingravi H, Vela P (2013) A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst Appl 40:200–210. https://doi.org/10.1016/j.eswa.2012.07.021

Chen G, Liu H, Zhang J, Liu P, Dong S (2012) Factors affecting summer maize yield under climate change in Shandong Province in the Huanghuaihai Region of China. Int J Biometeorol 56:621–629. https://doi.org/10.1007/s00484-011-0460-3

Chou C, Chen C-A, Tan P-H, Chen KT (2012) Mechanisms for global warming impacts on precipitation frequency and intensity. J Clim 25(9):3291–3306. https://doi.org/10.1175/JCLI-D-11-00239.1

De Amorim RC (2016) A survey on feature weighting based K-means algorithms. J Classif 33:210–242. https://doi.org/10.1007/s00357-016-9208-4

De Amorim RC, Mirkin B (2012) Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering. Pattern Recogn 45(2012):1061–1075. https://doi.org/10.1016/j.patcog.2011.08.012

De Martonne E (1926) Une nouvelle function climatologique: L'indice d'aridité. La. Meteorologie 2:449–458

Deneubourg J-L, Gross S, Franks NR, Sendova-Franks A, Detrain C, Chretien L (1991) The dynamics of collective sorting: robot-like ants and ant-like robots. In: Meyer J-A, Wilson S (eds) Proc. The First International Conference on Simulation of Adaptive Behavior. From Animals to Animals J. MIT Press, Cambridge MA, pp 356–363

Dorigo M, Maniezzo V, Colorni A, 1991. The ant system: an autocatalytic optimizing process. Technical Report, Politecnico diMilano, Italy 91–106

Eyshi Rezaie E, Bannayan M (2012) Rainfed wheat yields under climate change in northeastern Iran. Meteorol Appl 19:346–354. https://doi.org/10.1002/met.268

FAO, Statistical Pocketbook (2015) Food and Agriculture Organization of the United Nations. Rome, Italy

Halkidi M, Batistakis Y, Vazirgiannis M (2001) 2001. On clustering validation techniques. Intell Inf Syst J 17(2–3):107–145

Handl J, Meyer B (2002) Improved ant-based clustering and sorting in a document retrieval interface, proceedings of the 7th International Conference on Parallel Problem Solving from Nature. LNCS 2439:913–923. https://doi.org/10.1007/3-540-45712-7_88

Hertz A, Kobler D (2000) A framework for the description of evolutionary algorithms. Eur J Oper Res 126(1):1–12. https://doi.org/10.1016/S0377-2217(99)00435-X

Jing-Song S, Guang-Sheng Z, Xing-Huaa S (2012) Climatic suitability of the distribution of the winter wheat cultivation zone in China. Eur J Agron 43:77–86. https://doi.org/10.1016/j.eja.2012.05.009

Kao YT, Zahara E, Kao IW (2008) A hybridized approach to data clustering. Expert Syst Appl 34(3):1754–1762. https://doi.org/10.1016/j.eswa.2007.01.028

Kettlewell PS, Sothern RB, Koukkari WL (1999) U.K. wheat quality and economic value are dependent on the North Atlantic Oscillation. J Cereal Sci 29:205–209 Article No. jcrs.1999.0258, available online at http://www.idealibrary.com

Kim KJ, Ahn H (2008) A recommender system using GA K-means clustering in an online shopping market. Expert Syst Appl 34:1200–1209. https://doi.org/10.1016/j.eswa.2006.12.025

Krishna K, Narasimha Murty M (1999) Genetic K-means algorithm. IEEE Trans Syst Man Cybernet B 29(3):433–439. https://doi.org/10.1109/3477.764879

Kuo RJ, Wang HS, Hu T-L, Chou SH (2005) Application of ant K-means on clustering analysis. Computers & Mathematics with Applications 50 (10-12):1709–1724

Laszlo M, Mukherjee S (2007) A genetic algorithm that exchanges neighboring centers for k-means clustering. Pattern Recogn Lett 28(16):2359–2366. https://doi.org/10.1016/j.patrec.2007.08.006

Li S, Wheeler T, Challinor A, Lind E, Ju H, Xu Y (2010) The observed relationships between wheat and climate in China. Agric Forest Meteor 150:1412–1419. https://doi.org/10.1016/j.agrformet.2010.07.003

Liu Y, Li Z, Xiong H, Gao X, Wu J (2010) Understanding of internal clustering validation measures. IEEE Int Conf Data Min 2010:911–916. https://doi.org/10.1109/ICDM.2010.35

Lobell DB, Field CB (2007) Global scale climate-crop yield relationships and the impact of recent warming. Environ Res Lett 2(1):1–7. https://doi.org/10.1088/1748-9326/2/1/014002

Luo QY, Bellotti W, Williams M, Bryan B (2005) The potential impact of climate change on wheat yield in South Australia. Agric For Meteorol 132(3–4):273–285. https://doi.org/10.1016/j.agrformet.2005.08.003

Machnik L (2006) ACO documents clustering—details of processing and results of experiments. Annales UMCS Informatica AI 5:279–289 http://www.annales.umcs.lublin.pl/

Ministry of Jihad-e-Agriculture (Iran). 2009. Crop statistics. [2009-04-03]. http://dpe.agri-jahad.ir/portal/File/ShowFile.aspx?ID=bd799699-4e89-437f-8a30-5e15a014d332. (In Persian)

Mirkin B, 2011. Choosing the number of clusters. John Wiley & Sons, Inc. WIREs Data Min Knowl Discov 1: 252–260. DOI:https://doi.org/10.1002/widm.15

Mualik U, Bandyopadhyay S (2000) Genetic algorithm based clustering technique. Pattern Recogn 33(9):1455–1465. https://doi.org/10.1016/S0031-3203(99)00137-5

Niknam T, Amiri B (2010) An efficient hybrid approach based on pso, aco, and k-means for cluster analysis. Appl Soft Comput 10(1):183–197. https://doi.org/10.1016/j.asoc.2009.07.001

Niknam T, Taherian Fard E, Pourjafarian N, Rousta A (2011) An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering. Eng Appl Artif Intell 24:306–317. https://doi.org/10.1016/j.engappai.2010.10.001

Olgun M, Okan Onarcan A, Özkan K, Isik S, Sezer O, Özgisi K, Gözde Ayter N, Budak Basçiftçi Z, Ardiç M, Koyuncu O (2016) Wheat grain classification by using dense SIFT features with SVM classifier. Comput Electron Agric 122:185–190. https://doi.org/10.1016/j.compag.2016.01.033

Rahimi J, Khalili A, Bazrafshan J (2014) Estimation of effective precipitation for winter wheat in different regions of Iran using an extended soil-water balance model. Desert. 19(2):91–98

Rendón E, Abundez I, Arizmendi A, Quiroz EM (2011) Internal versus external cluster validation indexes. IntJ Comp Commun 1(5):27–34

Romero JR, Roncallo PF, Akkiraju PC, Ponzoni I, Echenique VC, Carballido JA (2013) Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. Comput Electron Agric 96:173–179. https://doi.org/10.1016/j.compag.2013.05.006

Rostami Khaleghi M, Mohseni Saravi M, Hesami D, Rashidpour M, Salmani H (2014) Evaluation of groundwater quality in Mashhad city, using geostatistical methods in drought and wet periods. J Appl Hydrol 1(1):49–57

Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20(1):53–65

Salehnia N, Alizadeh A, Sanaeinejad H, Bannayan M, Zarrin A, Hoogenboom G (2017) Estimation of meteorological drought indices based on AgMERRA precipitation data and station-observed precipitation data. Journal of Arid Land 9(6):797–809. https://doi.org/10.1007/s40333-017-0070-y

Shelokar PS, Jayaraman VK, Kulkarni BD (2004) An ant colony approach for clustering. Anal Chim Acta 509:187–195. https://doi.org/10.1016/j.aca.2003.12.032

Sung CS, Jin HW (2000) A tabu-search-based heuristic for clustering. Pattern Recogn 33(5):849–858. https://doi.org/10.1016/S0031-3203(99)00090-4

Tabari H, Talaee PH (2011a) Analysis of trends in temperature data in arid and semi-arid regions of Iran. Glob Planet Chang 79:1–10. https://doi.org/10.1016/j.gloplacha.2011.07.008

Tabari H, Talaee PH (2011b) Temporal variability of precipitation over Iran: 1966–2005. J Hydrol 396(3):313–320. https://doi.org/10.1016/j.jhydrol.2010.11.034

Tabari H, Shifteh Somee B, Rezaeian Zadeh M (2011) Testing for long-term trends in climatic variables in Iran. Atmos Res 100(1):132–140. https://doi.org/10.1016/j.atmosres.2011.01.005

Tsai CF, Tsai CW, Wu HC, Yang T (2004) ACODF: a novel data clustering approach for data mining in large databases. J Syst Softw 73(1):133–145. https://doi.org/10.1016/S0164-1212(03)00216-4

USDA Foreign Agricultural Service. 2010. Iran: crop progress report. FAS—Office of Global Analysis (OGA), United States Department of Agriculture (USDA). International Operational Agriculture Monitoring Program. https://www.pecad.fas.usda.gov/pdfs/Iran/Iran_December_28_2009.pdf

Vardhan B, Ramesh D, Chander Goud O (2014) Density based clustering technique on crop yield prediction. Int J Electron Electr Eng 2(1): 56–59. https://doi.org/10.12720/ijeee.2.1.56-59

Wheeler TR, Craufurd PQ, Ellis RH, Porter JR, Vara Prasad PV (2000) Temperature variability and the annual yield of crops. Agric Ecosyst Environ 82:159–167. https://doi.org/10.1016/S0167-8809(00)00224-3

Xu S, Bing Z, Lina Y, Shanshan L, Lianru G, 2010. Hyperspectral image clustering using ant colony optimization (ACO) improved by K-means algorithm, 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)

Zhang X, Wang J, Wu F, Fan Z, Li X (2006) A novel spatial clustering with obstacles constraints based on genetic algorithms and K-medoids, Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications. IEEE. 1:605–610. https://doi.org/10.1109/ISDA.2006.75