Caspian Journal of Environmental Sciences
Online ISSN: 1735-3866
Print ISSN: 1735-3033

Uncovering the hidden patterns of fire risks: A cluster analysis approach (K-Medoids and FCM) for Hyrcanian Forest in Iran

Shaghayegh Zolghadri*[1], Mehrdad Ghodskhah Daryaei[1], Kamran Nasirahmadi[2], Esmaeil Ghajar[1]

1. Department of Forestry, Faculty of Natural Resources, University of Guilan, Iran
2. Assistant Professor, Faculty of Chemistry, University of Science and Technology, Mazandaran

* *Corresponding author's E-mail: Shaghayeghzolghadrii@gmail.com*

ABSTRACT

Forest fires have become a significant environmental concern in the Hyrcanian Forest, causing extensive loss of vegetation and posing a threat to biodiversity. Accurate prediction of high-risk fire locations is crucial for effective forest management. In this study, we developed and evaluated a clustering-based model using a multilayer perceptron artificial neural network with an error backpropagation training procedure to model fire risk potential in Saravan Forest, Guilan Province, North Iran. To optimize generalization, the model utilized two unsupervised clustering-specific procedures, namely Fuzzy C-Means and k-Medoids. The primary focus of our study was on the model's ability to predict potential fire risk locations, which is essential for forest fire prevention and control. The input criteria included recorded fire incidents, distances to farmland, roads, rivers, air pressure, solar radiation, slope, aspect, wind speed, and percentage of canopy cover density. The results showed that the procedure of the two algorithms used in this study in allocating potential fire hazard points is highly similar, differing mainly in the methodology employed for data center allocation. According to the results, the RMSE, R^2, and MSE for the model used in this study are respectively equal to 0.2861, 99.38, and 0.01919, which indicates the reliability of the model. Moreover, according to the Confusion matrix analysis table's results, FCM was slightly better than K-medoids in terms of its predictive accuracy. This model demonstrated high accuracy in predicting fire hazards, showing promising potential for forest fire prediction using clustering-based models. Additionally, our model exhibited superior performance compared to other clustering techniques for identifying potential fire hazard sites. Our developed clustering-based model provides valuable insights for forest managers to identify locations at fire risk, enabling more efficient resource allocation and preventative measures. This approach can significantly improve forest fire management and reduce ecological damage caused by wildfires.

Keywords: Forest fire modeling, Hyrcanian Forest, Unsupervised learning, Partitioning clustering.
Article type: Research Article.

INTRODUCTION

Artificial intelligence has been applied in wildfire science and management since the 1990s, with early applications including neural networks and expert systems. Since then, the field has rapidly progressed congruently with the wide adoption of machine learning (ML) methods in the environmental sciences (Jain *et al.* 2020). Forest fires are a major environmental concern, with devastating effects on the forest ecosystem. Climate change is one of the main drivers of forest fires in the 21st century, with rising temperatures, declining rainfall, and prolonged drought seasons exacerbating their incidence (Argañaraz *et al.* 2015; Littell *et al.* 2016). In addition to natural factors such as topography, biology, and climate, human activities like forest roads, settlements, agriculture, and recreation also contribute to the occurrence of forest fires (Eskandari 2015). Iran's forest areas, for instance, experience an annual destruction of over 5,000 ha due to forest fires (Adab Kanniah *et al.* 2013). The Saravan forests, as a part of the

Caspian Journal of Environmental Sciences, Vol. 23 No. 2 pp. 481-493 Received: Oct. 19, 2024 Revised: Jan. 16, 2025 Accepted: Feb. 21, 2025
DOI: 10.22124/CJES.2024.7977

Publisher: University of Guilan,

Hyrcanian forests inscribed as a UNESCO World Heritage Site located in Rasht, Guilan Province, North Iran represent a significant challenge in this regard. Despite recurring annual forest fires in the area and other parts of Iran, statistical reports indicate that insufficient control measures have been taken (Eskandari & Chuvieco 2015). Irrespective of the number of fires, the Global Forest Watch website registered the loss of 31 ha of forest in Rasht between 2001 and 2017, most of which were related to forest fires (Zarekar *et al.* 2013; Global Forest Watch 2018). The protection of Saravan forests is thus crucial, as it represents a global challenge. To mitigate the risks associated with forest fires, the development of effective fire management guidelines is fundamental. Machine learning-based simulations can aid in understanding the behavior and impact of forest fires. This study aims to investigate the ability of clustering algorithms, including FCM and K-medoids, to predict forest fire occurrence. Few studies have explored the application of these two algorithms in modeling forest fires, hence the focus of this study (Khatami *et al.* 2015; Jafarzadeh *et al.* 2017; Khatami *et al.* 2017; Giwa & Benkrid 2018). Based on previous studies, the factors considered in the model include three primary criteria, including environmental, climatic, and human factors, with ten sub-criteria. The model's parameters are selected according to established literature (Eskandari *et al.* 2013; Eskandari & Chuvieco 2015; Zhong *et al.* 2017; Tien Bui *et al.* 2018). The study hypothesizes that there is no significant difference between FCM and K-medoids algorithms' performance in predicting forest fires. The results of this study will pave the way for further research on clustering algorithms' application to improve forest fire prediction studies and better preparation for potential fires based on empirical evidence. In conclusion, effective strategies and advanced technologies are necessary to mitigate the threats posed by forest fires. Forest fires' devastating impact on ecosystems necessitates the development of robust fire management guidelines and machine learning-based simulations to comprehend their complex dynamics and design better approaches to manage them. By investigating the effectiveness of clustering algorithms in forest fire modeling, we can improve our preparedness and reduce the damage caused by forest fires.

MATERIALS AND METHODS

Sampling and analysis of samples

Saravan forest is located at 37°5′35″ north latitude and 49°24′29″ east longitude. It encompasses the outskirts of Rasht City and the Rudbar forestry area in Iran (Fig. 1). This area falls within the humid to very humid climatic zone. The characteristic soil profile of this region exhibits significant depth, well-developed horizons, a high organic matter content, and a neutral to acidic pH, spans over 8937 ha with an altitude ranging from 50 m to 600 m above sea level (Farahi *et al.* 2012). This forest is a crucial natural resource for the region, but it is under constant threat from forest fires. To better understanding the factors contributing to forest fires in this area, we developed a model using various data sources. Historic records of forest fire data were obtained from the Department of Natural Resources and Watershed Guilan, while meteorological data such as air pressure and wind direction were collected from the Guilan Meteorological Organization. These data sets covered the years 2007 to 2017. Additionally, using GPS from Guilan Environmental Protection Agency, we identified roads, rivers, and agricultural areas in the region. Despite the importance of this area, there was no suitable information layer available in the Environmental Protection and Natural Resources Organization of Guilan Province. As a result, all required layers were produced. Using ArcGIS and ENVI software, we created these layers and prepared the criteria that would affect the likelihood of forest fires. These criteria were then categorized into three parts: human factors (such as agriculture, roads, and villages), environmental factors (including slope, slope direction, and rivers), and climate factors (air pressure, radiation, and wind speed). To incorporate these criteria into our model, we converted the resulting maps of the study area into a matrix that could be inputted into MATLAB software. Despite the challenges faced during data collection and processing, we believe that our model will provide valuable insights into the factors that contribute to forest fires in the Saravan forest and aid in developing strategies for its conservation and protection. In the first stage of data mining, the correlation coefficient between the assumed features was extracted. The Pearson correlation coefficient was used in this study, which measures the relationship between two quantitative variables. The coefficient value ranges from 1 to -1, with higher values indicating a stronger correlation between the variables being considered. A value close to 1 or -1 is indicative of a high correlation in the positive or negative direction, respectively. Meanwhile, a value close to 0 implies that there is no significant correlation between the variables. The Pearson correlation coefficient has been widely utilized in different fields, including finance, economics, and social sciences (Hunt 1986). By quantifying the extent of the relationship between different variables, it becomes possible to identify the most relevant features for inclusion in

the model and exclude redundant information. Thus, this coefficient serves as a crucial tool for effective data preprocessing in machine learning applications.

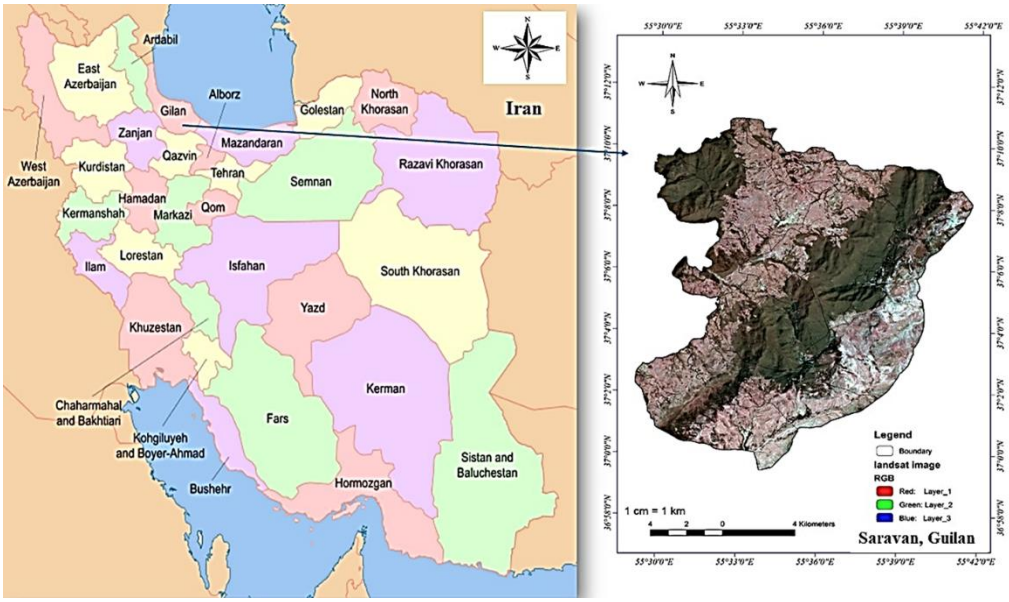


Fig. 1. Map of sampled stations.

In summary, through the use of the Pearson correlation coefficient, we can gain valuable insights into the strength and nature of the relationship between different variables and improve the overall accuracy of the model, as follows:

$$P(x, y) = r(xy)\,(\mathrm{cov}(x, y))/(\sigma x\,\sigma y) \quad \textbf{(1)}$$

where: cov is the covariance, σx is the standard deviation X and σy is the standard deviation of Y. Then, two clustering methods - Fuzzy C-mean and K-medoids - were employed to model fire risk prediction in the Saravan region, using the inputs identified in earlier stages. Before this, however, it was necessary to normalize the input data due to the heterogeneous nature of the variables considered. In cases where there is a large discrepancy between the maximum and minimum values of the data, distance criteria can become inefficient in clustering algorithms, and the size of the data can have a significant impact on overall accuracy. Given that our input criteria consisted of variables with different definitions and widely varying intervals, pre-processing was deemed essential. To achieve this, we utilized the base 10 logarithms to make the data more homogeneous across a specific range. The polyfit command was then applied to standardize all the data within the range of 0 and 1, which was completed through the use of the polyval command or linear mapping, a function that preserves the actions of addition and scalar multiplication. All of these operations were carried out within the MATLAB (version R2019a) software environment, providing a reliable platform for pre-processing the data before clustering. This step was crucial in ensuring the accuracy and effectiveness of the subsequent clustering methods employed. By utilizing both Fuzzy C-mean and K-medoid clustering techniques, we were able to model fire risk prediction in the Saravan region with an improved level of accuracy, benefiting from the standardized input data processed using the methods outlined above. In the following stage of this study, clustering learning methods were employed to effectively model fire risk potential. This method represents one branch of multivariate statistical analysis and unsupervised learning in artificial neural networks. One of the primary benefits of utilizing these methods is that it allows for learning without pre-existing labels and minimal human supervision. The process of clustering divides society into several sub-communities, known as clusters, where samples are classified based on their similarity to others within the same cluster, while also being dissimilar to those outside the cluster (Kaufman *et al.* 2005; Karimov *et al.* 2015; Littell *et al.* 2016). These clusters are classified based on the relationships between them, a crucial aspect of clustering analysis (Jiawei *et al.* 2001). The multilayer perceptron (MLP) model of artificial neural network was utilized in conjunction with an error backpropagation training procedure to model fire risk potential effectively. In spatial analysis, MLPs, self-organizing maps (SOMs), and radial basis functions (RBFs) have been primarily relied upon due to their practical capabilities. Among these options, MLPs remain a popular choice. Our ANN model

consisted of three layers, with 10 input nodes, five nodes in the middle layer, and one node in the output layer. To train the ANN model effectively, we utilized both unsupervised clustering methods, such as the fuzzy c-means (FCM) algorithm, and supervised algorithms, like the k-medoids algorithm. This approach enabled us to leverage the strengths of both methodologies to produce more accurate and reliable models for predicting fire risk potential in Saravan region.

K-Medoids algorithm

The study utilizes an improved k-medoids algorithm as the first clustering technique. This algorithm is similar to the k-means algorithm, but instead of using the mean, it uses the actual sample itself as a representation of the cluster. The main objective of the algorithm is to minimize the sum of differences between the points in a cluster and the center point of the cluster. Each medoid represents the most central data point of a cluster. Noteworthy, this algorithm is generally less sensitive to data outside the cluster when compared to other clustering algorithms (Krishnapuram *et al.* 1999). To implement the k-medoids algorithm, we first randomly select the initial representatives for the k clusters. We then identify the nearest representative for each sample and create a similarity matrix (n-k). Samples are then grouped into one of the k clusters. To assess the quality of the clusters obtained, we replace a sample with one of the representative samples and calculate the cost of this replacement. If the cost is negative, meaning the new assignment is a better fit, the transfer takes place. This step is repeated until the center point of the clusters remains constant in two consecutive iterations (Kaufman & Rousseeuw 2005). The PAM (Partitioning Around Medoids) algorithm is a widely used and effective clustering method in K-medoids analysis. However, its high iteration requirement makes it unsuitable for large datasets (Kaufman & Rousseeuw 2005). To address this issue, K-medoid clustering also includes CLARA and CLARANS algorithms. CLARA is designed specifically for larger datasets. It randomly selects samples from the dataset and applies the PAM algorithm to them, generating the best clustering output. The rest of the database components are then assigned to their nearest cluster based on the generated output (Wei *et al.* 2000). In contrast, CLARANS is based on off-center data. In this study, we employed the PAM algorithm as it is one of the most commonly used and powerful techniques in K-medoids clustering. When calculating data spacing, the Euclidean distance criterion was applied, which is a standard approach in K-medoids clustering (Dunn 1973; Krishnapuram *et al.* 1999; Nayak *et al.* 2015). Specifically, we calculated and stored the distances between each datum and its corresponding center in a designated variable, consistent with established practice in clustering research.

FCM algorithm

The current study employs the Fuzzy c-means algorithm as a second approach. This algorithm employs a membership function to assign the degree of membership to each data point for each cluster. The algorithm then iteratively updates the membership values and cluster centroids until convergence, resulting in the final clusters. The FCM algorithm is particularly suitable for cases where the boundaries between clusters are not well-defined, as it allows for overlapping clusters and soft assignments of data points to multiple clusters (Bezdek 1981). The Fuzzy c-means algorithm follows these steps:

Random initialization of cluster membership values μij.

Calculation of the cluster centers.

Updating the membership values based on the following equation:

Calculation of the objective function Jm.

Repeating steps 2-4 until the value of Jm stops improving beyond a specified threshold or after reaching a predetermined number of iterations.

For this particular study, five clusters and corresponding centers were selected, and a threshold level of 1 km + 0.5 was employed. To compare the performance of both algorithms used in the research, a confusion matrix was used (Fawcett 2006).

Model validation

To assess the accuracy of our model, we employed two widely-used statistical measures: the root means square error (RMSE), and the value of the coefficient of determination (R^2; Shahin *et al.* 2008). R^2 value indicates how much variation in the dependent variable can be accounted for by changes in the independent variable(s), while RMSE quantifies the average error between predicted and observed values. In other words, R^2 determines what percentage of changes in the dependent variable are affected by the corresponding independent variable, and

RMSE compares the prediction errors of a data set. The higher the RMSE and R^2 values, the better the data fit together. The formulas RMSE and R^2 are as follows (Lewis-Beck & Skalaban 1990; Chai & Draxler 2014).

$$R^2 = \sum_{i=1}^{n}(O_i - \overline{O}_i)(P_i - \overline{P}_i)^2 / (P_i - \overline{P}_i)^2 \, (P_i - \overline{P}_i)^2 \quad \textbf{(2)}$$

$$\sqrt{\sum_{i=1}^{n}(O_i - P_i)/n} \quad \textbf{(3)}$$

where Oi is the observational data, Pi is the simulated data, P is the simulated data by the model, and n is the number of data.

The Mean Squared Error (MSE) is a commonly used metric for measuring the accuracy of a regression model. It is calculated as the average squared difference between the predicted and actual values in the test dataset. The formula for calculating MSE is as follows (Mood *et al.* 2013):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \textbf{(4)}$$

where the averaging operation is performed with $\sum 2i = 1$ and $(yi\text{-}y\tilde{i})^2$ calculates the square value of the error of each data (Mohamed *et al.* 2008). Eventually, to objectively compare the performance of two algorithms used for cluster selection in the clustering method, a Confusion Matrix was employed. This evaluation technique, introduced by Fawcett (2006), provides a tabular representation of the true positive, false positive, true negative, and false negative rates of the algorithms, allowing for an in-depth analysis of their effectiveness. By assessing metrics such as accuracy, precision, recall, and F_1 score derived from the confusion matrix, one can gain insights into the suitability of each algorithm for generating accurate and meaningful clusters. Thus, this approach facilitates an unbiased comparison of the algorithms and enables the selection of the optimal one for the specific task at hand.

RESULTS

Fig. 2 presents a map of various environmental variables that were analyzed in the study area. The map is divided into ten layers, each representing a different variable. The first layer, labelled "Aspect," indicates the direction in which each portion of the landscape is facing. The second, "Slope," shows the angle of the terrain. The third, "Wind Speed," displays the average wind speed at each location.

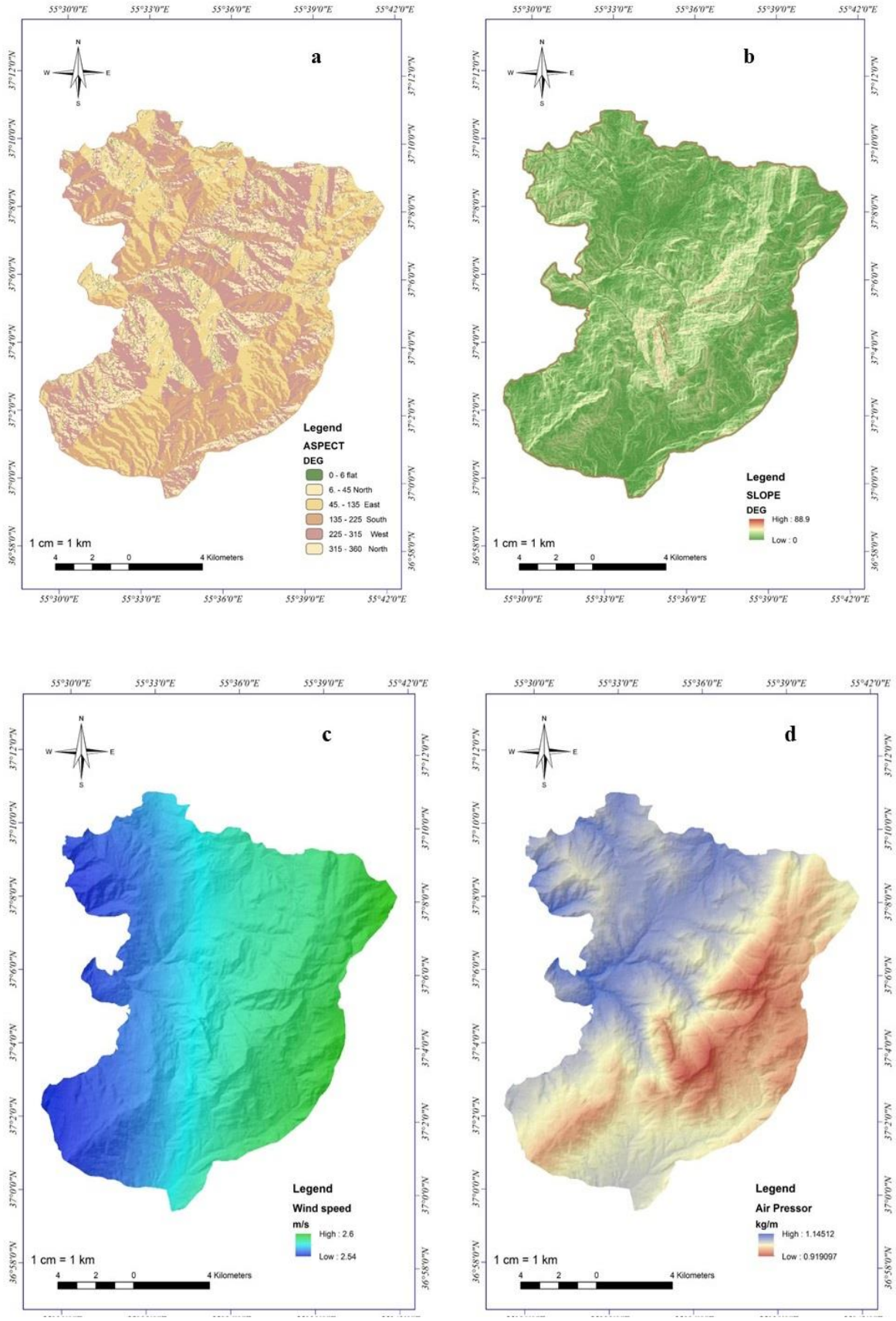
The fourth, "Air Pressure," indicates the air pressure at each point on the map. The fifth, sixth, and seventh layers exhibit the distance to the nearest road, river, and agricultural area, respectively. The eighth, "Solar Radiation," displays the amount of solar radiation that each location receives. The ninth, "Canopy Cover Density," indicates the amount of vegetation cover at each location. Finally, the tenth, "Forest Type," shows the different types of forests in the study area. The present study outlines the findings of our investigation on fire estimation models. Our analysis of the correlation coefficient revealed a relatively high degree of similarity between forest type and canopy cover density, as well as between air pressure, distance to road, and forest type. This finding is not surprising given the strong influence of forest texture and air pressure on climate. The correlation coefficient analysis table enables designers and analysts to streamline their decision-making process by identifying and eliminating variables that exhibit significant redundancy. In Table 1, the primary diameter is marked with "1", while highly similar variables are denoted by an absence of a rounded square. Overall, our findings suggest that data collection was appropriately conducted, as evidenced by the relatively low degree of similarity in correlation coefficients across various variables.

General procedure for allocating potential fire hazard points in K-Medoids and FCM

The steps taken in both the K-medoids and FCM algorithms are highly similar, differing mainly in the methodology employed for data center allocation. Specifically, while the K-medoids algorithm utilizes medoid data centers, the FCM algorithm employs a centroid data center evaluated fuzzily (Wei *et al.* 2000). In both cases, we have five clusters and corresponding cluster centuries, with a threshold level of 1 km deemed desirable for each cluster (as demonstrated in Tables 2-3). Both algorithms evaluate data based on its distance from the center, with the distance between each data point and its corresponding center calculated accordingly.

By assuming a maximum of 15% of the maximum distance between the accident point and the center of its corresponding cluster, the high probability space for fire can be determined. Adjusting this distance up or down will correspondingly decrease or increase the likelihood of fire. Noteworthy, the aforementioned approaches rely heavily on distance calculations to determine the probability space for fire. As such, fine-tuning the maximum

allowable distance between the accident point and the cluster center through further experimentation may refine the accuracy of the resulting predictions.



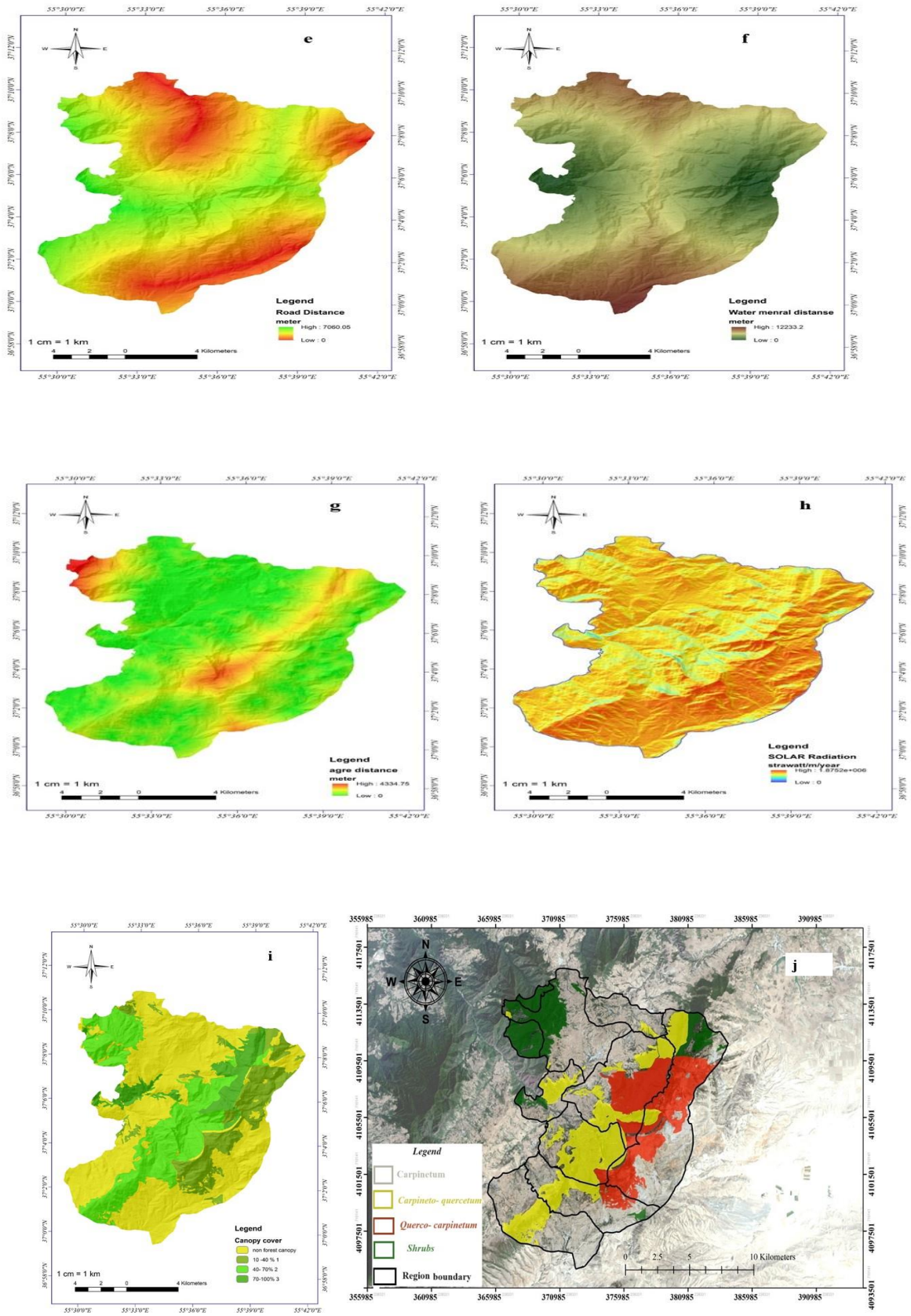


Fig. 2. Map of A) Aspect, B) Slope, C) wind speed, D) Air pressure, E) Distance to the road, F) Distance, to the river, G) Distance to agriculture, H) Solar radiation, I) Canopy cover density, J) forest type in the study.

Table 1. Pearson correlation coefficient analysis table.

	Percentage of canopy density	**Aspect**	**Air pressure**	**Distance to agriculture**	**Distance to rivers**	**Distance to roads**	**slope**	**Solar radiation**	**Forest type**	**Wind speed**
Percentage of canopy cover density	**1.0000**	0.1416	0.5139	0.0506	0.1775	0.5424	-0.0679	-0.3392	0.8787	-0.1179
Aspect	0.1416	**1.0000**	-0.0251	0.0135	-0.0306	0.0588	-0.0859	-0.0101	0.1526	-0.0350
Air pressure	0.5139	-0.0251	**1.0000**	0.1711	0.0353	0.6146	0.0207	-0.3929	0.6121	-0.2598
Distance to agriculture	0.0506	0.0135	0.1711	**1.0000**	-0.0025	0.1246	-0.0791	0.0818	0.0643	0.1795
Distance to rivers	0.1775	-0.0306	0.0353	-0.0025	**1.0000**	-0.0572	-0.3865	0.1266	-0.0742	0.0637
Distance to roads	0.5424	0.0588	0.6146	0.1246	-0.0572	**1.0000**	0.0644	-0.3269	0.7234	-0.2613
Slope	-0.0679	-0.0859	0.0207	-0.0791	-0.3865	0.0644	**1.0000**	-0.4828	0.0941	-0.0723
Solar radiation	-0.3392	-0.0101	-0.3929	0.0818	0.1266	-0.3269	-0.4828	**1.0000**	-0.4578	0.2057
Forest type	0.8787	0.1526	0.6121	0.0643	-0.0742	0.7234	0.0941	-0.4578	**1.0000**	-0.2779

Validation

According to the results, the RMSE, R^2, and MSE for the model used in this study are respectively equal to 0.2861, 99.38, and 0.01919, which indicates the reliability of the model.

Table 2. Clustering of studied parameters by K-Medoids.

Threshold level	**Centre**	**percentage of canopy**	**Aspect**	**Air pressure**	**Distance to Agriculture**	**Distance to river**	**Distance to road**	**Slope**	**Solar radiation**	**Forest type**	**Wind speed**
0.802	1	0.000	0.473	0.329	0.460	0.443	0.281	0.807	0.952	0.000	0.229
0.733	2	0.667	0.598	0.617	0.703	0.798	0.990	0.469	0.813	1.000	0.307
0864	3	0.802	0.767	0.189	0.677	0.692	0.508	0.651	0.703	1.000	0.776
0641	4	0.667	0.480	0.719	0.630	0.620	0.882	0.823	0.257	1.000	0.623
1.299	5	0.333	0.598	0.309	0.700	0.540	0.432	0.515	0.816	0.333	0.802

Map of clustering with FCM and K-Medoids

In this step, the fire-prone areas maps based on FCM and K-Medoids for forecasting fire-prone areas wereproduced (Fig. 3). The results of comparing FCM and K-Medoids are shown in Table 4. The blue columns correspond to the prediction similarity percentage of both algorithms. Orange cells are cases of fire by the FCM algorithm that were not confirmed by the K-medoids algorithm.

Table 3. Clustering of studied parameters by FCM.

Threshold level	Centre	percentage of canopy	Aspect	Air pressure	Distance to Agriculture	Distance to river	Distance to road	Slope	Solar radiation	Forest type	Wind speed
0.713	1	0.329	0.593	0.692	0.624	0.517	0.684	0.786	0.341	0.000	0.716
1.324	2	0.722	0.589	0.707	0.685	0.829	0.587	0.705	0.923	1.000	0.501
1.087	3	0.128	0.329	0.647	0.639	0.533	0.684	0.783	0.144	0.000	0.658
0.812	4	0.052	0.228	0.627	0.681	0.532	0.682	0.809	0.059	1.000	0.700
0.630	5	0.660	0.688	0.673	0.591	0.888	0.777	0.425	0.961	0.333	0.499

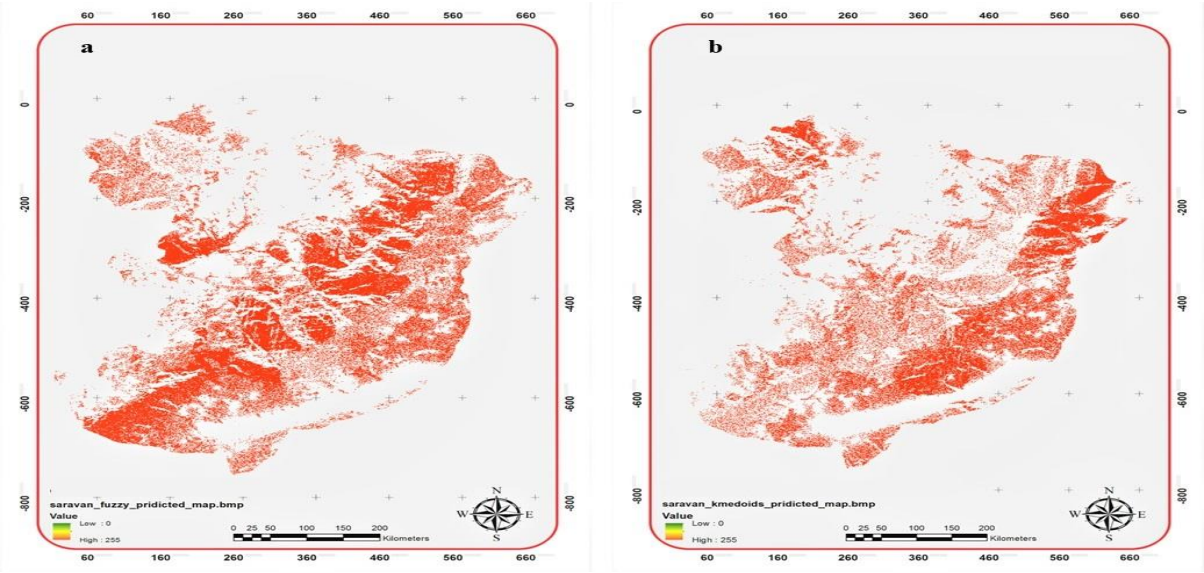
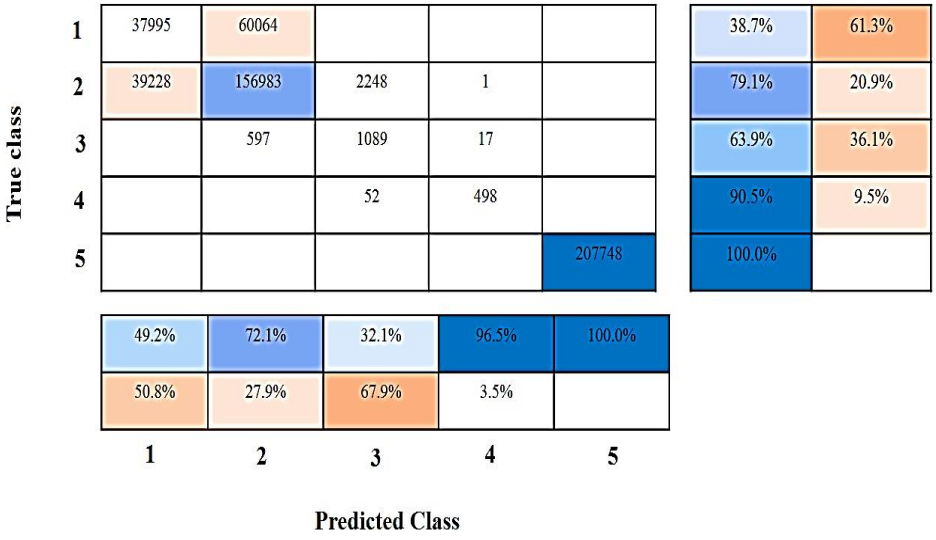


Fig. 3. Final map for forecasting fire-prone areas with a) FCM, b) K-Medoids.

Table 4. Confusion matrix analysis table for FCM and K-medoids.



DISCUSSION

The Saravan Forest Park in Rasht, Guilan Province, represents a vital old-growth forest region renowned for its rich biodiversity and unique plant species. Despite its ecological significance, the area has been subject to several

challenges, including being utilized as a landfill site by neighbouring cities (Piruz *et al.* 2010), clearcutting activities aimed at making room for power towers (Yaghmaeiyan Mahabadi *et al.* 2017), and the proliferation of dense shrubs - all of which have contributed to an alarming frequency of fires each year (Frahi *et al.* 2012). This study compared the performance of two clustering algorithms, Fuzzy C-Means and K-medoids, in modelling fire distribution within Saravan Forest Park. The purpose was to identify high-risk areas for wildfire. The findings showed that clustering-based methods are essential for improving forest fire prediction models. Specific input criteria such as recorded fire locations, distance from farmland, road proximity, river proximity, air pressure, solar radiation, slope, aspect, wind speed, and percentage of canopy cover density are key predictors of fire risk. Incorporating these variables increases the accuracy of the model, resulting in more effective predictions. This study highlights the importance of using advanced analytical tools and relevant input criteria to predict and manage wildfires in ecologically sensitive regions such as Saravan Forest Park. Moreover, this study showed the issue of forest fires and the importance of using advanced analytical tools and relevant input criteria to predict and manage wildfires in ecologically-sensitive regions. Our findings align with the results of another authors (Sathishkumar *et al.* 2023).

The effective prediction and management of forest fires are critical challenges for forest managers worldwide. This study highlights the potential of machine learning techniques, particularly clustering-based methods, in addressing these challenges. Specifically, the successful application of such methods in Hyrcanian Forests underscores their practical utility. Two algorithms were employed in this research to analyse fire probability within the forest. The resulting matrix analysis table revealed five distinct classes of fire probability. The first class, called the hot spot floor, includes areas located within 1 km of the nearest cluster centre. The second, identified as the first-class high-risk floor, corresponds to areas situated between 1-5 km from the nearest cluster centre. The third, known as the second-class high-risk floor, encompasses areas located 5-10 km from the cluster centre. The fourth, referred to as the first-class low-risk floor, covers areas positioned from 10-50 km from the centre of the cluster. Finally, the fifth, designated as the second-class low-risk floor, consists of areas situated between 50-255 km from the cluster centre. By identifying these five classes of fire probability, this study offers valuable insights to forest managers, enabling them to develop effective strategies in order to prevent or mitigate forest fires. The current study utilized a confusion matrix to evaluate the performances of two algorithms for predicting forest fires. The results were presented in three tables: (i) a large table containing pixel information for the study area, (ii) a right-hand table displaying rows of pixel surfaces for the study area using the FCM algorithm, and (iii) a bottom table showing pixel levels in the study area with priority given to the K-medoids algorithm. Using linear analysis, blue columns were created to represent the percentage of similarities in fire predictions made by both algorithms. The analysis revealed that 38.7% of pixels fell under the hotspot category in both algorithms, while 79.1%, 63.9%, 96.5%, and 100% of pixels were categorized as first-degree high-risk, second-class high-risk, first-class low-risk, and second-class low-risk, respectively. The orange cells present in the table indicate fire hazard predictions made by the FCM algorithm that was not confirmed by the K-medoids algorithm. These discrepancies reflect differences between the two models. In particular, 61.3% of pixels were classified as hotspots by the FCM algorithm, while 20.9%, 36.1%, and 9.5% were identified as first-class high-risk, second-class high-risk, and first-class low-risk categories, respectively. A column analysis was also performed with a priority given to the K-medoids algorithm. This analysis demonstrated that 49.2% of pixels were identified as hotspots by both algorithms, while 72.1%, 32.1%, 96.5%, and 100% of pixels were categorized as first-class high-risk, second-class high-risk, first-class low-risk, and second-class low-risk, respectively (as indicated in blue). Orange cells in this table represent fire hazard predictions made only by the K-medoids algorithm that where not confirmed by the FCM algorithm. Specifically, 50.8% of pixels were categorized as hotspots, while 27.9%, 67.9%, and 3.5% fell under the high-risk, second-class high-risk, and low-risk categories, respectively. In this study, we compared the performance of two algorithms in predicting fire risk locations. Observations revealed an elevation in commonality in higher classes of analysis for both row and column analyses of the study area, suggesting a decrease in algorithm sensitivity as one move away from the clusters' centres. Overall, the fuzzy algorithm demonstrated marginally superior performance over the K-medoids algorithm, with an overall subscription rate of 372.2% compared to 349% for the K-medoids algorithm. These findings are consistent with previous studies demonstrating the effectiveness of clustering algorithms in wildfire risk modelling. For instance, Bharany *et al.* (2022) reported that clustering algorithms exhibited good performance in predicting wildfire risk. Moreover, studies applying the FCM algorithm to cluster data in neural network training have shown its effectiveness in

improving the input-output relationship, thus increasing the likelihood of predictive accuracies. For example, Xu & Wunsch (2005) and Esakar & Chaudhari (2013) reported that the FCM algorithm was effective in improving the accuracy of predictive models. Rakshit *et al.* (2021) in their study, used the machine learning method to predict the risk of forest fires. They focused on using meteorological data while their paper considered a broader range of factors such as distances to farmland, roads, and rivers, air pressure, solar radiation, slope, aspect, wind speed, and percentage of canopy cover density. Additionally, the other paper aimed to predict the depth of risk for specific areas, while their paper focuses on identifying locations at risk of fire. In one of the most recent articles, Sathishkumar *et al.* (2023) similar to our study, used the machine learning method to predict the risk of forest fires. They used different methodologies to address different forest fire-related challenges. Also, in contrast to their result, our study demonstrates high accuracy in predicting fire hazards and exhibits superior performance compared to other clustering techniques for identifying potential fire hazard sites. However, their results are consistent with our results in showing the potential of machine learning algorithms to improve forest fire management and reduce the environmental damage caused by forest fires.

CONCLUSIONS

The increasing occurrence of wildfires has become a global concern, and the Saravan Forest Park in Guilan Provine, North Iran has also experienced multiple fire outbreaks. To address this issue, the study aimed to understand the causes of wildfires in the area and develop models using clustering algorithms for assessing fire risks. The findings of the study suggest that it is necessary to have a good understanding of the reasons behind the occurrence of wildfires to devise effective prevention strategies (Rakshit *et al.* 2021; Shreya *et al.* 2022). In addition, the study's framework can be used as a prototype model that can be customized by changing input parameters and algorithms. This approach allows fire prediction models to be tailored to specific regions where wildfire incidence is on the rise. The study evaluated two clustering algorithms, Fuzzy C-means (FCM) and K-medoids, for their ability to identify high-risk areas. The results indicate that both algorithms can predict high-risk areas effectively. However, FCM was slightly better than K-medoids in terms of its predictive accuracy. Noteworthy, the accuracy of clustering algorithms drops, by elevation in the distance from the fire cluster centre. Overall, the study highlights the potential of clustering algorithms in predicting fire risks and provides useful insights into managing fire hazards. The findings offer relevant information that could inform land management policies, such as prescribed fires and resource allocation for firefighting activities. The implications of this study go beyond the Saravan Forest Park and contribute to the broader field of forest fire management. Future research aimed at enhancing the accuracy and efficiency of wildfire risk models could build on the study's findings. This study provides valuable insights into the effectiveness of clustering algorithms in wildfire risk modelling. However, several limitations should be considered when interpreting the findings. First, our analysis was limited to the Saravan Forest Park in Guilan Province, North Iran, and our results may not be generalizable to other regions with different climates, ecosystems or topographical features. Thus, further research is required to determine the extent to which these algorithms can be applied in other settings. Second, we only evaluated the performance of two clustering algorithms and did not compare them with other modelling approaches, such as machine learning or deep learning algorithms. Future research could expand on these findings by comparing clustering algorithms with other machine learning methods and evaluating their potential to predict wildfire risk locations. Third, further research is needed to determine the most effective methods for validating predictive models, such as the quantification of predictive probability accuracy. Future studies could employ ground-truthing exercises to validate the predictive power of these models. Ground-truthing involves collecting data from the forest floor, such as leaf litter depth, fuel load, and vegetation density, to verify the accuracy of the predictive models. This approach has been employed in previous studies and could offer a reliable validation method for wildfire risk prediction models. Despite these limitations, this study provides an important contribution to the field of wildfire risk modelling and management. By demonstrating the potential of clustering algorithms such as FCM and K-medoids, this study offers promising avenues for developing proactive strategies to mitigate the risk of catastrophic wildfires. Future research could build on these findings by exploring other clustering algorithms or comparing clustering with other machine learning and deep learning models.

ACKNOWLEDGEMENTS

We thank Dr Abdul Reza Alavi Qarebagh for his assistance and comments that improved the paper.

REFERENCES

Adab, H Kasturi Kanniah, KD & Solaimani, K 2013, Modeling forest fire risk in the northeast of Iran using remote sensing and GIS techniques. *Natural Hazards* (*Dordr*) 65: 1723-1743, DOI:10.1007/s11069-012-0450-8.

Argañaraz, JP Pizarro, GG Zak, M Landi, MA & Bellis, LM 2015, Human and Biophysical Drivers of Fires in Semiarid Chaco Mountains of Central Argentina. Sci. *Total Environment*, 520: 1-12, https://doi.org/10.1016/j.scitotenv.2015.02.081.

Bharany, S Sharma, S Frnda, J Shuaib, M Khslid, MI Hussain, S Iqbal, J & Uliah, SS 2022, Wildfire monitoring based on energy efficient clustering approach for FANETS, *Drones*, 6: 193.

Bezdek, J 1981, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, DOI: 10.1007/978-1-4757-0450-1.

Chai, T & Draxler, RR 2014, Root Mean Square Error (RMSE) Or Mean Absolute Error (MAE)? *Geoscientific Model Development (GMD) & Discussions*, 7: 1525-1534, DOI: 10.5194/gmdd-7-1525-2014.

Dunn, JC 1973, A Fuzzy Relative of the ISODATA Process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3: 32-57, DOI: 10.1080/01969727308546046.

Esakar, S & Chaudhari, M 2013, A review of clustering algorithms, www.ijcst.com 4.

Eskandari, S 2015, Investigation of relation between climate change and fire in the forests of Golestan Province, IJFRPR. 13.

Eskandari, S & Chuvieco, E 2015, Fire danger assessment in Iran based on geospatial information. *International Journal of Applied Earth Observation and Geoinformation*, 42: 57-64, DOI: 10.1016/j.jag.2015.05.006.

Eskandari, S Oladi, J Jalilvand, H & Saradjian, MR 2013, Role of human factors on fire occurrence in district three of Neka Zalemroud Forests, Iran, *World Applied Sciences Journal*, 27, DOI: 10.5829/idosi.wasj.2013.27.09.708.

Frahi, E Ghodskhahdaryaei, M Mohamadi Samani, K & Amlashi, M 2012, Review of fire sensitive areas with emphasis on drought impact with the joint use of DSI, AHP and GIS (Case study: Forest Saravan, Guilan Province), *Forest and Range Protection Research*. 10: 83-101, https://www.sid.ir/en/Journal/ViewPaper.aspx?ID=529534.

Fawcett, T 2006, An introduction to ROC analysis, *Pattern Recognition Letters*, 27 (8): 861-74, DOI: 10.1016/j.patrec.2005.10.010.

Giwa, O & Abdsamad, B 2018, Fire detection in a still image using color information. https://doi.org/10.48550/arXiv.1803.03828.

Global Forest Watch 2018 Tree cover loss in Rasht, Guilan, Iran, https://www.globalforestwatch.org.

Hunt, RJ 1986, Percent agreement, Pearson's correlation, and Kappa as Measures of Inter-Examiner reliability, *Journal of Dental Research*, 65: 128-30, DOI: 10.1177/00220345860650020701.

Jafarzadeh, A Mahdavi, A & Jafarzadeh, H 2017, Evaluation of forest fire risk using the Apriori Algorithm and Fuzzy C-Means Clustering, DOI: 10.17221/7/2017-JFS.

Jain, P Cogan, SCP Subramanian, SG Crowley Taylor, S & Flannigan, MD 2020, A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28: 478-505. https://doi.org/10.1139/er-2020-0019.

Jiawei, H Kamber, M & Tung, A 2001, Spatial clustering methods in data mining: A survey. *Data Mining and Knowledge Discovery*, *DATAMINE*.

Karimov, J Ozbayoglu, M & Dogdu, N 2015, K-means performance improvements with centroid calculation heuristics both for serial and parallel environments. In: 2015 IEEE International Congress on Big Data, 444–451, DOI: 10.1109/BigDataCongress.2015.72.

Kaufman, L & Rousseeuw, PJ 2005, Finding groups in data: An Introduction to cluster analysis. Wiley series in probability and mathematical statistics, Hoboken, NJ: Wiley-Intercedence. http://catdir.loc.gov/catdir/description/wiley033/89031460.html.

Khatami, A Mirghasemi, S Khosravi, A Lim, CP & Nahavandi, S 2017, A new PSO-Based approach to fire flame detection using K-Medoids Clustering, *Expert Systems with Applications*, 68: 69-80, DOI: 10.1016/j.eswa.2016.09.021.

Khatami, A Mirghasemi, S Khosravi, A & Nahavandi, S 2015, An efficient hybrid algorithm for fire flame detection. In 2015 International Joint Conference on Neural Networks (IJCNN), edited by IEEE Staff, 1-6, Piscataway: IEEE.

Krishnapuram, R Joshi, A & Liyu, Y 1999, A Fuzzy Relative of the K-Medoids algorithm with application to web document and Snippet Clustering. In FUZZ-IEEE'99, 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No.99CH36315), 1281-1286 Vol. 33.

Lewis-Beck, MS & Skalaban, A 1990, The R-squared: Some straight talk. *Political Analysis*, 2: 153-171, DOI:10.1093/pan/2.1.153.

Littell, JDL Peterson, Riley, KL Yongquiang, L & Luce, CH 2016, A review of the relationships between drought and forest fire in the United States. *Global Change Biology*, 22 (7): 2353-2369.

Mood, Al Franklin, M Graybill, A & Duane, CB 2013, Introduction to the theory of statistics. 3. ed., McGraw Hill Education (India) ed., 13. reprint. New Delhi: McGraw-Hill Education (India).

Nayak, J Naik, B & Behera, HS 2015, Fuzzy C-Means (FCM) clustering algorithm: A decade review from 2000 to 2014, In: *Computational Intelligence in Data Mining*, 2: 133-149: Springer, New Delhi. https://link.springer.com/chapter/10.1007/978-81-322-2208-8_14.

Mohamed, SH Jaksa, M & Maier, H 2008, State of the art of artificial neural networks in geotechnical engineering, *Electronic Journal of Geotechnical Engineering.*

Piruz, B Razdar, B Bagherzadeh, A & Kavianpour, M 2010, Assessment of the damage caused by the dumping of waste from Rasht city in Saravan Forest area located in Gilan Province, National Conference on Man, Environment and Sustainable Development.

Rakshit, P Sarkar, S Khan, S Saha, P Bhattacharyya, S Dey, ... & Pal, S 2021, Prediction of forest fire using machine learning algorithms: The search for the better algorithm. In 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA), pp. 1-6. IEEE.

Sathishkumar, VE, Cho, J Subramanian, M & Naren, O 2023, A forest fire and smoke detection using deep learning-based learning without forgetting. *Fire Ecology*, 19: 1-17.

Shahin, MA Jaksa, MB Holger, MR 2008, State of the art of artificial neural networks in geotechnical engineering, *Electronic Journal of Geotechnical Engineering*, 8: 1-26.

Shreya, M Rai, R & Shukla, S 2022, Forest fire prediction using machine learning and deep learning techniques. In: Computer networks and inventive communication technologies: Proceedings of Fifth ICCNCT 2022, pp. 683-694, Singapore: Springer Nature Singapore.

Tien Bui, D van Le, H & Hoang, ND 2018, GIS-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method ecological informatics, 48: 104-116, DOI: 10.1016/j.ecoinf.2018.08.008.

Wei, Ch P Lee, YH Hsu Che, M 2000, Empirical comparison of fast clustering algorithms for large data sets. In Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, edited by Ralph H. Sprague, 10: IEEE Computer Society, DOI:10.1109/HICSS.2000.926655.

Xu, R & Wunsch, D 2005, Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16: 645-678. DOI: 10.1109/TNN.2005.845141.

Yaghameiyan mahabadi, N Khosroabadi, M & Asadi, H 2017, The effect of afforestation and topography on some physicochemical characteristics affecting soil quality in Saravan region of Gilan. *Soil Research (Soil and Water Sciences)*, 31: 277-290.

Zareka, A Zamani, B Ghorbani, S Moalla, M & Jafari, H 2013, Mapping spatial distribution of forest fire using MCDM and GIS (Case study: Three forest zones in Guilan Province). *Irianin Journal of forest and polar research*, 21:218–30. DOI:10.22092/ijfpr.2013.3854.

Zhong, Zh Huang WLi, S & Zeng, Y 2017, Forest fire spread simulating model using cellular automaton with extreme learning machine. *Ecological Modelling*. 348: 33-43, DOI: 10.1016/j.ecolmodel.2016.12.022.

Bibliographic information of this paper for citing:

Zolghadri, S, Ghodskhah Daryaei, M, Nasirahmadi, K, Ghajar, E 2025, Uncovering the hidden patterns of fire risks: A cluster analysis approach (K-Medoids and FCM) for Hyrcanian Forest in Iran, Caspian Journal of Environmental Sciences, 23: 481-493.