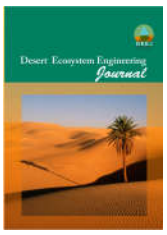
DEEJ (2023) 12 (10) 1-14

University of

Desert Ecosystem Engineering Journal

Journal homepage: http://deej.kashanu.ac.ir



Identifying Dust Storm-Prone Areas Using Google Earth Engine Data and Classified Variable Data Mining Methods: A Case Study of Yazd Province, Iran

Mohammad Kazemi[1]*, Atefeh jafarpoor[2], Reza Naderi Samani[3]

Received: 2023/06/06 Accepted: 2023/09/14

Abstract

As complex climatic events, dust storms could be managed by considering their nature and attributes. Therefore, this study sought to investigate interactions between the aerosol optical depth index and climatic land surface characteristics using data mining and zoning techniques in dust-prone regions of the Yazd province, Iran. To this end, the required data was collected from several climatic products of the University of Idaho and Modis Sensor for the 2000–2017 period using Google Earth Engine.

Moreover, the image of the maximum dust was processed using AOD Modis and ENVI 5.1 software. Then, the underlying correlation between the variables was identified through various data mining techniques. In addition, the ROC curve was used for cross-validation, and different metrics were applied to assess the model, including Square Root of Error, Absolute Normalized Error, Classification Error, Absolute Error, and Crucial Class Fraction Ratio. Finally, the best data mining approach was used to determine the location and zoning of dust-prone regions.

The findings of the study indicated that the decision tree outperformed the Bayesian theory with 89.53% accuracy and that it performed better than the nearest neighbor with an accuracy of 61.3% and 81.31%, respectively. As for the validation of the models, the decision tree methods, nearest neighbor search, and Bayesian network theory showed 74.21%, 64.39%, and 55.42% values, respectively.

Moreover, in regions with crucial harvest and dust ranges, wind speed and soil surface moisture were found to have the most significant role. On the other hand, the zoning of dust-prone regions revealed that 888,067067 Km^2 of areas were covered with the highest concentrations of dust, most of which were located in the central and eastern parts of Yazd province, with the AOD values being 0.465, 0.309, 0.162, and 0.065 for the ranges of 0-0.036, 0.036-0.072, 0.072-0.107, and 0.107-0.3, respectively.

Keywords: Dust, Optical Depth, Climatic Variables, Data Mining Models, Machine Learning.

1. Hormoz Studies and Research Center, University of Hormozgan, Bandar-Abbas, Hormozgan, Iran, (Corresponding author) mohamad.kazemi86@gmail.com, M.kazemi@hormozgan.ac.ir
2. Former Ph.D. Student, Department of Watershed Management Engineering, Faculty of Natural Resources and Marine Sciences, Tarbiat Modares University, Noor 46417-76489, Iran. E-mails: atefeh.jafarpoor@modares.ac.ir
3. Researcher Soil Conservation and Watershed Management Research Department, Isfahan Agricultural and Natural Resources, Research and Education Center, AREEO, Isfahan, Iran; rezanadderi9343@gmail.com

1. Introduction

Dust storms result from the co-existence of persistent dryness, parched topsoil, and gusty winds (O'Liongsigh et al., 2014). Moreover, the wind acts as a mediator of such climatic anomalies, considering the presence of temperature, humidity, and pressure variations in the horizontal direction of the atmosphere (Liu et al., 2020). On the other hand, while dust particles move a hundred miles per hour, sand particles move far less. Consequently, the movement of spinning systems through the higher layers of the atmosphere and the subsequent cold air loss, together with radiant heat at the land surface, generate such circumstances that cause airflow imbalances. Therefore, it could be argued that land use, geology, slope, altitude, precipitation, and vegetation variations significantly contribute to the creation of dust accumulation hotspots (Lee et al., 2012).

Dust storms are considered a kind of environmental occurrence that can hurt public health and the economy (Barbulescu and Nazzal, 2020). Thus, it is essential to identify and track the sources of dust generation in cases when wind storms occur (Gholami et al., 2020). Furthermore, as dust storms are currently classified among the world's major challenges, it is crucial to assess original data to enhance management power.

Addressing fundamental interconnections between a system's components is required for a comprehensive understanding of environmental hazards. It is also necessary for data analysis as a critical component in conceptualizing the dynamics behind natural occurrences (Gibert et al., 2018). Therefore, investigating all climatic and surface properties of the soil in dusty areas could help assess the dusty anomalies of a specific location (Boroughani et al., 2020). In this regard, several techniques have been suggested by researchers for surveying dust, including remote sensing (Nabavi et al., 2017), sediment streams (Gholami et al., 2020), microwave observation, and modeling techniques (Fernández et al., 2019), and numerical methods (Beegum et al., 2018).

Hybrid methodologies, most of which are primarily focused on machine learning, data mining, ArcGIS, and spatial modeling are among the most recent advancements in dust studies (Gholami et al., 2020). Accordingly, data mining is the process of identifying unknown concepts involved in and investigating the relationships existing between statistical information to obtain a more accurate overall understanding (Gibert et al., 2018). In other words, data mining is a technique generally used for extracting potentially innovative information from a specific set of input data, encompassing a variety of prediction models.

Moreover, the optical depth of a tiny dust can be estimated using optical depth data (a non-dimensional variable retrieved from Modis observational data). For instance, Aloysius et al. (2009), Retalis & Hadjimitsis (2010), and Nafarzadegan et al. (2021) found 0.9, 0.86, 0.97, and 0.83 correlation coefficients between AOD accuracy values measured by Modis satellite and the ones measured by surface ground station measuring PM_{10} particles.

In addition, Gholami et al. (2020) used twelve climate variables, soil and land surface attributes, eight data mining algorithms, and GIS to generate geographical maps of the dust sources in the Khuzestan province, reporting that the EM (Expectation Maximization) model presented the most accurate prediction in locating the sources of dust and that the model included wind speed as the most important determining element.

On the other hand, Namdari et al. (2018) examined the movements of dust storms in the Middle East from 2005 to 2016 using the aerosol optical depth index and monthly meteorological data on precipitation, temperature, wind speed, and air pressure, finding a strong relationship between temperature and the aerosol's optical depth.

Furthermore, Wang et al. (2023) used Machine a learning-based model to predict the sources of sand and dust storms in arid areas of Central Asia using the Google Earth Engine (GEE) platform to select fourteen meteorological and terrestrial factors involved in controlling the susceptibility of the sources and modeling process. They found that the random forest (RF) algorithm, gradient boosting tree (GBT), maximum entropy

(MaxEnt) model, and support vector machine (SVM) performed well in predicting the sources of sand and dust storms, and that wind speed and normalized difference vegetation index (NDVI) played the most important role in this regard.

Continuously exposed to dust storms, Yazd province experiences substantial damage in certain periods (Parivar et al., 2020). Therefore, pinpointing those zones that are vulnerable to dust storms and their detrimental effects might assist in mitigating the phenomenon's adverse consequences. However, no study has so far been conducted on the zoning and mapping of dust-sensitive foci in Iran through the application of decision tree data mining algorithms, Bayesian networks, nearest neighbor K, multiple variables, and large satellite data. Thus, the current study sought to investigate dust-prone regions in the Yazd province using data mining techniques and zoning.

2. Materials and Methods

2.1. Case Study

Covering an area of 131575 km^2, Yazd province is a vegetation-free area located in Central Iran's dry and desert region between latitudes 29˚ and 52́, 33˚ and 27 north, and 52˚ and 55, and 56˚ and 37′ east. Characterized by relatively cold winters and scorching dry summers (Taghizadeh-Mehrjardi et al., 2020), the province comprises deserts and low-water plains that are exposed to wind erosion (Khosravi et al., 2017).

On the other hand, vertical air masses contribute to atmospheric instability in the province and thus bring about severe sandstorms, the majority of which occur in May. Furthermore, the presence of a prospective storm source in arid and saline parts of the province results from the Gavkhoni wetlands and neighboring deserts (Mohammadpour Penchah et al., 2020). Fig. 1 illustrates the location of Yazd province and sampling spots for data mining models.

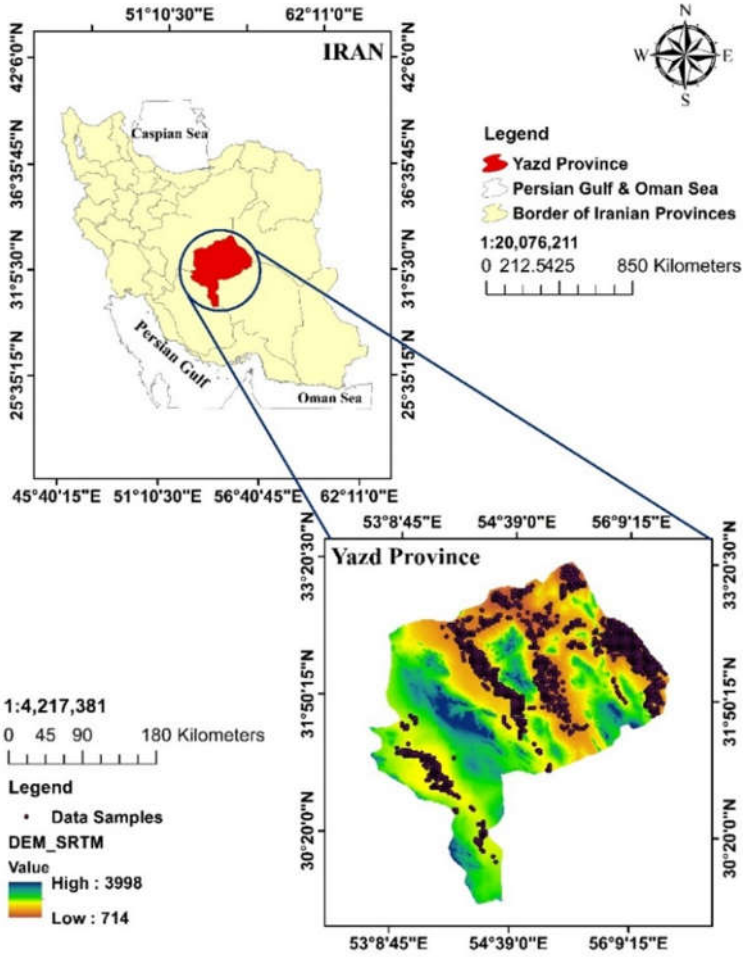


Figure (1): The location of sample points

2.2 Methodology

This study collected satellite images of climate products and surface specifications from the University of Idaho and the Modis satellite using Google Earth Engine. To this end, the required codes were written via JavaScript based on the shape of the Yazd province's boundary file, followed by the application of innovative techniques in the field of remote sensing. It should be noted that this form of computation is merely conducted through programs such as ENVI, which take months to perform.

Since satellite images should first be corrected geometrically and atmospherically to decrease the proportion of errors, each computation requires a substantial set of variables and a series of images with large volumes that should be generated and categorized to acquire a final output. Therefore, the operation is highly time-consuming and needs a significant volume of large-scale storage memory for each extracted image.

The intended images were collected in May of each year from 2000 to 2017, which were then entered into the GIS. The parameters used in this regard included wind speed, deep soil moisture, cumulative precipitation, Palmer drought index, exposed vegetation index, and reference evapotranspiration that were calculated using the Pemman-Mantis algorithm, soil surface moisture, ground surface short wavelength radius, minimum and maximum air temperature, vapor pressure, lack of steam pressure, landforms and land cover, lithology, percentage of clay and sand, and soil dryness (Fig 2, 3, 4, Table 1). Accordingly, the climatic parameters included Palmer drought index data, Precipitation accumulation, maximum temperature, minimum temperature, wind speed, and other parameters that were related to the characteristics of the land surface and soil. The unclassified data were then entered into the modeling process in a raster form.

As for the year parameter, the maximum values of the aerosol optical depth index were averaged in May and then computed for the considered period using Modi's dust observational service and the ENVI5.3.1 add-on toolkit (MCTK). Moreover, the MOD04 product with the dark algorithm was imported daily in Modi's aerosol products (from https://worldview.earthdata.nasa.gov) (Sun et al., 2021).

Table (1): The variables that have been considered in this study

Variable name	Unit	Min	Max	Average	Standard deviation
Actual evapotranspiration	millimeter	0.729	5.753	2.859	0.789
Aspect	-	-1.000	357.274	175.951	105.217
Soil bulk density	Kg per cubic meter	136.000	154.000	145.763	1.896
Soil organic carbon content	percent	0.000	2.000	0.005	0.081
Clay content	percent	4.000	26.000	11.950	4.259
Climate water deficit	millimeter	169.894	241.378	223.857	12.756
Elevation	meter	724.000	2421.000	1186.739	283.154
Geology	-	2.000	145.000	110.572	23.497
Landform	-	0.000	41.000	30.169	5.510
NDVI	dimensionless variable	0.026	0.270	0.074	0.021
Land cover	-	3.000	33.000	12.397	9.786
Palmer Drought Severity Index	dimensionless variable	-3.863	-1.688	-2.677	0.421
Reference evapotranspiration	millimeter	175.106	243.494	221.588	12.327
Precipitation accumulation	millimeter	0.000	5.000	2.116	0.741
Sand content	percent	41.000	91.000	72.416	10.638
Slope	percent	0.000	43.267	3.502	5.518
Soil moisture	millimeter	0.000	2.344	0.306	0.335

Variable name	Unit	Min	Max	Average	Standard deviation
Soil PH	-	75.000	83.000	79.150	1.047
Downward surface shortwave radiation	Watts per square meter	270.900	281.078	276.726	2.177
SRTM-derived topographic diversity	dimensionless variable	0.000	1.000	0.131	0.154
Minimum temperature	centigrade	9.311	21.383	18.474	2.372
Maximum temperature	centigrade	26.567	36.922	34.687	1.955
Vapor pressure	KPa	0.539	0.971	0.830	0.071
Vapor pressure deficit	KPa	1.797	3.521	3.048	0.392
Wind-speed	Meters per second	2.574	3.457	3.206	0.145
Aerosol Optical Depth	dimensionless variable	0.000	0.143	0.049	0.032

Three common Bayesian data mining techniques were used to assess the relationship between the dust event factors, including the nearest neighbor, decision tree, and regression tree. On the other hand, 70% of the dataset was used for training and 30% of which for validation (Nabipour et al., 2020). The Simple Bayesian approach is a simplified categorization based on Bayesian theory which directs the process of learning.

Providing a mechanism for computing the secondary probability P (h | D) of p (h), P (D), and P (D | h), the Bayesian theorem provides a framework for Bayesian learning techniques. Equation 1 shows a simplified Bayesian calculation algorithm.

$$P\ (h \mid D) = \frac{P\ (h \mid D)\ P\ (h\ )}{P\ \ D)} \quad (1)$$

Where P - (h) indicates starting probability, P (D) shows the initial probability of witnessing training data and D, P (D | h) represents the probability of witnessing D in the presence of the valid hypothesis h.

The second probability h (confirmation of hypothesis h upon observation of an instructive example D) is sometimes attributed to the secondary probability h. Equation 2 shows the h_{ML} calculating operation:

$$h_{ML} = arg^{max}_{\epsilon} H^{\mathrm{P\ (D \mid h)}} \quad (2)$$

Accordingly, a learner considers a collection of hypotheses such as H and selects the most likely one (H) based on the teaching instances D. The probability of D data for h is symbolized by P (D | h), and any hypothesis that maximizes P (D | h) is expressed by h_{ML}, i.e., the maximum likelihood (Llorente et al., 2022).

A decision tree is a system used to predict and visually express probable outcomes based on observed data, where each node comprises two branches, performing a thorough analysis of all available variables and value branches of each option. Moreover, the tree is optimized for fragmentation based on Equation 3. Thus, if ∅(s ⏐ t) is the suitable fragmentation criteria for the variable s in node t, the formula would be as follows:

$$\emptyset(\mathrm{s|t}) = 2P_L P_R \sum_{j=1}^{\neq class} \left| P\left({}^{j}/_{t_L}\right) - p({}^{j}/_{t_R}) \right| \quad (3)$$

Where t_L shows left-handed branch t, t_R represents right-handed branch t, P_L is the ratio of the number of observations in t_L to the total number of observations in educational data, P_R is the ratio of the number of observations in t_R to the total number of observations in educational data, P (j/t_L) stands for the ratio of the number of classes j in t_L to the number of observations in node t, and P (j/ t_R) shows the ratio of the number of classes j in t_R to the number of observations in node t.

In the above-mentioned formula, the ideal point is reached when the preceding criteria are maximized for all potential crushing scenarios in a particular node (Llorente et al., 2022). Furthermore, K is used to approximate the density distribution function of educational data and categorize the testing data based on

educational trends (Carrasco et al., 2020).

The primary aim of such a data mining technique is to identify and estimate the features of a sequence of unlabeled data based on their degree of resemblance to known data (Esteban et al., 2022). Accordingly, a sample is classified by the majority vote of its neighbors, depending on the closest features of the other members of the collection (training samples) (Weinberger and Saul, 2009). As a positive value in such a procedure (properties most similar to the target sample), if k = 1, the sample is simply found in the class immediately next to it. It is worth mentioning that K should be individual so that identical votes are not repeated.

Moreover, the intended target data is categorized in a data mining model based on its proximity to the training examples. On the other hand, the Euclidean distance is typically adopted in this model to identify the relationship between the testing dataset and the data that is being tested. Equation 4 shows the formula for calculating Euclidean distance:

$$(X.Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

Where X stands for the educational data with particular parameters (x1 to xn) and Y shows educational data with the same number of specific parameters (y1 to yn).

This study used the receiver operating characteristic (ROC) curve to assess the outcomes of modeling dust-prone regions. The curve yields consistent results when assessing and comparing the performance of models (Boroughani et al., 2020) and applies to a wide variety of spatial modeling applications (Naghibi et al., 2017; Chen et al., 2018).

The ROC curve is a visual and graphical depiction of the degree of dependability of the modeling findings when 70% of the data is used, and the values of the area under the curve vary from zero to one. Accordingly, 0.9-1, 0.8-0.9, 0.7-0.8, 0.7-0.6, and 0.6-0.5 indicate excellent, very good, good; medium, and poor accuracy rates, respectively (Nandi and Shakoor, 2009; Nabipour et al., 2020).

3. Results

The modelers were evaluated using validation, classification error, kappa coefficient, absolute error, normalized absolute error, and the sum of the residual squares of the error. As shown in Table 2, the nearest neighbor technique is the most effective modeler, followed by the decision tree. In this regard, a decision tree with a kappa coefficient of 0.70 or higher produced acceptable simulation results. However, the basic Bayesian model failed to offer a satisfactory performance.

Table (2): Indicators of model assessment as a result of modeling

Index name	K-Nearest Neighborhood	Simple Bayesian model	Decision Tree model
Model accuracy	89.53 %	61.03 %	81.31 %
Classification error	10.47 %	38.97 %	18.69 %
Kappa	0.844 %	0.422 %	0.702 %
Absolute error	0.105 ± 0.306	0.393 ± 0.440	0.262 ± 0.299
Normalized absolute error	1.4	5.257	3.503
Root mean squared error	0.324	0.590	0.397

Following the analysis of the modelers, it was necessary to evaluate the models' predictive ability for various classes of optical dust depth. Tables 3, 4, and 5 show the validation results found for the models used in the current study. Accordingly, the findings revealed that decision tree algorithms and nearest neighbor accurately predicted the second class of aerosol optical depth values with the greatest frequency by reality. Moreover, basic Bayesian network models provided one of the highest frequency values of aerosol optical depth.

Table (3): Simple Bayesian model validation

Prediction. Layer	Correct Layer 1 (0 _ 0.036)	Correct Layer 2 (0.072 _ 0.036)	Correct Layer 3 (0.107 _ 0.072)	Correct Layer 4 (0.3 _ 0.107)	Class accuracy
1 (0 _ 0.036)	369	112	61	27	64.85 %
2 (0.072 _ 0.036)	5	116	24	5	77.33 %
3 (0.107 _ 0.072)	4	210	94	9	29.65 %
4 (0.3 _ 0.107)	1	9	10	14	41.18 %

Table (4): Validation of the decision tree model

Prediction. Layer	Correct Layer 1 (0 _ 0.036)	Correct Layer 2 (0.072 _ 0.036)	Correct Layer 3 (0.107 _ 0.072)	Correct Layer 4 (0.3 _ 0.107)	Class accuracy
1 (0 _ 0.036)	359	74	20	0	79.25 %
2 (0.072 _ 0.036)	20	367	123	18	69.51 %
3 (0.107 _ 0.072)	0	5	39	8	75 %
4 (0.3 _ 0.107)	0	1	7	29	78.38 %

Table (5): Validation of the k-nearest neighbor model

Prediction. Layer	Correct Layer 1 (0 _ 0.036)	Correct Layer 2 (0.072 _ 0.036)	Correct Layer 3 (0.107 _ 0.072)	Correct Layer 4 (0.3 _ 0.107)	Class accuracy
1 (0 _ 0.036)	301	60	23	9	76.59 %
2 (0.072 _ 0.036)	49	299	80	11	68.11 %
3 (0.107 _ 0.072)	19	80	70	16	37.84 %
4 (0.3 _ 0.107)	10	8	16	19	35.85 %

The Simple Bayesian model, decision tree, and nearest neighbor predicted aerosol optical depth by 55.42% ± 3.76%, 74.21% ± 3.55%, and 64.39% ± 4.41%, respectively. Based on the findings regarding performance validation, it could be argued that the decision tree data mining approach delivered the best performance in forecasting aerosol optical depth values despite being the best model of the nearest neighbor. Figure 2 shows the ROC curve for the decision tree model. As illustrated in Fig.2 and Table 5, the model's maximum performance is related to the decision tree with 99.81% area under the curve. Therefore, the tree is placed at the top class of model performance accuracy (Yesilnacar, 2005).

Table (6): Performance evaluation results for the models

Model	Upper limit	Under limit	Receiver Operating Characteristic (ROC)
Decision tree	1	0.75	99.81 %
K-Nearest Neighborhood	1	0.54	92 %
Simple Bayesian	1	0.08	66.35 %

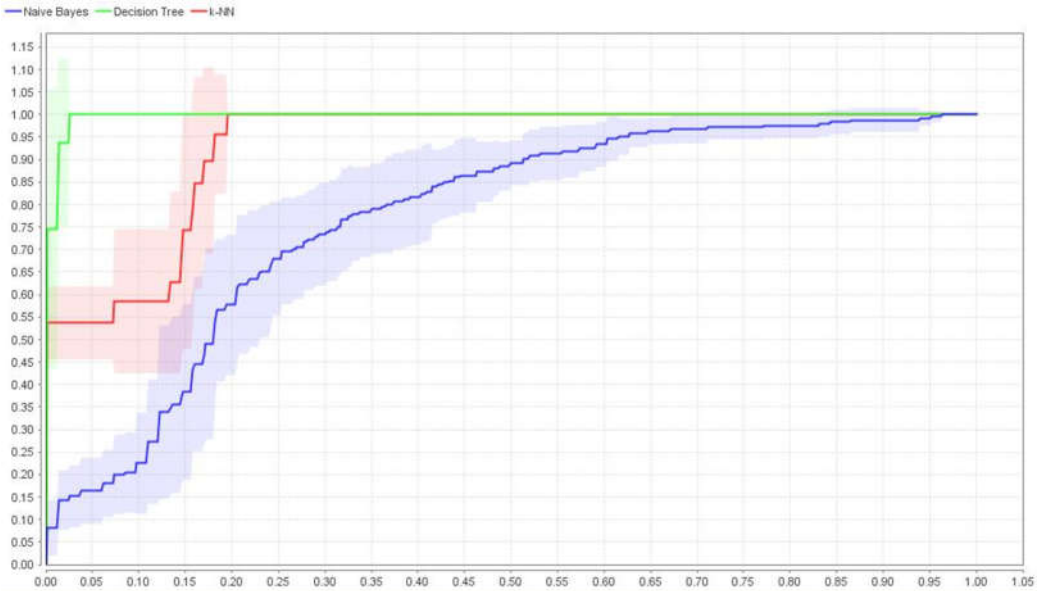


Figure (2): ROC curves for the models employed (green: decision tree, red: k-nearest neighbor, blue: Simple Bayesian)

To support this approach, the decision tree model was used to generate a confidence map for various aerosol optical depth classes to identify areas susceptible to dust storms using the model's branches and nodes. Figure 3 illustrates the model's predicted map, dust-prone region, and dust hotspots.

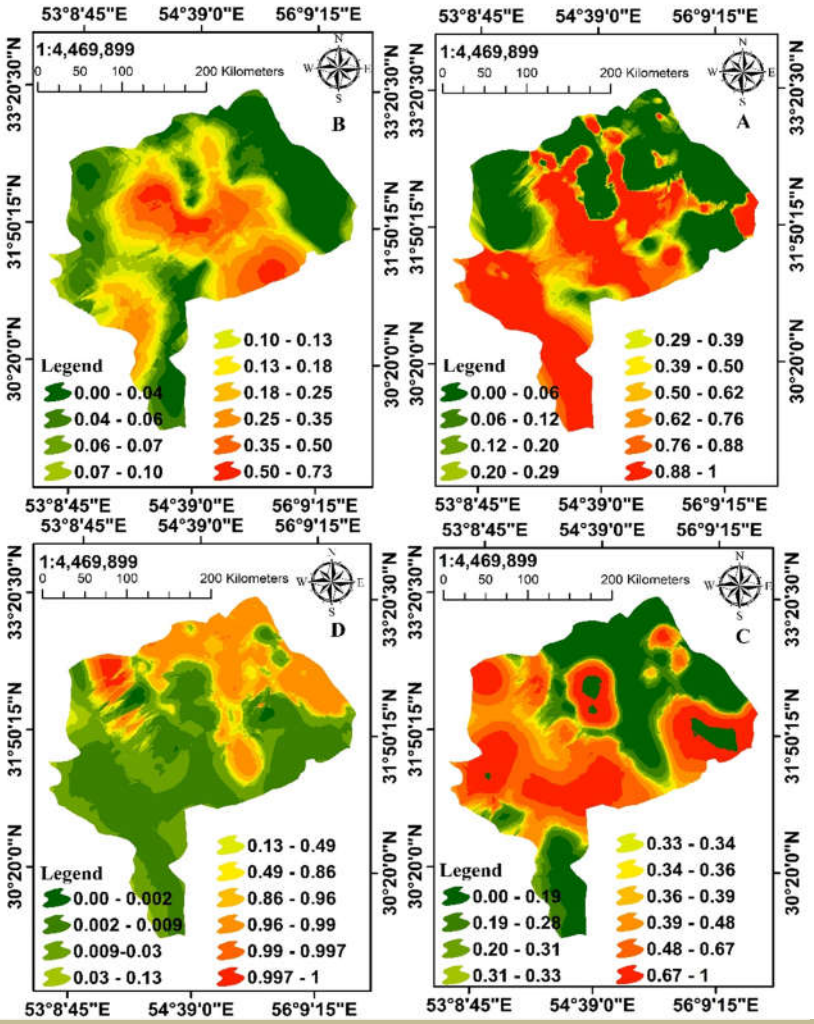


Figure (3): Reliability maps for estimating the optical depth of aerosols using decision tree data mining

As shown in Figure 3, a dust susceptibility model was generated across four levels. The purpose of this study was to assess high-harvesting locations and the regions vulnerable to dust storms, the majority of which are located in the center, west, and southwest of Yazd province. On the other hand, northern, eastern, and southeastern regions of the province possess less vulnerable areas to dust storms. The extent of dust hotspots in the province is estimated to be roughly 7972 square kilometers, approximately 6512 km of which (equivalent to 81.68%) falls under a 0.88-1 confidence level (Fig. 3A).

The magnitude of the intended zone will decrease if the range of the aforementioned class (0.88-1) becomes more limited and precise, thus increasing the accuracy of the measurement. The abovementioned technique, which excluded field references and relied only on the AOD index, the aforementioned factors, and remote sensing, demonstrated acceptable precision in regions susceptible to dust storms. Moreover, as illustrated in Figure 4, the technique selected the main dust-generating regions by finding the branches and logic of the tree. Accordingly, the zones were found to have a high degree of overlap with the dust foci found during the Geological Survey of Iran's field visits.

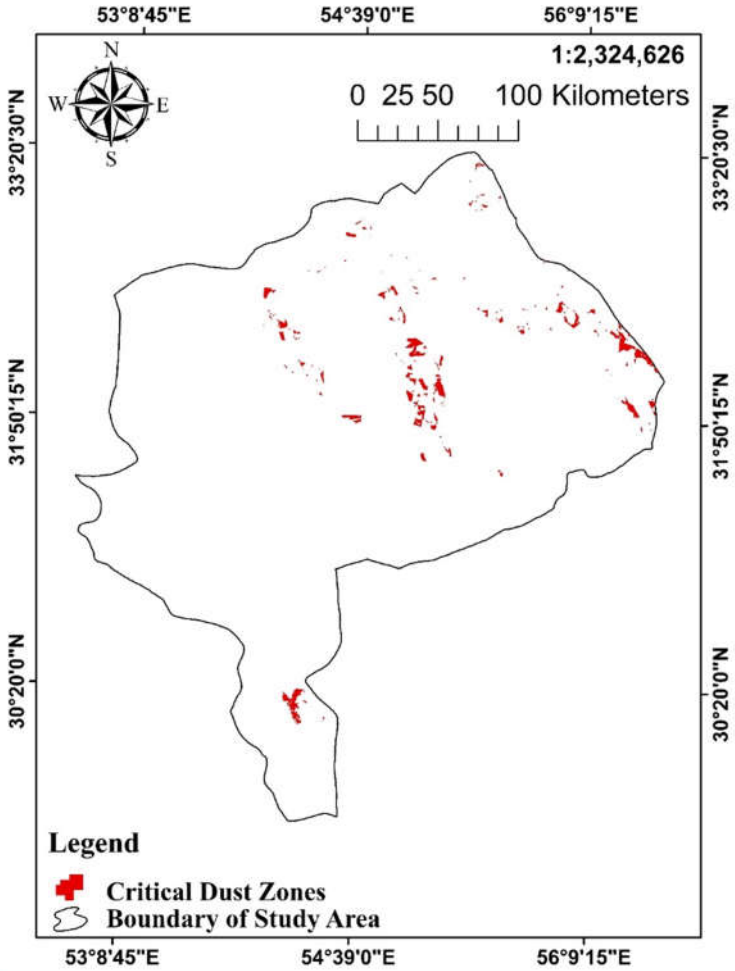


Figure (4): Key zones prone to dust based on a decision tree data mining methodology

The decision Tree algorithm provided the best accuracy for predicting AOD. Therefore, this model can be the best choice for preparing susceptibility maps of dust-generating sources investigated in this study. As seen in Table 4, the critical class of the AOD index (0.107-0.3) enjoys the highest precision (78.38% accuracy) compared to the other two models. In other words, the values of classification accuracy for the class 0.107-0.3 were 41.18%, 78.38%, and 35.85% in the simple Bayesian, Decision Tree, and Nearest neighbor models, respectively.

The class prior to this range, i.e., the class 0.072-0.107, also demonstrated the highest classification accuracy of 75% in the Decision Tree model.

Therefore, compared to other models, the Decision Tree performed better in estimating the critical classes of dust storms. Moreover, the Decision Tree algorithm delivered a good performance in producing susceptibility maps of dust-generating sources, resulting from the interaction between the three statistical-based machine learning algorithms in the study area. It should be noted that although a model or classifier may achieve more accurate results at the initial assessment of the training dataset, it does not necessarily guarantee higher classification accuracy for the classified categories in the validation dataset that reflects real-world performance.

In this study, while the k-nearest Neighbor classifier outperformed the decision tree method in the initial evaluation, the final classified categories found through the application of the decision tree model offered higher validation accuracy and better alignment with reality in the assessment of the model. Moreover, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) ensured 99.81% accuracy, and the validation accuracy was found to be %74.21± %3.55

4. Discussion

The current study used decision tree, basic Bayesian, and nearest neighbor data mining models to zone dust-sensitive regions for the first time in Iran. Broghani et al. (2020) analyzed the performance of the two models using the surface curve underneath the ROC curve, concluding that the stochastic forest model outperformed the logistic regression model. Moreover, they applied factors such as soil, lithology, slope, vegetation index, distance from the river, geomorphological categories, and land use in their simulation model.

Boroughani et al. (2020) used a wide variety of f factors in their study, including soil, rock properties, elevation, vegetation index, proximity to the river, geologic divisions, and land use. Wind speed (at a height of ten meters above the ground), soil moisture, cumulative precipitation, Palmer drought index, exposed vegetation index, soil dryness (water scarcity), and reference evapotranspiration, all of which were calculated using the Pemman-Mantis method.

The current study employed soil surface moisture, visible soil weight, surface-ground short wavelength radius, minimum and maximum air temperature rates, vapor pressure, lack of steam pressure, and percentages of sand and clay. The results indicated that the decision tree technique delivered a better performance than the other two models, with 99.81% area under the receiver operating characteristic curve, which is consistent with the findings reported by Gholami et al. (2020).

On the other hand, seven assessment factors were used to evaluate the modelers. Accordingly, it was found that the most appropriate model bore the highest value and the largest area under the ROC curve, which is compatible with the findings suggested by Gholami et al. (2020). The results also showed that while the nearest neighbor model was the greatest modeler, it did not deliver the best performance in terms of assessment and validation. However, the decision tree technique performed better when modeling restrictions were acceptable. Therefore, it is crucial to provide the model developers' assessment findings with their evaluation results.

Boroughani et al. (2020) prepared a potential map of dust-generating sources using the data mining technique of the amplified regression tree. Moreover, they mapped dust hotspots from 2005 to 2016 using dust detection criteria. On the other hand, the

aerosol optical depth index was used in their study to detect dust across a 17-year period. The results of their study suggested that the area under the curve for the data mining technique was 79.68%, which is an acceptable grade. However, they did not report the findings of their modeling assessment regarding the testing data.

This study tried to elaborate on the modeling assessment in terms of the testing data through zoning and identifying dust-sensitive classes using the findings of fieldwork studies, empirical surveys, and the map of dust-generating centers published by Iran's Geological Survey and Mineral Exploration. It should be noted that as there is no consensus over which data mining technique yields the desired findings for environmental data, each study uses the technique it finds the best based on the type of data it collects for investigation.

Nafarzadegan et al. (2021) sought to assess the decision tree model's efficiency in predicting precipitation in central regions of Iran, indicating that the extremely randomized trees (ERT) model was more accurate in predicting wind erosion-sensitive areas. Moreover, while Boroughani et al. (2020) and Gholami et al. (2020) considered random forest (RF) and maximal expectation (EM) as the best data mining models, respectively, the current study found the decision tree as the optimal model for data mining of dust events. The study also found that the decision tree model produced more accurate verification results with respect to the reality and the evaluation of the model.

5. Concoction

This study attempted to identify dust-prone areas in Yazd province using Google Earth Engine data and classified variable data derived from mining methods. The results of the study suggested that wind speed and soil moisture played an important role in the formation of dust storms. Also, the central and eastern regions of the province suffered from the highest amount of dust. In general, the use of new approaches such as data mining proves useful in identifying climatic components.

Finally, it is recommended that readily accessible satellite data be combined with ready-made products and data mining techniques to investigate various correlations between environmental data and the map of dust-generating regions. The conclusion and recommendations of the study can be used to improve the management of dust-prone regions.

References

1. Aloysius M, Mohan M, Suresh Babu S, Parameswaran K, Moorthy K K. 2009. Validation of MODIS-derived aerosol optical depth and an investigation on aerosol transport over the South East Arabian Sea during ARMEX-II. Annales Geophysicae, 27 (6), 2285–2296. https://doi.org/10.5194/angeo-27-2285-2009
2. Barbulescu A, Nazzal Y. 2020. Statistical analysis of dust storms in the United Arab Emirates. Atmospheric Research, 231, 104669. https://doi.org/10.1016/j.atmosres.2019.104669
3. Beegum S N, Gherboudj I, Chaouch N, Temimi M, Ghediram H. 2018. Simulation and analysis of synoptic-scale dust storms over the Arabian Peninsula. Atmospheric Research, 199, 62–81. https://doi.org/10.1016/j.atmosres.2017.09.003
4. Boroughani, M., Pourhashemi, S., Hashemi, H., Salehi, M., Amirahmadi, A., Asadi, M. A. Z., & Berndtsson, R. (2020). Application of remote sensing techniques and machine learning algorithms in dust source detection

and dust source susceptibility mapping. Ecological Informatics, 56, 101059. https://doi.org/10.1016/j.ecoinf.2020.101059

5. Carrasco, J., García, S., Rueda, M. M., Das, S., & Herrera, F. (2020). Recent trends in the use of statistical tests for comparing swarm and evolutionary computing algorithms: Practical guidelines and a critical review. Swarm and Evolutionary Computation, 54, 100665. https://doi.org/10.1016/j.swevo.2020.100665
6. Chen W, Xie X, Peng J, Shahabi H, Hong H, Bui DT, Duan Z, Li S, Zhu A-X. 2018. GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical-based random forest method. Catena, 164: 135-149.
7. Esteban, A., Zafra, A., & Ventura, S. (2022). Data mining in predictive maintenance systems: A taxonomy and systematic review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(5), e1471. https://doi.org/10.1002/widm.1471
8. Fernández A J, Sicard M, Costa M J, Guerrero-Rascado J L, Gómez-Amo J L, Molero F, … Bedoya-Velásquez A E. 2019. Extreme, wintertime Saharan dust intrusion in the Iberian Peninsula: Lidar monitoring and evaluation of dust forecast models during the February 2017 event. Atmospheric Research, 228, 223–241. https://doi.org/10.1016/j.atmosres.2019.06.007
9. Gholami H, Mohamadifar A A, Collins A L. 2020. Spatial mapping of the provenance of storm dust: Application of data mining and ensemble modeling. Atmospheric Research, 233, 1-17. https://doi.org/10.1016/j.atmosres.2019.104716
10. Gibert K, Izquierdo J, Sanchez-Marre M, Hamilton S H, Rodriguez-Roda I, Holmes G. 2018. Which method to use? An assessment of data mining methods in environmental data science. Environmental Modelling and Software. 110, 3–27. https://doi.org/10.1016/j.envsoft.2018.09.021
11. Khosravi, H., Haydari, E., Shekoohizadegan, S., & Zareie, S. (2017). Assessment of the effect of drought on vegetation in the desert area using Landsat data. The Egyptian Journal of Remote Sensing and Space Science, 20, S3-S12. https://doi.org/10.1016/j.ejrs.2016.11.007
12. Lee J, Baddock M, Mbuh M, Gill T. 2012. Geomorphic and land cover characteristics of aeolian dust sources in West Texas and eastern New Mexico, USA. Aeolian Research, 3(4): 459-466. https://doi.org/10.1016/j.aeolia.2011.08.001
13. Liu, Y., Wang, G., Hu, Z., Shi, P., Lyu, Y., Zhang, G., ... & Liu, L. (2020). Dust storm susceptibility on different land surface types in arid and semiarid regions of northern China. Atmospheric research, 243, 105031. https://doi.org/10.1016/j.atmosres.2020.105031
14. Llorente, F., Martino, L., Curbelo, E., López-Santiago, J., & Delgado, D. (2022). On the safe use of prior densities for Bayesian model selection. Wiley Interdisciplinary Reviews: Computational Statistics, e1595. https://doi.org/10.1002/wics.1595
15. Mohammadpour Penchah M, Memarian MH, Mirrokni SM. 2015. Modeling and Analysis of Dust Storms of Yazd Province Using Numerical Models. Journal of Geography and Environmental Hazards. 3 (4): 67-83. (In Persian). https://doi.org/10.22067/GEO.V3I4.34323
16. Nabavi S O, Haimberger L, Samimi C. 2017. Sensitivity of WRF-chem predictions to dust source function specification in West Asia. Aeolian Research. 24, 115–131.

https://doi.org/10.1016/j.aeolia.2016.12.005

17. Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access, 8, 150199-150212. https://doi.org/10.1109/ACCESS.2020.3015966
18. Nafarzadegan, A. R., Ebrahimi-Khusfi, Z., & Kazemi, M. (2021). Spatial characterization of dust emission-prone arid regions using feature extraction and predictive algorithms. Applied Geography, 133, 102495. https://doi.org/10.1016/j.apgeog.2021.102495
19. Naghibi S A, Moghaddam D D, Kalantar B, Pradhan B, Kisi O. 2017. A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. Journal of Hydrology. 548: 471–483. https://doi.org/10.1016/j.jhydrol.2017.03.020
20. Namdari S, Karimi N, Sorooshian A, Mohammadi G, Sehatkashani S. 2018. Impacts of climate and synoptic fluctuations on dust storm activity over the Middle East. Atmospheric Environment, 173, 265-276. https://doi.org/10.1016/j.atmosenv.2017.11.016
21. Nandi A, Shakoor A. 2009. A GIS-based landslide susceptibility evaluation using bivariate and multivariate statistical analyses. Engineering Geology, 110, 11–20. https://doi.org/10.1016/j.enggeo.2009.10.001
22. O'Liongsigh T, McTainsh G H, Tews E K, Strong C L, Leys J F, Shinkfield P, Tapper N J. 2014. The Dust Storm Index (DSI): A method for monitoring broad-scale wind erosion using meteorological records. Aeolian Research, 12(1): 29-40. https://doi.org/10.1016/j.aeolia.2013.10.004
23. Parivar, P., Quanrud, D., Sotoudeh, A., & Abolhasani, M. (2021). Evaluation of urban ecological sustainability in arid lands (case study: Yazd-Iran). Environment, Development and Sustainability, 23(2), 2797-2826. https://doi.org/10.1007/s10668-020-00637-w
24. Retalis A. Hadjimitsis D G. 2010. Comparison of aerosol optical thickness with in situ visibility data over Cyprus. Natural Hazards and Earth System Sciences, 10, 421–428. https://doi.org/10.5194/nhess-10-421-2010
25. Sun, Z., Wei, J., Zhang, N., He, Y., Sun, Y., Liu, X., ... & Sun, L. (2021). Retrieving High-Resolution Aerosol Optical Depth from GF-4 PMS Imagery in Eastern China. Remote Sensing, 13(18), 3752. https://doi.org/10.3390/rs13183752
26. Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., ... & Scholten, T. (2020). Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. Remote Sensing, 12(7), 1095. https://doi.org/10.3390/rs12071095
27. Wang, W., Samat, A., Abuduwaili, J., De Maeyer, P., & Van de Voorde, T. (2023). Machine learning-based prediction of sand and dust storm sources in arid Central Asia. International Journal of Digital Earth, 16(1), 1530-1550. DOI: 10.1080/17538947.2023.2202421
28. Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 10 (2).

29. Yesilnacar E K. 2005. The Application of Computational Intelligence to Landslide Susceptibility Mapping in Turkey. PhD Thesis, Department of Geometrics the University of Melbourne.