Gholami, G., Mohamadifar, A. and Collins, A. L. 2020. Spatial mapping of the provenance of storm dust: Application of data mining and ensemble modelling. *Atmospheric Research.* 233 (104716).

Atmospheric Research 233 (2020) 104716

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

Spatial mapping of the provenance of storm dust: Application of data mining and ensemble modelling

Hamid Gholami[a,*], Aliakbar Mohamadifar[a], Adrian L. Collins[b,*]

[a] Department of Natural Resources Engineering, University of Hormozgan, Bandar-Abbas, Hormozgan, Iran
[b] Sustainable Agriculture Sciences Department, Rothamsted Research, North Wyke, Okehampton, Devon EX20 2SB, UK

ARTICLE INFO

Keywords:
Dust provenance
Spatial modelling
Data mining algorithms
Multicollinearity
Receiver operating characteristic
Ensemble modelling
R software

ABSTRACT

Spatial modelling of storm dust provenance is essential to mitigate its on-site and off-site effects in the arid and semi-arid environments of the world. Therefore, the main aim of this study was to apply eight data mining algorithms including random forest (RF), support vector machine (SVM), bayesian additive regression trees (BART), radial basis function (RBF), extreme gradient boosting (XGBoost), regression tree analysis (RTA), Cubist model and boosted regression trees (BRT) and an ensemble modelling (EM) approach for generating spatial maps of dust provenance in the Khuzestan province, a main region with active sources for producing dust in south-western Iran. This study is the first attempt at predicting storm dust provenance by applying individual data mining models and ensemble modelling. We identified and mapped in a geographic information system (GIS), 12 potential effective factors for dust emissions comprising two for climate (wind speed, precipitation), five soil characteristics (texture, bulk density, Ec, organic matter (OM), available water capacity (AWC)), a normalized difference vegetation index (NDVI), land use, geology, a digital elevation model (DEM) and land type, and used a mean decrease accuracy measure (MDAM) to determine the corresponding importance scores (IS). A multi-collinearity test (including the variance inflation factor (VIF) and tolerance coefficient (TC)) was applied to assess relationships between the effective factors, and an existing map of dust provenance was randomly categorized into two groups consisting of training (70%) and validation (30%) data. The individual data mining models were validated using the area under the curve (AUC). Based on the TC and VIF results, no collinearity was detected among the 12 effective factors for dust emissions. The prediction accuracies of the eight data mining models and an EM assessed by the AUC were as follows: EM (with AUC = 99.8%) > XGBoost > RBF > Cubist > RF > BART > SVM > BRT > RTA (with AUC = 79.1%). Among all models, the EM was found to provide the highest accuracy for predicting storm dust provenance. Using the EM, areas classified as being low, moderate, high and very high susceptibility for storm dust provenance comprised 36, 13, 23 and 28% of the total mapped area, respectively. Based on MDAM results, the highest and lowest IS were obtained for the wind speed (IS = 23) and geology (IS = 6.5) factors, respectively. Overall, the modelling techniques used in this research are helpful for predicting storm dust provenance and thereby targeting mitigation. Therefore, we recommend applying data mining EM approaches to the spatial mapping of storm dust provenance worldwide.

1. Introduction

Dust storms are one consequence of wind erosion that are a meteorological phenomenon with negative impacts for economic activities and human health (Barbulescu and Nazzal, 2020) or the dispersal of allergens (Almasi et al., 2014) as well as environment impacts (Goudie and Middleton, 2006). Therefore, identifying and mapping the provenance of the dust redistributed by wind storms is necessary to target mitigation.

Different techniques and tools such as conventional sediment source fingerprinting (Gholami et al., 2017; Dahmardeh Behrooz et al., 2019; Gholami et al., 2019a,b), geochemical characteristics (Zarasvandi et al., 2011), remote sensing (Schepanski et al., 2012; Nabavi et al., 2017), Lidar monitoring and forecast models (Fernández et al., 2019), numerical modelling (Beegum et al., 2018; Péré et al., 2018) and meteorological data (Li et al., 2019; Yang et al., 2019) have been employed for studying atmospheric dust.

Since dust as an environmental problem is one of the most

* Corresponding authors.
E-mail addresses: hgholami@hormozgan.ac.ir (H. Gholami), adrian.collins@rothamsted.ac.uk (A.L. Collins).

https://doi.org/10.1016/j.atmosres.2019.104716
Received 14 August 2019; Received in revised form 14 October 2019; Accepted 19 October 2019
Available online 24 October 2019

important challenges of our world today, there remains a need to explore the utility of new methods for using data to improve our capacity to manage the problem (Gibert et al., 2018). Data Science (DS) is a research field for better understanding of the complex mechanisms driving environmental phenomena (Gibert et al., 2018). Data mining, a component of DS, is a generic term for a wide range of models for providing predictions (Witten et al., 2011).

In the specific case of prediction, available machine learning algorithms include: support vector machine (SVM) (Shadman Roodposhti et al., 2017; Sachindra et al., 2018); ensemble-ANFIS (Ali et al., 2018); cubist (Houborg and McCabe, 2018); random forest (RF) (Nashwan and Shahid, 2019); radial basis function (RBF) (Frank, 2014); neural networks (NN)(Meyer et al., 2016); multivariate adaptive regression spline (MARS)(Gomez-Gutierrez et al., 2009; Pourghasem and Rossi, 2016); extreme gradient boosting (XGBoost) (Chen and Guestrin (2016); bayesian additive regression trees (BART) (Kapelner and Bleich, 2014), and Bayesian networks (BNs) (Bui et al., 2018). Some works has integrated data mining methods into ensemble models (Lazri and Ameur, 2018; Arabameri et al., 2019).

Spatial modelling using machine learning has been applied to different environmental fields and problems including: digital soil mapping (Heung et al., 2014); statistical downscaling of precipitation (Sachindra et al., 2018); forecasting multi-scalar standardized precipitation index (Ali et al., 2018); gully erosion mapping (Pourghasemi et al., 2017); soil pollution (Boente et al., 2019); prediction of aerosol optical depth (Nabavi et al., 2018); digital mapping of soil carbon fractions (Keskin et al., 2019); quantifying suspended sediment loads (Khosravi et al., 2018); mapping of drought (Shadman Roodposhti et al., 2017); land subsidence (Rahmati et al., 2019); forecasting of wind power (Demolli et al., 2019), and landslide risk mapping (Dickson and Perry, 2016). However, to date, data mining has not been applied to the spatial modelling of storm dust provenance.

In west Asia, several regions including Iraq, Kuwait, the western parts of Khuzestan in southwestern Iran and some parts of the Arabian Peninsula are affected by the Shamal dust storm (Middleton, 1986). During recent decades, Khuzestan province has suffered severely from the environmental problems, especially air pollution, resulting from dust storms (Zarasvandi et al., 2011). Ahvaz, the capital of Khuzestan province, is one of the dustiest cities in the world and, consequently, has been the focus of numerous previous investigations (e.g., Maleki et al., 2016; Karimi et al., 2019; Heidari Farsani et al., 2018; Hashemimanesh and Matinfar, 2012; Naimabadi et al., 2016). These studies focused on different aspects of the wind erosion such as PM_{10} variability in Ahvaz, but, to date, no work has examined spatial modelling of dust provenance using data mining algorithms. Accordingly, here we apply and compare the performance of eight data mining techniques (RF, SVM, BART, RBF, XGBoost, RTA, cubist model and BRT) individually, and in an ensemble model (EM) approach, for the spatial modelling of dust provenance in Khuzestan province in southwestern Iran.

2. Study area

Khuzestan province (48 to 49.5°E and 31 to 32°N; Fig. 1), between Iran and Iraq, has a population of ~4,274,979 and occupies an area of 64,016 km^2. The climate the Khuzestan province is arid to humid (Zarasvandi et al., 2011). The dominant directions of wind in the study area are west to east and northwestern to southeastern. Annual mean temperature varies from 9 °C (in March) to 50 °C (in July) and annual mean rainfall varies from 150 to 256 mm in southern areas to 995–1100 mm in the northern parts of the province (Zarasvandi et al., 2011). Desertification is a major environmental threat in Khuzestan province, where sandy landscapes cover > 20.2% (about 13,000 km^2) of the study area (Hashemimanesh and Matinfar, 2012). The capital of Khuzestan province, Ahvaz, with a population of ~1.112 million, is the most polluted city on the basis of reporting for PM_{10} annual averages (Goudie, 2014).

3. Materials and methods

3.1. Effective factors driving dust emissions

Dust emission as one consequence of wind erosion is affected by several key physical factors including climate, soil, vegetation cover and landform (Goudie and Middleton, 2006). Here, we used 12 factors including two for climate (wind speed and precipitation), five soil characteristics (texture, bulk density, Ec, organic matter (OM) and available water capacity (AWC)), a normalized difference vegetation index (NDVI), land use, geology, a digital elevation model (DEM) and land type. The sections below provide further detail on the data sources for each of these factors.

3.1.1. Climatic factors

Wind speed and precipitation are two effective climatic factors affecting wind erosion. These parameters were used as a local wind erosion climatic factor (C) in the Chepil wind erosion equation (WEQ; Chepil et al., 1962 and McTainsh et al., 1990). More specifically, we assembled data for daily average wind speed and average annual precipitation totals for the period 1998–2018 from 21 meteorological stations located across Khuzestan province (Fig. 2).

3.1.2. Soil factors

Soil characteristics such as AWC, bulk density, organic carbon content (OC), Ec and texture are important controls on a soil erodibility (Saadoud et al., 2018). In this study, all soil factors were extracted from the soil world map produced by the IUSS-WRB (2015). For the mapping of soil factors, 586 points (Fig. 3) in the study area were randomly selected and then, overlain with the soil world map. Finally, values of the five soil factors mentioned above were extracted from the soil world map. For generating maps of soil factors, the inverse distance weighting (IDW) interpolation method (Golla et al., 2019), a commonly used technique for interpolation in environmental sciences (Li and Heap, 2011), was applied in ArcGIS using the extracted values for the individual 586 points.

3.1.3. Vegetation cover

Vegetation cover (V) as one of main controls on wind erosion is used in the WEQ (Goudie and Middleton, 2006). Here, the normalized difference vegetation index (NDVI) for the study area was extracted from Landsat 1 satellite images. After downloading four Landsat image frames, these were processed into mosaic using ENVI5.3/LLC. Radiometric and atmospheric corrections on the images used the Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) algorithm. The NDVI (Eq. (1)) was calculated from the red (R) and near infrared (NIR) bands (Wessels et al., 2007):

$$NDVI = \frac{NIR - R}{NIR + R} \quad (1)$$

3.1.4. Topography

The shuttle radar topography mission (SRTM) images with 30*30 m resolution was used for preparing the digital elevation model (DEM). The DEM data was downloaded from https://earthexplorer.usgs.gov. The lowest and highest elevations in the study area ranged between −35 and 3715 m.

3.1.5. Other environmental factors

Land use and landform type maps were generated by the Iran Forest, Rangeland and Watershed Management Organization (IFRWMO). In some cases, land cover and land use and land type maps provided by IFRWMO contain errors (e.g., a land use was classified incorrectly). To remove such errors and correct the map, we used Google Earth images

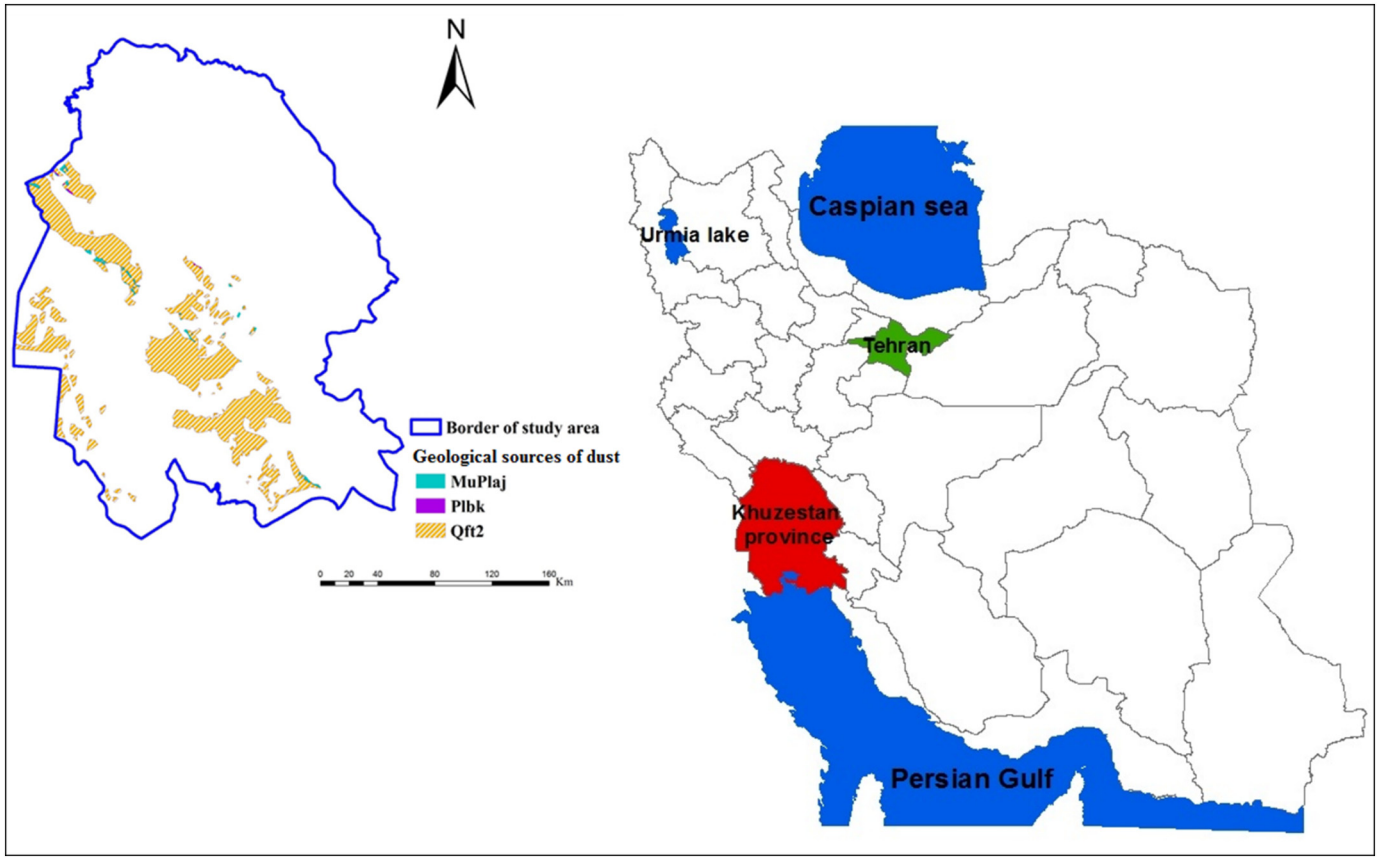

Fig. 1. The location of Khuzestan province in Iran and the distribution of geological sources for storm dust in the study area. MuPlaj, Plbk and Qft2 respectively indicate: brown to grey, calcareous, feature-forming sandstone and low weathering, gypsum-veined, red marl and siltstone (Aghajari formation); alternating hard consolidated, massive, feature forming conglomerate and low-weathering cross-bedded sandstone (Bakhtyari formation), and; Low level piedmont fan and valley terrace deposits. MuPlaj, Plbk and Qft2 represent 2.2, 0.3 and 97.5% of the mapped sources of storm dust, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

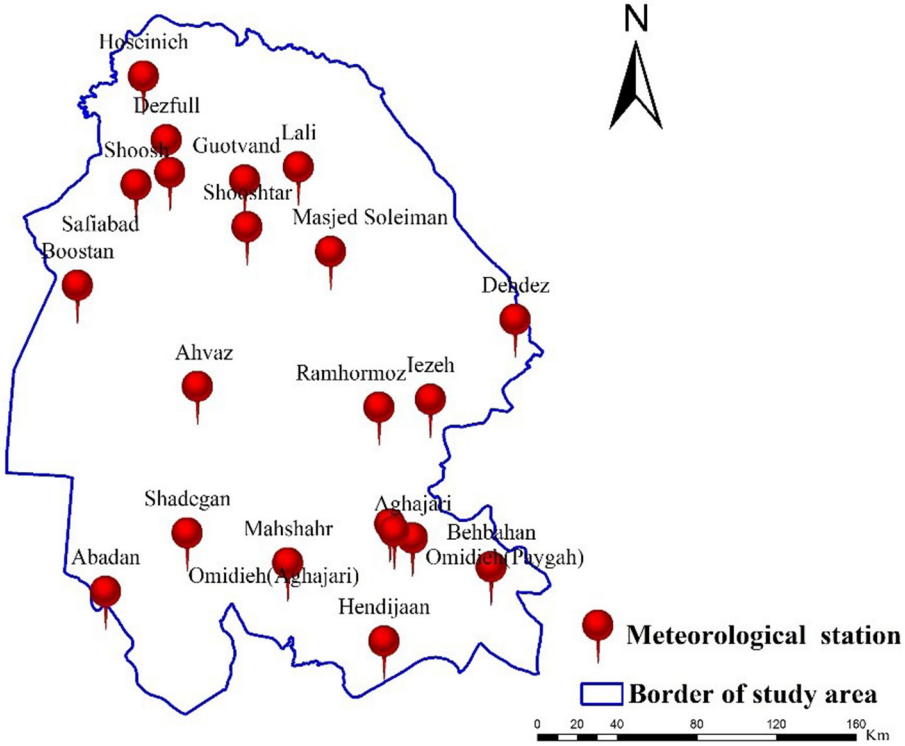

Fig. 2. Location of meteorological stations in the study area.

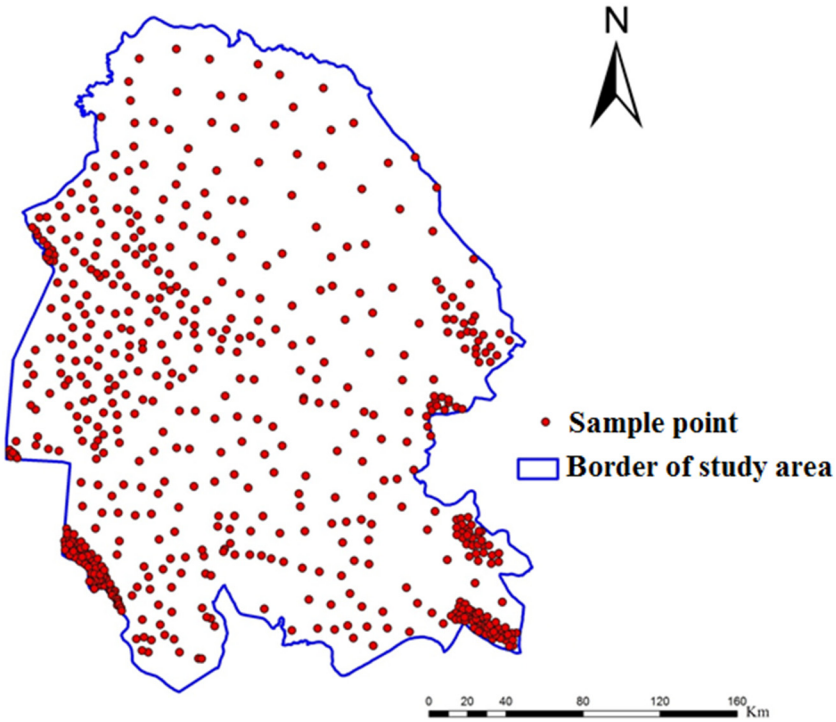

Fig. 3. Sampling points used to extract data from the soil world map for the mapping of soil factors across the study area.

Table 1
Descriptions and corresponding areas of the geological units in the study area.

Geology unit	Description	Area (Km^2)	Area (%)
Ekn	Tine-bedded argillaceous limestone and calcareous shale (Kandavan Shale)	102	0.2
EMas-sb	Undivided Asmari and Shahbazan Formation	519	0.8
JKkgp	Undivided Khami Group, consist of massive thin - bedded limestone comprising the following formations: Surmeh, Hith Anhydrite, Fahlian, Gadvan and Dariyan	1174	2
K1 l	Massive to thick - bedded Orbitolina limestone	10	0.02
Kbgp	Undivided Bangestan Group, mainly limestone and shale, Albian to Companian, comprising the following formations: Kazhdumi, Sarvak, Surgah and Ilam	2096	3
KEpd-gu	Grey and brown, medium - bedded to massive fossiliferous limestone (Kazhdumi formation)	1461	2.29
Kgu	Bluish grey marl and shale with subordinate thin - bedded argillaceous -limestone (Gurpi formation)	492	0.77
Klsol	Grey thick - bedded to massive Orbitolina limestone	29	0.05
Ktb	Massive, shelly, cliff - forming partly anhydrite limestone (Tarbur formation)	0.94	0.001
Mgs	Anhydrite, salt, grey and red marl alternating with anhydrite, argillaceous limestone and limestone (Gachsaran formation)	3084	4.8
Mmn	Low weathering grey marls alternating with bands of more resistant shelly limestone (Mishan formation)	1081	1.7
MuPlaj	Brown to grey, calcareous, feature-forming sandstone and low weathering, gypsum- veined, red marl and siltstone (Aghajari formation)	9239	14.5
OMas	Cream to brown - weathering, feature - forming, well - jointed limestone with intercalations of shale (Asmari formation)	2470	3.9
PeEpd	Blue and purple shale and marl interbedded with the argillaceous limestone (Pabdeh formation)	380.5	0.6
Plbk	Alternating hard of consolidated, massive, feature forming conglomerate and low -weathering cross -bedded sandstone (Bakhtyari formation)	3848	6.1
Qft1	High level piedmont fan and valley terrace deposits	91	0.14
Qft2	Low level piedmont fan and valley terrace deposits	37,514	59

and the ArcBruTile plugin. A geology map (1:100,000 scale) produced by the Geological Survey of Iran was used for this factor. Tables 1–3 describe lithology, land uses and types in the study area, respectively.

3.1.6. Spatial mapping of the dust emission controlling factors

Figs. 4 and 5 present spatial maps for each of the potential controlling factors affecting dust emissions in the study area.

3.2. Dust emission effective factors importance scores

For determining importance scores (IS) of the effective factors (Figs. 4 and 5) influencing dust emissions, we applied a mean decrease accuracy measure (MDAM) (Pourghasemi et al., 2017) in the R Rattle software package.

3.3. Dust provenance inventory mapping

Mapping of existing understanding of dust sources is required to assess the relationship between the distribution of source regions and the effective factors (Figs. 4 and 5). Here, we used a map (Fig. 6) of dust provenance for Khuzestan province produced by Heydarian et al. (2018). Based on this inventory map of dust provenance, seven sources of dust cover are identified with an area of ~699,231 ha. Ahvaz County is identified as having 76,372 ha (about 21.8%) and is therefore the most expansive dust source in the study area. The seven sources of dust (Fig. 6) in the Khuzestan province include the south west of Hoveyzeh, the north and east sides of Khorramshahr, eastern Ahvaz, the south and south eastern parts of Ahvaz, Bandar Imam Khomeini-Omidiyeh, Bandar Mahshahr-Hendijan and the east and south eastern sides of

Table 2
Summary information on the different land uses in the study area.

Land use	Area (km^2)	Area (%)
Agricultural lands and gardens	19,469	30.74
Airport	10	0.016
Aquifer	4	0.006
Bareland	5372	8.46
Fisherypool	118	0.19
Forest	6907	10.9
Mangro	1	0.002
Masil	127	0.2
Range	23,948	37.73
Rock	33	0.052
Saltland	2168	3.42
Sanddune	520	0.82
City	605	0.95
Water	527	0.9
Wetland	3290	5.18
Woodland	358	0.56

Table 3
Summary information on the land use types in the study area.

Land type	Area (Km^2)	Area %
Alluvial plains	18,060	28.47
Alluvial Fan	837	1.32
Flood Plain	1703	2.68
Hill	9686	15.27
Lowland	6068	9.56
Miscellaneous Land	3649	5.75
Mixed Land	522	0.82
Mountain	14,539	22.92
Piedmont	4072	6.45
Plateau and upper terrace	3689	5.82
Residential and industrial areas	596	0.94

Hendijan (Heydarian et al., 2018). For building models from the 787 pixels marked as being the location of a source of storm dust, 551 (70%) and 236 (30%) were randomly selected for the training and validation of the EM, respectively (Fig. 7) (Phillips et al., 2006).

3.4. Multicollinearity test on the effective factors for dust emissions

Two statistical indicators comprising the tolerance coefficient (TC) and variance inflation factor (VIF) (Pourghasemi et al., 2017) were applied to examine the relationship between the effective factors (independent variables) for dust emissions. The TC (Eq. (2)) and VIF (Eq. (3)) are defined as follows:

$$TC = 1\text{–}R^2J \tag{2}$$

$$VIF = \left[\frac{1}{TC}\right] \tag{3}$$

where, R^2J indicates the regression coefficient of determination of variable J. If the TC is < 0.1 and the VIF is > 10, both coefficients show a collinearity problem (Bui et al., 2012). These tests were performed in SPSS22.

3.5. Spatial modelling of dust provenance using data mining algorithms

3.5.1. Random forest (RF)

RF as an effective tool for the analysis of large datasets (Lawrence et al., 2006), builds a number K of regression trees (Rodriguez-Galiano et al., 2015) and based on these classification trees provides a prediction with high accuracy, without over-fitting, for spatial modelling of environmental phenomena (Breiman, 2001). The number of trees (K) and factors (x) are taken as input variables for the RF (Pourghasemi and Rahmati, 2018). After K such trees $\{T(x)\}_1^K$ are grown, the RF regression predictor is expressed as (Rodriguez-Galiano et al., 2015):

$$f_{rf}^K(x) = \frac{1}{K}\sum_{K=1}^{K} T(x) \tag{4}$$

3.5.2. Support vector machine (SVM)

SVM proposed by Vladimir and Vapnik (1995), a classifier based on statistical learning theory, is designed to construct an optimal separating hyper-plane between various classes (Hastie et al., 2009). Different types of classification functions including exponential kernel (Chen et al., 2014); radial basis function (RBF) kernel (Gayen et al., 2019) and linear kernel, polynomial kernel or sigmoid kernel (Amiri et al., 2019) were used can be used. Here, we applied a SVM with an exponential kernel function proposed by Chen et al. (2014). This function is expressed as (Kandola et al., 2003):

$$K(\lambda) = K_0 \exp(\lambda K_0) \tag{5}$$

where K_0indicates the kernel matrix of the BoW kernel and λ (λϵ [0, $+\infty$)) is a decay factor.

3.5.3. Bayesian additive regression trees (BART)

BART, a sum-of-trees ensemble Bayesian approach, estimates a non-parametric function using a fully Bayesian probability model. In this approach, for estimating the f unknown function, regression trees depend on the recursive binary portioning of predictor space into a set of hyperrectangles. The BART model can be expressed as (Kapelner and Bleich, 2014):

$$Y = f(X) + \varepsilon \approx A_1^K(X) + A_2^K(X) + \ldots + A_m^K(X) + \varepsilon,\ \varepsilon \sim Nn(0, \sigma 2In) \tag{6}$$

where Y is the $a \times 1$ vector of responses, X is the $a \times g$ design matrix, ε is the $a \times 1$ vector of noise, and m is regression trees. The A^K comprises an entire tree.

3.5.4. Radial basis function (RBF)

The RBF classifier as a radial base function network (RBFN) trains a model for classification problems (Wu et al., 2008). The Gaussian function as one typical RBF for classification can be expressed as (Frank, 2014):

$$f(x_1, x_2, \ldots, x_m) = g\left(w_0 + \sum_{i=1}^{b} w_i \exp\left(-\sum_{j=1}^{m} \frac{a_j^2 (x_j - c_{i,j})^2}{2\sigma_{i,j}^2}\right)\right) \tag{7}$$

where, x_1, x_2, …, x_m is the vector of attribute values for the instance concerned, $g(.)$ is the activation function, b is the number of basis functions, w_i is the weight for each basis function, a_j^2 is the weight of the jth attribute, and $c_{i,.}$ and $\sigma_{i,.}^2$ are the basis function centers and variances, respectively.

3.5.5. Extreme gradient boosting (XGBoost)

The XGBoost algorithm proposed by Chen and Guestrin (2016), based on the concept of boosting (Fan et al., 2018b), is a useful technique for K classification and regression trees. This model combines all the predictions of a set of weak learners.

3.5.6. Classification and regression tree (CART)

The CART, proposed by Breiman et al. (1984), constructs predictive relationships from input variables (Choubin et al., 2018) and generates a sequence of sub-trees by growing a large tree instead of using stopping rules (Pham et al., 2017). The CART has two different sub-models including classification tree analysis (CTA) and regression tree analysis (RTA). Here, we applied RTA to the mapping of storm dust provenance.

3.5.7. Cubist model

Cubist, a non-parametric model (Houborg and McCabe, 2018) is

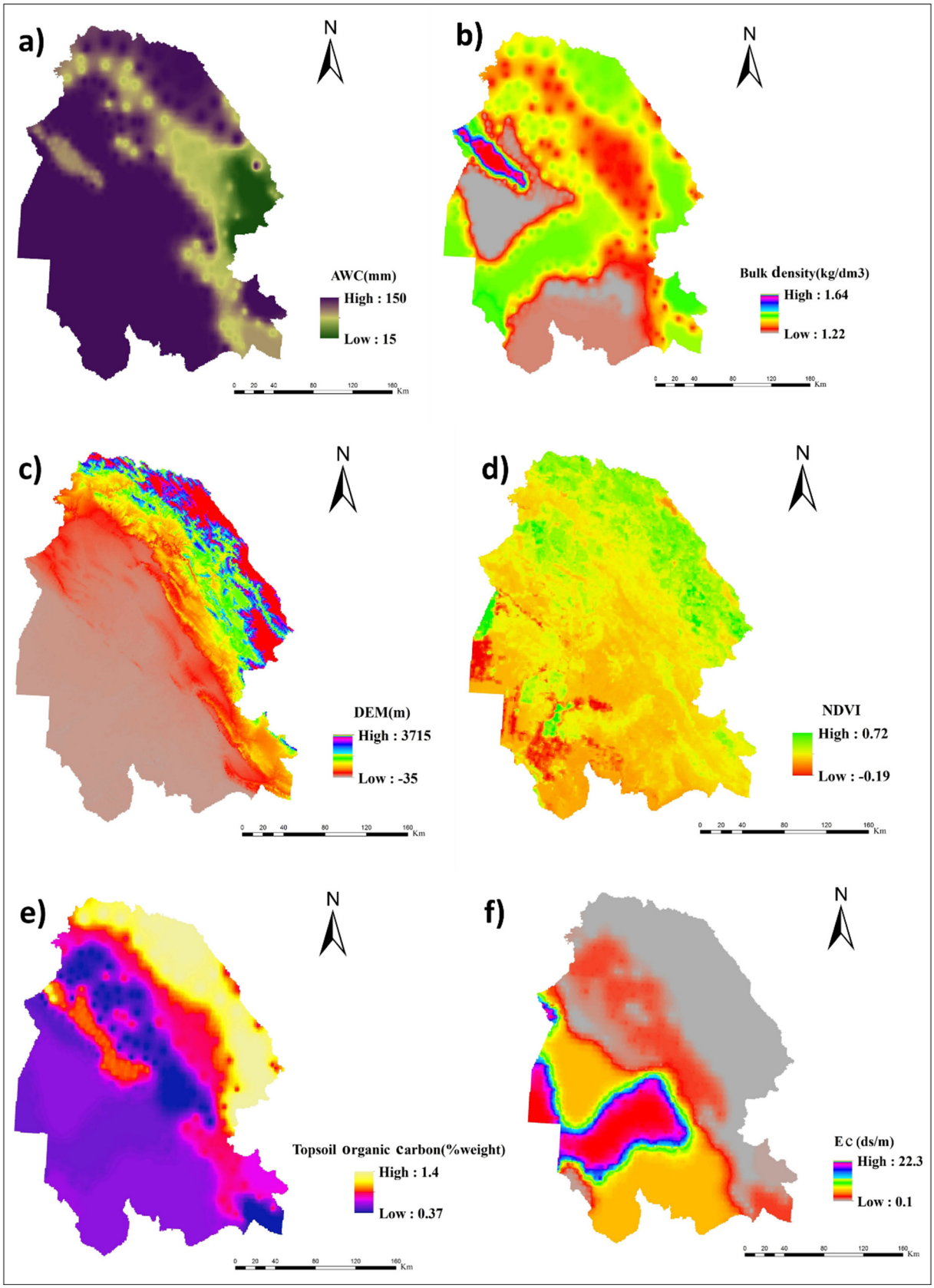

Fig. 4. Maps of effective factors for dust emissions. a) Soil available water capacity (AWC); b) soil bulk density; c) DEM; d) NDVI; e) organic carbon (OC) and f) soil EC.

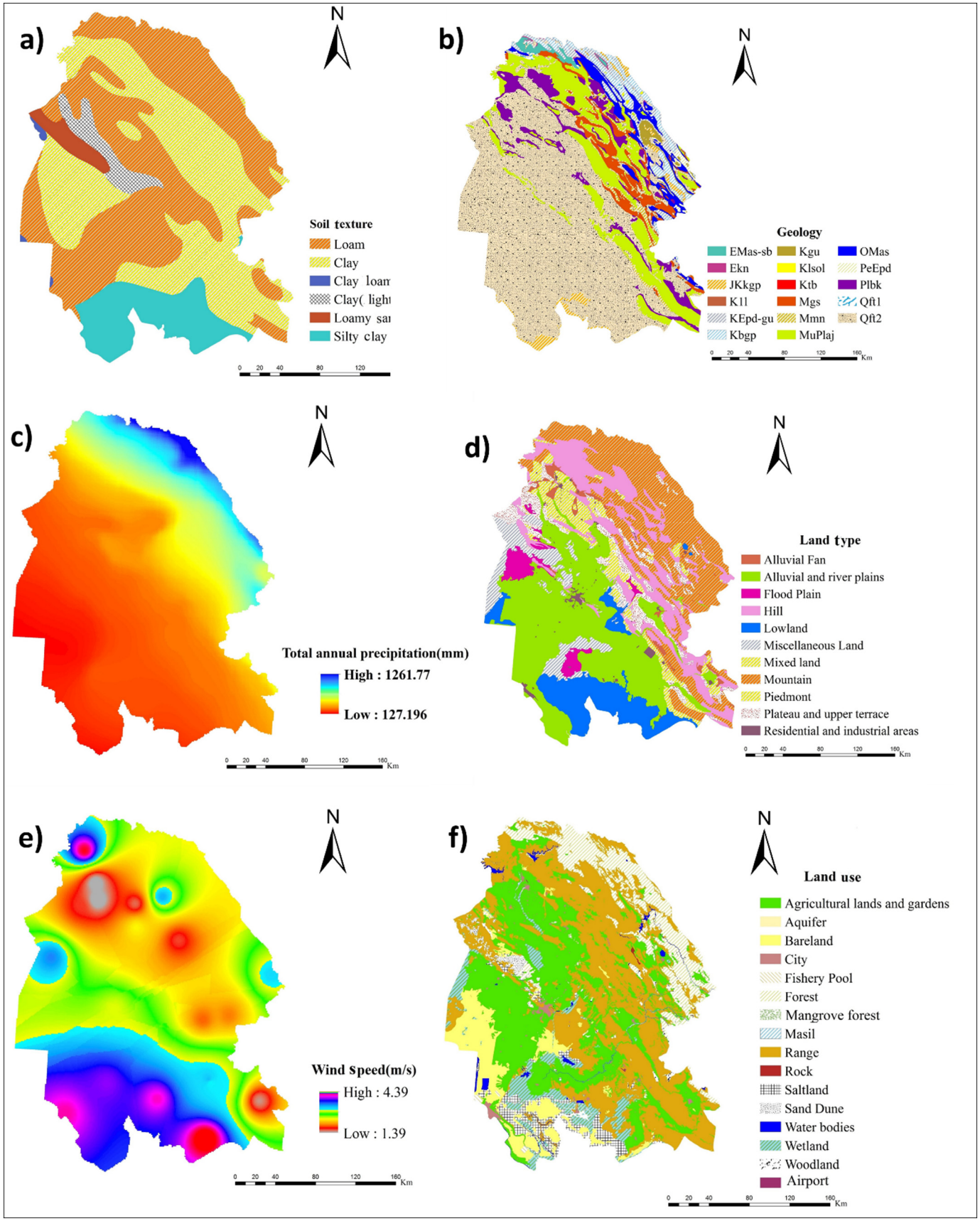

Fig. 5. Maps of effective factors for the dust emissions. a) Soil texture; b) geology; c) total annual precipitation; d) land type; e) wind speed and f) land use.

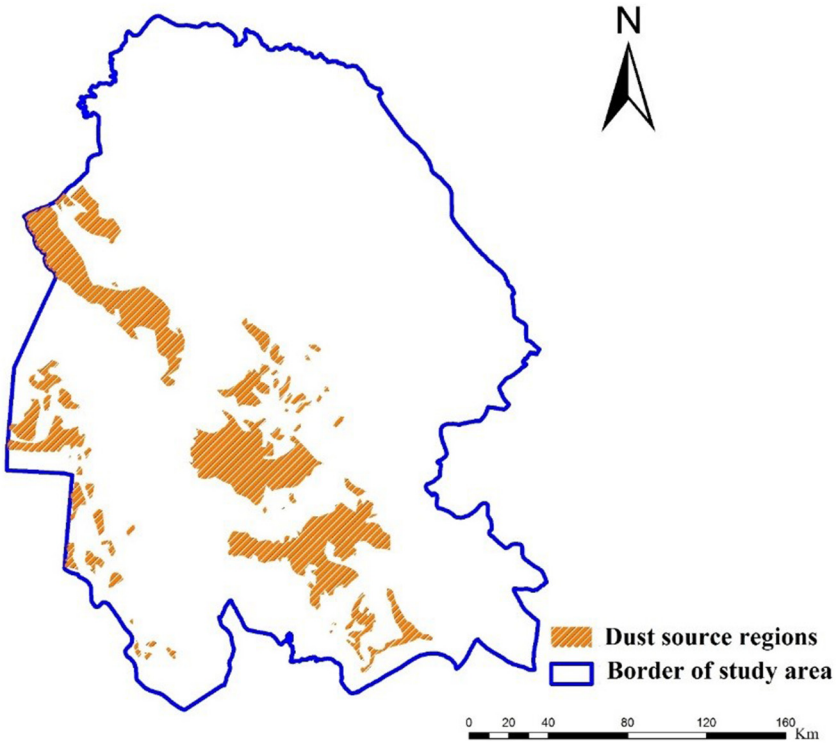

Fig. 6. The spatial distribution of dust provenance in the study area (Heydarian et al., 2018).

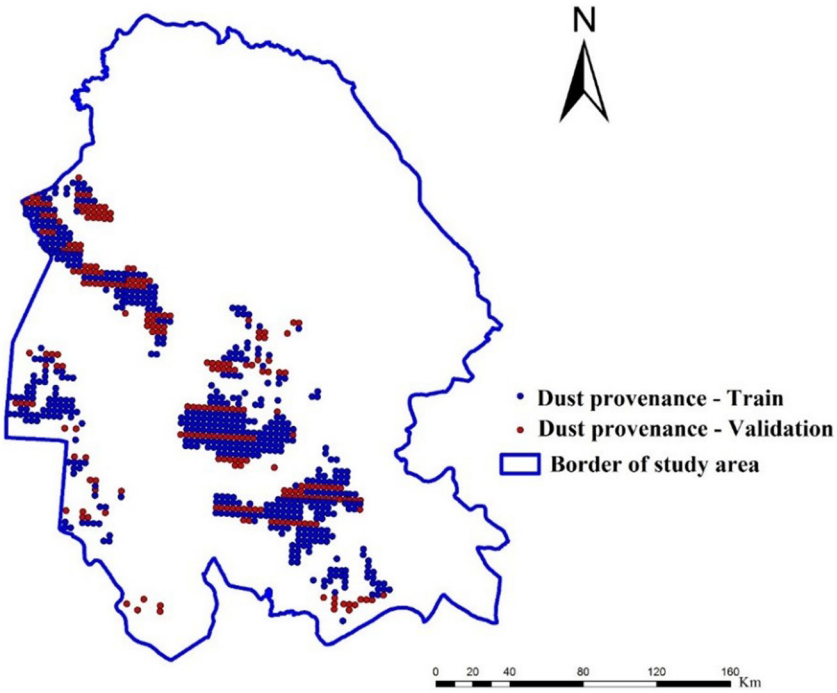

Fig. 7. The spatial distribution of 551 training (70%) and 236 validation (30%) pixels in the study area.

based on the M5 theory (Quinlan, 1992), generates multiple linear regression models in the terminal nodes of a tree. For improving model accuracy, a prediction resulted from the model is combined with predictions made from nearest-neighbour nodes (Xu et al., 2018).

3.5.8. Boosted regression trees (BRT)

The BRT (Elith et al., 2008) synthesizes the regression trees and boosting algorithms. To optimize predictions, this approach combines many simple tree models adaptively (Elith et al., 2008). This model has reduced sensitivity to over-fitting and high speed in the analysis of large datasets (Amiri et al., 2019).

3.6. Ensemble modelling (EM)

An ensemble modelling (EM) approach can be applied to combine different models into an overall synthesized model to improve accuracy (Sajedi-Hosseini et al., 2018). The ensemble modelling (EM) used:

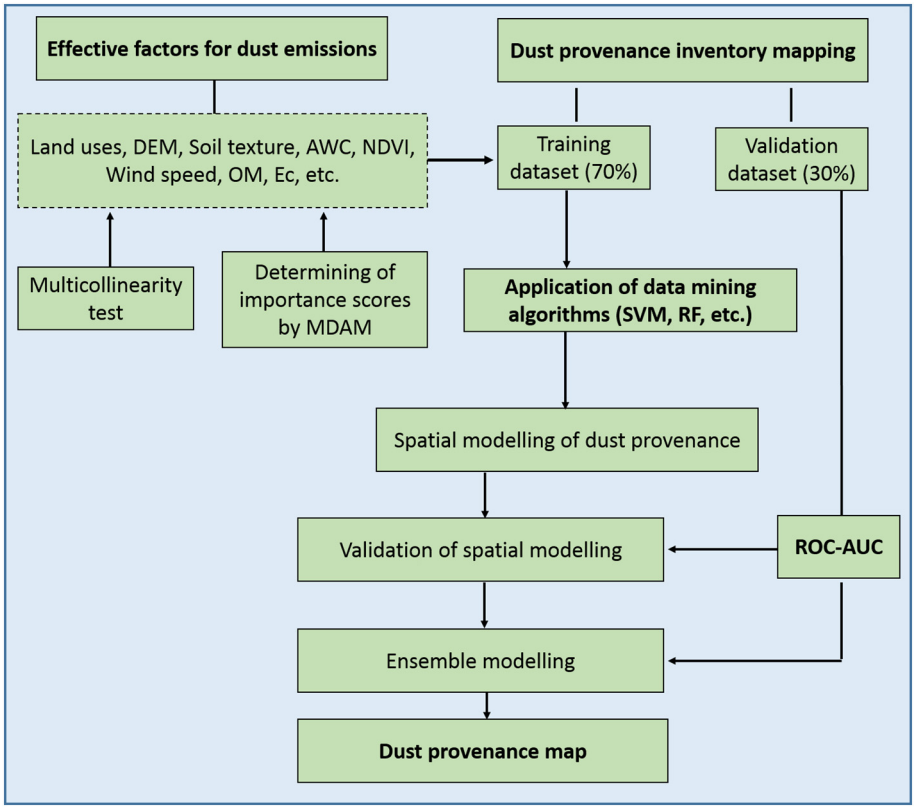

Fig. 8. Flowchart of the key stages in the spatial modelling of storm dust provenance using data mining algorithms and EM.

$$EM = \frac{\sum_{i=1}^{n} (AUC_i * M_i)}{\sum_{i=1}^{n} AUC_i} \quad (8)$$

where, AUC_i is the value of area under of curve for the ith single model (M_i).Models with AUC > 80% were used in the integration process (Choubin et al., 2019; Sajedi-Hosseini et al., 2018).

3.7. Evaluation of data mining and ensemble models

The receiver operating characteristic (ROC) curve, suggested by Hong et al. (2016), was applied for the validation of the dust provenance maps predicted by the data mining algorithms and the EM. The area under curve (AUC) was used to determine the accuracy of the storm dust provenance predictions provided by the different models. Values of AUC vary from 0.5 (a random prediction) to 1 (an excellent prediction) (Park, 2011). Fig. 8 shows a flowchart summarizing key steps in the spatial modelling of storm dust provenance using the above data mining algorithms and the resulting EM.

4. Results and discussion

4.1. Multicollinearity analysis

For calculating TC and VIF as collinearity statistics, values of 1 and 0 were assigned to pixels with and without dust provenance, respectively. Based on Table 4, the lowest and highest values of TC were calculated as 0.106 and 0.918, respectively, and these estimates were related to the DEM and land use. The respective lowest and highest estimates of VIF were 1.090 (for land use) and 9.426 (for the DEM). The lowest TC equated with the highest VIF (Table 4). Generally speaking, the results for TC and VIF revealed that multicollinearity among the 12 effective factors for storm dust emission was not a problem.

Table 4
The results of multicollinearity tests for the effective factors for storm dust emission.

Effective factors	Collinearity statistics	
	TC	VIF
Wind speed	0.427	2.342
Soil texture	0.602	1.662
AWC	0.750	1.334
Soil bulk density	0.548	1.826
Dem	0.106	9.426
Soil Ec	0.497	2.012
Geology units	0.559	1.788
Land type	0.707	1.414
Land use	0.918	1.090
NDVI	0.413	2.419
OM	0.109	9.149
Rainfall	0.274	5.711

4.2. Dust provenance maps

Figs. 9 and 10 present the storm dust provenance maps generated by the eight data mining algorithms. Table 5 illustrates the corresponding areas of dust provenance susceptibility classes predicted by the eight models. Using the RF model, 14 and 40% of the study area were classified as high and very high susceptibility classes, whereas 36 and 10% of the total area were assigned to low and moderate susceptibility classes, respectively (Fig. 9a). Based on the map of dust provenance susceptibility generated by the SVM (Fig. 9b), 33, 9, 17 and 41% of the total study area was classified as low, moderate, high and very high susceptibility, respectively. For the BART model, the very high susceptibility class covered 4% of the study area, whereas the low, moderate and very high susceptibility classes covered 65, 13 and 18%, respectively (Fig. 9c). According to the RBF model (Fig. 9d), low, moderate, high and very high susceptibility classes represented 37, 12,

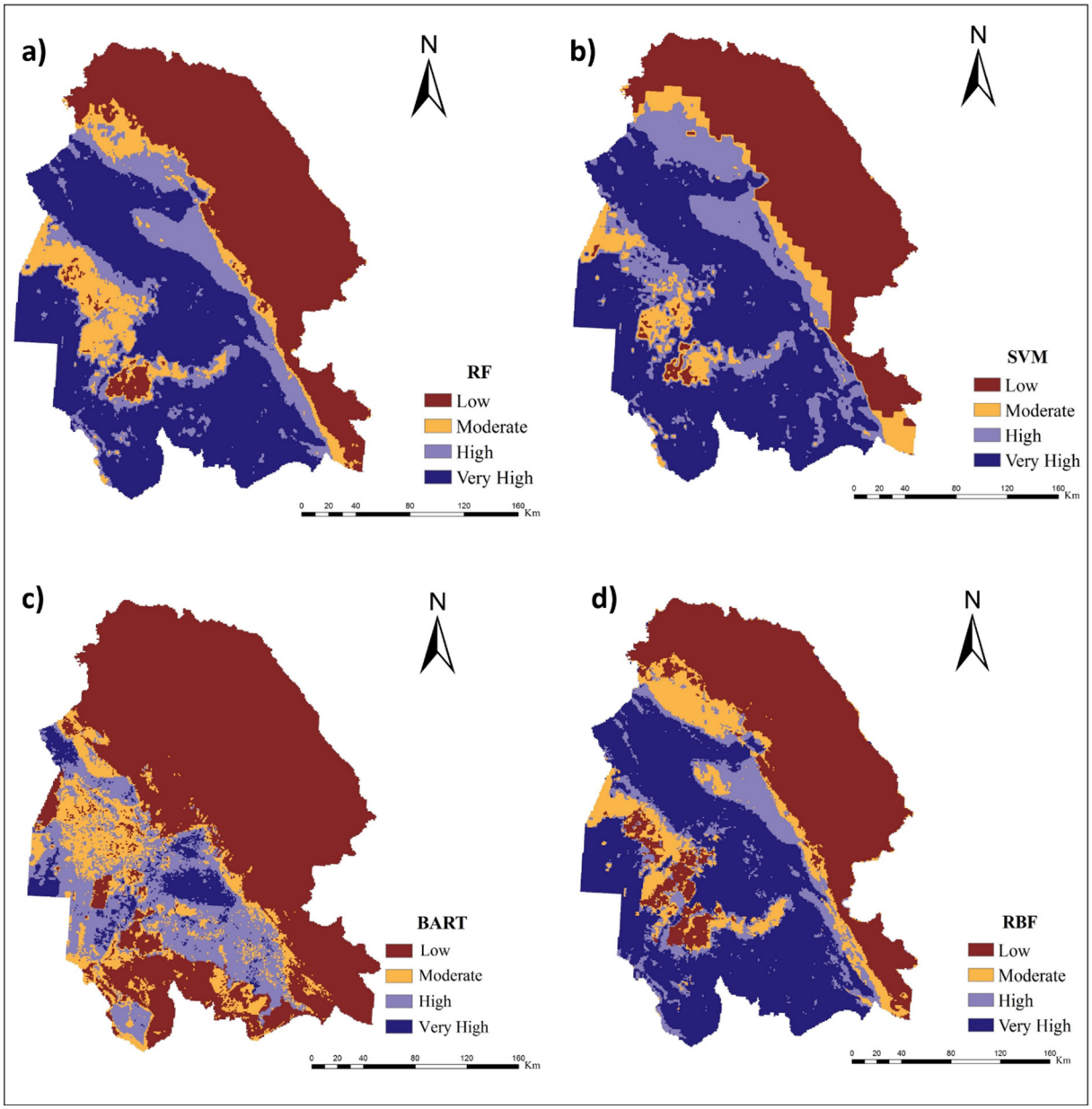

Fig. 9. Dust provenance maps generated by: (a) RF; (b) SVM; c) BART, and; d) RBF.

11 and 40%, respectively.

The results of the XGBoost model predicted that 37, 7, 13 and 43% of the total area belonged to the low, moderate, high and very high susceptibility classes, respectively (Fig. 10a). In the case of the RTA model (Fig. 10b), 35% of the area was classified as very high susceptibility, whereas 63, 2 and 0% were classified as low, moderate and high susceptibility. The map generated by the cubist model suggested that 8 and 49% of the study area were classified as high and very high susceptibility, whereas 39 and 4% of the area were classified as low and moderate susceptibility classes, respectively (Fig. 10c). Finally, in the case of the BRT model, the results classified 16, 25, 34 and 25% of the study area as low, moderate, high and very high susceptibility respectively (Fig. 10d).

4.3. Validation of the storm dust provenance maps using ROC-AUC

ROC-AUC has been employed in conjunction with classification problems in machine learning and in the evaluation of species distribution models (Phillips et al., 2006). Here, the model with the highest AUC value is selected as the fittest model. Based on the AUC value, model prediction accuracy can be classified to five categories (Yesilnacar, 2005): poor (50–60%), moderate (60–70%), good (70–80%), very good (80–90%) and excellent (90–100%). Here, the results of using ROC-AUC for evaluating the eight data mining models used to generate the storm dust provenance maps are presented in Figs. 11 and 12; and Table 6.

Among the eight models, six (RF, SVM, BART, RBF, XGBoost and Cubist) were confirmed as having excellent (with AUC > 90%) performance in the prediction of storm dust provenance (Fig. 11a–d; Fig. 12a, c). Based on Yesilnacar (2005), the BRT (AUC = 87.9%) and RTA (AUC = 79.1%) models were judged to be very good and good, respectively (Table 6; Figs. 12b,d).

In comparison with SVM, the XGBoost model has been shown to provide better performance for training data, is more stable and has much higher computational speed (Fan et al., 2018a). Chang et al. (2018) also reported that the XGBoost model, in comparison with other single-stage classifiers such as the SVM, is a superior tool for the development of risk models. Cubist, RF and XGBoost have been reported

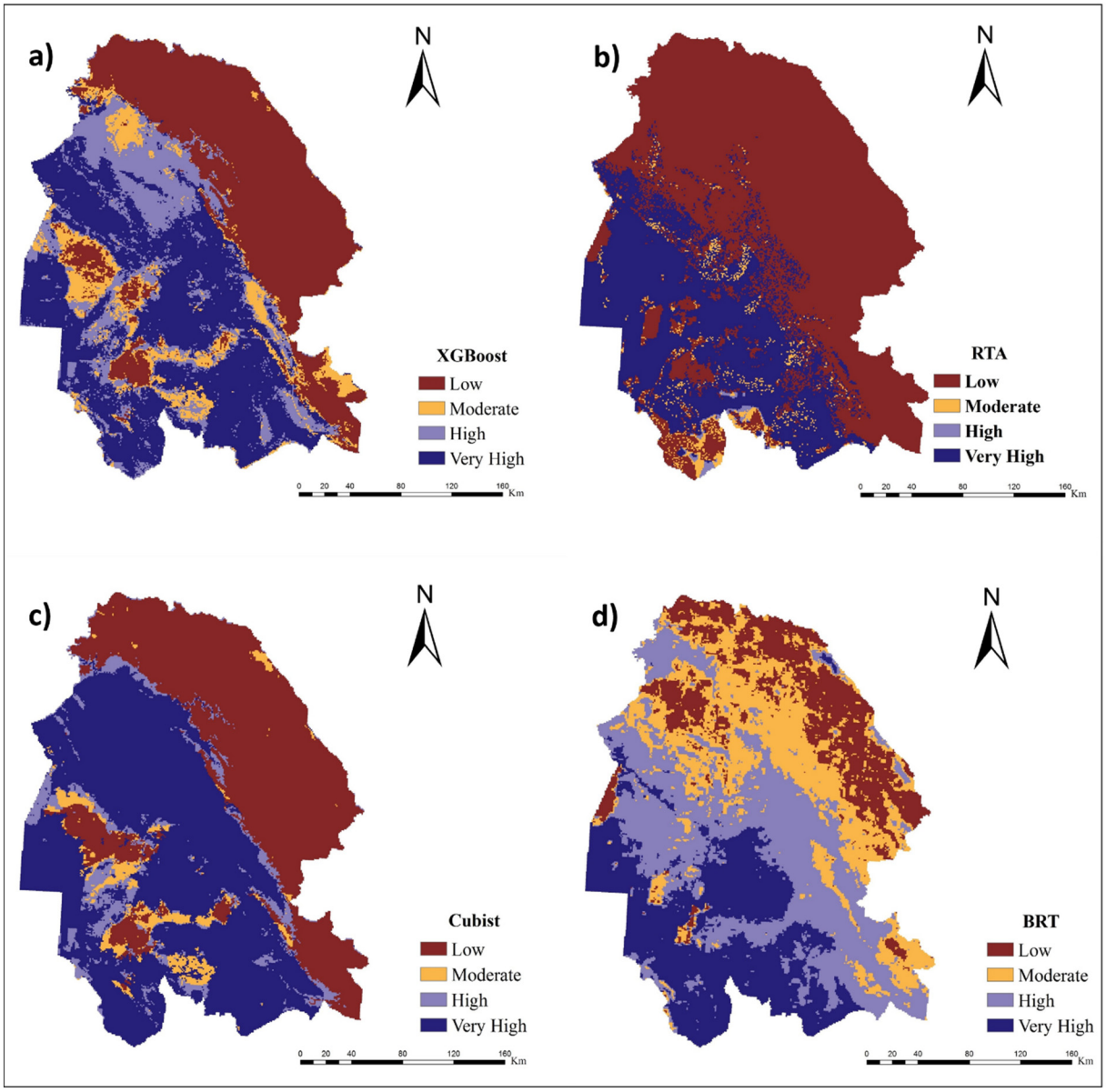

Fig. 10. Dust provenance maps generated by: (a) XGBoost; b) RTA; c) cubist and; d) BRT.

Table 5
The area of dust provenance susceptibility classes predicted by the eight data mining models.

Model	Class							
	Low		Moderate		High		Very high	
	Area (Km^2)	Area (%)	Area (Km^2)	Area (%)	Area (Km^2)	Area (%)	Area (Km^2)	Area (%)
RF	22,833	36	6595	10	8978	14	25,611	40
SVM	21,113	33	5646	9	10,944	17	26,316	41
BART	39,656	65	8088	13	10,811	18	2671	4
RBF	23,997	37	7445	12	6940	11	25,667	40
XGBoost	23,638	37	4311	7	8654	13	27,409	43
RTA	38,830	63	1194	2	–	–	21,211	35
Cubist	25,002	39	2527	4	4823	8	31,671	49
BRT	10,469	16	15,726	25	21,787	34	16,030	25

as having better performance relative to other machine learning algorithms, and the cubist model has been reported as the best for estimating monthly $PM_{2.5}$ (Xu et al., 2018). Both the cubist and RF models are capable of generating better predictions in comparison with ANN and conventional regression models (Zhou et al., 2019). Previous work has reported that the RBF classifier is a useful technique for the spatial modelling of landslides across the world (He et al., 2019).

In our work reported here, the AUC of the RF model (93.5%)

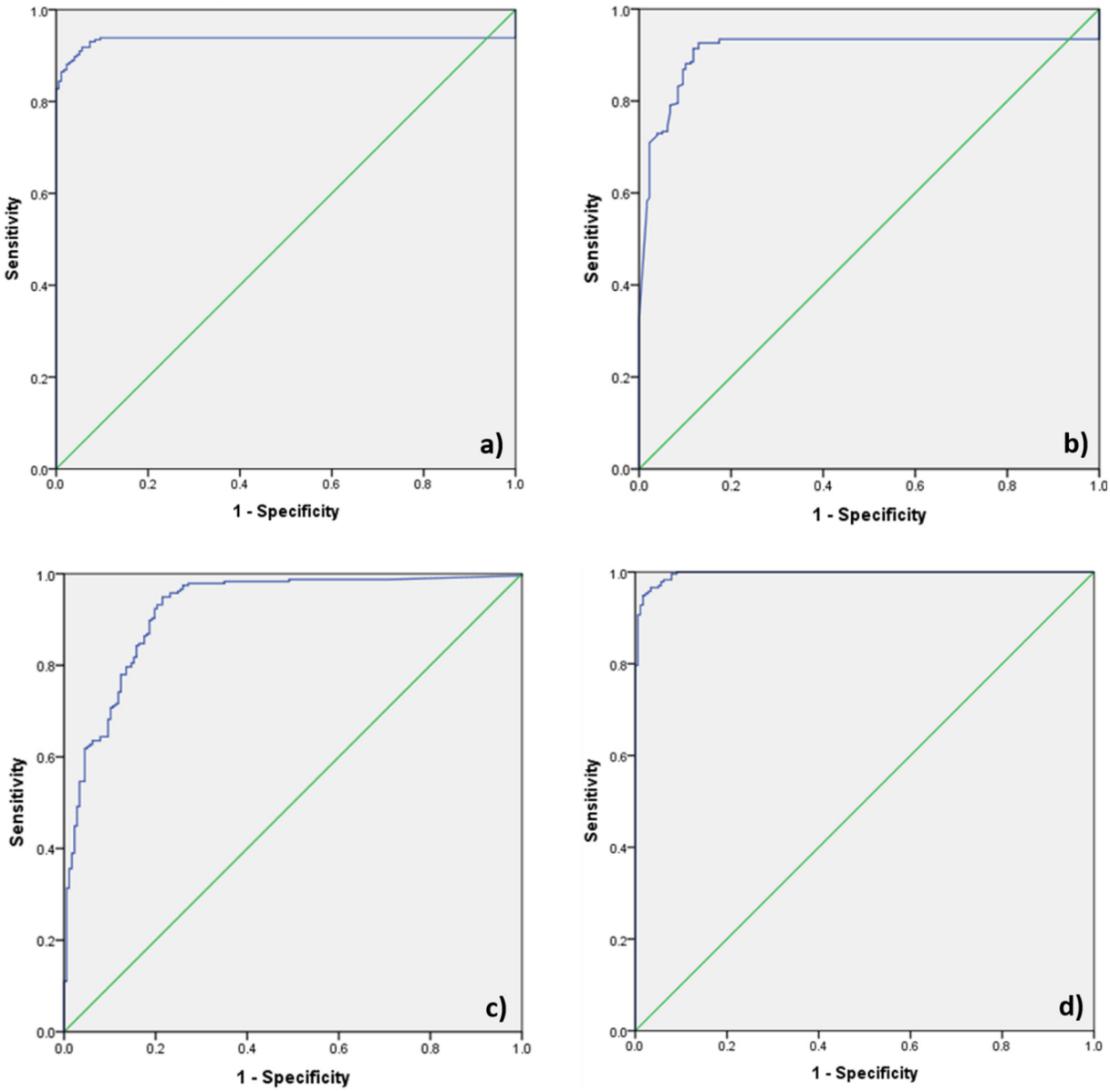

Fig. 11. Validation of the storm dust provenance maps generated by: (a) RF; (b) SVM; (c) BART, and; (d) RBF using the ROC-AUC.

(Fig. 11a) exceeded the corresponding values estimated for four other models comprising SVM (AUC = 91%; Fig. 11b), BART (AUC = 92.1%; Fig. 11c), RTA (AUC = 79.1%; Fig. 12d) and BRT (AUC = 87.9%; Fig. 12b), indicating that the RF model provides a better prediction map for storm dust provenance. Previously, work has reported that among three models (RF, CART and logistic model tree (LMT)), RF has the best capability for the spatial prediction of landslide susceptibility (Chen et al., 2017a). Similarly, in comparison with the RBDT, BRT and CART models, RF has previously been reported as generating the lowest predictive error during the mapping of landslide hazards (Rahmati et al., 2019). Gayen et al. (2019) and Amiri et al. (2019) also reported that the RF model in comparison with MARS, FDA, BRT and SVM, has the highest prediction accuracy for the assessment of gully erosion susceptibility.

Among all eight models, RTA (with AUC = 79.1%) (Fig. 12b) had the lowest accuracy for generating the storm dust provenance maps. The corresponding value for the BART model was 92.1% (Fig. 11c). Gomez-Gutierrez et al. (2009) also reported a lower efficiency for BART in comparison with MARS for predicting gullying. Similarly, in comparison with other models, the low prediction accuracy of BART has also been reported in conjunction with the digital mapping of soil carbon (Keskin et al., 2019), prediction of flood susceptibility (Choubin et al., 2019), and the spatial prediction of landslide susceptibility (Chen et al., 2017a,b).

4.4. Final storm dust provenance map produced by the EM

The individual models with AUC > 80% (see Table 6) were included in the EM process (Fig. 13 and Table 7). The EM results indicated that, 23 and 28% of the total study area were classified into high and very high susceptibility classes, whereas the low and moderate susceptibility classes accounted for 36 and 13%, respectively. Validation of the EM storm dust provenance map returned an AUC value of 99.8% (Fig. 14) which exceeded the corresponding AUC for the eight individual data mining models. Previous work has reported that EM have higher predictive accuracy than individual data mining models for gully erosion (Arabameri et al., 2019; Pourghasemi et al., 2017), suspended sediment load (Khosravi et al., 2018), flood susceptibility (Choubin et al., 2019), landslide susceptibility (Pham et al., 2017), and groundwater contamination risk (Sajedi-Hosseini et al., 2018).

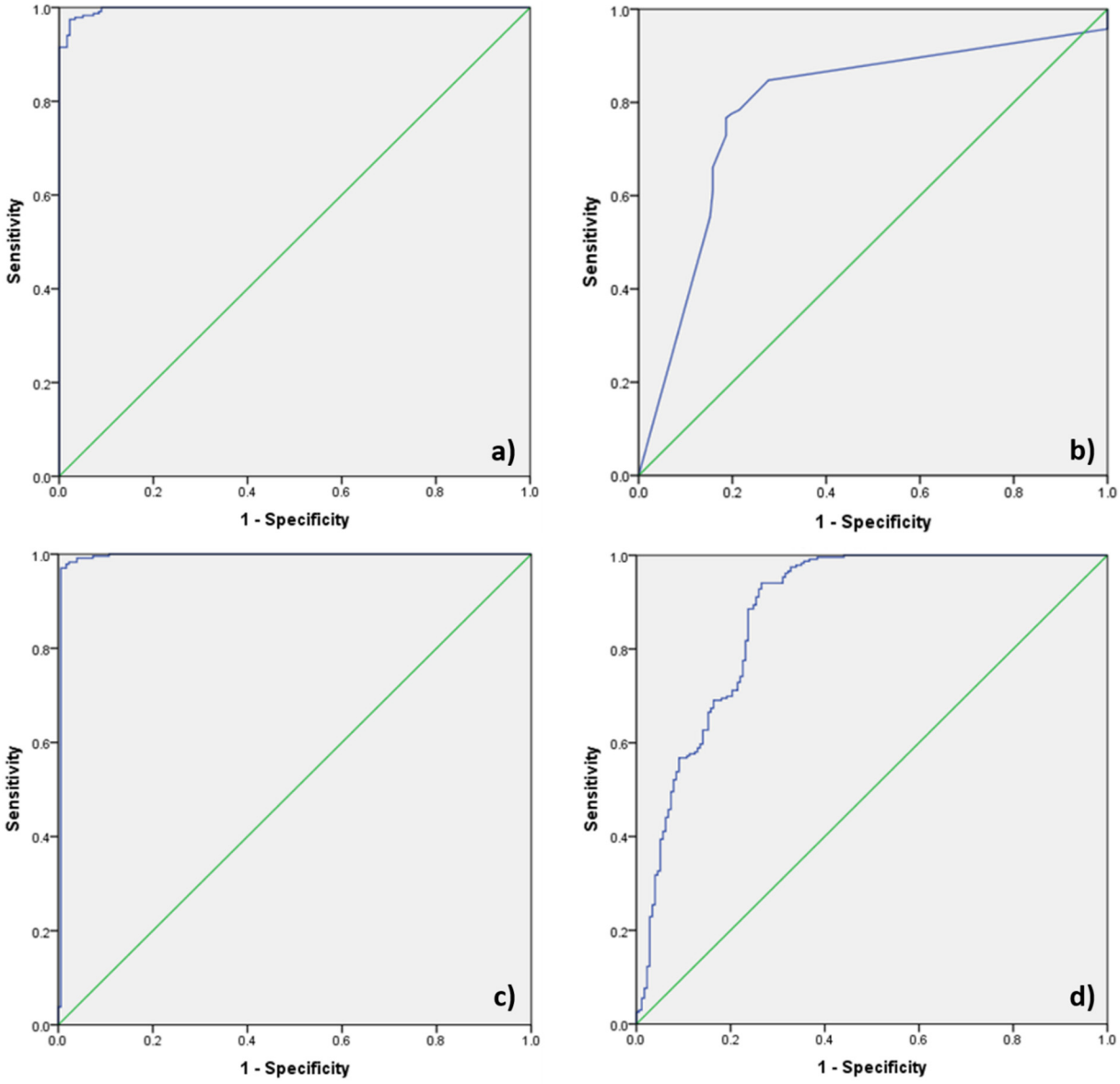

Fig. 12. Validation of the storm dust provenance maps generated by: (a) XGBoost; (b) RTA; (c) cubist, and; (d) BRT using the ROC-AUC.

Table 6
The values of AUC for data mining models applied to generating the map of storm dust provenance susceptibility.

Data mining models	AUC (%)
RF	93.5
SVM	91
BART	92.1
RBF	99.6
XGBoost	99.7
RTA	79.1
Cubist	99.3
BRT	87.9

4.5. Importance scores (IS) for effective factors for storm dust emissions

Storm dust emissions are controlled by many effective factors such as topographic conditions, land surface characteristics, soil moisture content, geology, soil characteristics (organic matter, AWC, etc), vegetation and atmospheric conditions (wind speed and rainfall) (Rashki et al., 2017; Saadoud et al., 2018; Ge et al., 2016; Shao, 2008). For our work, the importance scores (IS) for effective factors controlling storm dust emissions were determined by a mean decrease in accuracy measure (MDAM) (Fig. 15) (Chen et al., 2017a; Gayen et al., 2019). The MDAM results indicated that wind speed (IS = 23) is the most effective factor for dust emissions, followed by land use (IS =18), DEM (IS = 15.5), NDVI (IS = 14.5), precipitation (IS = 13), soil bulk density (IS = 12), soil Ec (IS = 11), soil AWC (IS = 9.8), OM (IS = 9.7), land type (IS = 8), soil texture (IS = 8) and geology (IS = 6.5). The wind speed (the most important factor) and rainfall (5th effective factor) as climatic variables play a major role in storm dust emissions. However, climate change isn't the only factor that will affect dust storms in the future; it is necessary to examine other future environmental changes caused by land use and land cover modifications in conjunction with human activities (Mahowald and Luo, 2003).

4.6. Limitations and advantages of data mining for mapping storm dust provenance

Data mining can help better understanding of the complex mechanisms behind environmental phenomena and in recent years, these

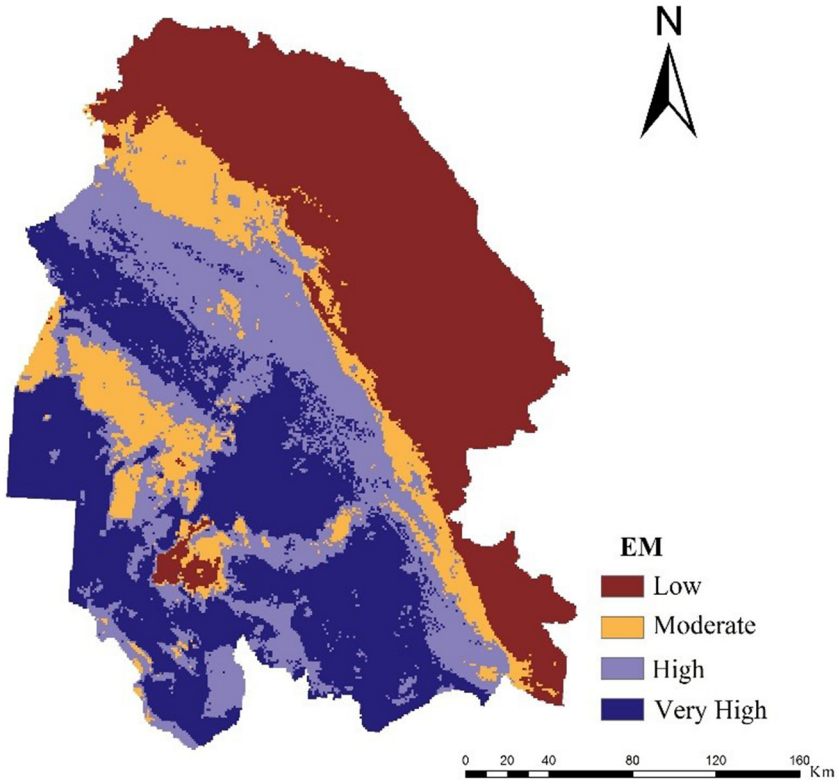

Fig. 13. Storm dust provenance susceptibility map generated by the EM.

techniques have been increasingly applied to environmental hazards (Gibert et al., 2018). The input variables are key controls to modelling processes and can influence model prediction accuracy. Therefore, preparing and mapping information on effective dust emission factors is vital and here it is important to be comprehensive in considering all the main variables affecting dust emissions. Different factors such as the spatial resolution of images, scale and number of measurement years can all affect the mapping of the effective factors for storm dust emissions such as NDVI, the digital elevation model, climatic variables (wind speed and rainfall) and soil characteristics.

Data mining models can help produce new validated and transferable knowledge by integrating various strands of existing knowledge and expertise. However, selecting the best machine learning model for an environmental hazard is not simple, and there is currently a lack of guidelines and criteria for guiding data scientists and environmental experts in the application of the available algorithms (Gibert et al., 2018). In the context of building upon the work reported here, we recommend future research examines the efficiency of other data mining models (e.g., boosted linear model, boosted generalized additive model, neural networks) for predicting dust provenance, so that progress can be made towards the development of guidelines for end-users.

5. Conclusion

The study reported herein is the first attempt at spatial modelling of storm dust provenance using eight individual data mining tools and an EM approach. Generally speaking, the AUC values for the eight data mining models and the EM ranged between 79.1 and 99.8% and on this basis it could be concluded that seven individual models (EM, RF, SVM, BART, RBF, XGBoost and Cubist) returned excellent, very good or good performances for the prediction of storm dust provenance. Overall, the EM (with AUC = 99.8%) returned the highest prediction accuracy for generating the storm dust provenance map. In comparison with other techniques for studying aeolian sediment provenance (such as geochemical source tracing methods), data mining models are inexpensive, do not require resource intensive field sampling and laboratory analyses, and critically, provide a basis for mapping provenance over large spatial scales. We therefore recommend applying data mining algorithms in conjunction with EM approaches to the spatial mapping of storm dust provenance worldwide.

Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgements

The authors would like to thank the Faculty of Agriculture and Natural Resources, University of Hormozgan, Iran for supporting this

Table 7
The area of storm dust provenance susceptibility classes predicted by the EM.

Model	Class							
	Low		Moderate		High		Very high	
	Area (Km^2)	Area (%)	Area (Km^2)	Area (%)	Area (Km^2)	Area (%)	Area (Km^2)	Area (%)
EM	21,937	36	8033	13	13,132	23	17,123	28

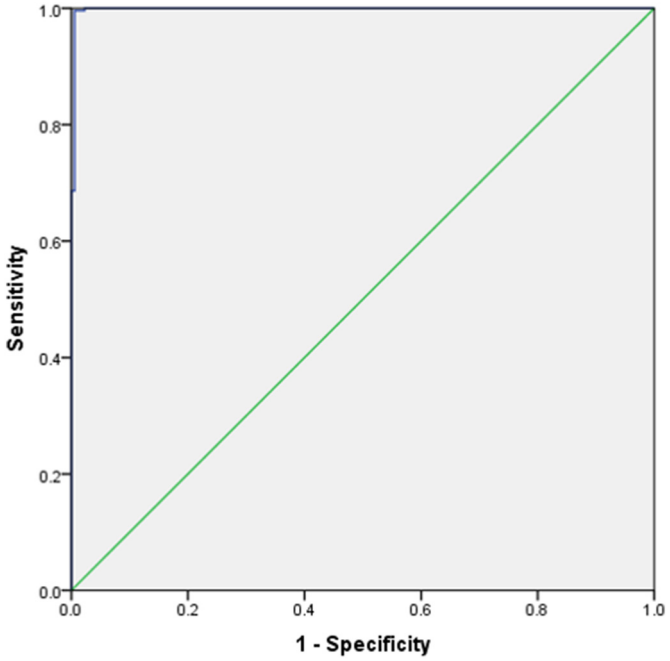

Fig. 14. Validation of the storm dust provenance map generated by the EM using the ROC-AUC.

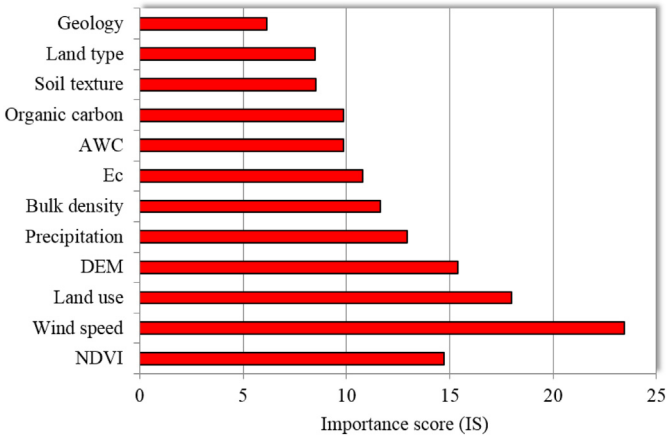

Fig. 15. The importance scores for effective factors for storm dust emissions using the MDAM.

joint research project. The contribution of ALC to this work was funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC) through grant award BBS/E/C/000I0330 (the Soil to nutrition institute strategic programme work package 3).

References

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018. An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index. Atmos. Res. 207 (15), 155–180. https://doi.org/10.1016/j.atmosres.2018.02.024.

Almasi, A., Mousavi, A.R., Bakhshi, S., Namdari, F., 2014. Dust storms and environmental health impacts. J. Mid. East Appl. Sci. Technol. (JMEAST) 8, 353–356.

Amiri, M., Pourghasemi, H.R., Ghanbarian, G.A., Afzali, S.F., 2019. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. Geoderma 340, 55–69. https://doi.org/10.1016/j.geoderma.2018.12.042.

Arabameri, A., Pradhan, B., Rezaei, K., 2019. Gully erosion zonation mapping using integrated geographically weighted regression with certainty factor and random forest models in GIS. J. Environ. Manag. 232, 928–942. https://doi.org/10.1016/j.jenvman.2018.11.110.

Barbulescu, A., Nazzal, Y., 2020. Statistical analysis of dust storms in the United Arab Emirates. Atmos. Res. 231, 104669. https://doi.org/10.1016/j.atmosres.2019.104669.

Beegum, S.N., Gherboudj, I., Chaouch, N., Temimi, M., Ghediram, H., 2018. Simulation and analysis of synoptic scale dust storms over the Arabian Peninsula. Atmos. Res. 199, 62–81. https://doi.org/10.1016/j.atmosres.2017.09.003.

Boente, C., Albuquerque, M.T.D., Gerassis, S., Rodriguez-Valdes, E., Gallego, J.R., 2019. A coupled multivariate statistics, geostatistical and machine-learning approach to address soil pollution in a prototypical Hg-mining site in a natural reserve. Chemosphere 218, 767–777. https://doi.org/10.1016/j.chemosphere.2018.11.172.

Breiman, L., 2001. Random forests. Mach. Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. In: The Wadsworth Statistics/Probability Series. Chapman and Hall, New York.

Bui, D.T., Pradhan, B., Lofman, O., Revhaug, I., Dick, O.B., 2012. Spatial prediction of landslide hazards in Vietnam: a comparative assessment of the efficacy of evidential belief functions and fuzzy logic models. Catena 96, 28–40. https://doi.org/10.1016/j.catena.2012.04.001.

Bui, B.T., Shahabi, H., Shirzadi, A., Chapi, K., Pradhan, B., Chen, W., Khosravi, k, Panahi, M., Ahmad, B., Saro, L., 2018. Land subsidence susceptibility mapping in South Korea using machine learning algorithms. Sensors 18, 2464. https://doi.org/10.3390/s18082464.

Chang, Y.C., Chang, K.H., Wu, G.J., 2018. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Appl. Soft Comput. 73, 914–920. https://doi.org/10.1016/j.asoc.2018.09.029.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

Chen, J., Zhong, J., Xie, Y., Cai, C., 2014. Text classification using SVM with exponential kernel. Appl. Mech. Mater. 519–520, 807–819. https://doi.org/10.4028/www.scientific.net/AMM.519-520.807.

Chen, W., Pourghasemi, H.R., Naghibi, S.A., 2017a. Prioritization of landslide conditioning factors and its spatial modeling in Shangnan County, China using GIS-based data mining algorithms. Bull. Eng. Geol. Environ. 77, 611–629. https://doi.org/10.1007/s10064-017-1004-9.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z., Ma, J., 2017b. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 151, 147–160. https://doi.org/10.1016/j.catena.2016.11.032.

Chepil, W.S., Siddoway, F.H., Armbrust, D.V., 1962. Climate factor for estimating wind erodibility of farm fields. J. Soil Water Conserv. 17, 162–165.

Choubin, B., Zehtabian, G., Azareh, A., Rafiei-Sardooi, E., Sajedi-Hosseini, F., Kişi, Ö., 2018. Precipitation forecasting using classification and regression trees (CART) model: acomparative study of different approaches. Environ. Earth Sci. 77, 314. https://doi.org/10.1007/s12665-018-7498-z.

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., Mosavi, A., 2019. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci. Total Environ. 651, 2087–2096. https://doi.org/10.1016/j.scitotenv.2018.10.064.

Dahmardeh Behrooz, R., Gholami, H., Telfer, M.W., Jansen, J.D., Fathabadi, A., 2019. Using GLUE to pull apart the provenance of atmospheric dust. Aeolian Res. 37, 1–13. https://doi.org/10.1016/j.aeolia.2018.12.001.

Demolli, H., Sakir Dokuz, A., Ecemis, A., Gokcek, M., 2019. Wind power forecasting based on daily wind speed data using machine learning algorithms. Energy Convers. Manag. 198 (15), 111823. https://doi.org/10.1016/j.enconman.2019.111823.

Dickson, M.E., Perry, G.L.W., 2016. Identifying the controls on coastal cliff landslides using machine-learning approaches. Environ. Model. Softw. 76, 117–127. https://doi.org/10.1016/j.envsoft.2015.10.029.

Elith, j, Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Fan, M., Hu, J., Cao, R., Ruan, W., Wei, X., 2018a. A review on experimental design for pollutants removal in water treatment with the aid of artificial intelligence. Chemosphere 200, 330–343. https://doi.org/10.1016/j.chemosphere.2018.02.111.

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., Xiang, Y., 2018b. Comparison of support Vector Machine and Extreme Gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. Energy Convers. Manag. 164, 102–111. https://doi.org/10.1016/j.enconman.2018.02.087.

Fernández, A.J., Sicard, M., Costa, M.J., Guerrero-Rascado, J.L., Gómez-Amo, J.L., Molero, F., Barragán, R., Basart, S., Bortoli, D., Bedoya-Velásquez, A.E., Utrillas, M.P., Salvador, P., Granados-Muñoz, M.J., Potes, M., Ortiz-Amezcua, P., Martínez-Lozano, J.A., Artíñano, B., Muñoz-Porcar, C., Salgado, R., Román, R., Rocadenbosch, F., Salgueiro, V., Benavent-Oltra, J.A., Rodríguez-Gómez, A., Alados-Arboledas, L., Comerón, A., Pujadas, M., 2019. Extreme, wintertime Saharan dust intrusion in the iberian Peninsula: lidar monitoring and evaluation of dust forecast models during the february 2017 event. Atmos. Res. 228, 223–241. https://doi.org/10.1016/j.atmosres.2019.06.007.

Frank, E., 2014. Fully Supervised Training of Gaussian Radial Basis Function Networks in WEKA.

Gayen, A., Pourghasem, H.R., Saha, S., Keesstra, S., Bai, S., 2019. Gully erosion susceptibility assessment and management of hazard prone areas in India using different machine learning algorithms. Sci. Total Environ. 668, 124–138. https://doi.org/10.1016/j.scitotenv.2019.02.436.

Ge, Y., Abuduwaili, J., Ma, L., Wu, N., Liu, D., 2016. Potential transport pathways of dust emanating from the playa of Ebinur Lake, Xinjiang, in arid Northwest China. Atmos. Res. 178–179, 196–206. https://doi.org/10.1016/j.atmosres.2016.04.002.

Gholami, H., Telfer, M.W., Blake, W.H., Fathabadi, A., 2017. Aeolian sediment fingerprinting using a Bayesian mixing model. Earth Surf. Process. Landf. 42, 2365–2376.

https://doi.org/10.1002/esp.4189.
Gholami, H., Dolat Kordestani, M., Li, J., Telfer, M.W., Fathabadi, A., 2019a. Diverse sources of aeolian sediment revealed in an arid landscape in southeastern Iran using a modified Bayesian un-mixing model. Aeolian Res. 41, 100547. https://doi.org/10.1016/j.aeolia.2019.100547.
Gholami, H., Jafari TakhtiNajad, E., Collins, A.L., Fathabadi, A., 2019b. Monte Carlo fingerprinting of the terrestrial sources of different particle size fractions of coastal sediment deposits using geochemical tracers: some lessons for the user community. Environ. Sci. Pollut. Res. 26 (22), 23206. https://doi.org/10.1007/s11356-019-05443-0.
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. Environ. Model. Softw. 110, 3–27. https://doi.org/10.1016/j.envsoft.2018.09.021.
Golla, V., Arveti, N., Etikala, B., Sreedhar, Y., Narasimhlu, K., Harish, P., 2019. Data sets on spatial analysis of hydro geochemistry of Gudur area, SPSR Nellore district by using inverse distance weighted method in Arc GIS 10.1. Data. Brief. 22, 1003–1011. https://doi.org/10.1016/j.dib.2019.01.030.
Gomez-Gutierrez, A., Schnabel, S., Francisco Lavado Contador, J., 2009. Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. Ecol. Model. 220, 3630–3637. https://doi.org/10.1016/j.ecolmodel.2009.06.020.
Goudie, A.S., 2014. Desert dust and human health disorders. Environ. Int. 63, 101–113. https://doi.org/10.1016/j.envint.2013.10.011.
Goudie, A.S., Middleton, N.J., 2006. Desert Dust in the Global System. Springer, pp. 287.
Hashemimanesh, M., Matinfar, H., 2012. Evaluation of desert management and rehabilitation by petroleum mulch base on temporal spectral analysis and field study (case study: Ahvaz, Iran). Ecol. Eng. 46, 68–74. https://doi.org/10.1016/j.ecoleng.2012.04.038.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. Springer Series in Statistics. Springer, New York.
He, Q., Shahabi, H., Shirzadi, A., Li, A., Chen, W., Wang, N., Chai, H., Bian, H., Ma, J., Chen, Y., Wang, X., Chapi, K., Ahmad, B.B., 2019. Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF classifier, and RBF Network machine learning algorithms. Sci. Total Environ. 663, 1–15. https://doi.org/10.1016/j.scitotenv.2019.01.329.
Heidari Farsani, M., Shirmardi, M., Alavi, N., Maleki, H., Sorooshian, A., Babaei, A., Asgharnia, H., Bagherian Marzouni, M., Goudarzi, G., 2018. Evaluation of the relationship between PM10 concentrations and heavy metals during normal and dusty days in Ahvaz, Iran. Aeolian Res. 33, 12–22. https://doi.org/10.1016/j.aeolia.2018.04.001.
Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214–215, 141–154. https://doi.org/10.1016/j.geoderma.2013.09.016.
Heydarian, P., Salehi, H., Fath Tabar, S., 2018. Dust Sources and Aeolian Sand Sheets in the Khuzestan Province Separated Based on Cities. Ministry of Industry, Mine and Business, Geological Survey and Mineral Explorations Country, Southwestern Region Organization, (Ahvaz) (In Farsi).
Hong, H., Pradhan, B., Bui, D.T., Xu, C., Youssef, A.M., Chen, W., 2016. Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China). Geomatics Nat. Hazards Risk 8 (2), 544–569. https://doi.org/10.1080/19475705.2016.1250112.
Houborg, R., McCabe, M.F., 2018. A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning. ISPRS J. Photogramm. Remote Sens. 135, 173–188. https://doi.org/10.1016/j.isprsjprs.2017.10.004.
IUSS-WRB, 2015. World reference base for soil resources 2014, update 2015. In: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports no. 106, Rome: FAO.
Kandola, J., Shawe-Taylor, J., Cristianini, N., 2003. Learning semantic similarity. In: Advances in Neural Information Processing Systems. 15. pp. 657–664.
Kapelner, A., Bleich, J., 2014. bartMachine: Machine Learning with Bayesian Additive Regression Trees. Cornell University.
Karimi, A., Shirmardi, M., Hadei, M., Tahmasebi Birgani, Y., Neisi, A., Takdastan, A., Goudarzi, G., 2019. Concentrations and health effects of short- and long-term exposure to PM2.5, NO2, and O3 in ambient air of Ahvaz city, Iran (2014–2017). Ecotoxicol. Environ. Saf. 180, 542–548. https://doi.org/10.1016/j.ecoenv.2019.05.026.
Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58. https://doi.org/10.1016/j.geoderma.2018.12.037.
Khosravi, K., Mao, L., Kisi, O., Yaseen, Z.M., Shahid, S., 2018. Quantifying hourly suspended sediment load using data mining models: case study of a glacierized andean catchment in chile. J. Hydrol. 567, 165–179. https://doi.org/10.1016/j.jhydrol.2018.10.015.
Lawrence, R.L., Wood, S.D., Sheley, R.L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman cutler classifications (random Forest). Remote Sens. Environ. 100, 356–362. https://doi.org/10.1016/j.rse.2005.10.014.
Lazri, M., Ameur, S., 2018. Combination of support vector machine, artificial neural network and random forest for improving the classification of convective and stratiform rain using spectral features of SEVIRI data. Atmos. Res. 203 (1), 118–129. https://doi.org/10.1016/j.atmosres.2017.12.006.
Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol. Inform. 6, 228–241. https://doi.org/10.1016/j.ecoinf.2010.12.003.
Li, Y., Song, Y., Kaskaoutis, D.G., Chen, X., Mamadjanov, Y., Tan, L., 2019. Atmospheric dust dynamics in southern central Asia: implications for buildup of Tajikistan loess sediments. Atmos. Res. 229, 74–85. https://doi.org/10.1016/j.atmosres.2019.06.013.
Mahowald, N., Luo, C., 2003. A less dusty future? Geophys. Res. Lett. 30 (17). https://doi.org/10.1029/2003GL017880.
Maleki, H., Soroshian, A., Goudarzi, G., Nikfal, A., Baneshi, M.M., 2016. Temporal profile of PM10 and associated health effects in one of the most polluted cities of the world (Ahvaz, Iran) between 2009 and 2014. Aeolian Res. 22, 135–140. https://doi.org/10.1016/j.aeolia.2016.08.006.
McTainsh, G.H., Lynch, A.W., Burgess, R.C., 1990. Wind erosion in eastern Australia. Aust. J. Soil Res. 28, 323–339.
Meyer, H., Kuhnlein, M., Appelhans, T., Nauss, T., 2016. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. Atmos. Res. 169 (B1), 424–433. https://doi.org/10.1016/j.atmosres.2015.09.021.
Middleton, N., 1986. Dust storms in the Middle East. J. Arid Environ. 10, 83–96.
Nabavi, S.O., Haimberger, L., Samimi, C., 2017. Sensitivity of WRF-chem predictions to dust source function specification in West Asia. Aeolian Res. 24, 115–131. https://doi.org/10.1016/j.aeolia.2016.12.005.
Nabavi, S.O., Haimberger, L., Abbasi, R., Samimi, C., 2018. Prediction of aerosol optical depth in West Asia using deterministic models and machine learning algorithms. Aeolian Res. 35, 69–84. https://doi.org/10.1016/j.aeolia.2018.10.002.
Naimabadi, A., Ghadiri, A., Idani, E., Babaei, A.A., Alavi, N., Shirmardi, M., Khodadadi, A., Bagherian Marzouni, M., Ahmadi Ankali, K., Rouhizadeh, A., Goudarzi, G., 2016. Chemical composition of PM10 and its in vitro toxicological impacts on lung cells during the Middle Eastern Dust (MED) storms in Ahvaz, Iran. Environ. Pollut. 211, 316–324. https://doi.org/10.1016/j.envpol.2016.01.006.
Nashwan, M.S., Shahid, S., 2019. Symmetrical uncertainty and random forest for the evaluation of gridded precipitation and temperature data. Atmos. Res. 230, 104632. https://doi.org/10.1016/j.atmosres.2019.104632.
Park, N.W., 2011. Application of Dempster-Shafer theory of evidence to GIS-based land slide susceptibility analysis. Environ. Earth Sci. 62, 367–376. https://doi.org/10.1007/s12665-010-0531-5.
Péré, J.C., Rivellini, L., Crumeyrolle, S., Chiapello, I., Minvielle, F., Thieuleux, F., Choël, M., Popovici, I., 2018. Simulation of African dust properties and radiative effects during the 2015 shadow campaign in Senegal. Atmos. Res. 199, 14–28. https://doi.org/10.1016/j.atmosres.2017.07.027.
Pham, B.T., Prakash, I., Bui, D.T., 2017. Spatial prediction of landslides using hybrid machine learning approach based on random subspace and classification and regression trees. Geomorphology 303, 256–270. https://doi.org/10.1016/j.geomorph.2017.12.008.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.
Pourghasem, H.R., Rossi, M., 2016. Landslide susceptibility modeling in a landslide prone area in Mazandarn province, North of Iran: a comparison between GLM, GAM, MARS, and M-AHP methods. Theor. Appl. Climatol. 130 (1–2), 609–633. https://doi.org/10.1007/s00704-016-1919-2.
Pourghasemi, H.R., Rahmati, O., 2018. Prediction of the landslide susceptibility: which algorithm, which precision? Catena 162, 177–192. https://doi.org/10.1016/j.catena.2017.11.022.
Pourghasemi, H.R., Yousefi, S., Kornejady, A., Cerda, A., 2017. Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. Sci. Total Environ. 609, 764–775. https://doi.org/10.1016/j.scitotenv.2017.07.198.
Quinlan, R., 1992. Learning with continuous classes. In: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, pp. 343–348 (Hobart, Australia, 16–18 November 1992).
Rahmati, O., Falah, F., Naghibi, S.A., Biggs, T., Soltani, M., Deo, R.C., Cerda, A., Mohammadi, F., Bui, D.T., 2019. Land subsidence modelling using tree-based machine learning algorithms. Sci. Total Environ. 672, 239–252. https://doi.org/10.1016/j.scitotenv.2019.03.496.
Rashki, A., Arjmand, A., Kaskaoutis, D.G., 2017. Assessment of dust activity and dust-plume pathways over Jazmurian Basin, Southeast Iran. Aeolian Res. 24, 145–160. https://doi.org/10.1016/j.aeolia.2017.01.002.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001.
Saadoud, D., Hassani, M., Peinado, F.J.M., Guettouche, M.S., 2018. Application of fuzzy logic approach for wind erosion hazard mapping in laghouat region (Algeria) using remote sensing and GIS. Aeolian Res. 32, 24–34. https://doi.org/10.1016/j.aeolia.2018.01.002.
Sachindra, A.D., Ahmed, K., Mamunur Rashid, M.D., Shahid, S., Perera, B.J.C., 2018. Statistical downscaling of precipitation using machine learning techniques. Atmos. Res. 212 (1), 240–258. https://doi.org/10.1016/j.atmosres.2018.05.022.
Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., Pradhan, B., 2018. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci. Total Environ. 644, 954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054.
Schepanski, K., Tegen, I., Macke, A., 2012. Comparison of satellite based observations of Saharan dust source areas. Remote Sens. Environ. 123, 90–97. https://doi.org/10.1016/j.rse.2012.03.019.
Shadman Roodposhti, M., Safarrad, T., Shahabi, H., 2017. Drought sensitivity mapping using two one-class support vector machine algorithms. Atmos. Res. 193, 73–82. https://doi.org/10.1016/j.atmosres.2017.04.017.
Shao, Y., 2008. Physics and modelling of wind erosion. In: Atmospheric and Oceanographic Sciences Library. vol. 37. pp. 459.
Vladimir, V.N., Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-

Verlag, New York, pp. 314.
Wessels, K.J., Prince, S.D., Malherbe, J., Small, J., Frost, P.E., VanZyl, D., 2007. Can human-induced land degradation be distinguished from the effects of rainfall variability? a case study in South Africa. J. Arid Environ. 68, 271–297. https://doi.org/10.1016/j.jaridenv.2006.05.015.
Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington.
Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14 (1), 1–37. https://doi.org/10.1007/s10115-007-0114-2.
Xu, Y., Ho, H.C., Wong, M.S., Deng, C., Shi, Y., Chan, C., Knudby, A., 2018. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level $PM_{2.5}$. Environ. Pollut. 242, 1417–1426. https://doi.org/10.1016/j.envpol.2018.08.029.
Yang, M., Zhu, X., Pan, H., Ai, W., Song, W., Pan, Y., 2019. Changes of the relationship between spring sand dust frequency and large-scale atmospheric circulation. Atmos. Res. 226 (15), 102–109. https://doi.org/10.1016/j.atmosres.2019.04.004.
Yesilnacar, E.K., 2005. The Application of Computational Intelligence to Landslide Susceptibility Mapping in Turkey. Ph.D Thesis. Department of Geomatics the University of Melbourne, pp. 423.
Zarasvandi, A., Carranza, E.J.M., Moore, F., Rastmanesh, F., 2011. Spatio-temporal occurrences and mineralogical–geochemical characteristics of airborne dusts in Khuzestan Province (southwestern Iran). J. Geochem. Explor. 111, 138–151. https://doi.org/10.1016/j.gexplo.2011.04.004.
Zhou, J., Li, E., Wei, H., Li, C., Qiao, Q., Armaghani, D.J., 2019. Random Forests and cubist algorithms for predicting shear strengths of rockfill materials. Appl. Sci. 9 (1621), 1–16. https://doi.org/10.3390/app9081621.