Geoscience Frontiers 12 (2021) 101232

HOSTED BY

Contents lists available at ScienceDirect

Geoscience Frontiers

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

Research Paper

APG: A novel python-based ArcGIS toolbox to generate absence-datasets for geospatial studies

Seyed Amir Naghibi [a,*], Hossein Hashemi [a], Biswajeet Pradhan [b,c,d,e]

[a] Department of Water Resources Engineering & Center for Advanced Middle Eastern Studies, Lund University, Lund, Sweden
[b] Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, New South Wales, Australia
[c] Department of Energy and Mineral Resources Engineering, Sejong University, Choongmu-gwan, 209 Neungdong-ro Gwangjin-gu, Seoul 05006, Republic of Korea
[d] Department of Meteorology, King Abdulaziz University, P. O. Box 80234, Jeddah 21589, Saudi Arabia
[e] Earth Observation Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 27 December 2020
Revised 12 May 2021
Accepted 19 May 2021
Available online 20 May 2021
Handling Editor: V.O. Samuel

Keywords:
Absence-dataset
Classification
Python
Machine learning algorithms
GIS
Groundwater
Hydrogeology

ABSTRACT

One important step in binary modeling of environmental problems is the generation of absence-datasets that are traditionally generated by random sampling and can undermine the quality of outputs. To solve this problem, this study develops the Absence Point Generation (APG) toolbox which is a Python-based ArcGIS toolbox for automated construction of absence-datasets for geospatial studies. The APG employs a frequency ratio analysis of four commonly used and important driving factors such as altitude, slope degree, topographic wetness index, and distance from rivers, and considers the presence locations buffer and density layers to define the low potential or susceptibility zones where absence-datasets are generated. To test the APG toolbox, we applied two benchmark algorithms of random forest (RF) and boosted regression trees (BRT) in a case study to investigate groundwater potential using three absence datasets i.e., the APG, random, and selection of absence samples (SAS) toolbox. The BRT-APG and RF-APG had the area under receiver operating curve (AUC) values of 0.947 and 0.942, while BRT and RF had weaker performances with the SAS and Random datasets. This effect resulted in AUC improvements for BRT and RF by 7.2, and 9.7% from the Random dataset, and AUC improvements for BRT and RF by 6.1, and 5.4% from the SAS dataset, respectively. The APG also impacted the importance of the input factors and the pattern of the groundwater potential maps, which proves the importance of absence points in environmental binary issues. The proposed APG toolbox could be easily applied in other environmental hazards such as landslides, floods, and gully erosion, and land subsidence.

1. Introduction

Sustainable development cannot be achieved without proper management of natural resources and environmental hazards such as water/groundwater (GW) resources, floods, landslides, and gullies. To produce valuable information for environmental degradation, many researchers have employed several algorithms ranging from simple models e.g., frequency ratio (FR) and weights-of-evidence to more complicated ones like machine learning algorithms (MLAs) such as random forest (RF), boosted regression trees (BRT), and support vector machines, and even ensemble approaches for assessing GW potential (Rahmati et al., 2016; Guru et al., 2017; Benjmel et al., 2020), or susceptibility to floods (Talukdar et al., 2020; Pham et al., 2021; Towfiqul Islam et al., 2021), gully erosions (Lei et al., 2020; Pourghasemi et al., 2020), landslides (Trigila et al., 2015; Lagomarsino et al., 2017; Park and Kim, 2019; Akinci et al., 2020; Segoni et al., 2020) as well as forest studies (Xu et al., 2020). In the case of groundwater potential assessment, data scarcity is the main reason that forces researchers to use machine learning algorithms compared to hydraulic models. This stems from a lack of piezometric, hydrodynamic, and hydrogeological data, particularly in developing countries, which makes such geospatial analysis applications important.

The majority of studies on geospatial issues have been merely focusing on enhancing the efficacy of the classification algorithms. Albeit there is an extent of uncertainty in those studies comprising the provision of the input factors, modeling components, i.e., cross-validation schemes, parameter optimization methods, as well as the delineation of absence-datasets that are fed into the MLAs.

* Corresponding authors.
E-mail address: seyed_amir.naghibi@tvrl.lth.se (S.A. Naghibi).

https://doi.org/10.1016/j.gsf.2021.101232

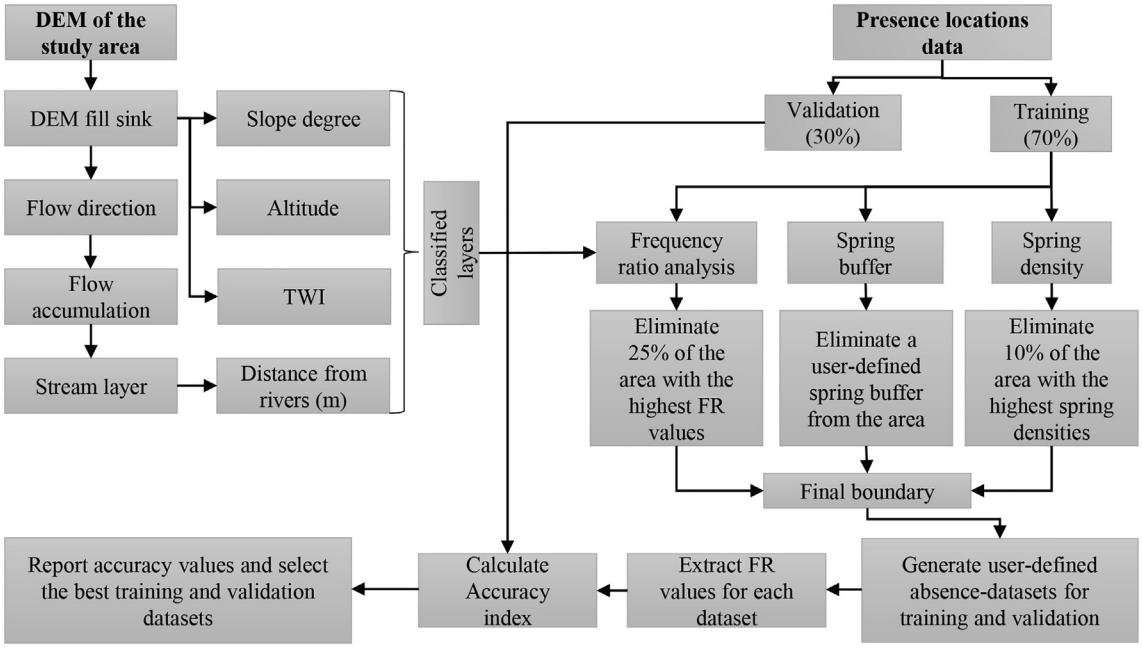


Fig. 1. APG toolbox schematic workflow.

The locations springs, floods, landslides, and gullies, are determined based upon on-site surveys and are typically more trustable than absence-datasets (Rahmati et al., 2019). A frequently conducted way for generating absence-dataset is the random generation of points in a given region. Nevertheless, this strategy cannot be the best option since some absence-points may randomly fall in the proximity of the presence-dataset. Apart from this, the random method does not examine any other criteria to connect the locations of the absence-dataset to driving factors.

According to the literature, Rahmati et al. (2019) developed a toolbox named selection of absence samples (SAS) which enhances the efficacy of the algorithms relying on the location (e.g., spring) and magnitude (e.g., discharges) of the presence data and confirmed the positive impact on the algorithm performances. Besides, Zhu et al. (2019) introduced a similar type of method for generating absence datasets for landslide studies. They computed the reliability of potential absence points according to dissimilarity in geospatial conditions among the absence and presence points. They reported that the best dissimilarity threshold was 0.5 while increasing the threshold to 0.9 decreased the performance of the algorithms. These methods follow different approaches, for instance, Rahmati et al. (2019) generated absence points only based on the geographic space, while Zhu et al. (2019)'s framework depends on the feature space.

Regarding the influence of absence-dataset on the efficiency of MLAs, this research intends to develop a novel framework called Absence Point Generation (APG) in the Python environment as an automated ArcGIS toolbox to generate absence-dataset for geospatial studies. The APG, on the other hand, integrates both aforementioned approaches in generating absence points to improve the performance of the algorithms. From the geographic space point of view, the APG considers a user-defined presence locations buffer and a pre-defined presence density function. Also, from the feature space point of view, the APG considers the important driving factors in GW, flood, landslide, and gully erosion studies, i.e., altitude, slope degree, distance from rivers, and topographic wetness index (TWI), all obtainable from a Digital Elevation Model (DEM), to calculate FR and remove high potentiate or susceptible zones from the final boundary for absence point generation. Overall, the APG provides an efficient and easy-to-use framework for generating absence points that requires a few input layers and factors and implements geographical and feature spaces to limit the area for generation of the absence points.

Table 1
Description of the inputs, user-defined thresholds, and outputs of the APG toolbox.

Model	Factor	Description
Inputs	Layers and folders	Presence locations, digital elevation model, and output folder.
	User-defined thresholds	Presence locations buffer to be erased from the boundary which could be selected from100, 200, 300, 400, and 500 m. Number of absence-datasets to create and select the best dataset with five options of 5, 10, 20, 25, and 50.
Outputs	Final_boundary.shp	A part of the boundary that remains after being erased by FR, and presence locations buffer and density layers; is used in the generation of absence-datasets.
	fr.rst	FR map obtained by the statistical analysis of driving factors, i.e., altitude, slope degree, distance from rivers, and TWI.
	Tr_x.shp and vl_y.shp	Selected training and validation datasets based on accuracy, where x and y show the number of selected datasets for training and validation, respectively.
	training.xlsx and validation.xlsx	The calculated accuracy values for each absence-dataset based on the FR map.

To test the effect of the APG toolbox on the classification efficiency of MLAs, we selected RF and BRT algorithms. RF and BRT have many advantages over other MLAs including the capability to handle a huge amount of input driving factors, producing highly accurate maps, and being resistant to outliers and overfitting (Mitchell, 2011). RF and BRT have proven to be robust geospatial estimators (Moghaddam et al., 2020). Thus, the objectives of the current study are: (i) developing the APG toolbox to reduce uncertainty in binary modeling of environmental problems through improving absence-dataset, (ii) determining the effect of the APG absence points on the efficacy of RF and BRT algorithms comparing to the random and SAS methods, and (iii) determining the impact of the APG toolbox on driving factors contribution.

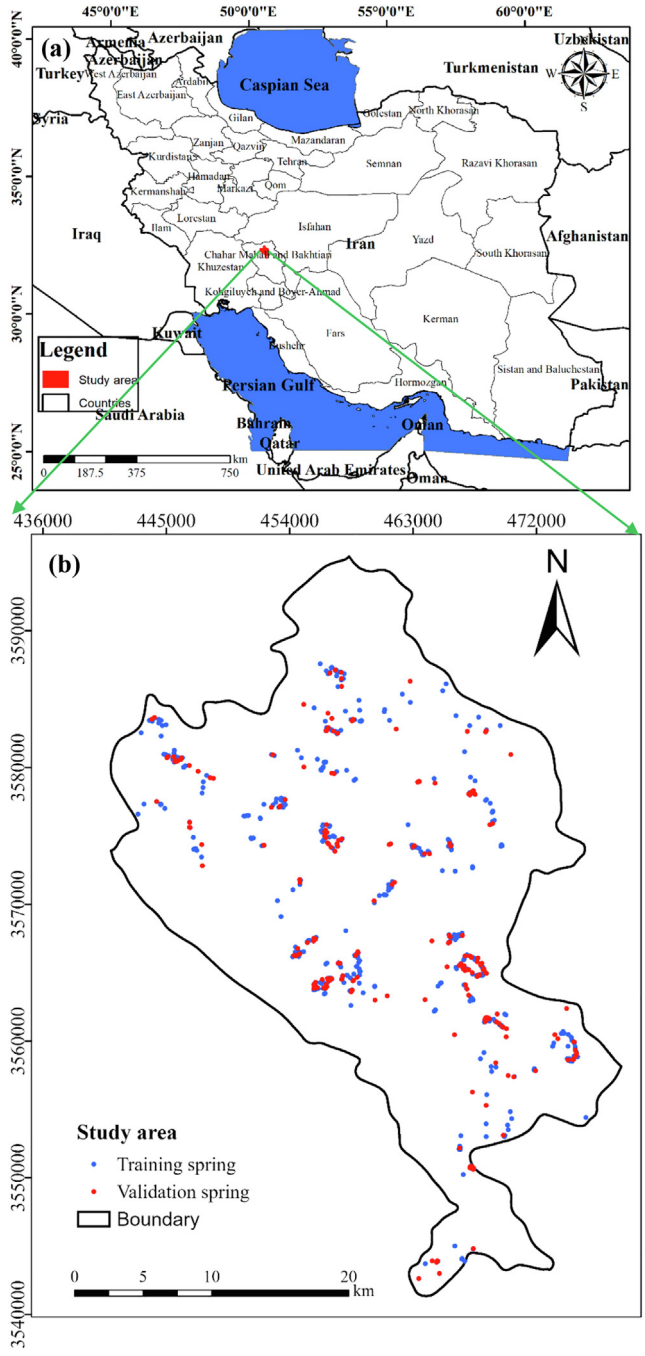


Fig. 2. (a) Location of the Farsan area in Iran, and (b) the training and validation spring locations in the Faresan region.

2. Materials and methods

In this section, we initially demonstrate the framework of the APG toolbox regarding its concepts, methods, inputs, user-defined and pre-defined thresholds, and outputs. Then, the APG absence dataset is used in a GW potential assessment and compared with the SAS and random datasets. This is done utilizing RF and BRT algorithms in a case study in Iran.

2.1. Absence point generation (APG) toolbox

The APG is a Python-based toolbox in ArcGIS relying on a statistical analysis of geospatial driving factors through FR analysis, and consideration of presence locations buffer and density to generate the final boundary in which the absence-dataset is generated. The entire process of the APG toolbox is illustrated in Fig. 1.

2.1.1. Selection, calculation, and classification of the driving factors

According to an extensive investigation of the geospatial research on GW potential and susceptibility to flood, gully, and landslide, we selected altitude, slope degree, TWI, and distance from rivers for FR analysis because they were repeatedly used as driving factors and proved to be major factors affecting the performance of classification algorithms (Kim et al., 2018; Razavi Termeh et al., 2018; Bui et al., 2020; Wang et al., 2020; Yousefi et al., 2020). The APG first fills the sinks in the provided DEM, calculates flow direction and flow accumulation layers, and then creates a stream layer. The APG calculates the slope degree of the study area by using the filled DEM. It calculates distance from the rivers layer, by using the "*Euclidean distance*" function on the stream layer. The APG calculates TWI as follows (Moore et al., 1991):

$$\mathrm{TWI} = \ln(\beta/\tan\alpha) \tag{1}$$

where β shows the upslope area and tanα points to the local slope radians.

For the FR analysis, altitude is categorized into five classes through an equal classification scheme to reveal the changes of presence data within various altitudes. Concerning slope degree classification, we considered three circumstances of maximum slope degree to extract as much useful information as possible from the dataset. If the maximum slope degree is over 40°, the classes will be from the minimum slope to 10°, 10° –20°, 20°–30°, and > 30°. If the maximum slope degree is between 20° and 40°, the slope is classified into the minimum slope to 5°, 5°–15°, 15°–30°, and > 30°. If the maximum slope is < 20°, slope classes will be from the minimum slope to 5°, 5°–10°, 10°–15°, and > 15°. TWI is classified into 0°–8°, 8°–12°, and > 12°, respectively, to represent the spatial pattern of the soil moisture under the best circumstance (Kanwal et al., 2017). For distance from rivers, five classes of 0–100, 100–200, 200–300, 300–400, and > 400 m were defined based on the literature (Pourghasemi and Beheshtirad, 2015).

2.1.2. Frequency ratio

FR is a commonly used statistical analysis approach in geospatial studies produced by Bonham-Carter (1994) with interpretable outputs. FR depicts the possibility of the presence of a certain feature, e.g., spring, flood, etc. It is in reliance on the incidence of the phenomenon in each class of factors. FR can be computed as below:

$$\mathrm{FR} = \frac{s}{S}/\frac{a}{A} \tag{2}$$

where, s refers to the number of presences in various classes, S depicts the total number of presences, a represents the area of each class, and A shows the total area of the region. FR has a positive value where higher and lower values depict higher and lower potential or susceptibility, respectively. The APG toolbox computes FR for all classes of the driving factors. Then, it sums the FR values of the four factors to generate the final potential or susceptibility map. To apply this method, the APG toolbox divides the presence locations into two groups for training and validation with a 70:30 ratio. Then, the APG uses training data for calculating FR and generating the FR potential or susceptibility map.

2.1.3. Presence locations buffer and density layers

Geospatial phenomena are not discrete and cannot be represented by just a pixel. For instance, proximities to flood, landslide, and gully locations tend to be more susceptible, and closer areas to springs are more likely to have higher potential. In the same manner, zones having greater presence densities tend to be more potentiate or susceptible to the investigated phenomenon. Thus,

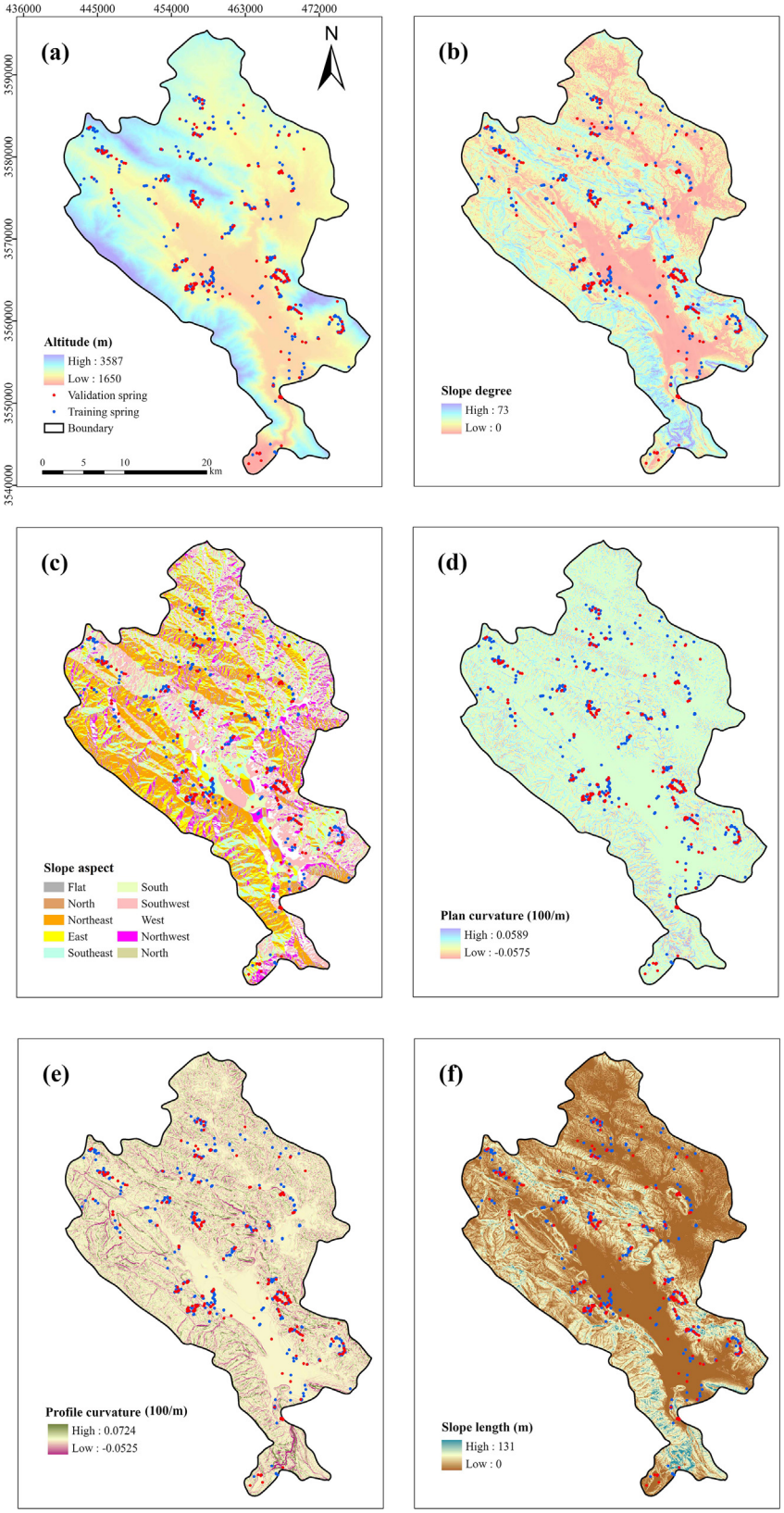


Fig. 3. GW spring potential driving factors in the Farsan area including (a) altitude, (b) slope degree, (c) aspect, (d) plan curvature, (e) profile curvature, (f) slope length.

the APG considers two other criteria relied on presence locations i.e., presence factor buffer and density eg., spring buffer and density. The APG toolbox considers these layers in obtaining the final boundary.

2.1.4. Generation of absence-datasets, accuracy assessment, and outputs

First, the boundary of the study area is eliminated by the three shapefiles produced based on FR, presence locations buffer, and

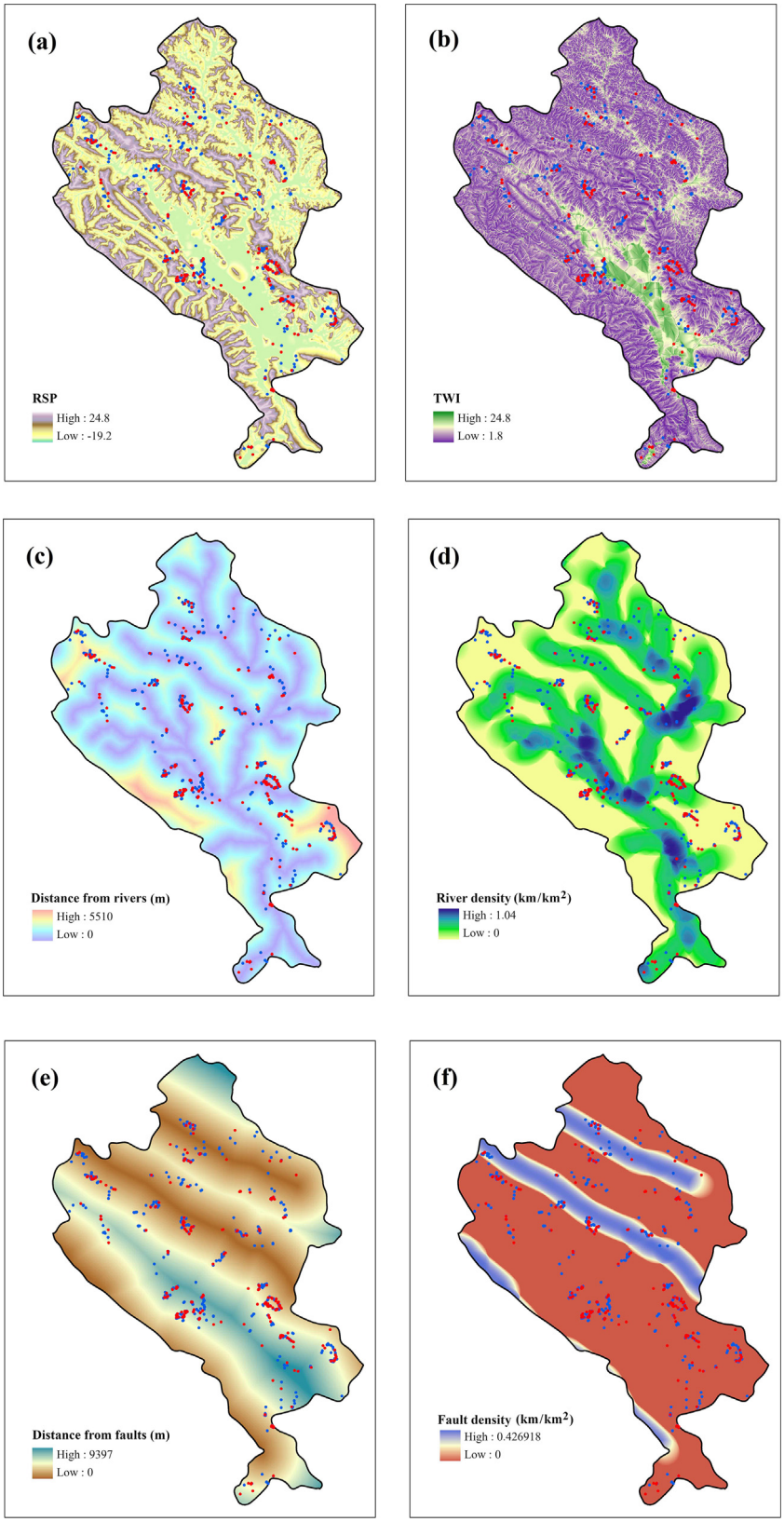


Fig. 4. GW spring potential driving factors in the Farsan area including (a) RSP, (b) TWI, (c) distance from rivers, (d) river density, (e) distance from faults, (f) fault density.

presence density layers considering pre-defined and user-defined thresholds. More specifically, 25% of the watershed has the greatest FR, and 10% of the area with the highest presence densities are eliminated. Further, a user-defined buffer from presence is created, to be eliminated from the boundary and form the final boundary. In the next step, *n* absence-datasets (user-defined) for both training and validation are randomly generated in the final boundary. Training and validation absence and presence-datasets

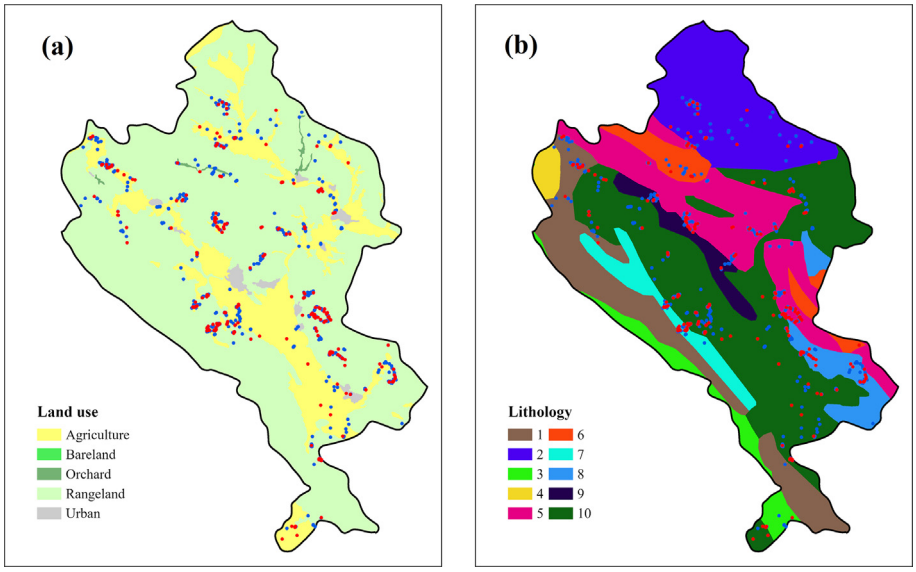


Fig. 5. GW spring potential driving factors in the Farsan area including (a) land use, and (b) lithology (see Table 2 for details).

Table 2
Lithological characteristics of the study region.

Class	Description	Age
1	Undivided Eocene rocks	Eocene
2	Grey, thick bedded, o'olitic, fetid limestone	Jurassic–Cretaceous
3	Undivided Bangestan Group, mainly limestone and shale, Albian to Companion, comprising the following formations: Kazhdumi, Sarvak, Surgah, and Ilam	Cretaceous
4	Grey and brown, medium bedded to massive fossiliferous limestone	Late Cretaceous
5	Grey thick bedded to massive orbitolina limestone	Early Cretaceous
6	Marl and calcareous shale with intercalations of limestone	Cretaceous
7	Cream to brown weathering, feature forming, well-jointed limestone with intercalations of shale	Miocene
8	Alternating hard of consolidated, massive, feature forming conglomerate and low weathering cross-bedded sandstone.	Pliocene
9	Polymictic conglomerate and sandstone	Pliocene
10	Low-level piedmont fan and valley terrace deposits	Quaternary

are merged, and *n* datasets of training and validation are produced. Then, the FR values for these sets are extracted. For determining the best training and validation datasets, the accuracy index is calculated for all datasets. The outputs of the APG toolbox are the final boundary, the best training, and validation datasets, accuracy results calculated for each training and validation dataset, and the FR map. Table 1 gives an overview of the APG toolbox.

2.2. Creating absence points by the SAS and Random methods

To test the APG performance, two previously implemented methods including random, and the SAS toolbox were used to generate absence points. The earlier was created by a random algorithm in ArcGIS. The latter was created using the SAS toolbox (Rahmati et al., 2019). This tool creates absence points with some limitations regarding presence locations buffer, hotspot buffer. The SAS also considers the average nearest neighbor to adjust the distribution pattern of the absence dataset.

2.3. Application of the APG toolbox on groundwater potential assessment

2.3.1. Study area and spring dataset

The Farsan area lies in Chaharmahal and Bakhtiari Province between 50°22′E and 50°46′E longitudes, and 32°00′N and 32°29′N latitudes occupying an area of 947 km^2 with a range of altitudes from 1650 to 3587 m above mean sea level (Fig. 2a). The average yearly rainfall in the Farsan area is measured to be 600 mm. The average temperature in the Faresan region is 12.1 °C with average 112 frosty days. The residents in the region are heavily dependent on agricultural activities. Spring locations data was acquired from the Chaharmahal and Bakhtiari Regional Water Authority (Chaharmahal and Bakhtiari Regional Water Authority CBRWA, 2019). The spring dataset includes 817 locations that are mainly consumed for the agriculture sector and drinking aim. The average measured pH and electrical conductivity of the springs are 7.67 and 641 μmhos cm^{-1}, respectively. The springs have an average discharge of about 2.65 L s^{-1}, with minimum and maximum values of 0.03 and 150 L s^{-1}. According to the literature, we considered a 70%:30% ratio for dividing the data for training (572 cases) and validation (245 cases) (Balogun et al., 2021; Fig. 2b).

2.3.2. Groundwater spring driving factors

To create the topographical factors of the Farsan area, a 20 m × 20 m DEM was implemented. There is a linkage between altitude and slope, which influences soil infiltration. Steeper slopes usually occur in mountainous regions where the altitude is greater. The altitude layer was obtained from the DEM of the Farsan area ranging from 1650 to 3587 m (Fig. 3a). The slope is another topographical factor that has an impact on flow velocity and soil infiltration. Low-slope regions own greater infiltration capacity and tend to be more GW productive (Daher et al., 2011). The slope in the Farsan area differs between 0° and 73.5° (Fig. 3b). The slope aspect is a high role driving factor greatly influenced by sunshine duration, which can affect soil moisture, water erosion, and subsequently soil infiltration (Fig. 3c). Plan and profile curvatures depict the curvature along the "*opposite and parallel to the slope direc-*

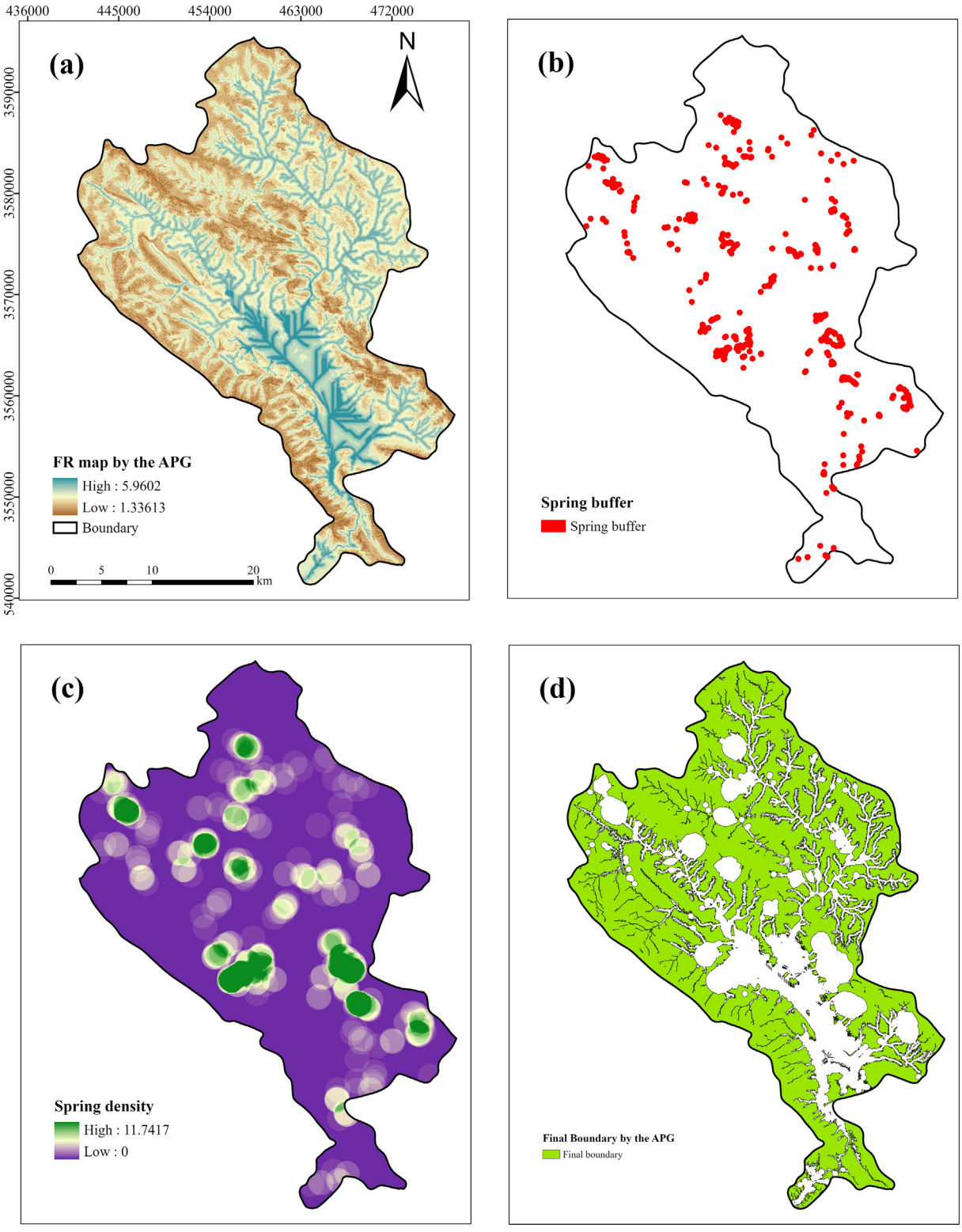


Fig. 6. Outputs of the APG toolbox including (a) the FR map, (b) the spring buffer, (c) the spring density, and (d) the final boundary.

tions", respectively (Ayalew and Yamagishi, 2005). These factors impact flow speed and erosion rate, and could subsequently impact soil infiltration (Fig. 3d, e). Flow accumulation at various portions of the hillside is a function of slope length that affects erosion, vegetation cover, and infiltration rate (Fig. 3f). Relative slope position (RSP) is regarded as the position of every point comparing to its surrounding points, including valleys and ridges. This factor was calculated using "System for Automatic Geoscientific Analyses (SAGA)" software (Fig. 4a). TWI is a proper indicator of soil moisture, which can also be associated with infiltration rate and GW potential (Raduła et al., 2018; Fig. 4b). Rivers are known as reliable GW recharge origins (Xi et al., 2010), especially in arid and semi-arid zones, because of intense rainfalls and floods that are the frequent events of these climates. Distance from rivers and river den-

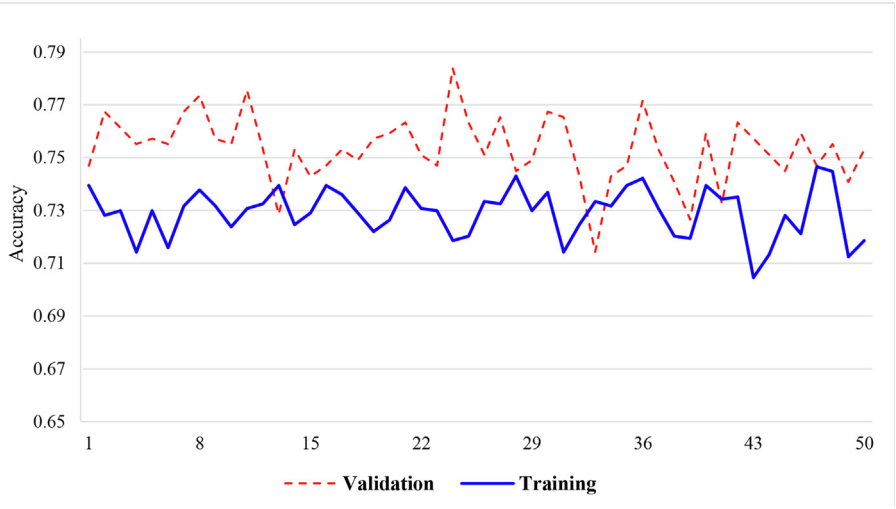


Fig. 7. Accuracy values calculated for the the training, and validation datasets by the APG toolbox.

sity layers are presented in Fig. 4c, d. Structural features have a critical function in GW recharge and movement as well as the appearance of the springs on the ground (Assatse et al., 2016). Since some parts of the Farsan basin are influenced by faults, this factor was included in the assessment through distance from faults and fault density (Fig. 4e, f). As per land use, forests and high vegetation cover areas are often regarded as high GW potential regions because of the plant's root development that facilitates surface water infiltration (Díaz-Alcaide and Martínez-Santos, 2019). The land-use layer is shown in Fig. 5a. Differences in lithological features often result in changes in rock strength and also soil infiltration, which can have significant impacts on GW potential (Geology Survey of Iran (GSI), 1997; Ozdemir, 2011) (Fig. 5b; Table 2).

2.3.3. Machine learning algorithms

RF is known as an MLA that generates many "*decision trees*" through bootstrapped subsets of the training set (Breiman, 2001; Loosvelt et al., 2012). The decision trees are created on distinct sets of data where the nodes are categorized by the best separating factor between m randomly chosen factors (Liaw and Wiener, 2018). This feature makes RF immune to "*overfitting*" and helps it to deal with hundreds of independent factors to estimate a target factor for both classification and regression issues (Breiman, 2001). In each step, RF uses 66% of the data to train, and what lefts is used to calculate the error which is named "*out-of-bag error*". The ultimate decision is made by simple voting of the trees' outputs. To map GW potential by RF, the "*randomForest*" script was implemented in R software.

BRT is an example of various algorithms developed to create a huge number of weak classifiers instead of setting up a single strong one (Schapire, 2003). BRT does not require the elimination of outliers and properly deals with missing values implementing surrogates and can also consider the interaction impacts among driving factors (Elith et al., 2008). It is capable of specifying the importance of the GW driving factors to GW potential assessment by the "relative influence" function. This can be done according to the number of times when every driving factor is chosen in constructing the trees (Shafizadeh-Moghadam et al., 2018). We implemented the "*gbm and caret scripts*" in the R software to model GW potential.

2.3.4. Validation of the GW potential maps

To compare the APG and random approaches for the generation of absence-datasets or non-springs, their impact on the RF and BRT outputs were assessed by various validation indices such as "receiver operating characteristics (ROC) curve, accuracy, kappa, sensitivity, and specificity" which were opted concerning several previous research on natural phenomena studies (Motevalli et al., 2019). The ROC curve plots the "*true positive rate*" against the "*false positive rate*" (Negnevitsky, 2005). The area under the curve of ROC varies from zero to one (Negnevitsky, 2005). Greater values show better performances of the algorithms, whereas lower values point to weaker performances. Sensitivity depicts the efficiency of the algorithms in predicting springs, and on the other hand, specificity denotes the performance of the algorithms in predicting non-springs. The accuracy denotes the proportion of truly categorized presence and absence cases to all cases. It is noteworthy that kappa is a quantitative assessment of the agreement among predicted and observed values. A kappa value of 1 represents the ideal model, which predicts all cases correctly (Viera and Garrett, 2005). Kappa and F1-score can be calculated as below:

$$\text{Kappa} = \frac{P_o/P_e}{1 - P_e} \tag{3}$$

$$P_o = \text{TP} + \text{TN}/n \tag{4}$$

$$P_e = (\text{TP} + \text{FN})(\text{TP} + \text{FP}) + (\text{FP} + \text{TN})(\text{FN} + \text{TN})/\sqrt{N} \tag{5}$$

$$\text{F1} - \text{score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \tag{6}$$

where n displays the ratio of springs that are appropriately labeled, N presents the total number of springs, TP shows "true positive", FP shows "false positive", TN is "true negative", FP points to "false positive".

3. Results

3.1. Training and validation datasets by the APG and random methods

For applying the APG toolbox and creating absence-datasets, we introduced the needed layers and entered two user-defined thresholds, including a spring buffer of 300 m, and 50 absence-datasets. The APG toolbox was applied and initially produced three output maps including the FR (Fig. 6a), spring buffer (Fig. 6b), and spring density (Fig. 6c). The APG eliminated 25% of the area having the highest FR values and 10% of the area having the highest spring

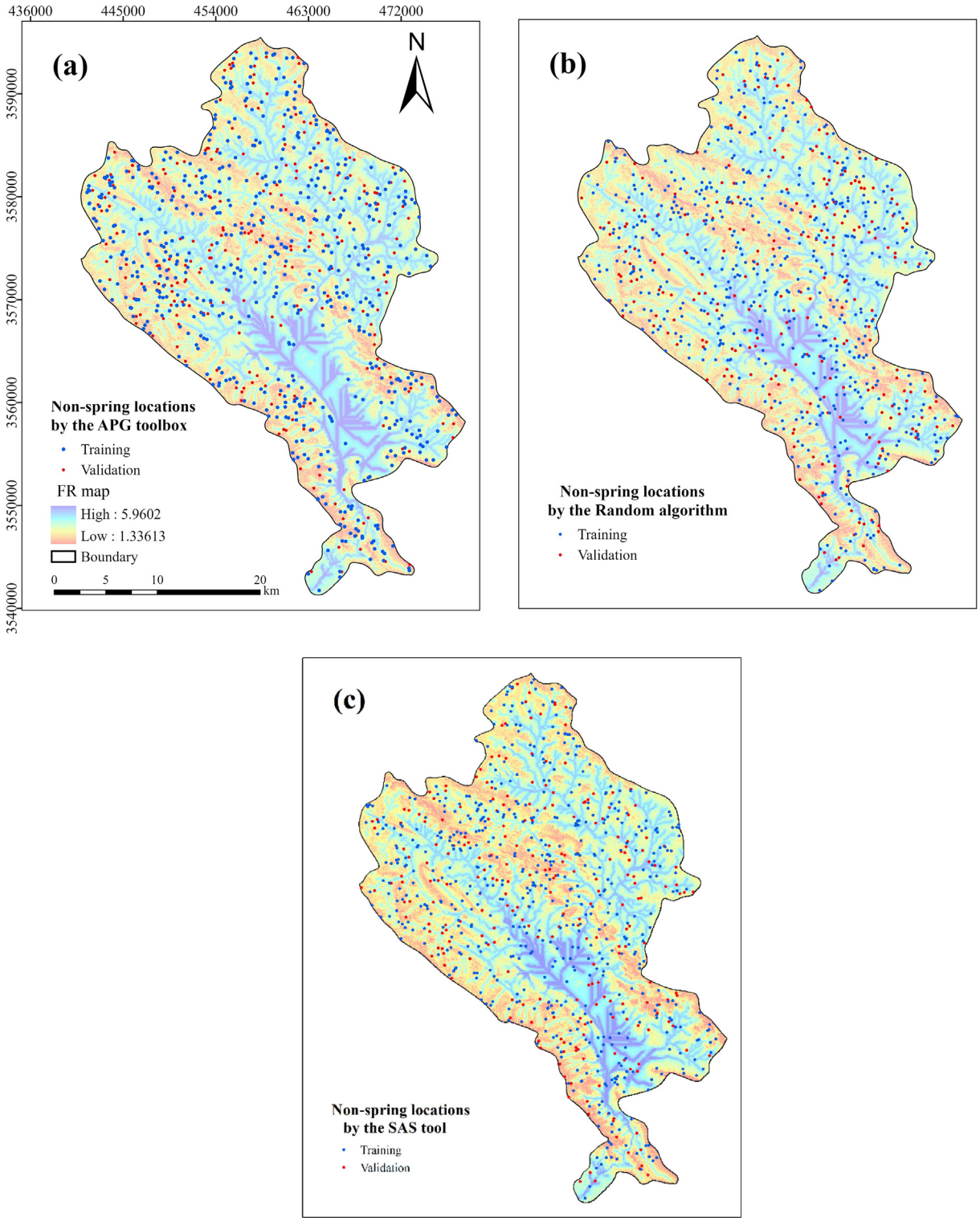


Fig. 8. The spatial distribution of the training and validation absence-datasets generated by (a) the APG toolbox, (b) the random algorithm, and the SAS method.

densities from the boundary based on the pre-defined threshold. Also, the APG eliminated a spring buffer of 300 m from the boundary of the Farsan area. The final boundary generated by the APG, which was used to create the random absence-datasets, is shown in Fig. 6d. Using the final boundary, 50 (user-defined) different absence-datasets for training and validation were randomly generated. Subsequently, 50 training and validation datasets were generated by integrating the training presence and training absence-dataset for the training dataset, and the validation presence and validation absence-datasets for the validation dataset.

To select the best datasets, the APG extracts the FR values and calculates the accuracy index for each training and validation dataset (Fig. 7). Accuracy values change from 0.704 to 0.746 for the training datasets and from 0.714 to 0.783 for the validation datasets. Eventually, the training and validation datasets with the highest accuracies were selected by the APG. We also created an

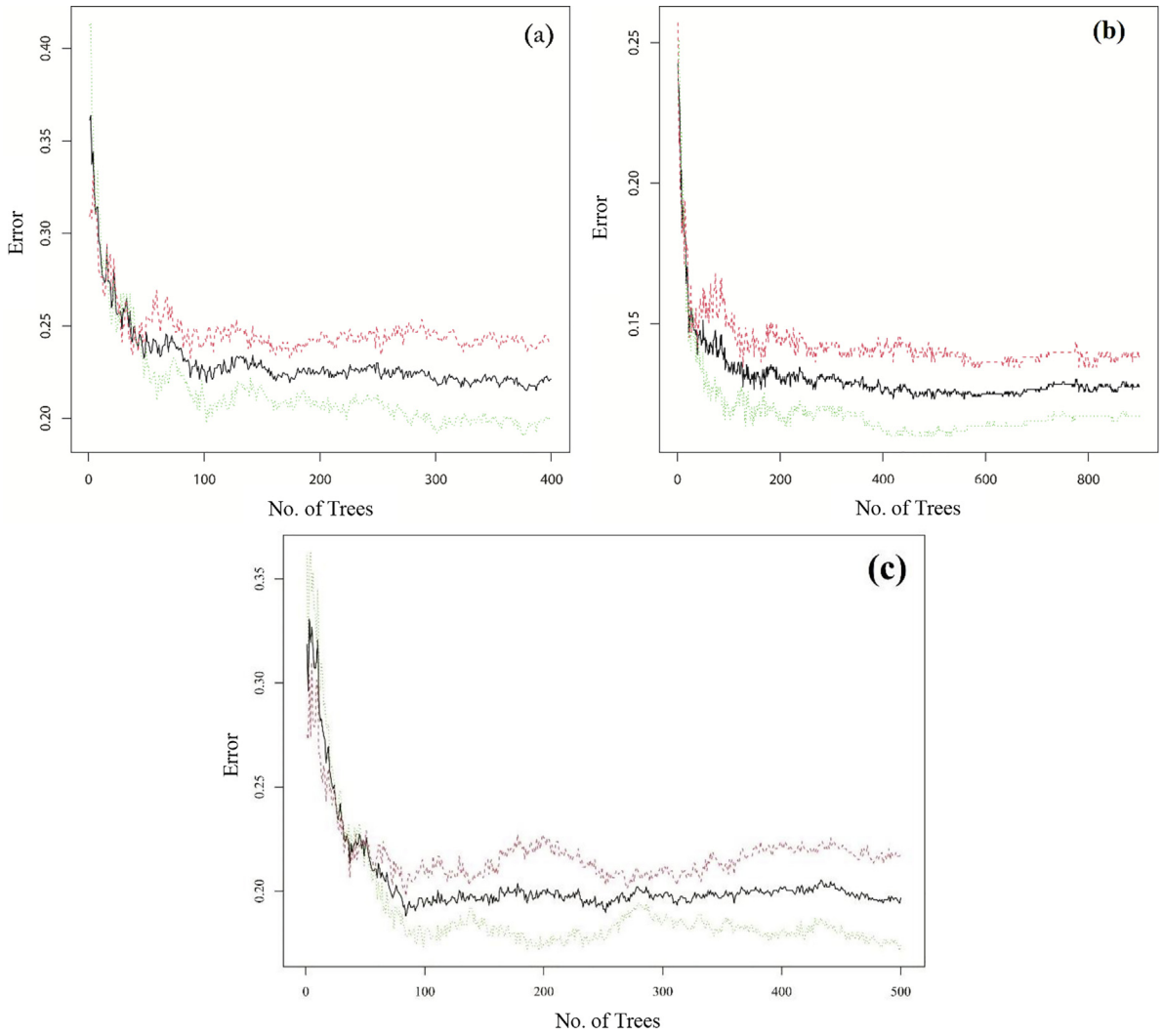


Fig. 9. The optimized number of trees according to out of bag predictions of the error rate by (a) the RF-Random, (b) the RF-APG, and (c) the RF-SAS. Red lines, green lines, and black lines represent the error for non-springs, springs, and total cases, respectively.

Table 3
Confusion matrix calculated for the training phase of the RF-APG, RF-Random, and RF-SAS.

Model/spring or non-spring	RF-APG			RF-Random			RF-SAS		
	Spring	Non-spring	Error rate	Spring	Non-spring	Error rate	Spring	Non-spring	Error rate
Spring	505	67	0.117	457	115	0.201	472	100	0.17
Non-spring	79	493	0.138	138	434	0.241	125	447	0.21

absence-dataset of the whole area using a random algorithm. The SAS method was applied with a 300 m buffer from springs, 500 m buffer from hotspots (based on the discharge values), and with the same number of absence points as springs to make it comparable with the APG dataset. The distribution of the random, APG and SAS absence-datasets in the Farsan area are shown in Fig. 8a–c, respectively. A visual investigation of the locations of the absence-points by the random method reveals that several points have fallen on the high GW potential zones obtained by the FR, while in the case of the APG dataset, points are located in the low to medium GW potential zones. Also, a ratio of the generated points by the SAS tool has been fallen into the high FR area.

3.2. RF optimization results by the random, APG and SAS datasets

Using the random dataset for the training of the RF algorithm, node size of 5, 3 factors at each node, and 400 trees were achieved with an error rate of 0.236. Whereas, by the APG dataset, the RF was optimized with a node size of 5, 3 factors at each node, and 900 trees with an error ratio of 0.143. Further, the RF-SAS was optimized with node size of 5, 2 factors at each node, and 500 trees with an error ratio of 0.21. Fig. 9a–c illustrates the errors for springs, non-springs, and total cases as a function of the number of trees for the RF-Random, RF-APG, and RF-SAS. It can be observed that the error percentages of the three algorithms decrease when the number of trees increases (Fig. 9). However, the RF-APG depicts relatively more stable results regarding error fluctuations between the prediction of springs and non-springs. Further, the RF-SAS and RF-APG have produced noticeably lower error rates and better performances comparing to the RF-Random at the training stage.

The confusion matrix of the RF algorithm is depicted in Table 3. As can be observed, the RF-Random correctly estimated 434 cases out of 572 non-springs and 457 out of 572 springs for the training dataset. The confusion matrix also depicts that the RF-APG success-

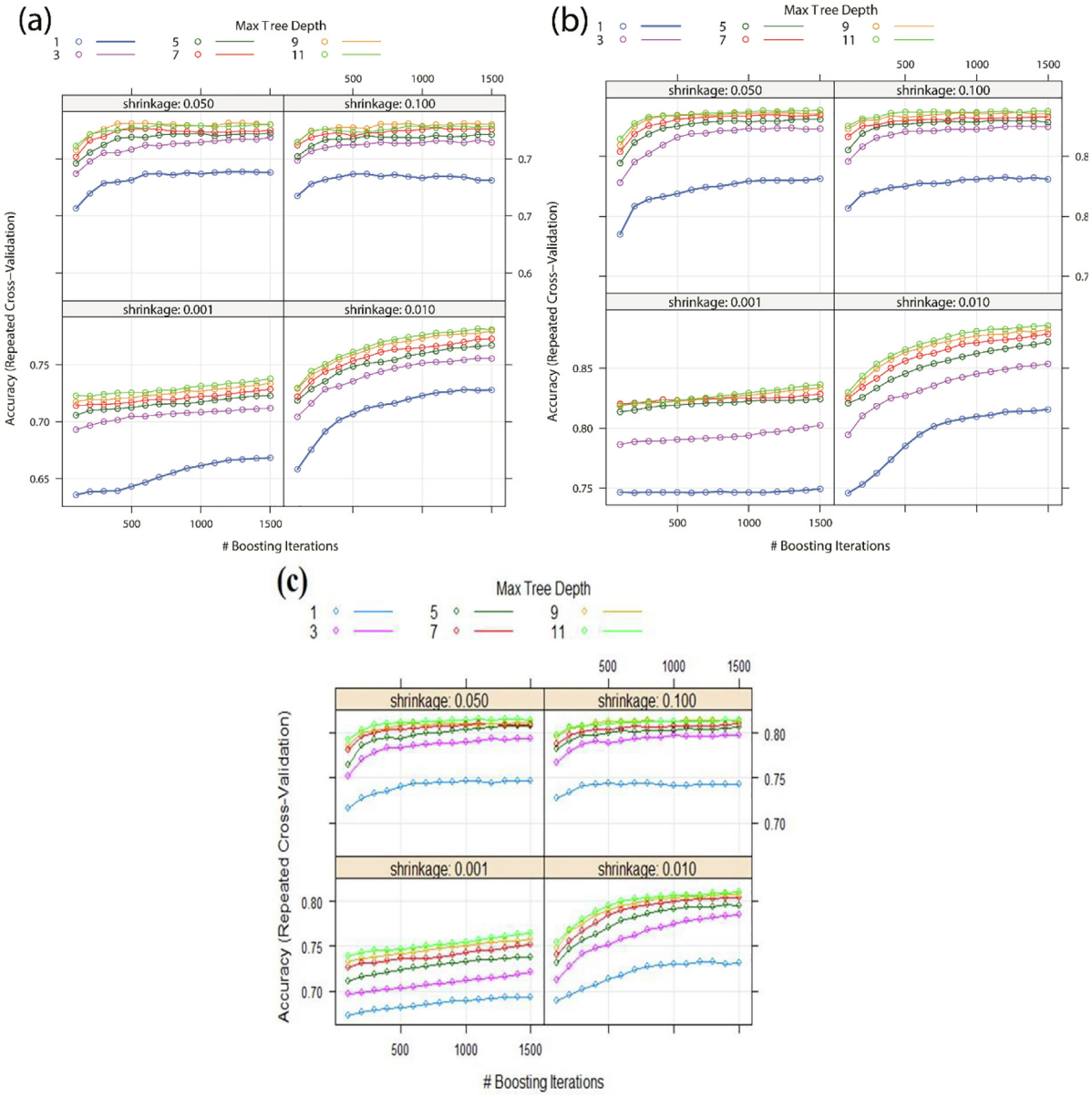


Fig. 10. Optimization results for (a) the BRT-Random, (b) the BRT-APG, and (c) the BRT-SAS.

fully predicted 493 out of 572 non-springs, as well as 505 out of 572 springs for the training dataset (Table 3). RF-APG had error rates of 0.117 and 0.138 for spring and non-spring prediction, while RF-Random had error rates of 0.201, and 0.241 for spring and non-spring prediction, respectively. The stated error rates indicate that the RF-APG predicted both springs and non-springs much better than the RF-Random. Further, RF-SAS had error rates of 0.17 and 0.21 for spring and non-spring prediction, respectively. This shows that RF-SAS predicted better than the random dataset and weaker than the APG dataset.

3.3. BRT optimization results by the random and APG datasets

To optimize the BRT algorithm with the random and APG datasets, we used a grid search scheme to select the number of trees between 100 and 1500 with 100-tree intervals, shrinkage values of 0.1, 0.01, 0.05, and 0.001, "interaction depths" of 1, 3, 5, 7, 9, and 11, and a fixed value of 20 for "*minimum terminal node size*" (Fig. 10). The BRT-Random was tuned with 1200 trees, interaction depth of 9, and shrinkage of 0.05 (Fig. 10a). The measured accuracy and kappa values for the BRT-random were 0.781 and 0.563, respectively. On the other hand, the BRT-APG was optimized with 1500 trees, interaction depth of 11, and shrinkage of 0.05 (Fig. 10b). The accuracy and kappa values measured for the BRT-APG were 0.888 and 0.777, respectively. The BRT-SAS was optimized with 1100 trees, interaction depth of 11, and shrinkage of 0.05 (Fig. 10c).

3.4. Validation of the RF and BRT algorithms optimized with different datasets

The validation outputs of the RF and BRT algorithms are displayed in Fig. 11. Based on the stated indices, the RF-APG significantly outperformed the RF-Random and had higher performance than the RF-SAS. Further, the results of the accuracy, kappa, sensitivity, specificity, AUC-ROC, and F1-score confirm the better performances of the models built on the APG comparing the SAS and random datasets. For instance, in the case of accuracy, RF-APG

had better performances than RF-SAS and RF-Random with differences of 0.061, and 0.082, respectively. In the case of kappa, RF-APG had higher efficiencies than RF-SAS and RF-Random with differences of 0.121, and 0.163, respectively. Also, in the case of AUC-ROC, it is observed that the RF-APG had better performances than RF-SAS and RF-APG with differences of 0.0485, and 0.063, respectively. Regarding the BRT algorithm, it was observed that BRT-APG has AUC differences of 0.0545, 0.084 with BRT-SAS and the BRT-Random.

3.5. GW potential maps generated by the RF and BRT algorithms

The classified GW potential maps generated by the RF-APG, RF-Random, RF-SAS, BRT-APG, BRT-Random, and BRT-SAS accompanied by the validation spring locations are depicted in Fig. 12a–f. The area percentages of each class of the GW potential maps are presented in Table 4. Based on Fig. 12a–f, it can be observed that the RF-APG and BRT-APG algorithms defined larger areas as "very high" potential class compared to the SAS and Random datasets. Further, it can be seen that there is a consistency between the very high GW potential classes defined by the RF-APG and BRT-APG relative to the random-based algorithms. The greater percentages of the "very high" GW potential classes predicted by the BRT-APG and RF-APG, compared to the medium and high classes, facilitates the decision-making process and assists the water professionals and decision-makers to arrive at appropriate land use and development plans.

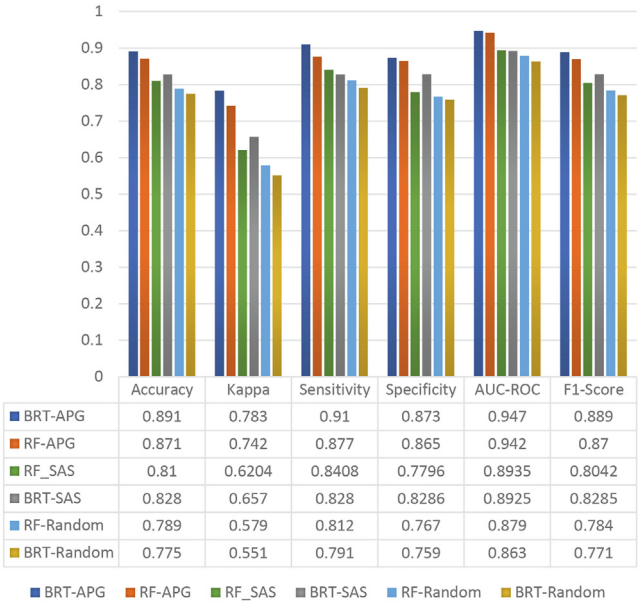
3.6. Impact of the APG in factors importance by the RF and BRT algorithms

The findings of assessing factor importance by "*mean decrease in Gini*" for the RF and "*relative influence*" for the BRT are illustrated in Fig. 13a, b. As illustrated, the contribution of the factors to GW potential mapping by the RF-APG, RF-SAS, and RF-Random follows a similar pattern, nevertheless, a more thorough investigation reveals some variations between the scores.

For instance, the RF-APG has defined the RSP, altitude, TWI, aspect, distance from faults, and slope degree as the five most important driving factors in the GW potential mapping. On the other hand, the first five important contributing factors defined by the RF-Random are altitude, RSP, TWI, distance from faults, and distance from rivers. The RF-SAS shows altitude, RSP, TWI, distance from faults and aspect as the most contributing factors. Additionally, the BRT-APG depicted a greater contribution of the RSP, altitude, TWI, aspect, and distance from faults to GW potential, while the BRT-Random depicted higher importance of the altitude, RSP, TWI, slope length, and distance from faults. The BRT-SAS depicted altitude, RSP, distance from faults, TWI, distance from rivers as the most contributing factors.

4. Discussion

The findings depict that both the BRT and RF performed significantly better than the random-dataset in modeling GW potential with the APG-dataset. Yesilnacar (2005) categorized the performance of classification algorithms into 5 classes of "poor" (0.5 < AUC < 0.6), "average" (0.6 < AUC < 0.7), "good" (0.7 < AUC < 0.8), "very good" (0.8 < AUC < 0.9), and "excellent" (AUC > 0.9). According to the stated criteria, the APG toolbox had a significant impact on the BRT and RF from "good" to "very good" predictors comparing to the Random and SAS datasets. The weaker performances of the RF-Random and BRT-Random could be associated with the uncertainties dealing with the generation of the absence or non-spring locations. For instance, absence-points could be generated



	Accuracy	Kappa	Sensitivity	Specificity	AUC-ROC	F1-Score
BRT-APG	0.891	0.783	0.91	0.873	0.947	0.889
RF-APG	0.871	0.742	0.877	0.865	0.942	0.87
RF_SAS	0.81	0.6204	0.8408	0.7796	0.8935	0.8042
BRT-SAS	0.828	0.657	0.828	0.8286	0.8925	0.8285
RF-Random	0.789	0.579	0.812	0.767	0.879	0.784
BRT-Random	0.775	0.551	0.791	0.759	0.863	0.771

Fig. 11. Validation results for the RF and BRT algorithms optimized by the APG and Random and SAS datasets.

in any existing pixel in the study region with the same probability. Further, absence-points could be located close to the springs where GW potential is deemed to be higher than other zones. Moreover, absence-points might be randomly located at higher spring density areas where GW potential is greater than the areas with lower or zero spring densities. Comparison of the BRT-SAS and RF-SAS with BRT-APG and RF-APG shows that the APG tool has notably improved the performance of the algorithms. The common feature between the APG and SAS tools is the absence-point filter based on the spring buffer (Rahmati et al., 2019). Nevertheless, they have different features; the SAS mainly focuses on the geographic space such as hotspot buffer and presence locations buffer, while the APG takes the feature space into account through the FR analysis of important driving factors, i.e., altitude, slope degree, TWI, and distance from rivers. This feature reduces the similarities between presence and absence points. The other difference that leads to better performance of the APG is the consideration of the presence points, i.e., spring density, and removal of the areas with the highest presence densities. Another supplementary upper hand of the APG toolbox is that it generates *n* absence-datasets and selects the best training and validations datasets based on the accuracy index. This is applied to reduce the uncertainties of the random generation of absence-points within the final boundary. The performance improvements of the BRT-APG and RF-APG can also be associated with the nature of decision trees that is to predict the new cases regarding the relationships between the target value and its driving factors. Proper generation of the absence-dataset assists the algorithms to detect relationships more robustly and arrive at highly accurate predictions. The stated reasons led to the greater efficacy of the RF and BRT in this research. The APG-dataset also impacted the contribution of the factors reported by the BRT and RF when they were constructed with the random, APG, and SAS datasets. This fact emphasizes the importance of absence point generation in such studies. With respect to model's performance, the BRT outperformed the RF, which is in agreement with the study carried out by Park and Kim (2019).

Considering the higher performance of the BRT-APG, the most effective driving factors in the current research are the RSP, altitude, TWI, aspect, and distance from faults depicting the great impact of topography on GW potential. TWI implies the chance

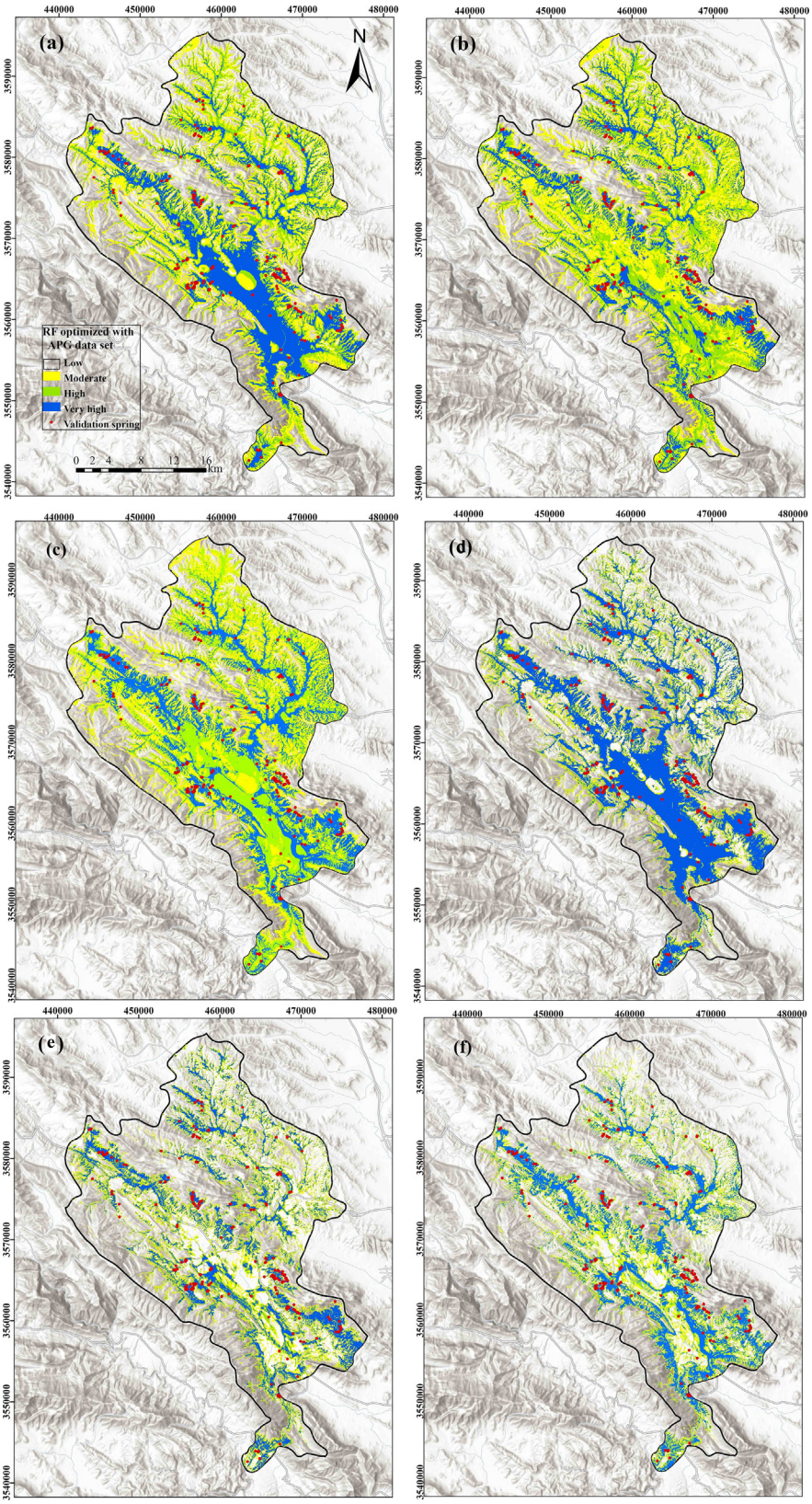


Fig. 12. GW potential maps produced by the (a) RF-APG, (b) RF-Random, (c) RF-SAS, (d) BRT-APG, and (e) BRT-Random, (f) BRT-SAS.

of water flow accumulation in various regions of the watershed and subsequently is a crucial factor in the GW modeling process. As per altitude, its influence can be associated with the impact it has on "slope degree, drainage system development, and flow velocity". The great role of aspect is related to the fact that it controls "sunshine duration, evapotranspiration", and snowmelt and ultimately affects GW conditions which its significant contribution agrees with Naghibi et al. (2020). Ozdemir (2011) stated that geol-

Table 4
Area percentage of each GW potential category obtained by the RF and BRT algorithms.

Model	Low	Moderate	High	Very high
RF-Random	30.3	28.9	26.6	14.2
RF-APG	35.4	24.5	20.3	19.8
RF-SAS	27.5	25.3	28.2	19.0
BRT-Random	60.5	14.8	10.5	14.2
BRT-APG	60	5.8	5.1	29.1
BRT-SAS	60.9	12.3	9.9	16.9

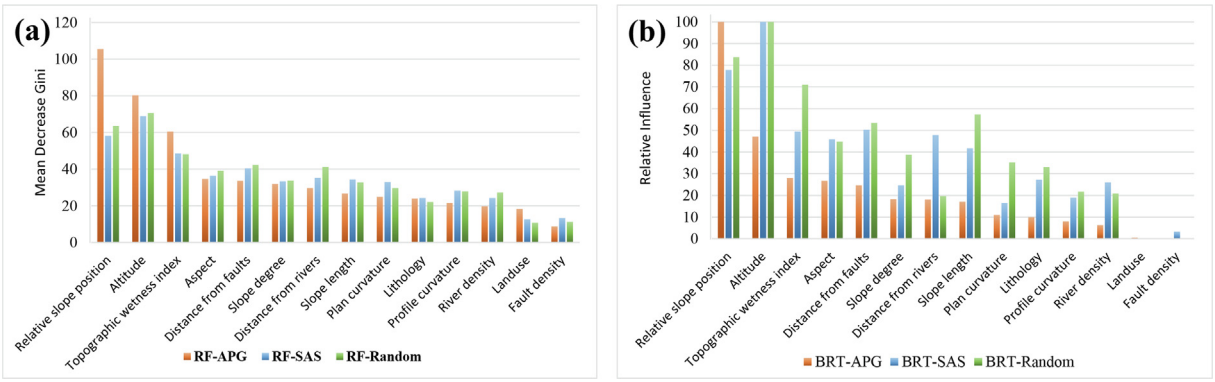


Fig. 13. Importance scores of the GW driving factors defined by the (a) RF-APG, RF-Random, and RF-SAS, and (b) BRT-APG and BRT-Random, and BRT-SAS algorithms.

ogy impacts the infiltration of soil and rock, therefore, owns a great influence on GW potential which could be the reason for the higher importance of the distance from faults in the current work. Moghaddam et al. (2020) declared that RSP and distance from faults were the major influencing variables on GW potential in their studies which coincides with our findings. Further, Naghibi et al. (2020) investigated the impact of topographical variables on GW potential and reported the higher impact of TWI and altitude, and lower importance of RSP. Comparing the importance of the factors obtained in the current study with other areas could provide a valuable estimate of the relative importance of the factors, nevertheless, these measures are region-based and controlled by the hydro-geological and topographical features of regions, hence cannot be deemed as a completely fixed set of input factors for future studies. Thus, it is critical to incorporate different effective factors in GW potential studies to assure finding as many useful relationships between the target and input factors as possible to construct the algorithms.

5. Conclusion

In classification algorithms, one of the major roots of uncertainty is the generation of absence-dataset which is usually done by random approaches though it cannot be the ideal method for spatial issues such as gully, flood, landslide, and spring. The current research develops a new ArcGIS toolbox based on Python language called APG that can generate absence points considering a statistical analysis of altitude, slope, TWI, and distance from rivers in addition to the presence of data locations. The application of the APG toolbox in a real-world case study on GW potential in Iran depicted that the benchmark algorithms, i.e., the BRT and RF performed much better with the APG dataset comparing to the random and SAS datasets. The outputs approved that the APG enhanced the efficiency of the RF and BRT to a considerable extent. Major differences of about 0.063 and 0.084 for AUC-ROC values were achieved for the RF and BRT algorithms respectively when considering the APG and random datasets. This substantial improvement in AUC approves the prosperous test of the APG toolbox. The APG dataset also impacted the importance of the driving factors in addition to creating different but more reliable GW potential maps. Obviously, highly accurate algorithms trained by the outputs of the APG toolbox could be used by the water sector decision-makers to come up with more effective plans on GW resources and achieve sustainable development goals. Based on the findings, the application of the APG toolbox in GW studies minimizes the impact of incorrectly chosen absence points. This, in turn, assists MLAs to extract as many useful patterns and information as possible from input factors. We recommend the implementation of the APG toolbox in other environmental issues such as landslides, gullies, and floods by different MLAs to test its impact. The more trustable maps generated by the APG dataset can help managers to have a better understanding of the environmental issues, allowing them to allocate money and efforts to areas with higher potential or susceptibility for exploitation and protection purposes, respectively. Future studies are suggested to focus on the feature space as well as involving a higher number of driving factors to generate more suitable absence points.

6. Software availability

Name of software: Absence Point Generation (APG)
Developer: Seyed Amir Naghibi
Software required: ArcGIS Desktop 10.4 (or later)
Program language: Python
Availability and cost: Freely available at Github (https://github.com/amir-naghibi/APG)

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the MECW research program and the Centre for Advanced Middle Eastern Studies, Lund University. We would like to appreciate the Iranian Department of Water Resources Management and Regional Water Authority of Chaharmahal and Bakhtiari at https://www.wrm.ir/index.php?l = EN and https://www.cbrw.ir/, which supplied the required datasets to conduct this research.

References

Akinci, H., Kilicoglu, C., Dogan, S., 2020. Random forest-based landslide susceptibility mapping in coastal regions of Artvin, Turkey. ISPRS Int. J. Geo-Inf. 9 (9), 553. https://doi.org/10.3390/ijgi9090553.

Assatse, W.T., Njandjock Nouck, P., Tabod, C.T., Akame, J.M., Nshagali Biringanine, G., 2016. Hydrogeological activity of lineaments in Yaoundé Cameroon region using remote sensing and GIS techniques. Egypt. J. Remote. Sens. Space Sci. 19 (1), 49–60.

Ayalew, L., Yamagishi, H., 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65 (1-2), 15–31.

Balogun, A.-L., Rezaie, F., Pham, Q.B., Gigović, L., Drobnjak, S., Aina, Y.A., Panahi, M., Yekeen, S.T., Lee, S., 2021. Spatial prediction of landslide susceptibility in western Serbia using hybrid support vector regression (SVR) with GWO. BAT and COA algorithms. Geosci. Front. 12 (3), 101104. https://doi.org/10.1016/j.gsf.2020.10.009.

Benjmel, K., Amraoui, F., Boutaleb, S., Ouchchen, M., Tahiri, A., Touab, A., 2020. Mapping of groundwater potential zones in crystalline terrain using remote sensing, GIS techniques, and multicriteria data analysis (case of the Ighrem region, western Anti-Atlas, Morocco). Water 12, 471.

Bonham-Carter, G.F., 1994. Geographic Information Systems for Geoscientists: Modeling with GIS. Paragon Press, Oxford, p. 398.

Breiman, L., 2001. Random forests. Mach. Learing 45, 5–32.

Bui, Q.-T., Nguyen, Q.-H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., Pham, V.-M., 2020. Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. J. Hydrol. 581, 124379. https://doi.org/10.1016/j.jhydrol.2019.124379.

Daher, W., Pistre, S., Kneppers, A., Bakalowicz, M., Najem, W., 2011. Karst and artificial recharge: Theoretical and practical problems. A preliminary approach to artificial recharge assessment. J. Hydrol. 408 (3-4), 189–202.

Díaz-Alcaide, S., Martínez-Santos, P., 2019. Review: Advances in groundwater potential mapping. Hydrogeol. J. 27 (7), 2307–2324.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77 (4), 802–813.

Chaharmahal and Bakhtiari Regional Water Authority (CBRWA), 2019. https://www.cbrw.ir/ (accessed December 2019).

Geology Survey of Iran (GSI)., 1997. Geological survey and mineral exploration of Iran. https://gsi.ir/fa. (accessed December 2019)

Guru, B., Seshan, K., Bera, S., 2017. Frequency ratio model for groundwater potential mapping and its sustainable management in cold desert, India. J. King Saud Univ. Sci. 29 (3), 333–347.

Kanwal, S., Atif, S., Shafiq, M., 2017. GIS based landslide susceptibility mapping of northern areas of Pakistan, a case study of Shigar and Shyok Basins. Geomat. Nat. Haz. Risk. 8 (2), 348–366.

Kim, J.-C., Lee, S., Jung, H.-S., Lee, S., 2018. Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. Geocarto Int. 33 (9), 1000–1015.

Lagomarsino, D., Tofani, V., Segoni, S., Catani, F., Casagli, N., 2017. A tool for classification and regression using random forest methodology: Applications to landslide susceptibility mapping and soil thickness modeling. Environ. Model. Assess. 22 (3), 201–214.

Lei, X., Chen, W., Avand, M., Janizadeh, S., Kariminejad, N., Shahabi, H., Costache, R., Shahabi, H., Shirzadi, A., Mosavi, A., 2020. GIS-based machine learning algorithms for gully erosion susceptibility mapping in a semi-arid region of Iran. Remote Sens. 12 (15), 2478. https://doi.org/10.3390/rs12152478.

Liaw, A., Wiener, M., 2018. Package 'randomForest.'. https://cran.r-project.org/web/packages/randomForest/index.html

Loosvelt, L., Peters, J., Skriver, H., Lievens, H., Van Coillie, F.M.B., De Baets, B., Verhoest, N.E.C., 2012. Random Forests as a tool for estimating uncertainty at pixel-level in SAR image classification. Int. J. Appl. Earth Obs. Geoinf. 19, 173–184.

Mitchell, M.W., 2011. Bias of the Random Forest Out-of-Bag (OOB) error for certain input parameters. Open J. Stat. 01 (03), 205–211.

Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5 (1), 3–30.

Moghaddam, D.D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Haghighi, A.T., Nalivan, O.A., Tien Bui, D., 2020. The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. Catena 187, 104421. https://doi.org/10.1016/j.catena.2019.104421.

Motevalli, A., Naghibi, S.A., Hashemi, H., Berndtsson, R., Pradhan, B., Gholami, V., 2019. Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. J. Clean. Prod. 228, 1248–1263.

Naghibi, S.A., Hashemi, H., Berndtsson, R., Lee, S., 2020. Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. J. Hydrol. 589, 125197. https://doi.org/10.1016/j.jhydrol.2020.125197.

Negnevitsky, M., 2005. Artificial Intelligence: A Guide to Intelligent Systems,. Pearson Education Canada, p. 504.

Ozdemir, A., 2011. Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). J. Hydrol. 405 (1-2), 123–136.

Park, S., Kim, J., 2019. Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. Appl. Sci. 9 (5), 942. https://doi.org/10.3390/app9050942.

Pham, B.T., Jaafari, A., Phong, T.V., Yen, H.P.H., Tuyen, T.T., Luong, V.V., Nguyen, H.D., Le, H.V., Foong, L.K., 2021. Improved flood susceptibility mapping using a best first decision tree integrated with ensemble learning techniques. Geosci. Front. 12 (3), 101105. https://doi.org/10.1016/j.gsf.2020.11.003.

Pourghasemi, H.R., Beheshtirad, M., 2015. Assessment of a data-driven evidential belief function model and GIS for groundwater potential mapping in the Koohrang Watershed, Iran. Geocarto Int. 30 (6), 662–685.

Pourghasemi, H.R., Sadhasivam, N., Kariminejad, N., Collins, A.L., 2020. Gully erosion spatial modelling: Role of machine learning algorithms in selection of the best controlling factors and modelling process. Geosci. Front. 11 (6), 2207–2219.

Raduła, M.W., Szymura, T.H., Szymura, M., 2018. Topographic wetness index explains soil moisture better than bioindication with Ellenberg's indicator values. Ecol. Indic. 85, 172–179.

Rahmati, O., Haghizadeh, A., Pourghasemi, H.R., Noormohamadi, F., 2016. Gully erosion susceptibility mapping: the role of GIS-based bivariate statistical models and their comparison. Nat. Hazards 82 (2), 1231–1258.

Rahmati, O., Moghaddam, D.D., Moosavi, V., Kalantari, Z., Samadi, M., Lee, S., Bui, D. T., 2019. An automated Python language-based tool for creating absence samples in groundwater potential mapping. Remote Sens. 11 (11), 1375.

Razavi Termeh, S.V., Kornejady, A., Pourghasemi, H.R., Keesstra, S., 2018. Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. Sci. Total Environ. 615, 438–451.

Schapire, R.E., 2003. The Boosting Approach to Machine Learning: An Overview. In: Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification. Springer, New York, NY, pp. 149–171.

Segoni, S., Pappafico, G., Luti, T., Catani, F., 2020. Landslide susceptibility assessment in complex geological settings: sensitivity to geological information and insights on its parameterization. Landslides 17 (10), 2443–2453.

Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., Shirzadi, A., 2018. Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. J. Environ. Manage. 217, 1–11.

Talukdar, S., Ghose, B., Shahfahad, Salam, R., Mahato, S., Pham, Q.B., Linh, N.T.T., Costache, R., Avand, M., 2020. Flood susceptibility modeling in Teesta River basin, Bangladesh using novel ensembles of bagging algorithms. Stoch. Environ. Res. Risk Assess. 34 (12), 2277–2300.

Towfiqul Islam, A.R.M., Talukdar, S., Mahato, S., Kundu, S., Eibek, K.U., Pham, Q.B., Kuriqi, A., Linh, N.T.T., 2021. Flood susceptibility modelling using advanced ensemble machine learning models. Geosci. Front. 12 (3), 101075. https://doi.org/10.1016/j.gsf.2020.09.006.

Trigila, A., Iadanza, C., Esposito, C., Scarascia-Mugnozza, G., 2015. Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). Geomorphology 249, 119–136.

Viera, A., Garrett, J., 2005. Understanding interobserver agreement: the kappa statistic. Fam Med. 37 (5), 360–363.

Wang, Y., Feng, L., Li, S., Ren, F.u., Du, Q., 2020. A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China. Catena 188, 104425. https://doi.org/10.1016/j.catena.2019.104425.

Xi, H., Feng, Q., Si, J.H., Chang, Z., Cao, S., 2010. Impacts of river recharge on groundwater level and hydrochemistry in the lower reaches of Heihe River Watershed, northwestern China. Hydrogeol. J. 18, 791–801.

Xu, Z., Huang, X., Lin, L.u., Wang, Q., Liu, J., Yu, K., Chen, C., 2020. BP neural networks and random forest models to detect damage by Dendrolimus punctatus Walker. J. For. Res. 31 (1), 107–121.

Yesilnacar, E.K., 2005. The application of computational intelligence to landslide susceptibility mapping in Turkey. PhD. thesis,. University of Melbourne, p. 200.

Yousefi, S., Sadhasivam, N., Pourghasemi, H.R., Ghaffari Nazarlou, H., Golkar, F., Tavangar, S., Santosh, M., 2020. Groundwater spring potential assessment using new ensemble data mining techniques. Measurement 157, 107652. https://doi.org/10.1016/j.measurement.2020.107652.

Zhu, A.-X., Miao, Y., Liu, J., Bai, S., Zeng, C., Ma, T., Hong, H., 2019. A similarity-based approach to sampling absence data for landslide susceptibility mapping using data-driven methods. Catena 183, 104188. https://doi.org/10.1016/j.catena.2019.104188.