A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of random forest approach

Mehrdad Daviran[1], Abbas Maghsoudi[✉,2], Reza Ghezelbash[2], Biswajeet Pradhan[3,4]

[1]*School of Mining, Petroleum and Geophysics Engineering, Shahrood University of Technology, Shahrood, Iran*

[2]*Faculty of Mining and Metallurgical Engineering, Amirkabir University of Technology, Tehran, Iran*

[3]*Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), University of Technology Sydney, Sydney, Australia*

[3]*Department of Energy and Mineral Resources Engineering, Sejong University, Choongmu-gwan, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Korea*

Abstract

Machine learning algorithms (e.g., random forest (RF)) have recently been performed in data-driven mineral prospectivity mapping. These methods are highly sensitive to hyperparameter values, since the predictive accuracy of them can significantly increase when the optimized hyperparameters are predefined and then adjusted to training procedure. The main goal of this contribution is to propose a hybrid genetic-based RF model, namely GRF, which is able to automatically adjust the optimized hyperparameters of RF with the excellent predictive accuracy. Therefore, three primary parameters of RF comprising NT, NS and d, were well-tuned employing genetic algorithm (GA) in establishing an efficient RF model. The proposed GRF model and also conventional RF were tested on mineralization-related geo-spatial dataset and the predictive models were generated for comparing the accuracy of the proposed GRF model with that of RF. The input dataset (e.g., multi-element geochemical signature, geological-structural layer and hydrothermal alteration evidences) which acquired from Feizabad district, NE Iran, were translated into mappable targeting criteria in the form of four predictor maps. In addition, the locations of 13 known Cu–Au deposits as prospect data and the locations of 13 randomly selected non-prospect data were used as target variables to train the models. Three authentic validation measures, K-fold cross-validation, confusion matrix and success-rate curves, were employed to evaluate the overall performance of two predictive models. Experimental results suggested the superiority of GRF model over the RF, as the favorable areas derived by GRF model occupy only 9% of the study area while predicting 100% of the known deposits.

Keywords: Mineral prospectivity mapping; Machine learning; Hyperparameter tuning; Genetic algorithm; GIS; Random forest.

✉ Corresponding author: Abbas Maghsoudi, a.maghsoudi@aut.ac.ir

1. Introduction

Mineral prospectivity mapping (MPM) is a multiple criteria decision-making procedure which strives to explore and prioritize favorable exploration targets for facilitating the exploration of undiscovered mineral deposits (cf. Bonham-Carter, 1994; Carranza, 2008; Ghezelbash et al., 2020a). Developing a suitable algorithm or methodology for processing, analyzing and integrating various geospatial dataset (e.g., geochemistry, geophysics, geology and remote sensing) is highly necessary for obtaining an efficient mineral prospectivity map in order to visualize areas with a high favorability to be discovered further. However, MPM is a challenging procedure plagued with stochastic and also systemic uncertainties from various sources (Kreuzer et al., 2008; Zuo et al., 2009; Ghezelbash et al., 2019a). Using inefficient as well as insufficient exploration datasets leads to stochastic uncertainty in MPM (McCuaig et al., 2010; Lisitsin et al., 2013). On the other hand, inaccurate elicitation of exploration criteria, sensitivity of prospectivity model to inefficient and less efficient ore-related evidence layers and improper selection or application of a methodology for establishing the interrelations between geospatial features and known mineral deposits may propagate systemic uncertainties to MPM (Ghezelbash et al., 2020a).

Data-driven mapping of mineral prospectivity, unlike knowledge-driven mapping, is well-suited to discover new exploration targets in moderate to well-explored areas (Carranza et al., 2008). In this kind of modeling, various range of mathematical methods can be used for quantifying the spatial association between different evidential features and training locations. Recently, machine learning algorithms (MLAs), e.g., support vector machines (SVMs) (Zuo and Carranza, 2011), artificial neural networks (ANNs) (Porwal et al., 2003; Ghezelbash et al., 2020a) and random forests (RF) (Rodriguez-Galiano et al., 2015; Parsa et al., 2018) have gained much reputation and popularity in MPM, because of not requiring conditional independence of input features as well as ability to handle nonlinear correlations between known mineral deposits and spatial evidential features. Moreover, capability of managing a huge amount of spatial features and also no need evidence layers to follow any special distribution have led to wide applicability for MPM.

Machine learnings are self-calibration predictive models which are generated using training dataset (Brown et al., 2000; Pradhan, 2013). In other words, MPM is considered as a classification problem in application of MLAs, as the study region is categorized as either prospect or non-prospect (Ghezelbash

et al., 2019a). Such classification strategy significantly affected by sufficiency of input training data, abundance of predictor variables and also accurately determined hyperparameters for training model (Carranza and Laborte, 2015). Thus, the method to find the ideal values for hyperparameters (or tuning the model to input training data) is a fundamental step in MPM. Traditionally, researchers mostly set hyperparameters to MLAs based only on their experiences to train models. Nevertheless, the optimal settings of hyperparameters will change with different dataset and, thus, prescribing the hyperparameter values according to the previous expertise may lead to the bias and also increase the systemic uncertainty in MPM. To address this, the hyperparameters of MLAs must be well-tuned using optimization algorithms such as genetic algorithm (GA) (Srinivas and Patnaik, 1994) which is much effective and accurate than the default values. GA is a reputable evolutionary approach which has been frequently and successfully applied for resolving optimization issues in order to derive optimum solutions (Ghezelbash et al., 2020a, b).

In this research, we proposed a novel genetic-based RF approach, namely GRF, for predictive modeling of mineral systems associated with Cu-Au mineralization in Feizabad district, NE Iran. The hyperparameters of GRF methodology were well-tuned and adjusted based on GA to train predictive model. To evaluate the performance of GRF model, the classical random forest (CRF) model was built for MPM based on the same training dataset to verify if GRF does indeed yield improved results.

2. Methods

2.1. Conventional random forest method

Conventional random forest (CRF) (for both classification and regression) (Breiman, 2001), which is a developed form of decision trees (DTs) (Breiman, 1984), is a highly robust machine learning algorithm (MLA) consists of a large number of DTs that operates as an ensemble. Indeed, CRF is a classifier in which the training process is carried out using "bagging" method (Liaw and Wiener, 2002). CRF generates multiple DTs and integrates them to derive an accurate and stable prediction. Each DT learns from a random sample of datasets. The samples are drawn with replacement, called bootstrapping, meaning that some samples will be utilized for a number of times in an individual DT. A sample is categorized into the class that can win the majority votes of overall DTs within the forest (Fig. 1). An unbiased estimation of the prediction accuracy can be obtained based on the training data

that at each bootstrap iteration roughly 1/e training samples are left out as out-of-bag (OOB) data (Liaw and Wiener, 2002).

The CRF commences with the purification of child nodes through splitting the target variable based on predictor variables from the parent node. The splitting successively iterates until a pre-defined stop criterion is reached. Through this process, every DT reaches to its simple regression or classification model. CRF then averages the results of various DTs to achieve the final model. In this way, there are three hyperparameters that can be adjusted as following (Breiman, 2001):

1. **Number of trees (N_T):** The N_T parameter determines the number of trees in the forest of the model. Additional DTs typically can improve the accuracy of model, as the predictions are carried out using a large number of votes from diverse DTs, although, the large number of DTs lead to increase the computing time.

2. Number of split (N_S): The N_S parameter can control the minimum number of samples needed to split an internal leaf node. Too large values may cause under-fitting, as the DTs will not be able to split enough times to reach node purity.

3. Depth (d): Each DT in CRF model produces numerous splits for isolating homogeneous classes of outcomes. Larger numbers of splits permitted the DTs to describe more variation in dataset, however, DTs with many splits may lead to over-fitting.

Adjusting the parameters utilized to train models is the primary stage to reach an accurate prediction. Although, there are no definitive rules for defining the optimum parameters with the aim of increasing the accuracy of a model. In many studies, some empirical trial and error procedures were recommended to optimally adjust the hyperparameters. K-fold cross-validation is one of these model validation methods by which the data are categorized into K equal subsets, namely fold. A single subset is maintained for validation and the training is done using the rest K-1 folds. Training happens K times until each subset has been utilized once for validation of dataset. This process can make a better representation of error across the entire data set, because all samples are contributed as both training and validation (Rodriguez et al., 2009).

2.2. Genetic-based hyperparameter tuning of random forest method

Optimization of hyperparameters for machine learning models is a key step in generating an accurate prediction. Hyperparameters specify the characteristics of model which can strongly affect the accuracy of a model as well as the computational efficiency (Wu et al., 2007). Indeed, accurate selection of hyperparameters is one of the complicated phases of MLAs (i.e. RF) for which there is no definitive rule and, thus, the hyperparameters are selected during a trial-and-error procedure (Rodriguez-Galiano et al., 2015). This can increase the systemic uncertainty in predictive modeling of mineral prospectivity, not because the data used is noisy, or the algorithm used is non-robust, but due to the inappropriate selection of hyperparameter values. To address this, the optimization algorithm such as GA can be incorporated to the MLAs to achieve a higher accuracy of prediction (Wu et al., 2007).

GA is a random-based evolutionary algorithm which has been successfully implemented to solve different optimization problems (Mühlenbein et al., 1991). GA is suitable to simultaneous manipulating of models with different resolution and structures which is able to search non-linear solution spaces and make slight changes to that solutions until reaching the best solution without needing a priori knowledge about model characteristics (Wu et al., 2007). GA starts by initializing a population of chromosomes applying default or random values. Then, a fitness function is used to form the structure of that population. The fittest chromosomes of the population are selected during a reproducing process using a reproduction function (crossover and mutation). Then, the reproducing process is repeated until passing a favorable number of iterations, and finally, deriving the best population based on the fitness function.

This study introduces a new genetic-based methodology, namely GRF, for optimizing the three RF parameters (N_T, N_S and d), simultaneously. In the proposed GRF methodology, the RF parameters are dynamically tuned by applying GA evolutionary process and the RF algorithm then executes to generate a predictive model based on optimized parameters values. The various stages of GRF are shown in Fig. 2 and are described as following stages (Kim and Han, 2000):

1) Chromosome representation: The three parameters N_T, N_S and d of RF were directly coded to create chromosome G which is denoted as $G=\{a_1, a_2, a_3\}$, where a_1, a_2 and a_3 represent the regularization parameter N_T, N_S and d, respectively.

2) Initial population: The process starts with a group of chromosomes, namely a population. Initializing the population is performed through a selection of the suitable number of chromosomes within it.

3) Fitness function: A fitness function which can evaluate the performance of each chromosome, should be outlined prior to starting to search the tuned hyperparameters of RF. In this study, the OOB error as well as 10-fold cross-validation has been applied to assess the predictive accuracy of a model.

4) **Selection:** The roulette wheel as a well-known selection method of GRF was used whereby a number of chromosomes were selected from the population in the mating pool. According to the previously computed fitness value, the best chromosomes with minimum fitness value were selected for reproduction.

5) Crossover: For each pair of parents to be mated, the crossover took place by choosing a random point in the chromosome and exchanging genes of parents result in generating new offspring chromosomes.

6) Mutation: In specific new offspring generated, some of their genes were subjected to mutation. Mutation is a background operator can be applied to preserve genetic variety in the population. Mutation operator can alter gene values from their primary condition.

In the final step, the genetic-based tuned hyperparameters N_T, N_S and d were contributed to create GRF algorithm using training dataset which was then executed to generate a predictive model of mineral prospectivity.

2.3. Model evaluation

In the field of machine learning, particularly in the statistical classification problems, a confusion matrix can be used for describing the performance of a classification model. In a 2-class confusion matrix, four possible results are summarized, including: (a) true positive (*TP*), where the model correctly predicts the prospect locations; (b) true negative (*TN*), where the model correctly predicts the non-prospect locations; (c) false positive (*FP*), where the model incorrectly predicts the prospect locations; and (d) false negative (*FN*), where the model incorrectly predicts the non-prospect locations. Classification accuracy of a trained model can be described and formulized as follows (Liu et al., 2005):

$$Sensitivity = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$F - measure = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \quad (5)$$

3. Study area and exploration dataset

3.1. General geological setting

The study area is located in northeastern of Iran between Central Iran and Lut Block structural zones (Fig. 3), which covers an area of roughly 2,500 km^2. The oldest rock units comprising thinly bedded limestone, sandy limestone, silt, olive-green silty shale, dolomite and dolomitic limestone of middle-to-upper Cambrian age. Extensive sedimentary facies of Tertiary rocks are exposed throughout the study area. Granodioritic and dioritic intrusions of Eocene-Oligocene age in related to volcanic rock units host numerous IOCG and vein-type Cu-Au deposits in northern parts of the study area. Andesitic-basaltic volcanic rock units which are covered by Quaternary sediments appeared mainly in central, northern and southern parts (Behroozi, 1987) (Fig. 4a).

Fault movements which have played a vital role in the formation of intrusive-related deposits in the Feizabad district are categorized into two major groups: (1) east-west-trending faults, specifically Doruneh sinistral strike–slip fault, which is strongly affect the emplacement of magma intrusion and mineralization, and (2) northeast- and northwest-trending faults/lineaments (Hu et al., 1995) (Fig. 4a).

3.2. Input data

The exploration dataset used were (a) the 1:100,000 scale geological map of Feizabad district from which the various geological units and also faults/fractures were digitized, (b) ASTER and Landsat ETM$^+$ data which were implemented to detect hydrothermal alterations using remote sensing

techniques, and (c) analyzed geochemical data derived from 1033 composite stream sediment samples collected over a regular network of sampling (Fig. 4b) (Ghezelbash et al., 2019b).

3.3. Target variable

The occurrence of mineral deposits can be represented as a prospect (locations of known deposits and occurrences with a labeled value 1) and non-prospect (locations of non-deposits with a labeled value 0). Since each occurrence (prospect or non-prospect) is occupied a single pixel in each geo-spatial map, we expanded the boundaries occurrences to surrounding pixels. In this study, the locations of 13 known Cu-Au occurrences were used as prospect sites and 10-15 pixels covering each occurrence were used as positive training data. For selection of non-prospect locations, following points were considered (Carranza and Laborte, 2015):

(1) The number of non-prospect sites must be equal to that of prospect sites, so that the number of positive and negative objectives could be balanced and as a result the overall performance of the MLAs could be enhanced.

(2) Non-prospect sites should be completely distal to prospect sites because spatial features of non-prospect locations should differ from those of prospect ones.

(3) Unlike the prospect sites that follow cluster distribution, non-prospect sites must be randomly selected.

For optimally selection of non-deposit sites in this study, 13 sterile sites as non-prospect were selected by integrating the available spatial feature maps and through stratified random sampling technique (Ding et al., 1997) from non-favorable locations. Then, 10-15 pixels around each site were used as negative training data.

4. Results and discussions

4.1. Generating input data for efficient evidence features

4.1.1 Geochemical factor

Determination of significant anomalies in geochemical perspectives using stream sediment geochemical data is crucial for generation of geochemical evidence layers to be used in MPM. For this, the concentration data of 12 elements (Ag, As, Au, Bi, Cr, Cu, Hg, Mo, Ni, Pb, Sb and Zn) were used.

More information about the preparation procedure of stream sediment samples and also their descriptive statistics could be found in Ghezelbash et al. (2019b). Then, as the stream sediment geochemical data are composite, the raw analytical data of 12 analyzed elements were transformed using isometric logratio (or ilr) with the aim of addressing the inherent closure problem (Aitchison, 1986).

For selection of efficient elements which are significantly associated with Cu-Au mineralization in the study area, a trustworthy method called success-rate curve was applied (Agterberg and Bonham-Carter, 2005). This method has been used in many studies for measuring the degree of efficiency of spatial evidence layers as well as mineral prospectivity maps. For constructing a success-rate curve, the portion of the studied region classified as favorable for mineralization occurrences in x-axis plots versus the portion of mineralization occurrences correctly categorized in y-axis (Ghezelbash et al., 2020b). Then, a diagonal line is drawn which is a measure for distinguishing the efficient evidence layers from inefficient ones and also determining the relative importance of each evidence layer. In this study, the success-rate curves for 12 selected elements with ilr-transformed values were plotted and the results are shown in Fig. 5. As can be seen, the Au, Cu, Mo, Pb, Sb and Zn are the most efficient geochemical elements as their success-rate curves lie above the diagonal line indicating that the distribution of these six elements strongly associated with the locations of Cu-Au occurrences in the study area. The Ag, As and Bi are the least efficient geochemical elements, because some parts of their success-rate curves lie below the diagonal line and some parts lie above that (Fig. 5). Thus, these elements could not significantly representative of Cu-Au mineralization in this region. Moreover, the Cr, Hg and Ni are inefficient geochemical elements which did not spatially associated with the locations of Cu-Au occurrences in the study area (Fig. 5). Accordingly, the raster maps of Au, Cu, Mo, Pb, Sb and Zn with ilr-transformed values (Fig. 6) were selected as the most efficient evidence geochemical layers to be used for generating the enhanced multi-element geochemical signature associated with Cu-Au mineralization in the Feizabad district.

To generate the enhanced multi-element geochemical signature reflecting the Cu-Au mineralization, the maps of the most efficient geochemical elements must be integrated. In this regard, the ilr-transformed values of Au, Cu, Mo, Pb, Sb and Zn (Fig. 6) should be transformed to a same domain (e.g., fuzzy values). For this purpose, a GIS-based fuzzy member function namely mean-standard deviation large (or MSLarge) was applied and the ilr-transformed values of six geochemical

elements were transformed to [0-1] range. Then, the fuzzified geochemical layers which consider as weighted fuzzy evidence layers were subjected to fuzzy GAMMA operator in order to generate multi-element geochemical layer. The fuzzy GAMMA operator is the most significant fuzzy operator among the other operators (e.g., OR, AND, SUM and PRODUCT) which is widely applied for producing the mineral prospectivity maps. Appropriate using of GAMMA operator requires the optimum selection of γ parameter. After a trial-and-error procedure, it was found that the γ=0.9 is more succeeded in restricting the high-favorable multi-element geochemical landscapes associated with Cu-Au mineralization in Feizabad district. Thus, the fuzzified multi-element geochemical layer (Fig. 7a) which is representative of Au, Cu, Mo, Pb, Sb and Zn mineralization was selected to contribute to Final MPM.

4.1.2 Heat source factor

The Cu-Au mineralization in the Feizabad district is spatially linked to the dioritic and granodioritic intrusions of Eocene-Oligocene age (Hu et al., 1995) which are the proxy indicators of magmatic-related processes that led to the generation of ore-forming materials. Therefore, their surface outcrops could be served as a substantial spatial feature (the heat source) for related-mineralization in the study area. The locations of intrusions on 1:100,000 scale geological map of the study area were digitized and a map of presence of these intrusions was generated (Fig. 7b).

4.1.3 Structural factor

Transportation of ore-bearing magmatic fluids through rocks were structurally controlled by faults and fractures especially in the intersection points of the faults and fractures. This is because the faults/fractures and their intersection points provided pathways to transport the ore-forming materials from mantle- or crustal-derived sources and concentrate them near the surface of the earth. Thus, the density of faults/fractures is considered as a key structural factor controlling the Cu-Au mineralization in the study area. Finally, we have generated the fault density map (Fig. 7c) to be used in final MPM.

4.1.4 Factors derived from remotely senses data

Combination of remote sensing data and image processing techniques lead to detection of the outcrops of the hydrothermal alterations (e.g., argillic, phyllic, propylitic and Fe-oxide) which can provide a distinctive assemblage of minerals and vary according to the position, degree and longevity

of flow processes (Simpson et al., 2001). Therefore, mineralization-related hydrothermal alterations can be utilized as main exploration clues. In this regard, argillic, phyllic, propylitic alterations interpreted from ASTER images as well as Fe-oxide alteration extracted from ETM^+ images (Daviran et al., 2019) (Fig. 8) were assembled and the evidence map of presence of hydrothermal alterations was generated (Fig. 7d).

4.2. Sensitivity analysis of data-driven prospectivity models

A cell size of 200 m was objectively selected to generate raster maps of (a) geochemical factor (Fig. 7a), (b) heat source factor (Fig. 7b), (c) structural factor (Fig. 7c) and remote sensing factor (Fig. 7d). All 4 evidence layers which were translated to predictor maps were used for extraction of test and training data. Based on the 200 m cell size, a total of 40848 pixel values were generated for each predictor map. Then, a 40848 × 4 matrix representing 40848 cells and 4 predictor maps were prepared as test dataset. In addition, using the locations of prospect and non-prospect sites, a total of 290 pixels including occurrence locations were extracted from 4 predictor maps and a 298 × 5 matrix were generated as training dataset. In this matrix, the first 4 columns represent the multi-attribute features of predictor maps derived from the locations of prospect and non-prospect sites while the fifth column is a target variable in which the score 1 was labeled for prospect cells and score 0 was labeled for non-prospect cells. These labeled data were then divided into two parts; three-fourth of them (75%) were utilized for training the CRF and GRF models, while the rest of them (25%) as OOB data were not involved in the training procedures and used to validate the results through OOB error.

In this study, CRF modeling was conducted in MATLAB software and the three relevant hyperparameters N_T, N_S and d were experimentally set prior to CRF modeling. In this regard, a range of 1-350, 1-10 and 1-5 were considered for the number of trees (N_T), number of splits (N_S) and depth (d), respectively, during a trial-and-error procedure. In this paper, CRF algorithm with 320 number of trees as optimum value was executed to model mineral prospectivity (Table 1) based on 10-fold cross-validation. It should be noted that the increasing the number of trees do not necessarily lead to decrease the error, but may only increase the computing time. Besides, $N_S = 6$ and $d = 2$ was selected (Table 1), although these parameters have lesser impact on performance of CRF. Finally, the CRF model of Cu-Au mineralization prospectivity was generated (Fig. 9a).

For hyperparameter-tuning of GRF model proposed in this study, the accurate number of populations (N_P) must be defined according to the minimum value of fitness (best fitness) with 10-fold cross validation as well as the appropriate number of iterations (N_i). The performance of GRF algorithm was tested by successive runs for 10 to 200 N_p, and for any population, the N_i varied between 25 and 500. As can be seen in Fig. 10, tuning of GRF parameters terminated after 225 iterations in N_p=200 with best fitness function of 0.08. Then, the tuned parameters of GRF model were derived. As a result, the GRF prospectivity model was generated (Fig. 9b) based on N_T=957, N_S=10 and d=3 (Table 1).

The graphical confusion matrices for CRF and GRF models are shown in Fig. 11 that used for evaluating the classification performance in both training and OOB dataset. As shown in Fig. 11a, the OOB error for GRF model was 4.16% and, thus, the model accuracy was 95.83%, while for CRF model (Fig. 11b) was 6.95% and, thus, the model accuracy was 93.05% %. Besides, it can be seen that the performance of GRF is much better than CRF in the training processes, reaching 97.93 % accuracy rate of classification (Fig. 11c); while the CRF reaches 95.17 % accuracy rate of classification (Fig. 11d). The classification accuracy evaluation indices (e.g., sensitivity, specificity, precision and F-measure) for CRF and GRF models that were calculated from confusion matrices described above are listed in Table 2. Both GRF and CRF models gained the highest and nearly highest possible value of sensitivity (100 % and 99.32 %, respectively), representing that both models were able to correctly classify the prospect cells to the prospect class (Table 2). On the other hand, the specificity of GRF reaches 90.6 % meaning that GRF was able to correctly classify 90.6 % of the non-prospect cells to the non-prospect class compared to CRF which was able to correctly classify only 85.9 % of the non-prospect cells to the non-prospect class (Table 2). Moreover, GRF model achieves 91.41 % of precision, representing that among the predicted cells that labeled as prospect, 91.41 % of them are actually true prospect locations (Table 2). While, 87.57 % of predicted cells that labeled as prospect, are truly prospect locations by CRF model (Table 2). Finally, F-measure which is the weighted average of precision and sensitivity and also takes both false positives (*FP*) and false negatives (*FP*) into account was used for measuring the classification accuracy. As can be seen in Table 2, the GRF model gained higher F-measure value (95.51 %) than CRF model (93.07 %) indicating that the GRF model reflects stronger correlation between the predictions and reality compared to CRF. Thus, the GRF model is more reliable for modeling the Cu-Au prospectivity in the study area.

Evaluation of prospectivity models of CRF (Fig. 9a) and GRF (Fig. 9b) were conducted through measuring the correlation between the prospectivity values and known mineral occurrences, for each, the success-rate curves were plotted. In this study, a 10-percentile interval was selected for constructing the success-rate curves of CRF and GRF model using P_a, the portion of the study area classified as favorable for Cu-Au occurrences, in horizontal axis, and P_o, the portion of Cu-Au occurrences correctly classified, in vertical axis. As shown in Fig. 12, the success-rate curves of CRF and GRF models lie above the gauge line, meaning that both models are perfectly plausible. However, the success-rate curve of GRF model is much better than CRF model, and thus, GRF model has performed over the CRF model and is more reliable for delineating high-favorable areas as well as discovering new deposits associated with Cu-Au mineralization.

Accurate interpretation of prospectivity models requires restriction of prospect areas in order to delineate high-favorable targets. For this purpose, a threshold value of 90% confidence interval was selected for discretizing the prospectivity scores of CRF and GRF models. Finally, the favorable targets derived from CRF and GRF models based on 90 % interval were delineated (Fig. 13). The favorable areas in the GRF model (Fig. 13b) captures 100% of known Cu-Au occurrences within only 9% of the study area, while the favorable areas of CRF model (Fig. 13a) contain same percent of known Cu-Au occurrences but within larger areas (14%).

5. Conclusion

This contribution in this work has pioneered on applying the genetic-based RF, namely GRF, to predictive modeling of mineral prospectivity. The main objective of this study is to modulate the exploration uncertainty to MPM and, thus, enhance the predictive accuracy of mineral exploration. Empirical achievements of this study revealed that the proposed GRF model is a highly promising hybrid RF model for predictive modeling of mineral prospectivity. The proposed GRF model was able to automatically adjust the optimized values of RF hyperparameters and exhibited more remarkable increasing in predictive accuracy in given multi-source geo-information than conventional RF. The GRF model was selected as the more applicable predictive model after comprehensive quantitative comparisons using validation techniques (e.g., confusion matrix and success-rate curves). The prospect targets including favorable areas occupy only 9% of the study area while estimating all of the known

Cu-Au deposits. This represents the capability of the proposed model not only for discovering new target explorations in the study region, but also for exploring undiscovered deposits in other promising areas.

References

Behroozi, A. (1987). Geological map of Iran 1: 100,000 series, Feizabad. *Geological Survey of Iran, Tehran*.

Bonham-Carter, G. F. (1994). Geographic information systems for geoscientists-modeling with GIS. *Computer methods in the geoscientists*, *13*, 398.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brown, W. M., Gedeon, T. D., Groves, D. I., & Barnes, R. G. (2000). Artificial neural networks: a new method for mineral prospectivity mapping. *Australian journal of earth sciences*, *47*(4), 757-770.

Carranza, E. J. M. (2008). *Geochemical anomaly and mineral prospectivity mapping in GIS*. Elsevier.

Carranza, E. J. M., & Laborte, A. G. (2015). Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). *Computers & Geosciences*, *74*, 60-70.

Carranza, E. J. M., Hale, M., & Faassen, C. (2008). Selection of coherent deposit-type locations and their application in data-driven mineral prospectivity mapping. *Ore geology reviews*, *33*(3-4), 536-558.

Daviran, M., Maghsoudi, A., Cohen, D. R., Ghezelbash, R., & Yilmaz, H. (2020). Assessment of Various Fuzzy c-Mean Clustering Validation Indices for Mapping Mineral Prospectivity: Combination of Multifractal Geochemical Model and Mineralization Processes. *Natural Resources Research*, *29*(1), 229-246.

Ding, C. S., Haieh, C. T., Wu, Q., & Pedram, M. (1996, November). Stratified random sampling for power estimation. In *Proceedings of International Conference on Computer Aided Design* (pp. 576-582). IEEE.

Ghezelbash, R., Maghsoudi, A., & Carranza, E. J. M. (2019a). Performance evaluation of RBF-and SVM-based machine learning algorithms for predictive mineral prospectivity modeling: integration of SA multifractal model and mineralization controls. *Earth Science Informatics*, *12*(3), 277-293.

Ghezelbash, R., Maghsoudi, A., & Carranza, E. J. M. (2020a). Sensitivity analysis of prospectivity modeling to evidence maps: enhancing success of targeting for epithermal gold, Takab district, NW Iran. *Ore Geology Reviews*, 103394.

Ghezelbash, R., Maghsoudi, A., & Carranza, E. J. M. (2020b). Optimization of geochemical anomaly detection using a novel genetic K-means clustering (GKMC) algorithm. *Computers & Geosciences*, *134*, 104335.

Ghezelbash, R., Maghsoudi, A., & Daviran, M. (2019b). Combination of multifractal geostatistical interpolation and spectrum–area (S–A) fractal model for Cu–Au geochemical prospects in Feizabad district, NE Iran. *Arabian Journal of Geosciences*, *12*(5), 152.

Hu, D., Liu, D., & Xue, S. (1995). Explanatory text of geochemical map of Feizabad (7760). *Geological Survey of Iran, Tehran*.

Kim, K. J., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with Applications*, *19*(2), 125-132.

Kreuzer, O. P., Etheridge, M. A., Guj, P., McMahon, M. E., & Holden, D. J. (2008). Linking mineral deposit models to quantitative risk analysis and decision-making in exploration. *Economic Geology*, *103*(4), 829-850.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Lisitsin, V. A., González-Álvarez, I., & Porwal, A. (2013). Regional prospectivity analysis for hydrothermal-remobilised nickel mineral systems in western Victoria, Australia. *Ore Geology Reviews*, *52*, 100-112.

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28*(3), 385-393.
McCuaig, T. C., Beresford, S., & Hronsky, J. (2010). Translating the mineral systems approach into an effective exploration targeting system. *Ore Geology Reviews*, *38*(3), 128-138.
Mühlenbein, H., Schomisch, M., & Born, J. (1991). The parallel genetic algorithm as function optimizer. *Parallel computing*, *17*(6-7), 619-632.
Parsa, M., Maghsoudi, A., & Yousefi, M. (2018). Spatial analyses of exploration evidence data to model skarn-type copper prospectivity in the Varzaghan district, NW Iran. *Ore Geology Reviews*, *92*, 97-112.
Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers & Geosciences*, *51*, 350-365.
Rodriguez, J. D., Perez, A., & Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), 569-575.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804-818.
Simpson, M. P., Mauk, J. L., & Simmons, S. F. (2001). Hydrothermal alteration and hydrologic evolution of the Golden Cross epithermal Au-Ag deposit, New Zealand. *Economic Geology*, *96*(4), 773-796.
Srinivas, M., & Patnaik, L. M. (1994). Adaptive probabilities of crossover and mutation in genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, *24*(4), 656-667.
Wu, C. H., Tzeng, G. H., Goo, Y. J., & Fang, W. C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert systems with applications*, *32*(2), 397-408.
Zuo, R., Cheng, Q., & Agterberg, F. P. (2009). Application of a hybrid method combining multilevel fuzzy comprehensive evaluation with asymmetric fuzzy relation analysis to mapping prospectivity. *Ore Geology Reviews*, *35*(1), 101-108.

Table captions:

Table 1. Tuned hyperparameter values used for construction of CRF and GRF models.

Table 2. The classification accuracy evaluation indices of CRF and GRF models.

Table 1. Tuned hyperparameter values used for construction of CRF and GRF models.

Model	N_T	N_S	d
CRF	320	6	2
GRF	957	10	3

450 Table 2. The classification accuracy evaluation indices of CRF and GRF models.

Indices	CRF	GRF
Sensitivity	99.32 %	100 %
Specificity	85.9 %	90.6 %
Precision	87.57 %	91.41 %
F-Measure	93.07 %	95.51 %

451
452

Figure captions:

Figure 1. Flowchart of classification using RF algorithm.

Figure 2. Different stages of GRF procedure.

Figure 3. Location of the study area in NE of Iran.

Figure 4. (a) Simplified geological map of Feizabad (1:100,000, modified after Behroozi, 1987), (b) location of the systematically collected stream sediment samples from the study area.

Figure 5. Success-rate curves for 12 selected elements including: (a) Ag, (b) As, (c) Au, (d) Bi, (e) Cr, (f) Cu, (g) Hg, (h) Mo, (i) Ni, (j) Pb, (k) Sb and (l) Zn.

Figure 6. Continuous-value maps for ilr-transformed values of 6 efficient elements including: (a) Au, (b) Cu, (c) Mo, (d) Pb, (e) Sb and (f) Zn.

Figure 7. Maps of (a) fuzzy-based multi-element geochemical signature, (b) presence of Eocene-Oligocene intrusive rocks, (c) fault density and (d) presence of hydrothermal alterations.

Figure 8. Dispersion patterns of different hydrothermal alterations derived from ASTER and Landsat ETM^{+} data.

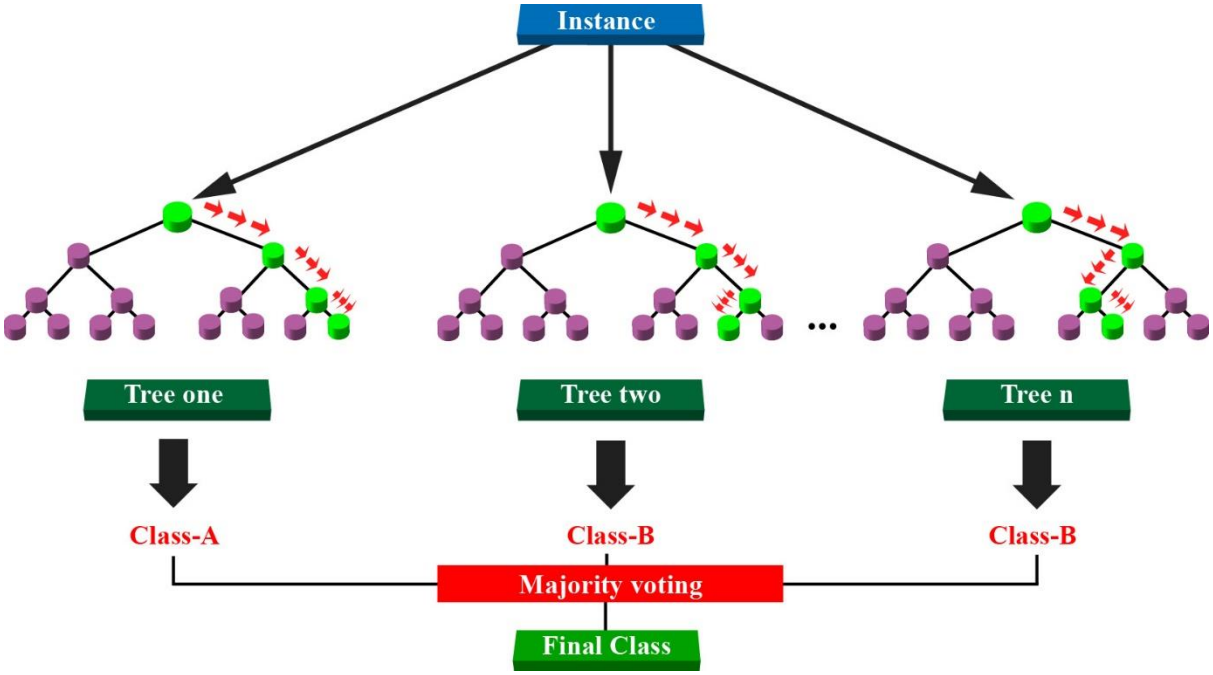
Figure 9. Predictive models of mineral prospectivity derived by (a) CRF and (b) GRF models.

Figure 10. 3D-plot of GRF procedure indicating the number of populations (N_P), number of iterations (N_i) and best fitness in x, y and z axes, respectively. Bold red square represents the optimized condition for calculation of GRF hyperparameters.

Figure 11. Graphical confusion matrices of (a) GRF model-OOB data, (b) CRF model-OOB data, (c) GRF model-training data and (d) CRF model-training data.

Figure 12. Success-rate curves for CRF and GRF predictive models of mineral prospectivity.

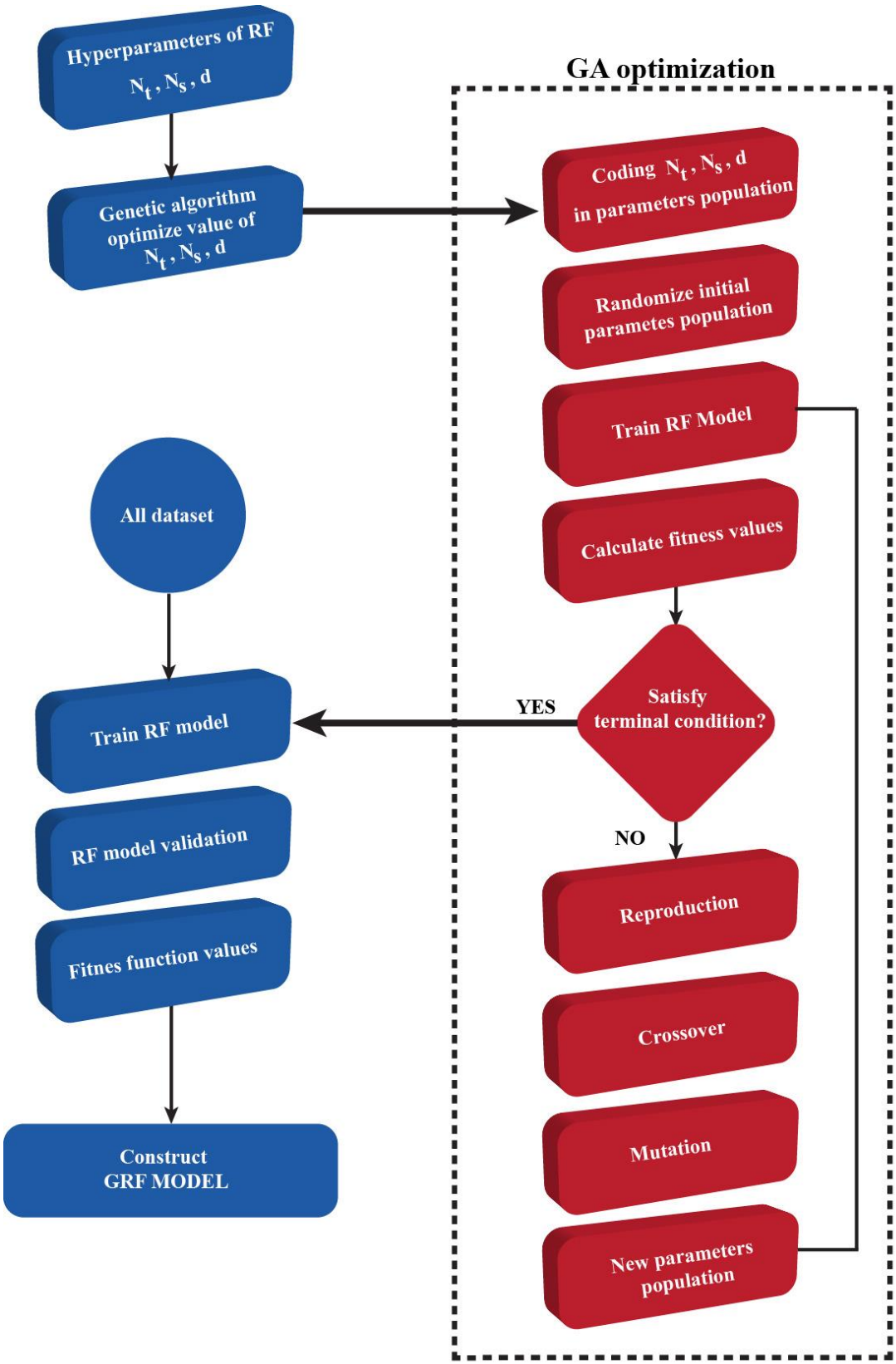
Figure 13. Predictive maps of (a) CRF and (b) GRF models showing favorable and non-favorable areas by a threshold value of 90% confidence interval.



476

477 Fig. 1

478



479

480 Fig. 2
481

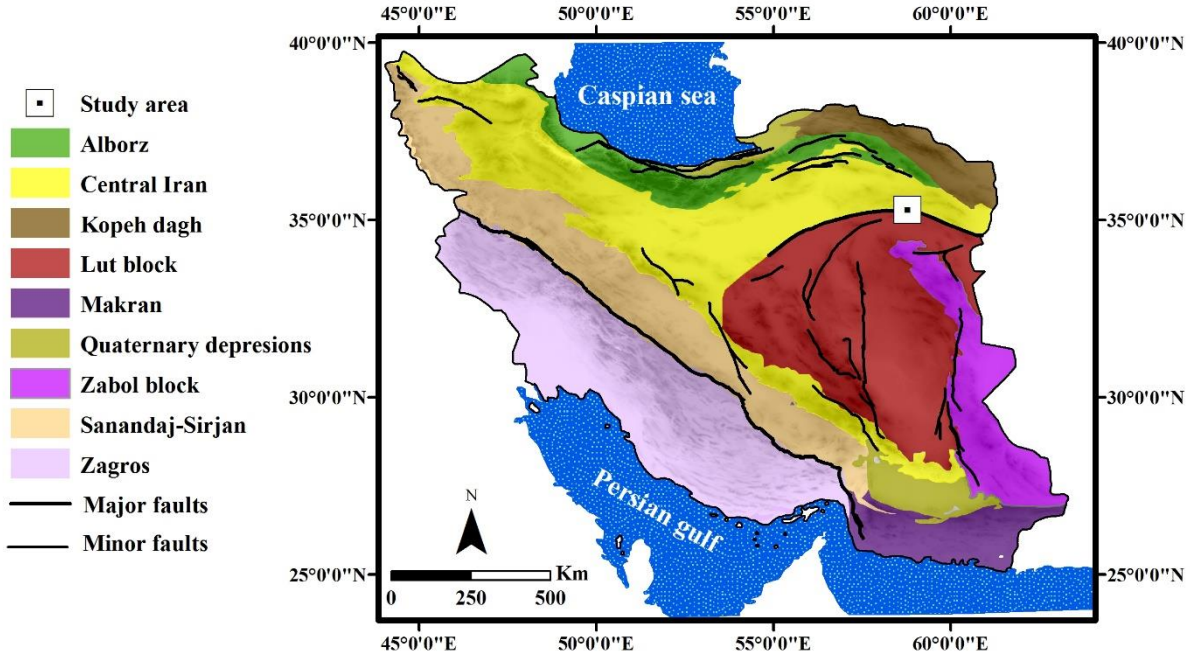

Study area
Alborz
Central Iran
Kopeh dagh
Lut block
Makran
Quaternary depresions
Zabol block
Sanandaj-Sirjan
Zagros
Major faults
Minor faults
45°0'0"E
50°0'0"E
55°0'0"E
60°0'0"E
40°0'0"N
35°0'0"N
30°0'0"N
25°0'0"N
Caspian sea
Persian gulf
N
Km
0
250
500

Fig. 3

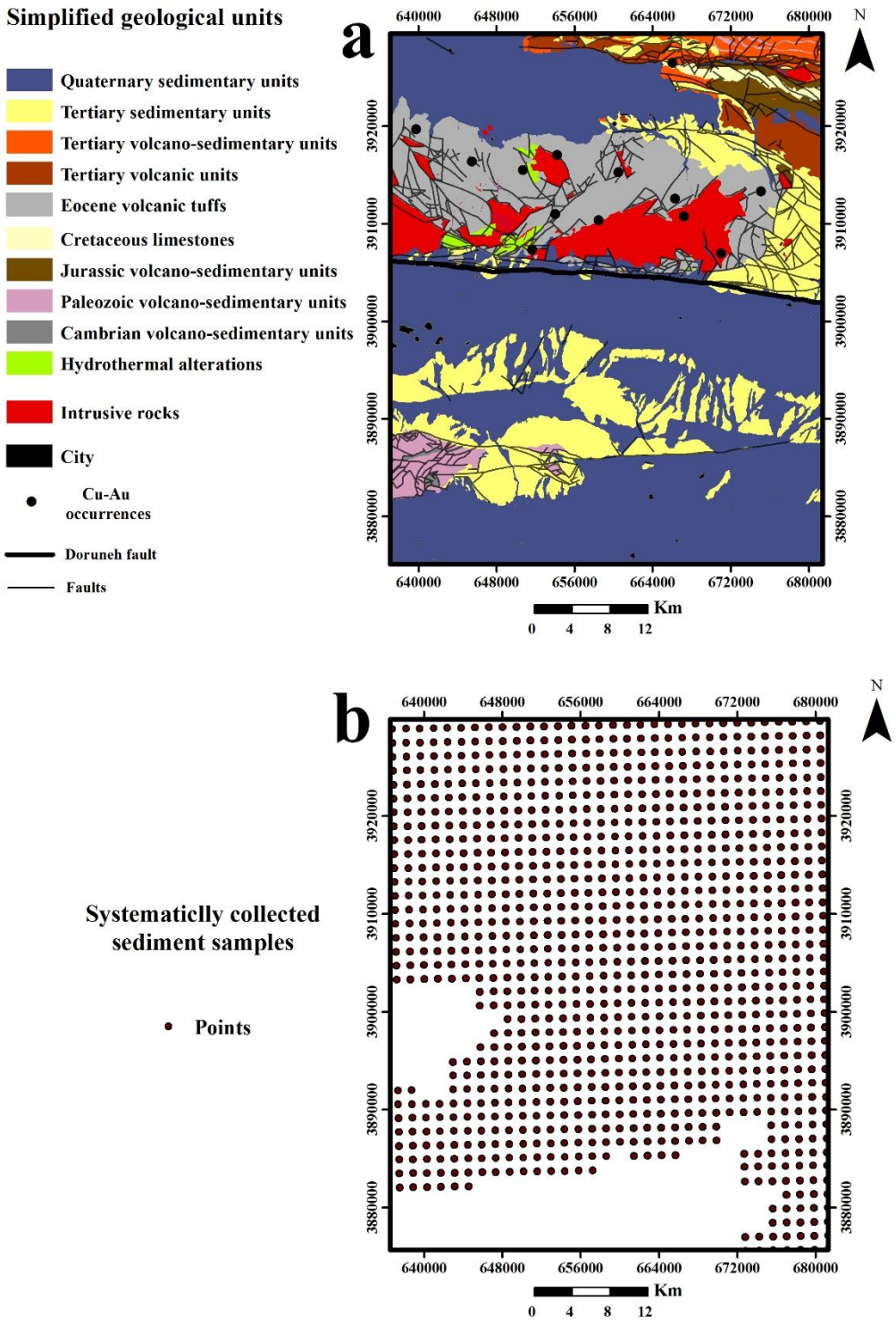

Simplified geological units
Quaternary sedimentary units
Tertiary sedimentary units
Tertiary volcano-sedimentary units
Tertiary volcanic units
Eocene volcanic tuffs
Cretaceous limestones
Jurassic volcano-sedimentary units
Paleozoic volcano-sedimentary units
Cambrian volcano-sedimentary units
Hydrothermal alterations
Intrusive rocks
City
Cu-Au occurrences
Doruneh fault
Faults
a
b
N
640000
648000
656000
664000
672000
680000
3920000
3910000
3900000
3890000
3880000
Km
0 4 8 12
Systematiclly collected sediment samples
Points

Fig. 4

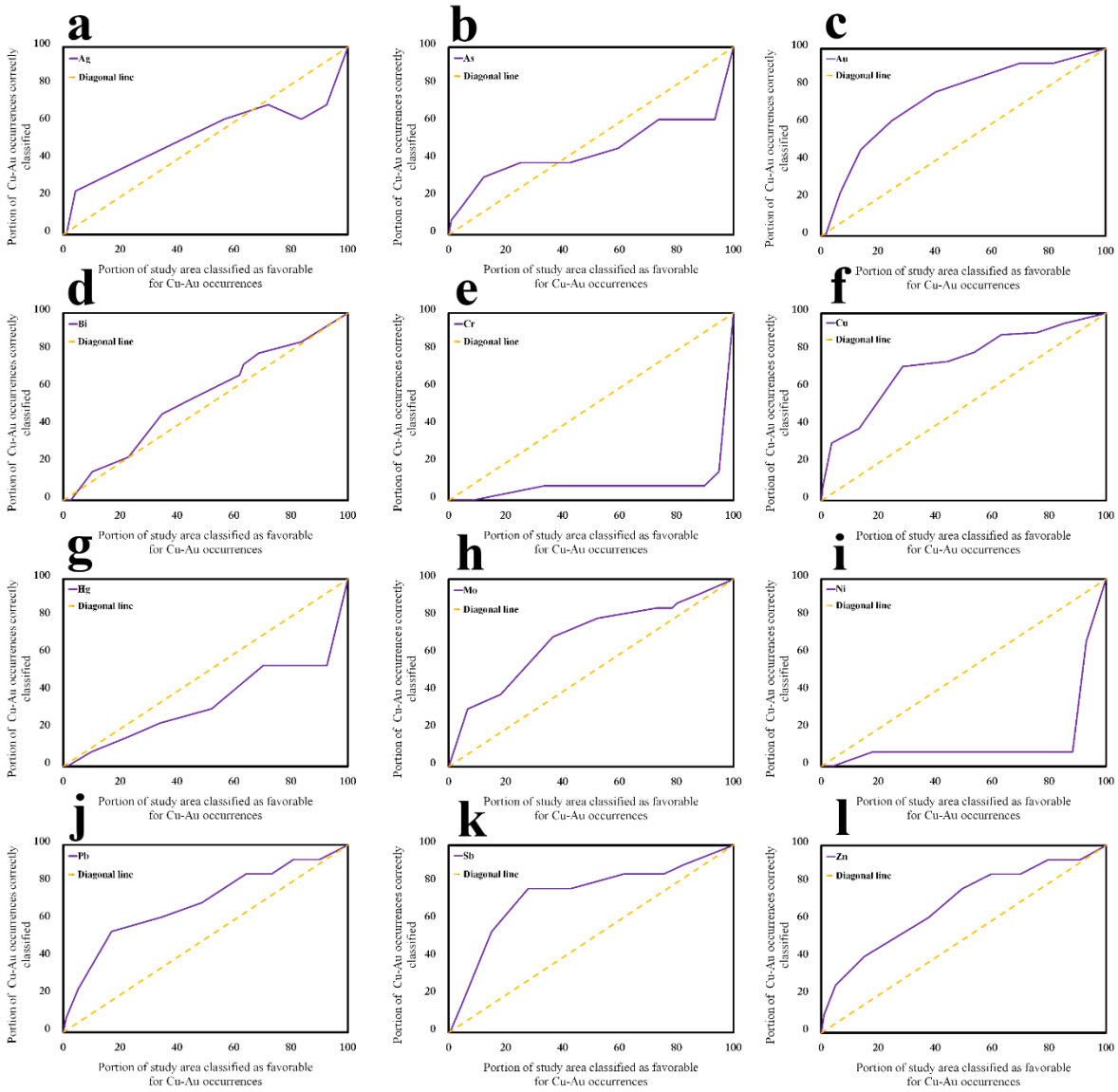

a
Ag
Diagonal line
b
As
Diagonal line
c
Au
Diagonal line
d
Bi
Diagonal line
e
Cr
Diagonal line
f
Cu
Diagonal line
g
Hg
Diagonal line
h
Mo
Diagonal line
i
Ni
Diagonal line
j
Pb
Diagonal line
k
Sb
Diagonal line
l
Zn
Diagonal line
Portion of Cu-Au occurrences correctly classified
Portion of study area classified as favorable for Cu-Au occurrences
0
20
40
60
80
100

Fig. 5

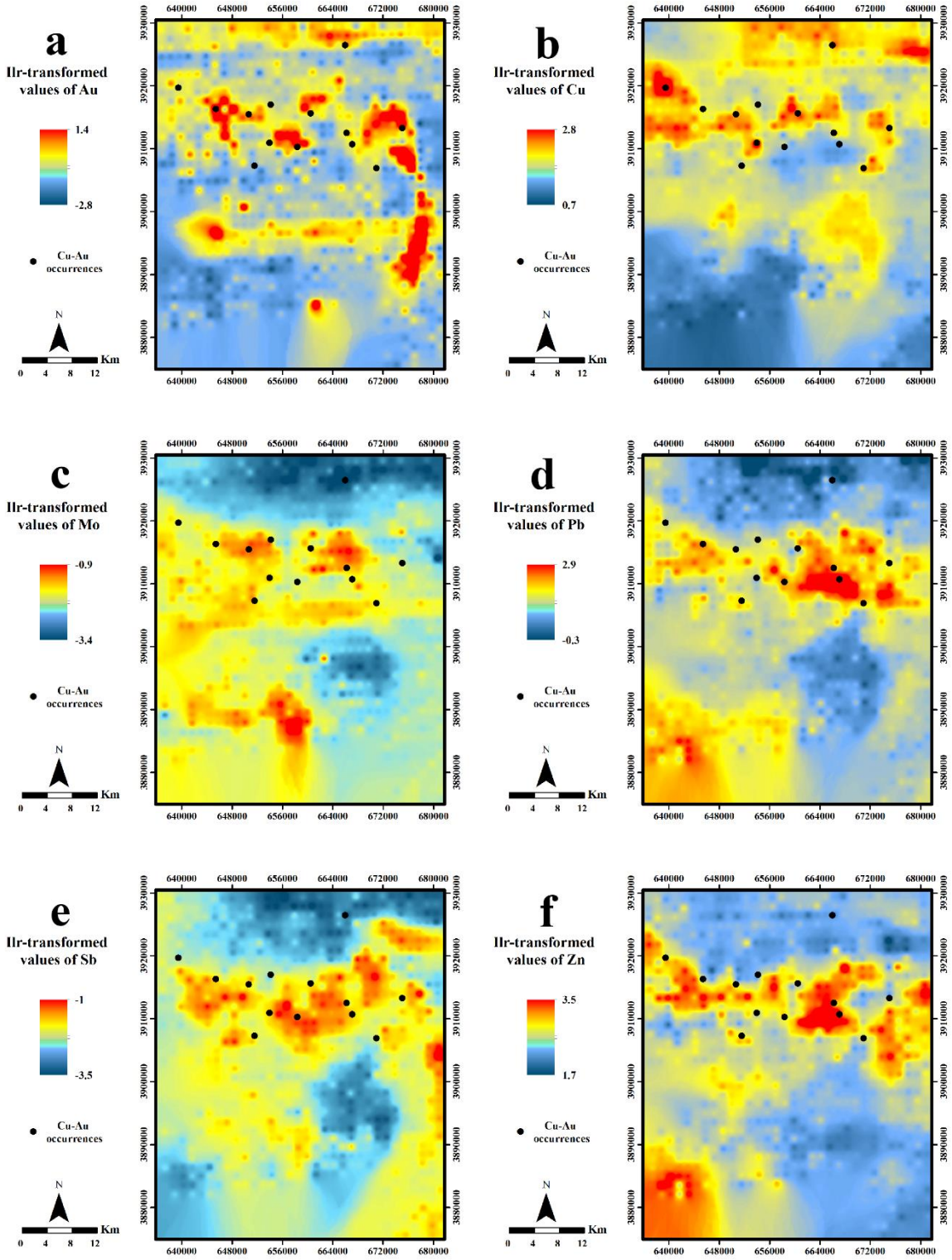

a
Ilr-transformed values of Au
1.4
-2.8
Cu-Au occurrences
N
Km
0 4 8 12
b
Ilr-transformed values of Cu
2.8
0.7
Cu-Au occurrences
c
Ilr-transformed values of Mo
-0.9
-3.4
Cu-Au occurrences
d
Ilr-transformed values of Pb
2.9
-0.3
Cu-Au occurrences
e
Ilr-transformed values of Sb
-1
-3.5
Cu-Au occurrences
f
Ilr-transformed values of Zn
3.5
1.7
Cu-Au occurrences
640000 648000 656000 664000 672000 680000
3930000 3920000 3910000 3900000 3890000 3880000

Fig. 6

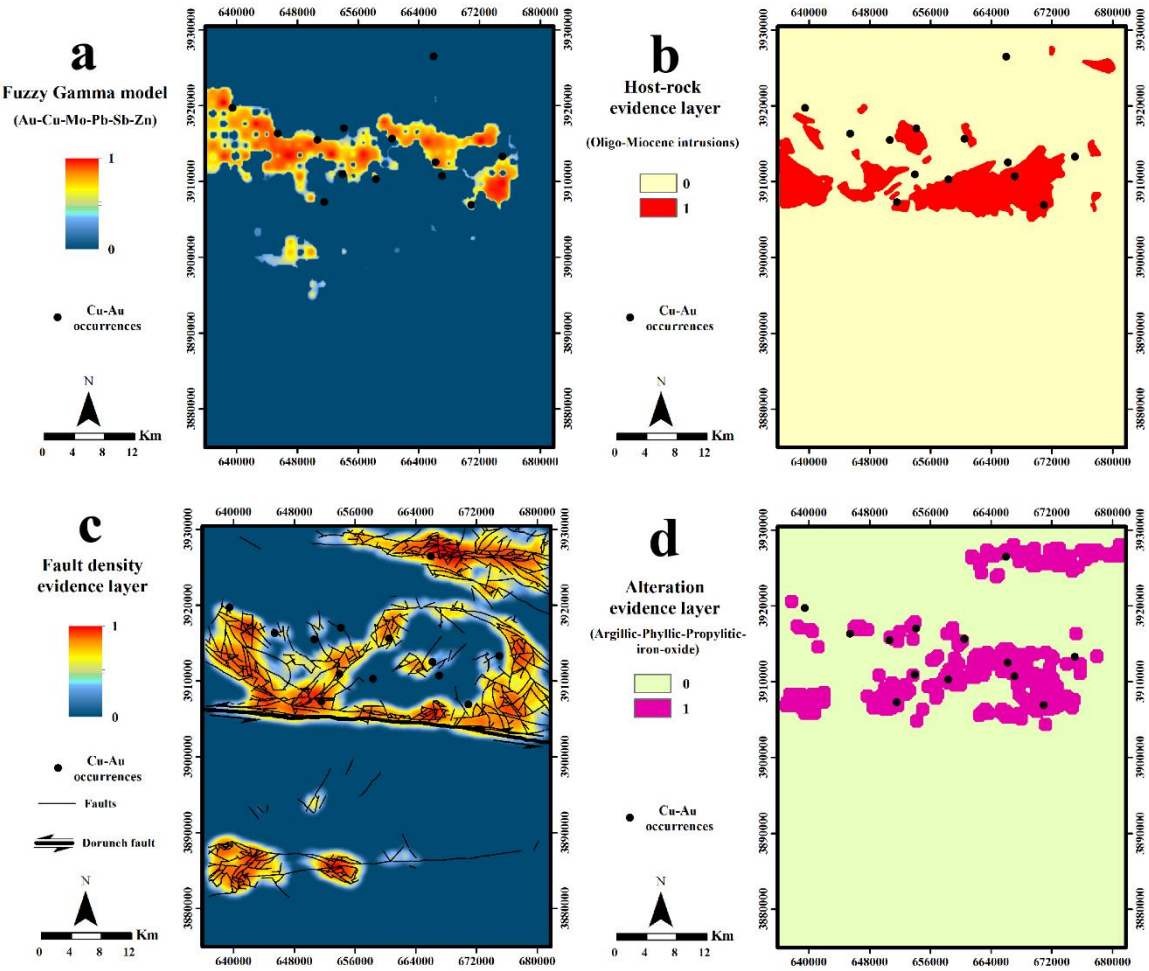
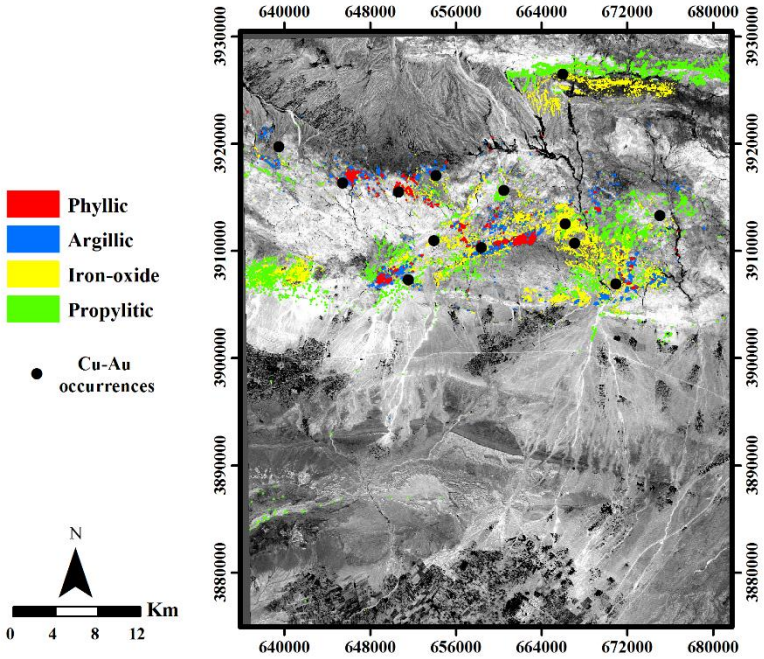

a
Fuzzy Gamma model
(Au-Cu-Mo-Pb-Sb-Zn)
1
0
Cu-Au occurrences
N
Km
0 4 8 12
b
Host-rock evidence layer
(Oligo-Miocene intrusions)
0
1
Cu-Au occurrences
c
Fault density evidence layer
Faults
Dorunch fault
d
Alteration evidence layer
(Argillic-Phyllic-Propylitic-iron-oxide)
640000 648000 656000 664000 672000 680000
3930000 3920000 3910000 3900000 3890000 3880000

Fig. 7



505
506
507
508 Fig. 8
509

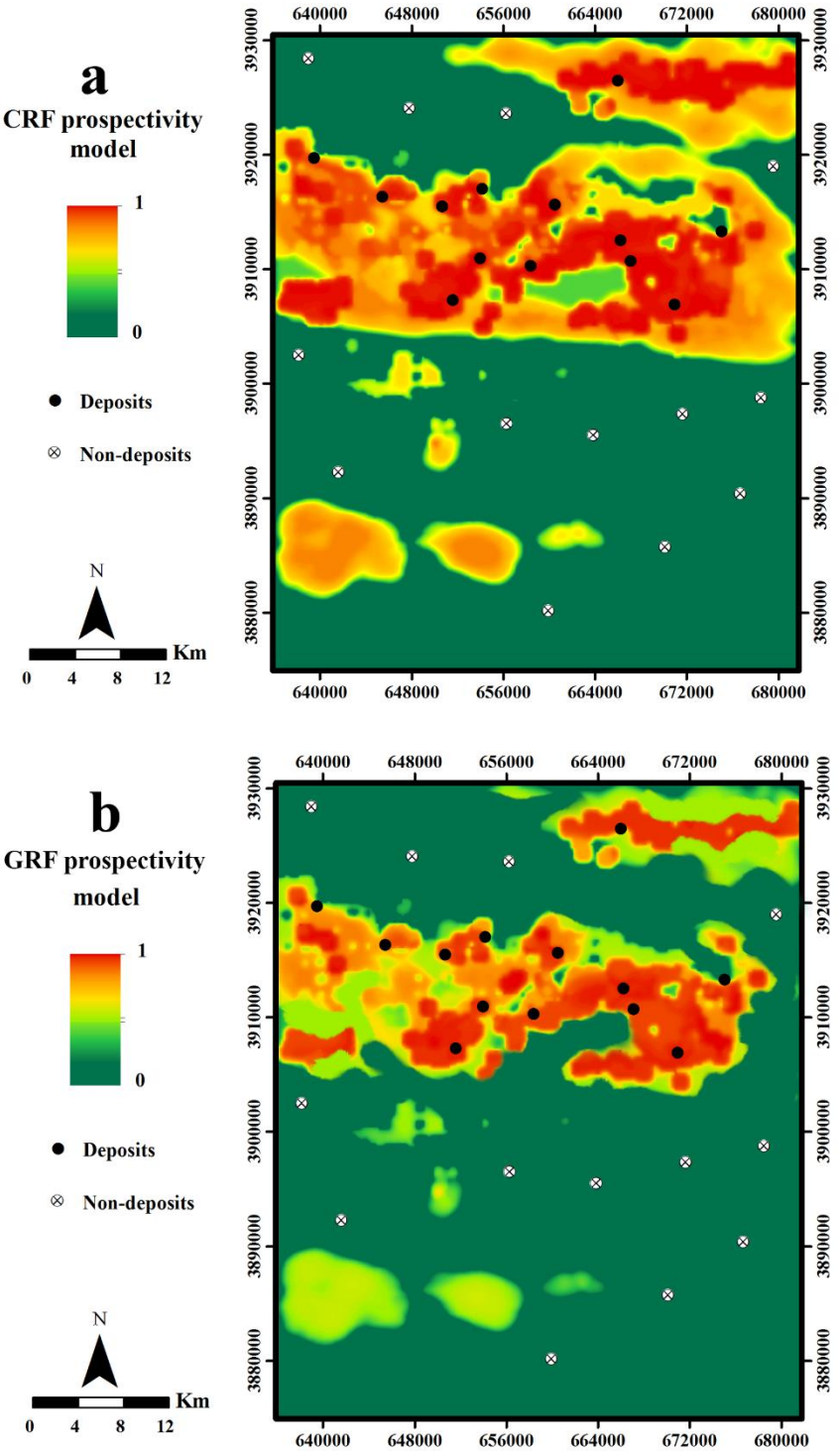

a
CRF prospectivity model
1
0
Deposits
Non-deposits
N
0 4 8 12 Km
640000 648000 656000 664000 672000 680000
3930000 3920000 3910000 3900000 3890000 3880000
b
GRF prospectivity model
1
0
Deposits
Non-deposits
N
0 4 8 12 Km
640000 648000 656000 664000 672000 680000
3930000 3920000 3910000 3900000 3890000 3880000

Fig. 9

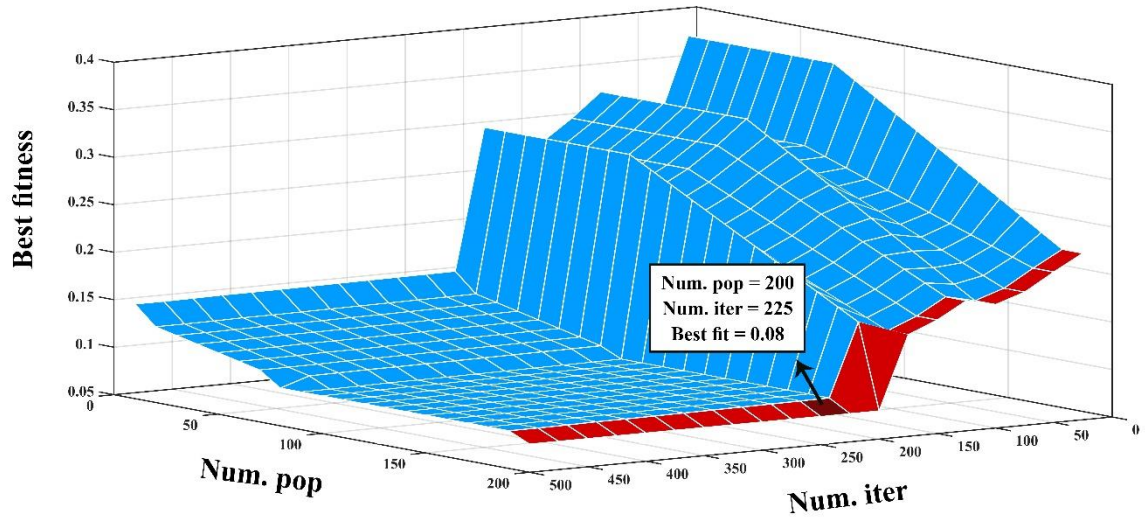
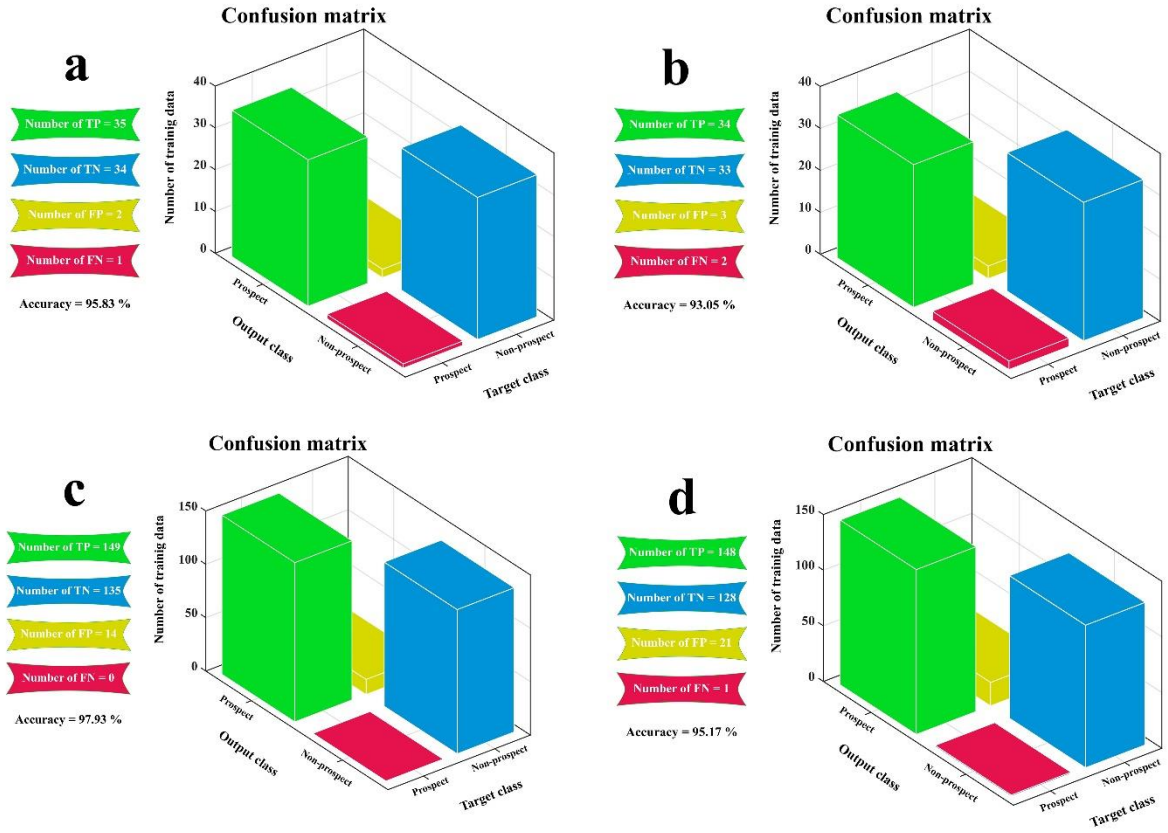
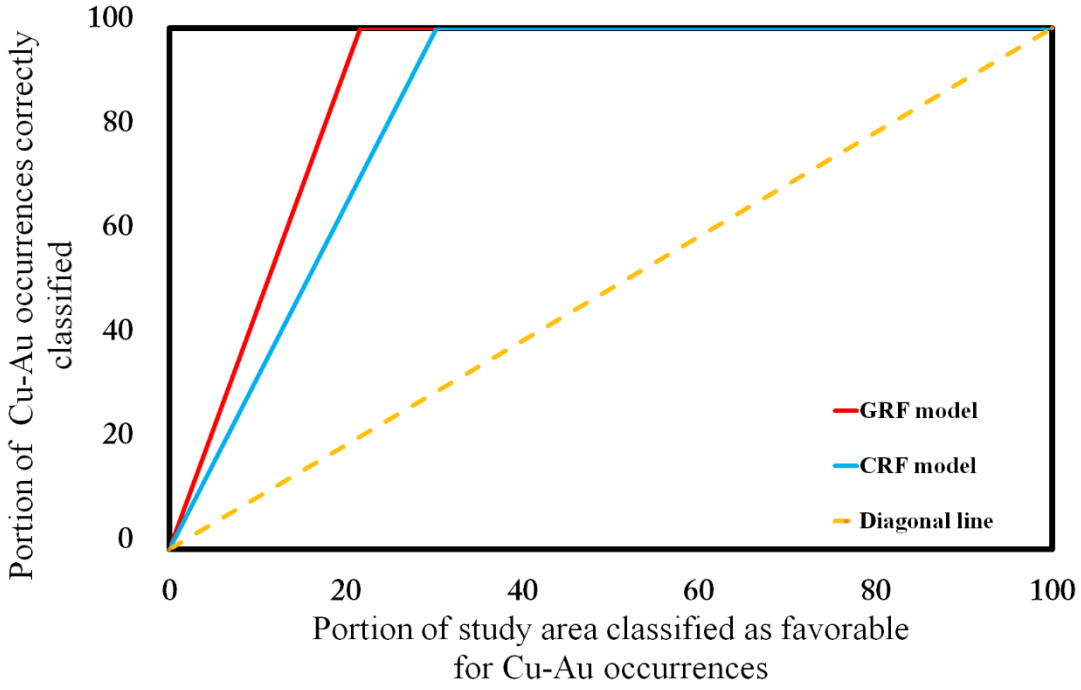

Best fitness
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
Num. pop = 200
Num. iter = 225
Best fit = 0.08
0
50
100
150
200
Num. pop
500
450
400
350
300
250
200
150
100
50
0
Num. iter

Fig. 10



518
519
520 Fig. 11
521



522
523 Fig. 12
524

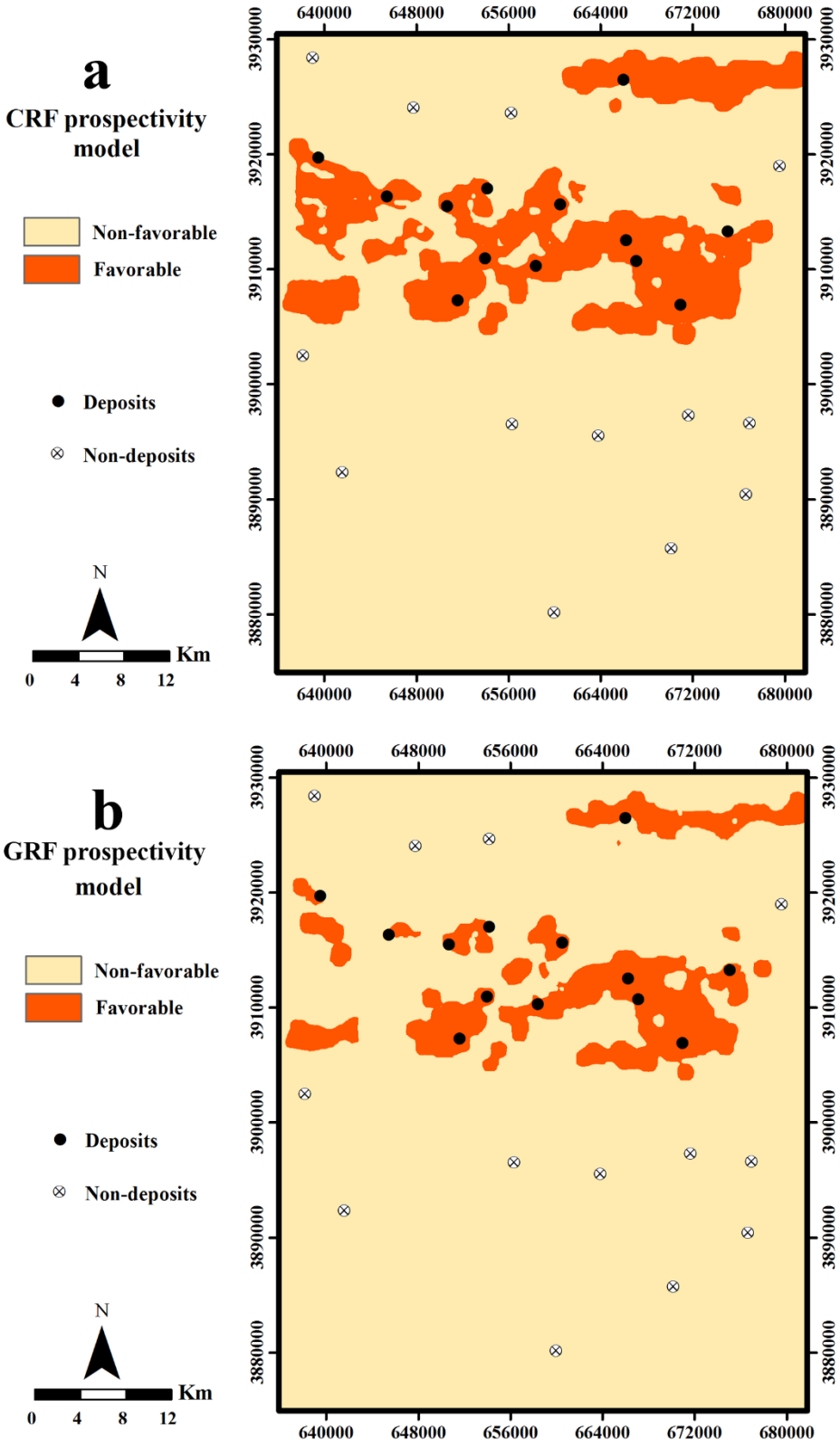

a
CRF prospectivity model
Non-favorable
Favorable
Deposits
Non-deposits
N
Km
0 4 8 12
640000 648000 656000 664000 672000 680000
3930000 3920000 3910000 3900000 3890000 3880000
b
GRF prospectivity model
Non-favorable
Favorable
Deposits
Non-deposits
N
Km
0 4 8 12
640000 648000 656000 664000 672000 680000
3930000 3920000 3910000 3900000 3890000 3880000

Fig. 13