Comparison of different approaches for modeling of heavy metal estimations

Parveen Sihag¹ · Ali Keshavarzi² · Vinod Kumar³

© Springer Nature Switzerland AG 2019

Abstract

Soil contamination by heavy metals is very important for environmental scientists due to the greater mobilization of metals and the possibility of contamination of groundwater. One of the most effective tools for evaluating the risk is the combination of experimentation with computer modeling. Modeling techniques are important in assessing the potential risks associated with heavy metals in the environment. Determination of models that can precisely evaluate the heavy metals in soils is an important need of agricultural researches, which could eradicate the weaknesses in the measurement of heavy metals in soils. The purpose of the present study is to test and compare different models according to their suitability for describing the estimation of heavy metals. The models used in this study were multilayer perceptron neural network (MLP), M5 model tree (M5) and bagging approach (BM5P). The data from 164 sampling sites from Neyshabur and Mashhad plains were taken in this study. The inputs combination according to feature selection-based correlation was used to feed the models. To model soil heavy metals, soil attributes, namely sand, silt, clay (as texture fractions), organic carbon, pH and available phosphorus, were entered in some models. To evaluate the performance of various techniques used in this study, several statistical indexes, including the correlation coefficient, root-mean-square error, Nash–Sutcliffe coefficient, Willmott's index (d) and mean absolute error, were assessed. Comparison of different models for Fe, Cu, Mn and Zn indicated that MLP is the most suitable method for estimations of Fe and Mn, whereas BM5P and M5P are the most suitable models for determinations of Cu and Zn, respectively. This study concluded that machine learning models can be successfully applied to the rapid prediction of soil heavy metals using soil variables.

Keywords BM5P · Heavy metals · Khorasan-e-Razavi Province · MLP · M5P

Abbreviations

MLP	Multilayer perceptron neural network
BM5P	Bagging approach
OC	Organic carbon
Ava. P	Available phosphorus
CC	Correlation coefficient
RMSE	Root-mean-square error
NSE	Nash–Sutcliffe coefficient
MAE	Mean absolute error
Cu	Copper
Zn	Zinc
Fe	Iron

Mn	Manganese
AAS	Atomic absorption spectrometer

1 Introduction

Soil comprises reduced and non-renewable resources that required being preserved, although they are polluted by various heavy metals [1–6]. Heavy metals accumulation in soil have negative impact on the environment and endanger the human's health and ecosystems via energy and material cycling [7]. The industrial, urban and agricultural

✉ Ali Keshavarzi, alikeshavarzi@ut.ac.ir; ✉ Vinod Kumar, vinodverma507@gmail.com | ¹National Institute of Technology, Kurukshetra, Haryana, India. ²Laboratory of Remote Sensing and GIS, Department of Soil Science, University of Tehran, P.O. Box: 4111, Karaj 31587-77871, Iran. ³Department of Botany, DAV University, Jalandhar, Punjab, India.

activities are the major causes for soil contamination [8, 9]. Reducing and controlling the pollution of heavy metals in soils is to find out the sources of contamination [10, 11]. On local scale, agricultural soils are being polluted by addition of heavy metals through natural and human activities [12–14]. Heavy metals in agricultural soils are mainly derived from weathering of parent materials, industrial emissions, disposal of high metal wastes, fertilizers, agrochemicals, irrigation water, atmospheric accumulation and pesticides [15, 16]. Each conclusion concerning the function of any assessment in soil quality and organization should be established on authentic data on the degree and causes of heavy metal contamination in a region [17]. Hence, the investigation and distribution of heavy metal pollution causes in agricultural soils on local level is important.

The elevated spatial diverseness of heavy metals in soils, the wide arrayness of pollution causes and improper long-term monitoring information are the issues for scientists to evaluate the multi-source pollution of heavy metals from agricultural soils on a local level; searching appropriate approaches to tackle this issue is important. For solving all these issues, modeling approaches serve as important techniques in investigation and apportionment of heavy metal source [18]. These techniques required a minute sample size, and consequently time and work are cost-efficient [19]. Diverse modeling approaches have been used for the determination of associations among spatial distribution of soil characteristics. Extrapolative mapping techniques such as geostatistical, linear and multiple regression, and neural networks have been effectively applied for soil mapping [20, 21]. In Iran, widespread usage of pesticides, fertilizers and fast expansion of industrial and urban development took place, but there is no appropriate data about agricultural soils of different land types in Iran and standard procedures regarding different modeling techniques. With this objective, the present research was made to develop models to determine its possibility to assess heavy metals such as Fe, Mn, Cu and Zn in agricultural soils of Northeastern Iran employing modeling.

2 Materials and methods

2.1 Study area

Neyshabur plain, Khorasan-e-Razavi Province, Northeast Iran, was chosen for the present study (Fig. 1). Neyshabur plain is situated between lat. (35°40'N–36°40'N) and long. (58°12'E–59°31'E) with altitude of 1256 m above mean sea level and Mashhad Plain is located between lat. (35°59'N–37°04'N) and long. (58°22'E–60°07'E), and 900–1500 m above sea level. The climate of Neyshabur

plain is semiarid with average annual precipitation of 233.7 mm and average annual temperature of 14.5 °C. The irrigated farming is the major land use approach in Neyshabur plain. Mashhad plain is described by semiarid climate with average annual temperature and precipitation of 15.8 °C and 222.1 mm, respectively. The primary soil types are Calcaric Fluvisols, Calcaric Cambisols, Calcaric Regosols and Gypsic Regosols, found in colluvial fans, pediment plains, plateau and upper terraces, respectively. Mashhad plain is mainly consisting of granitic and metamorphic rocks. The major portions of these rocks are covered with loess deposit. Soil texture changed from loam to sandy loam, and sandy loam to sandy clay loam.

2.2 Soil sampling and analysis

A digital elevation model (DEM) with 10 × 10 m grid size was prepared using ArcGIS software from topographic map (1:25,000 scale) with 10 m contour lines distance. A total of 164 soil samples were collected from 0 to 30 cm and 30 to 60 cm soil depth by stratified random sampling technique. The sampling points were prepared in such a way that indicate all the main soils and land use types. The collected soil samples were cleaned by removal of plant materials and other pebbles, air-dried, powdered and sieved by using 2-mm sieve. The soil pH was determined with the help of digital pH meter [22]. Soil textural properties [sand (0.05–2 mm), silt (0.002–0.05 mm) and clay (< 0.002 mm)] were estimated by the hydrometer method [23]. Soil organic carbon (SOC) content was estimated by following Walkley and Black [24]. Heavy metals (Fe, Mn, Cu and Zn) in the agricultural soils were estimated by atomic absorption spectrometer (AAS). The soil samples were digested by the aqua regia method by following the protocol of Kumar et al. [25]. The limits of detection of the instrument were: Fe (7.3 mg/l), Cu (1.2 mg/l), Mn (1.0 mg/l) and Zn (1.6 mg/l). For quality assurance and quality control (QA–AC), the standards and blanks were run after every five samples to check the 95% accuracy of instrument [1]. The 95–105% recovery rates for samples spiked with standards authenticated that the results were satisfactory [26].

2.3 Multilayer perceptron neural network (MLP)

McCulloch and Pitts [27] first time proposed the artificial neural network. In general, the layered perceptron neural network structure contains three layers: the input, hidden and target layers. Each layer includes the number of neurons (Fig. 2). The number of neurons in the input and output layers is determined by the nature of the problem under consideration, while the number of neurons in the hidden layers, as well as the number

Fig. 1 Location of study area and sampling points (up: Mashhad plain, down: Neyshabur plain)

of these layers, is determined by the trial and errors to reduce the amount of error by the user [28, 29]. Each of the neurons in the input layer is weighted whose value determines the effect of each variable on the input layer performance. Each neuron consists of two parts: In the first part, the weighted sum of the input values is computed, and in the second part, the output of the first part

is in form of a mathematical function and through which the output of the neuron is calculated. This mathematical function is referred to as the actuator function, the threshold function, or the transfer function which functions like a nonlinear filter and makes the output of the neuron in a number range [30].

Fig. 2 Simple configuration of multilayer perceptron neural network

2.4 M5 model tree (M5)

M5 tree, introduced by Quinlan [31], is a decision tree learner for regression problems. This tree algorithm assigns linear regression functions at the terminal nodes and fits a multivariate linear regression model to each subspace by classifying or dividing the whole data space into several subspaces. The M5 tree method deals with continuous class problems instead of discrete classes and can handle tasks with very high dimensionality. It reveals piecewise information of each linear model constructed to approximate nonlinear relationships of the dataset.

The information about the splitting criteria for the M5 model tree is gained based on the calculation of error at each node. The error is analyzed by the standard deviation of the class values that arrive at a node. The attribute that maximizes the expected error reduction resulted from the testing of each attribute at that node is chosen for splitting at the node. The standard deviation reduction (SDR) is calculated by:

$$\text{SDR} = \text{sd}(K) - \sum \frac{|K_i|}{|K|} \text{sd}(K_i), \quad (1)$$

where K indicates set of instances that attain the node; K_i indicates the subset of illustrations that have the i th product of the possible set; and sd indicates the standard deviation.

2.5 Bagging approach (BM5P)

Bagging is a machine learning ensemble meta-algorithm, which is designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Bagging is a special case of the model

averaging approach. Bagging is a process to develop a training dataset by randomly drawing with substitute X examples, where X is the size of the initial training set [32], or a randomly selected part of the training set will be used for the construction of individual trees for every feature/feature blend. In case of bagging (bootstrap sample), training set consists about 70% of data from the initial dataset; thus, the remaining 30% of the data are missing from every tree matured. These missing data are called out-of-bag (out of the bootstrap sampling). Bagging leads to “improvements for unstable procedures” [33] which include, for example, ANN, classification and regression trees, and subset selection in multi-linear regression.

2.6 Model evaluation criteria

To evaluate the performance of various techniques used in this study, several statistical indexes, including the correlation coefficient (CC), root-mean-square error (RMSE), Nash–Sutcliffe coefficient (NSE), Willmott’s index (d) and mean absolute error (MAE), are assessed.

The closer the CC to unity, the higher the agreement among actual and predicted values. The precision in the prediction is higher when the values of RMSE and MAE approach zero. The closer the NSE and d values to unity, the higher the accuracy in prediction. The considered performance evaluation parameters are defined as follows:

$$\text{CC} = \frac{a \sum kl - (\sum m)(\sum l)}{\sqrt{a(\sum k^2) - (\sum k)^2} \sqrt{n(\sum l^2) - (\sum l)^2}} \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{a} \left(\sum_{i=1}^a (l - k)^2 \right)} \quad (3)$$

Fig. 3 Correlation on input and target variables

Table 1 Inputs combination according to feature selection-based correlation ($n = 164$)

Target	Input
Fe (mg/kg)	Sand + silt + pH + Ava.P
Mn (mg/kg)	OC + clay + pH + Ava.P
Zn (mg/kg)	Sand + silt + clay + Ava.P
Cu (mg/kg)	Sand + silt + OC

$$d = 1 - \frac{\sum_{i=1}^a (l - k)^2}{\sum_{i=1}^a (|l - \bar{k}|) + (|k - \bar{k}|)^2} \quad (4)$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^a (l - k)^2}{\sum_{i=1}^a (k - \bar{k})^2} \quad (5)$$

$$\text{MAE} = \frac{1}{a} \left(\sum_{i=1}^a |l - k| \right), \quad (6)$$

where k are the observed values, l predicted values and a number of observations.

2.7 Model development

Total dataset consists of 164 observations of Fe, Mn, Zn, Cu, pH, OC, sand, silt, clay and available P (Ava. P). Out of 164 observations, 117 were selected for model preparation, 24 were selected for testing and rest 23 were used for validation purposes. Figure 3 indicates the correlation

among various variables, and input combination for model development for Fe, Mn, Zn and Cu is given in Table 1.

2.8 Implementation of machine learning methods

Five standard statistical measures, CC, RMSE, d , NSE and MAE, were selected to assess the performance of the machine learning methods. Numerous trials were carried out to find optimum value of primary parameters. Higher values of CC, d and NSE and lesser values of RMSE and MAE indicate better estimation accuracy of the models. In M5P, calibration of models was done by changing the value of number of instances allowed at each node (m), and iteration and number of instances are the primary parameters of Bagged M5P tree model. Hidden layers, number of neurons, learning rate, momentum and iteration are the primary parameters for the ANN model.

3 Results

3.1 Models for Fe (mg/kg)

3.1.1 Dataset

The preparation of training, testing and validation dataset is like Zn, Cu and Mn dataset. Descriptive statistics of total dataset for Fe (mg/kg) are represented in Table 2, in which pH, sand, silt and Ava. P are input parameters and Fe is the output. Descriptive statistics of datasets utilized for training ($n = 117$), testing ($n = 24$) and validation ($n = 23$) for Fe (mg/kg) are listed in Table 3. Figure 4 indicates the 3D surface plot of Fe (mg/kg) against silt (%) and pH.

3.1.2 Assessment of model's performance in Fe (mg/kg) estimation in soil

Performance evaluation parameters considered for model's performance evaluation are same as for implementing Zn, Cu and Mn models' evaluation. Figure 5 shows the scatter plot between actual and predicted Fe (mg/kg) in soil using MLP-, M5P- and Bagged M5P-based models for training, testing and validation stages. The prediction accuracy of MLP model (RMSE = 1.4625, MAE = 1.0473 for testing and RMSE = 1.2890, MAE = 0.9890 for validation) is found higher than M5P and BM5P models. Table 4 indicates that MLP model is most suitable model than M5P- and BM5P-based models for the estimation of Fe for this dataset.

Table 2 Descriptive statistics for Fe, Cu, Mn and Zn (mg/kg) ($n = 164$)

Variables	Mean	Min	Max	SD	CV	Skewness	Kurtosis
Fe	3.56	1.16	9.64	1.61	45.4	1.01	1.07
pH	8.01	7.5	8.40	0.18	2.29	−0.62	0.02
Sand	35.74	13.0	73.0	11.03	30.86	0.70	0.92
Ava. P	12.68	1.20	63.2	12.9	102.1	1.90	3.26
Silt	43.66	19.0	66.0	9.55	21.88	−0.09	−0.49
Cu (mg/kg) ($n = 164$)							
Cu	1.31	0.62	3.46	0.48	36.5	1.19	2.11
OC	0.58	0.11	1.61	0.31	53.4	1.36	2.04
Sand	35.74	13.0	73.0	11.03	30.8	0.70	0.92
Silt	43.66	19.0	66.0	9.55	21.8	−0.09	−0.49
Mn (mg/kg) ($n = 164$)							
Mn	6.80	1.64	21.0	3.29	48.4	1.51	3.23
pH	8.01	7.50	8.40	0.18	2.2	−0.62	0.02
OC	0.58	0.11	1.61	0.31	53.4	1.36	2.04
Clay	20.59	4.00	41.0	6.95	33.7	0.12	−0.29
Ava. P	12.68	1.20	63.2	12.9	102.1	1.90	3.26
Zn (mg/kg) ($n = 164$)							
Zn	1.63	0.20	13.9	2.68	163.9	3.04	8.88
Sand	35.74	13.0	73.0	11.03	30.8	0.70	0.92
Silt	43.66	19.0	66.0	9.55	21.8	−0.09	−0.49
Clay	20.59	4.0	41.0	6.95	33.7	0.12	−0.29
Ava. P	12.68	1.2	63.2	12.95	102.1	1.9	3.26

The data were significant at $P < 0.05$

Table 3 Descriptive statistics of datasets utilized for training, testing and validation

	Mean	Minimum	Maximum	SD	C.V.	Skewness	Kurtosis
Fe (mg/kg)							
Train ($n = 117$)	3.53	1.16	9.64	1.50	42.5	0.99	1.67
Test ($n = 24$)	3.75	1.38	8.76	2.03	54.2	0.99	0.15
Validation ($n = 23$)	3.51	1.38	7.72	1.75	50.1	0.99	0.22
Cu (mg/kg)							
Train ($n = 117$)	1.33	0.62	3.46	0.49	37.2	1.41	2.80
Test ($n = 24$)	1.30	0.66	2.16	0.49	37.8	0.40	−1.26
Validation ($n = 23$)	1.25	0.68	2.16	0.40	32.0	0.36	−0.64
Mn (mg/kg)							
Train ($n = 117$)	6.64	1.64	21.06	3.04	45.8	1.63	4.46
Test ($n = 24$)	6.63	2.02	18.08	3.65	55.1	1.59	3.14
Validation ($n = 23$)	7.80	2.16	18.2	4.04	51.8	1.02	1.22
Zn (mg/kg)							
Train ($n = 117$)	1.71	0.2	13.9	2.78	162.3	3.07	9.12
Test ($n = 24$)	1.28	0.2	10.1	2.36	184.2	3.26	10.14
Validation ($n = 23$)	1.59	0.3	10.46	2.52	158.3	2.82	7.70

The data were significant at $P < 0.05$

3.2 Models for Cu (mg/kg)

3.2.1 Dataset

The preparation of training, testing and validation dataset is like Zn dataset. Descriptive statistics of total dataset for

Cu (mg/kg) ($n = 164$) are represented in Table 2, in which organic carbon (OC), sand and silt are input parameters and Cu is the output. Descriptive statistics of datasets utilized for training ($n = 117$), testing ($n = 24$) and validation ($n = 23$) for Cu (mg/kg) are listed in Table 3. Figure 6

Fig. 4 3D surface plot of Fe (mg/kg) against pH and silt (%)

Fig. 5 Performance of MLP, M5P and BM5P for estimating Fe

Table 4 Performance evaluation parameters for MLP, M5P and BM5P techniques using Fe, Cu, Mn, Zn dataset

Approaches	CC	RMSE	NSE	<i>d</i>	MAE
Fe dataset					
Training dataset					
MLP	0.665	1.119	0.442	0.778	0.863
M5P	0.625	1.197	0.361	0.663	0.885
BM5P	0.685	1.153	0.407	0.683	0.850
Testing dataset					
MLP	0.693	1.462	0.461	0.812	1.047
M5P	0.615	1.691	0.280	0.562	1.181
BM5P	0.569	1.748	0.231	0.508	1.278
Validation dataset					
MLP	0.668	1.289	0.438	0.747	0.989
M5P	0.543	1.458	0.281	0.597	1.153
BM5P	0.643	1.398	0.339	0.603	1.111
Cu dataset					
Training dataset					
MLP	0.451	0.440	0.203	0.572	0.331
M5P	0.573	0.409	0.312	0.636	0.311
BM5P	0.603	0.403	0.330	0.633	0.306
Testing dataset					
MLP	0.689	0.359	0.450	0.747	0.324
M5P	0.599	0.395	0.332	0.650	0.356
BM5P	0.673	0.379	0.386	0.672	0.349
Validation dataset					
MLP	0.632	0.316	0.357	0.707	0.271
M5P	0.614	0.325	0.317	0.675	0.287
BM5P	0.678	0.306	0.398	0.692	0.253
Mn dataset					
Training dataset					
MLP	0.501	2.623	0.251	0.619	1.968
M5P	0.622	2.453	0.345	0.649	1.841
BM5P	0.627	2.424	0.360	0.666	1.806
Testing dataset					
MLP	0.698	2.744	0.412	0.695	2.079
M5P	0.554	3.067	0.266	0.579	2.429
BM5P	0.258	3.710	-0.073	0.476	2.891
Validation dataset					
MLP	0.343	3.936	0.010	0.507	2.934
M5P	0.047	4.242	-0.148	0.355	3.006
BM5P	0.153	4.131	-0.089	0.388	2.958
Zn dataset					
Training dataset					
MLP	0.607	2.202	0.367	0.708	1.289
M5P	0.605	2.278	0.323	0.614	1.339
BM5P	0.590	2.289	0.316	0.614	1.345
Testing dataset					
MLP	0.960	0.828	0.871	0.967	0.710
M5P	0.773	1.751	0.426	0.699	1.246
BM5P	0.785	1.697	0.461	0.733	1.223
Validation dataset					
MLP	0.338	2.361	0.088	0.297	1.391
M5P	0.552	2.115	0.268	0.535	1.188
BM5P	0.487	2.214	0.199	0.445	1.372

Table 4 (continued)

Coefficient of correlation (CC), root-mean-square error (RMSE), Nash–Sutcliffe coefficient (NSE), Willmott's index (d) and mean absolute error (MAE)

Bold italic values are indicated to the best suited model for the estimation of particular heavy metal

Fig. 6 3D surface plot of Cu (mg/kg) against OC (%) and sand (%)

indicates the 3D surface plot of Cu (mg/kg) against OC (%) and sand (%).

3.2.2 Assessment of model's performance in Cu (mg/kg) estimation in soil

Performance evaluation parameters considered for model's performance evaluation are same as for implementing Fe model evaluation. Figure 7 shows the scatter plot between actual and predicted Cu (mg/kg) in soil using MLP-, M5P- and Bagged M5P-based models for training, testing and validation stages. The prediction accuracy of BM5P model (RMSE=0.3794, MAE=0.3498 for testing and RMSE=0.3060, MAE=0.2537 for validation) is found higher than MLP- and M5P-based models.

3.3 Models for Mn (mg/kg)

3.3.1 Dataset

The preparation of training, testing and validation dataset is like Fe and Cu dataset. Descriptive statistics of total dataset for Mn (mg/kg) ($n=164$) are represented in Table 2, respectively, in which pH, organic carbon (OC), clay and Ava. P are input parameters and Mn is the output. Descriptive statistics of datasets utilized for training ($n=117$), testing ($n=24$) and validation ($n=23$) for Mn (mg/kg) are listed

Fig. 7 Performance of MLP, M5P and BM5P for estimating Cu **Fig. 8** 3D surface plot of Mn (mg/kg) against OC (%) and clay (%)

in Table 3. Figure 8 indicates the 3D surface plot of Mn (mg/kg) against clay (%) and OC (%).

3.3.2 Assessment of model's performance in Mn (mg/kg) estimation in soil

Performance evaluation parameters considered for model's performance evaluation are same as for implementing Fe and Cu models' evaluation. Figure 9 shows the scatter plot between actual and predicted Mn (mg/kg) in soil using MLP-, M5P- and Bagged M5P-based models for training, testing and validation stages. The prediction accuracy of MLP model (RMSE=2.7447, MAE=2.0795 for testing and RMSE=3.9365, MAE=2.9342 for validation) is found higher than M5P and BM5P models. Table 4 indicates that MLP model is most suitable model than M5P- and BM5P-based models for the estimation of Mn in soil for this dataset.

Fig. 9 Performance of MLP, M5P and BM5P for estimating Mn

3.4 Models for Zn (mg/kg)

3.4.1 Dataset

The whole dataset containing 164 observations from field was divided into three parts such as training, testing and validation, respectively. Training data involve 70% of the total data chosen randomly from whole dataset, while testing and validation data involve remaining 15% and 15% of the total dataset. Descriptive statistics of total dataset for Zn (mg/kg) ($n=164$) are given in Table 2, and descriptive statistics of datasets utilized for training ($n=117$), testing ($n=24$) and validation ($n=23$) for Zn (mg/kg) are represented in Table 3 in which sand, clay, silt and Ava. P are input parameters and Zn is the target. Figure 10 indicates the 3D surface plot of Zn (mg/kg) against sand (%) and silt (%).

3.4.2 Assessment of model's performance in Zn (mg/kg) estimation in soil

To compare the performance of models, CC, RMSE, NSE, d and MAE values were considered as performance evaluation parameters. Figure 11 illustrates the scatter plot between actual and predicted Zn (mg/kg) in soil using MLP-, M5P- and Bagged M5P-based models for training, testing and validation stages. Few predicted values of Zn by MLP are negative. MLP model is not suitable for the estimation of Zn in soil. The prediction accuracy of M5P model (RMSE=1.7513, MAE=1.2464 for testing and RMSE=2.1154, MAE=1.1885 for validation) is found higher than BM5P model (RMSE=1.6979, MAE=1.2230 for testing and RMSE=2.2140, MAE=1.3727 for validation). Table 4 indicates that M5P model is most suitable model than MLP and BM5P.

Fig. 10 3D surface plot of Zn (mg/kg) against sand (%) and silt (%)

Fig. 11 Performance of MLP, M5P and BM5P for estimating Zn

4 Discussion

Generally, a quick and suitable method for estimating heavy metal contents in soil is illustrated in this investigation. Actually, experimentally investigation of heavy metal contents in soil is very expensive, hard, laborious and time-consuming, whereas computer-based methods are easier and less time-consuming. These methods provide a valid tool for estimating heavy metal contents in soil near mining and suburban regions, thus facilitating the management and assignment of human settlements and other natural resources. Previous studies have reported that statistical methods and artificial neural networks can be successfully applied for estimating various other heavy metal contents in soils in various regions [34–37]. In this study, four popular and dominant soft computing techniques were used for the estimation of heavy metals in soil. However, only four heavy metals were investigated in the present work. The influences of other heavy metals on the prediction accuracy were not taken into consideration.

From the results, it was observed that MLP-based models worked better than M5P- and BM5P-based models for the prediction of both Fe and Mn. Figure 4 suggests that estimation values of Cu using BM5P are close to line of perfect agreement than the values of other modeling (independent variables: organic carbon (OC), sand and silt). Figure 11 indicates that M5P models work superior than other approaches for the estimation of Zn with better standard statistical values (CC, RMSE, d , NSE, MAE). MLP model predicts some negative values in estimation of Zn metal. Overall performance of discussed approaches is satisfactory for the estimation of heavy metals in soil.

5 Conclusions

The prediction of soil pollution examines a significant field of study with regard to the overall concern of environmental protection issues. Determination of models that can accurately estimate the heavy metals in soils is an important need of agricultural researches, which could eliminate the weaknesses in the measurement of heavy metals in soils. In this study, using MLP, M5P and BM5P models, heavy metals (Zn, Cu, Mn and Fe) were estimated and modeled for arid region of Iran. In this regard, different models have been developed for different metals using the above-discussed machine learning techniques. Finally, the accuracy of each model was investigated in terms of performance evaluation parameters. The experimental data collected from field samples were used in this study. The comparison of the performance of these models in predicting the heavy metals in soils showed that M5P model predicts the Zn content in soil and that BM5P model predicts the Cu content in soil with a much higher accuracy and less error than the other models. Outcomes of performance assessment parameters suggested that MLP model performed better than other models for predicting the Mn and Fe contents in soils. MLP Mn model included the combination of pH, organic carbon (OC), clay, Ava. P and MLP Fe model included the combination of pH, sand, silt and Ava. P in this study.

Compliance with ethical standards

Conflict of interest Authors declared that there is no conflict of interest.

References

1. Dogra N, Sharma M, Sharma A, Keshavarzi A, Minakshi, Bhardwaj R, Thukral AK, Kumar V (2019) Pollution assessment and spatial

- distribution of roadside agricultural soils: a case study from India. *Int J Environ Health Res* 18:1–4
2. Keshavarzi A, Kumar V (2019) Spatial distribution and potential ecological risk assessment of heavy metals in agricultural soils of Northeastern, Iran. *Geol Ecol Landsc*. <https://doi.org/10.1080/24749508.2019.1587588>
  3. Kumar V, Sharma A, Kaur P, Kumar R, Keshavarzi A, Bhardwaj R, Thukral AK (2019) Assessment of soil properties from catchment areas of Ravi and Beas rivers: a review. *Geol Ecol Landsc* 3(2):149–157
  4. Keshavarzi A, Kumar V (2018) Ecological risk assessment and source apportionment of heavy metal contamination in agricultural soils of Northeastern Iran. *Int J Environ Health Res* 11:1–7
  5. Kumar V, Sharma A, Bhardwaj R, Thukral AK (2016) Assessment of soil enzyme activities based on soil samples from the Beas river bed, India using multivariate techniques. *Malay J Soil Sci* 20:135–145
  6. He ZL, Yang XE, Stoffella PJ (2005) Trace elements in agroecosystems and impacts on the environment. *J Trace Elem Med Biol* 19(2–3):125–140
  7. Sun Y, Zhou Q, Xie X, Liu R (2010) Spatial, sources and risk assessment of heavy metal contamination of urban soils in typical regions of Shenyang, China. *J Hazard Mater* 174(1–3):455–462
  8. Moor C, Lymberopoulou T, Dietrich VJ (2001) Determination of heavy metals in soils, sediments and geological materials by ICP-AES and ICP-MS. *Microchim Acta* 136(3–4):123–128
  9. Kumar V, Sharma A, Kaur P, Sidhu GP, Bali AS, Bhardwaj R, Thukral AK, Cerda A (2019) Pollution assessment of heavy metals in soils of India and ecological risk assessment: a state-of-the-art. *Chemosphere*. <https://doi.org/10.1016/j.chemosphere.2018.10.066>
  10. Zhang XY, Lin FF, Wong MT, Feng XL, Wang K (2009) Identification of soil heavy metal sources from anthropogenic activities and pollution assessment of Fuyang County, China. *Environ Monit Assess* 154(1–4):439–449
  11. Lin YP, Cheng BY, Shyu GS, Chang TK (2010) Combining a finite mixture distribution model with indicator kriging to delineate and map the spatial patterns of soil heavy metal pollution in Chunghua County, central Taiwan. *Environ Pollut* 158(1):235–244
  12. Sidhu GP, Singh HP, Batish DR, Kohli RK (2017) Tolerance and hyperaccumulation of cadmium by a wild, unpalatable herb *Coronopus didymus* (L.) Sm. (Brassicaceae). *Ecotoxicol Environ Saf* 135:209–215
  13. Sidhu GP, Bali AS, Singh HP, Batish DR, Kohli RK (2018) Phytoremediation of lead by a wild, non-edible Pb accumulator *Coronopus didymus* (L.) Brassicaceae. *Int J Phytoremed* 20(5):483–489
  14. Sidhu GP, Bali AS, Singh HP, Batish DR, Kohli RK (2018) Ethylenediamine disuccinic acid enhanced phytoextraction of nickel from contaminated soils using *Coronopus didymus* (L.) Sm. *Chemosphere* 205:234–243
  15. Khan S, Cao Q, Zheng YM, Huang YZ, Zhu YG (2008) Health risks of heavy metals in contaminated soils and food crops irrigated with wastewater in Beijing, China. *Environ Pollut* 152(3):686–692
  16. Hu Y, Cheng H (2013) Application of stochastic models in identification and apportionment of heavy metal pollution sources in the surface soils of a large-scale region. *Environ Sci Technol* 47(8):3752–3760
  17. Zovko M, Romic M (2011) Soil contamination by trace metals: geochemical behaviour as an element of risk assessment. In: *Earth and environmental sciences*. IntechOpen, pp 437–456
  18. Wang Q, Xie Z, Li F (2015) Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale. *Environ Pollut* 206:227–235
  19. McBratney AB, Santos MM, Minasny B (2003) On digital soil mapping. *Geoderma* 117(1–2):3–52
  20. Arun PV, Katiyar SK (2013) An evolutionary computing framework toward object extraction from satellite images. *Egypt J Remote Sens Space Sci* 16(2):163–169
  21. Scull P, Franklin J, Chadwick OA (2005) The application of classification tree analysis to soil type prediction in a desert landscape. *Ecol Model* 181(1):1–5
  22. Thomas GW (1996) Soil pH and soil acidity. In: Page AL (ed) *Methods of soil analysis: part 2. Agronomy handbook 9*. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp 475–490
  23. Gee GW, Bauder JW (1986) Particle size analysis. In: Klute A (ed) *Methods of soil analysis: part 1. Agronomy handbook 9*. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp 383–411
  24. Walkley A, Black IA (1934) An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. *Soil Sci* 37:29–38
  25. Kumar V, Sharma A, Minakshi Bhardwaj R, Thukral AK (2018) Temporal distribution, source apportionment, and pollution assessment of metals in the sediments of Beas river, India. *Hum Ecol Risk Assess* 24(8):2162–2181
  26. Xiao R, Bai J, Huang L, Zhang H, Cui B, Liu X (2013) Distribution and pollution, toxicity and risk assessment of heavy metals in sediments from urban and rural rivers of the Pearl River delta in southern China. *Ecotoxicology* 22(10):1564–1575
  27. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 5(4):115–133
  28. Schalkoff RJ (1997) *Artificial neural networks*, vol 1. McGraw-Hill, New York
  29. Firat M, Gungor M (2009) Generalized regression neural networks and feed forward neural networks for prediction of scour depth around bridge piers. *Adv Eng Softw* 40(8):731–737
  30. Ghorbani MA, Zadeh HA, Isazadeh M, Terzi O (2016) A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. *Environ Earth Sci* 75(6):476
  31. Quinlan JR (1992) Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence, vol 92, pp 343–348
  32. Breiman L (1996) Bagging predictors. *Mach Learn* 24(2):123–140
  33. Breiman L (1999) Using adaptive bagging to debias regressions. Technical report 547, Statistics Department, UCB
  34. Kemper T, Sommer S (2002) Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. *Environ Sci Technol* 36(12):2742–2747
  35. Alizamir M, Sobhanardakani S, Taghavi L (2017) Modeling of groundwater resources heavy metals concentration using soft computing methods: application of different types of artificial neural networks. *J Chem Health Risks* 7(3):68–77
  36. Alizamir M, Sobhanardakani S (2017) A comparison of performance of artificial neural networks for prediction of heavy metals concentration in groundwater resources of toyserkan plain. *Avicenna J Environ Health Eng* 4(1):11792
  37. Alizamir M, Sobhanardakani S (2017) Predicting arsenic and heavy metals contamination in groundwater resources of Gha-havand plain based on an artificial neural network optimized by imperialist competitive algorithm. *Environ Health Eng Manag J* 4:225–231

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.