Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-018-3458-0

RESEARCH ARTICLE

A novel method for predicting cadmium concentration in rice grain using genetic algorithm and back-propagation neural network based on soil properties

Yi Xuan Hou[1] · Hua Fu Zhao[1,2] · Zhuo Zhang[1,2] · Ke Ning Wu[1,2]

Received: 15 June 2018 / Accepted: 12 October 2018

Abstract

Heavy metal pollution is a global ecological safety issue, especially in crops, where it directly threatens regional ecological security and human health. In this study, the back-propagation (BP) neural network optimized by the genetic algorithm (GA) was used to predict the concentration of cadmium (Cd) in rice grain based on influencing factors. As an intelligent information processing system, the GA-BP neural network could learn the laws of Cd movement in the soil-crop system through its own training and use the soil properties to predict the concentration of Cd in grain with high accuracy. The total soil Cd concentration, clay content, Ni concentration, cation exchange capacity (CEC), organic matter (OM), and pH have important impacts and interactions on Cd concentration in rice grain were selected as input factors of the prediction model based on Pearson's correlation analysis and GeoDetector. By using GA to optimize the initial weight, the prediction accuracy of the GA-BP neural network model was optimal compared with the BP neural network model and multiple regression analysis. Based on the Cd concentration predicted in grain by the model, human exposure and health risk can be assessed quickly, enabling measures to be taken in time to reduce the transfer of Cd from soil to the food chain.

Keywords Prediction model · Cadmium · Rice grain · Soil-rice system · GA-BP neural network · Soil properties

Introduction

Heavy metal pollution in agricultural soils has become a problem in many parts of the world because of their toxicity, persistence, and concealment (Li et al. 2014b; Ran et al. 2016; Yan et al. 2013). Although Cd is unnecessary for the growth and production of crops, it is readily absorbed by plants, after which it induces adverse effects such as slow growth, stunted growth, and reduced yield (Rizwan et al. 2016a; Rizwan et al. 2016b). Furthermore, the transfer of Cd from soils into the food chain via soil-crop systems has been considered the predominant exposure pathway for humans (Qian et al. 2010). Therefore, it is necessary to investigate the transfer and accumulation of Cd from soil to crops, especially when predicting Cd concentration in crops (Lu et al. 2017; Novotna et al. 2015; Ye et al. 2014).

As a reasonable tool to estimate potential dietary risks, crop heavy metal prediction models need to be constructed using reliable methods (Novotna et al. 2015). The mechanistic model, which is one of the most frequently used models, is established based on the mechanisms of heavy metal transport in soil-crop systems and commonly considers factors including soil properties, crop characteristics, and atmospheric environment (Rein et al. 2011; Sterckeman et al. 2004). Mechanistic models have been widely researched and include the CLEA (Contaminated Land Exposure Assessment) model developed by the UK Environment Agency (Martin and Jeffries 2008), Hough's FIAM (Free Ion Activity Model) (Hough et al. 2005), Legind's NMF (New Model Framework) (Legind and Trapp 2010), Rein's dynamic model (Rein et al. 2011), the Barber-

Responsible editor: Marcus Schulz

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11356-018-3458-0) contains supplementary material, which is available to authorized users.

✉ Hua Fu Zhao
huafuzhao@163.com

[1] School of Land Science and Technology, China University of Geosciences (Beijing), 29 Xueyuan Road, Beijing 100083, China

[2] Key Laboratory of Land Consolidation and Rehabilitation, Ministry of Land and Resources, 37 Guanyingyuanxi District, Beijing 100035, China

Published online: 24 October 2018

Cushman model (Barber and Cushman 1981; Sterckeman et al. 2004), and Ingwersen's process-oriented model (Ingwersen and Streck 2005). However, data describing critical parameters in models are difficult to obtain, resulting in limited application of these models (Legind and Trapp 2010; Ye et al. 2014). When compared to the mechanistic models, empirical models have a relatively simple structure and require fewer input data, making them particularly suitable to large-scale and practical applications (Ye et al. 2014). Empirical models are mostly in the form of multiple regression, in which the parameters evaluated are commonly soil properties (Novotna et al. 2015). The majority of models contain the total concentration of heavy metals in soil and soil pH, which are both important factors influencing crop absorption of heavy metals (Adams et al. 2004; McBride 2002; Ran et al. 2016; Tudoreanu and Phillips 2004; Wang 2002). Other important factors such as OM, clay content, CEC, and the concentration of other heavy metals in soil are also considered in many models (Bester et al. 2013; Chaudri et al. 2007; Chen et al. 2016; Lu et al. 2017; Novotna et al. 2015; Romkens et al. 2009; Ye et al. 2014). Although crop heavy metal prediction models have been widely researched and applied, there is still room for improvement in the prediction accuracy of these models.

The BP neural network, which is one of the most extensively used artificial neural network (ANN) models, consists of a multi-layer network that uses a gradient descent-based algorithm for weight training (Zhang et al. 2017). As an intelligent information processing system, the BP neural network can approximate any complex nonlinear function with high accuracy by learning and simulating some sort of algorithm or function in nature (Zhang et al. 2016). Because of its strong fitting ability, the BP neural network is suitable for application of internal complex mechanisms such as the migration of heavy metals in soil-plant systems. However, few studies have investigated the use of the BP neural network to predict the heavy metal concentrations in crops to date. Some studies have used remote sensing data as input data and achieved a relatively high accuracy (Jiang et al. 2016; Liu et al. 2011). The GA is a powerful stochastic algorithm that has wide applicability in optimization problems. It can be used to search the global minimum based on natural selection and genetic mechanisms (Tongle et al. 2016). Therefore, the GA can overcome the weaknesses of the BP neural network in areas that easily fall into local minima and improve the training speed and forecast accuracy of the network (Pang and Shi 2008). The combination of GA and BP neural network has been applied in many fields and achieved good predictions (Shen et al. 2007; Wang et al. 2016). However, there was no research that applied the GA-BP neural network in the prediction of heavy metal concentrations in crops. Here, we present a hypothesis that the application of GA-BP neural network to predict the Cd concentration in rice grain will achieve high prediction accuracy.

The Yangtze River Delta (YRD), one of the most economically developed regions in China, has accelerated the accumulation of heavy metals in soils because of its urbanization and industrialization (Chen et al. 2016; Chen 2007). As the main agricultural product in the YRD, rice has great ability to be enriched by Cd in soil, posing a health hazard to consumers of rice and local residents. Therefore, a thorough understanding of the factors influencing the concentration of Cd absorbed by rice and accurate prediction of Cd in grains is important for human exposure and health risk assessment, as well as taking measures to reduce the transfer of Cd from soil to the food chain (Hough et al. 2003; Legind and Trapp 2010; Ye et al. 2014; Zhao et al. 2010). In this study, we used the GA-BP neural network to predict the Cd concentration in rice grain based on influencing factors selected by correlation analysis and GeoDetector. The specific objectives of this study include (1) identification of the main factors influencing Cd concentration in rice grain for model construction, (2) development of a GA-BP neural network model for predicting the concentration of Cd in rice grain based on selected factors, and (3) development of a BP neural network model and multiple regression model for comparing the prediction accuracy of the models.

Materials and methods

Data preparation

This study was conducted in a typical industrial county in the city of Yixing, Jiangsu Province (N31.35°, E119.82°) (Fig. 1), which is located in the YRD and characterized by a warm and moist subtropical climate. The study area is a commercial grain base and rice is the main crop at the local level. However, wastes from industrial activities have brought excessive pressure on the environment and caused serious heavy metal pollution of agricultural land.

A total of 45 pairs of soil samples and corresponding rice grain samples were collected during November of 2015. Rice grains were sampled at the same time from the same locations as the soil samples. Sampling points were selected from soil contamination areas identified in an investigation of ecological geochemistry quality for cultivated land conducted in 2007. Figure 1 illustrates the locations from which soil and rice grain samples were collected. To ensure the accuracy of the measurement, three sub-samples were taken from each sample point. Samples were collected from a depth of 0–20 cm, after which debris and gravel in the soil were removed. A five-point sampling method was adopted to take soil from each sub-sample. The quarter-point method was used to retain 1 kg of soil samples into polyethylene bags. During the collection of rice grain samples, grains that were empty and or showed signs of insect diseases were avoided. All samples were sealed in sample bags, then air-dried in a pollution free place for future analysis.

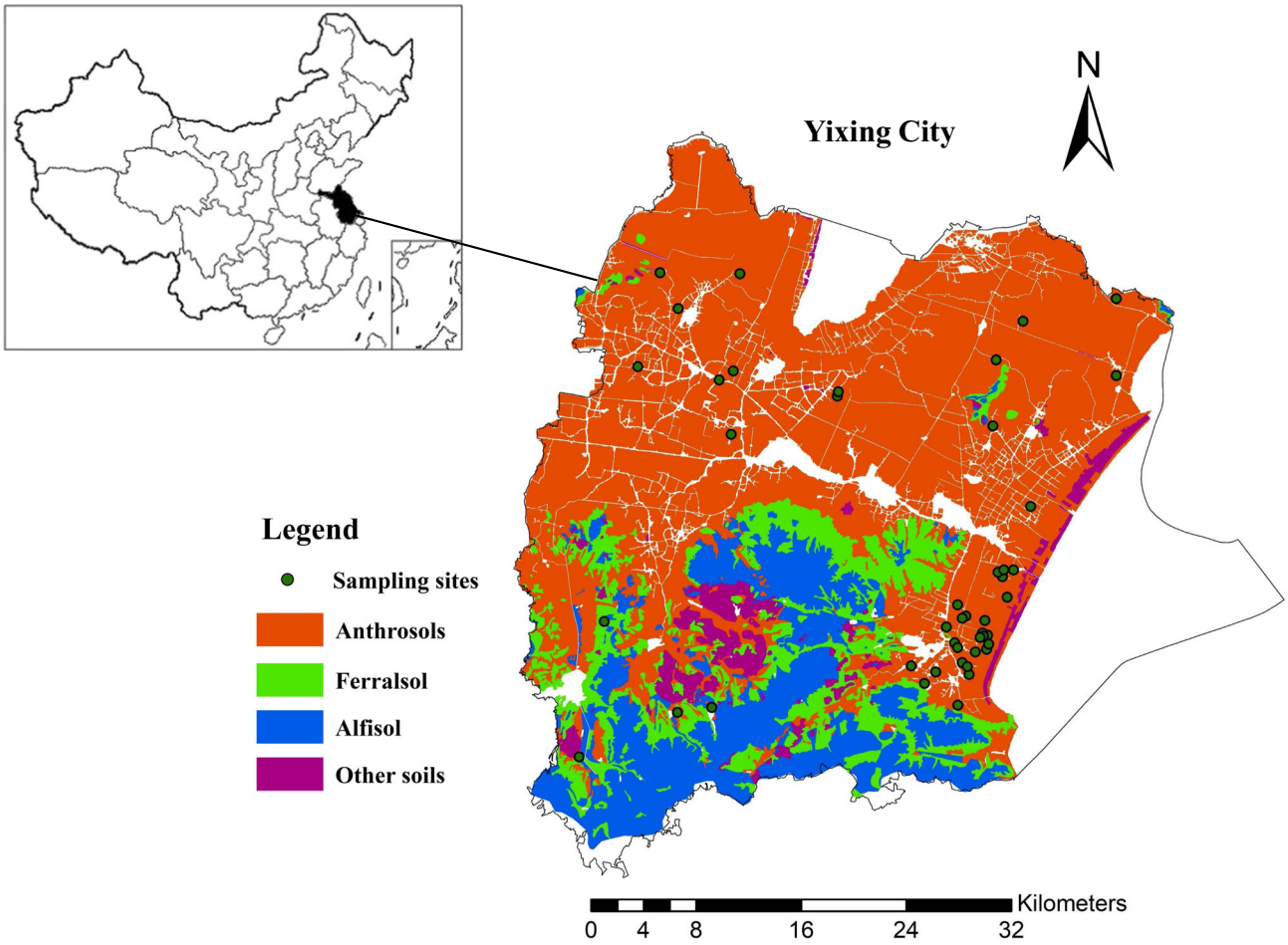

Fig. 1 Location of sampling sites and distribution of soil types in the study area

The pH, CEC, OM, clay content, heavy metal concentrations in soil, and Cd concentration in rice grain were determined. The Cd and Pb concentrations were determined by graphite furnace atomic absorption spectrometry (Optima 2100DV, Perkin Elmer, USA), while the Ni, Cu, and Zn concentrations in soil were determined by flame atomic absorption spectrometry (Optima 2100DV, Perkin Elmer, USA), the Hg and As concentration were determined by atomic fluorescence spectrometry (Primus-II, Rigaku Corporation, Japan), and the Cr concentration was determined by inductively coupled plasma atomic emission spectroscopy (Optima 2100DV, Perkin Elmer, USA). Soil pH was determined using an ion-selective electrode, while OM was measured by wet oxidation using $K_2Cr_2O_7$, and the CEC was determined with the ammonium acetate method using CH_3COONH_4 leaching. The clay content was measured by pipette and sieve analysis.

Correlation analysis

The SPSS 20.0 statistical package was used to analyze data. Specifically, Pearson's correlation coefficient was calculated to determine the relationships between different variables, while two-way analysis of variance (ANOVA) was conducted to identify differences among groups. Relationships were considered significant at $P < 0.05$ and $P < 0.01$.

Interaction analysis

Generally, the process of rice uptake of Cd from soil is influenced by many factors; therefore, the mechanism through which this process occurs is complicated and it is difficult to estimate the independent effects of factors (Dayton et al. 2006; Li et al. 2014b). As a result, interactions among factors should be considered. Geodetector has been widely used in a variety of fields to analyze the forces driving various phenomena as well as multi-factor interactions (Wang et al. 2010). This system employs a statistical method to detect spatial variability and reveal the driving forces behind this variability based on risk, factors, interactions and ecological detectors (Wang et al. 2010). Detecting the power determinant of a factor influencing Cd content in rice is mainly accomplished by comparing the total variance of each factor in different subareas with the total variance of crop Cd content in the entire study

area. The smaller the ratio, the stronger the determinant power. The general expression is

$$q = 1 - \frac{\sum_{k=1}^{M} N_k \sigma_k^2}{N\sigma^2}$$

where q indicates how much of the Cd concentration in rice grain is interpreted by the factor. $q \in [0, 1]$, where $q = 0$ indicates that there is no relationship between the Cd concentration in rice grain and the factor, and $q = 1$ indicates that the Cd concentration in rice grain is completely determined by the factor. $k = 1,\ldots, M$ indicates the number of strata in a factor, N_k and N are the number of the strata k and units of the Cd content in rice grain, respectively. σ_k^2 and σ^2 reflect the variance of the strata k and Cd concentration in the rice grain, respectively. The interaction detector can deal with the issue of whether two factors together have a stronger or weaker effect on the concentration of Cd in grain than they do independently. This index can quantify the interactive effect of two factors by stacking two layers (X1, X2) to form a new layer (X1 ∩ X2). The attribute of the new layer is determined by a combination of those two layers (Hu et al. 2011). By comparing the q-value of X1, X2, and X1 ∩ X2, the influence of the interaction between two soil factors on Cd concentration in grain can be determined (Huang et al. 2014).

Construction of GA-BP neural network

Data pre-processing phase

To eliminate the impact of different platforms on the results and improve the efficiency of network training, the samples data were normalized to [− 1, 1] using the following formula:

$$x_{norm} = (a-b) \times (x - x_{min})/(x_{max} - x_{min}) + b$$

where x, x_{norm}, x_{max}, and x_{min} are the actual value, normalized value, maximum value, and minimum value of the sample, respectively, and a and b are the maximum value and minimum value of the normalized interval, respectively.

Model building phase

To obtain high-precision prediction of the concentration of Cd in rice grain, a BP neural network model was established. The statistical learning algorithms of the network were inspired by the biological neural networks (Hooyberghs et al. 2005). The structure of the BP neural network consists of an input layer, hidden layer and output layer, which includes multiple neurons, and each layer is linked by connection weights. With the initial weights randomly set, neural networks are trained to modify all weights based on the back propagation algorithm until the errors between output data and desired data are within a predetermined range. When the weights among layers are decided, the neural network is determined and can be used for forecasting. The BP neural network has a strong nonlinear mapping ability and can achieve the final convergence of the network through training. However, it has weaknesses such as being prone to jumping into the local minima, which leads to training of the network being more sensitive to the initial network weights (Tongle et al. 2016). The GA can solve these problems effectively and is therefore widely applied in neural networks. Because the GA has strong global search ability in a complex, polymorphic, and discontinuous space, it can help optimize the structure and parameters of neural networks (Yu and Xu 2014). The use of GA to optimize the initial weights of the BP neural network can exert the advantages of the global search of GA and overcome the disadvantages of the BP neural network (Zheng 2017). The weights and thresholds of the network are initialized randomly and are connected and coded into individuals in order. The individual lengths are determined according to the network structure parameters, and n individuals are randomly generated within a given range to make up the initial population. The fitness function is taken as the reciprocal of the error in the neural network. After repeated selections, crossovers, and mutations, individuals with lower fitness are eliminated and a new population is obtained. The optimal initial weights and thresholds of the neural network are then obtained through decoding of individuals. The process of the GA-BP model is shown in Fig. 2.

Where d is the length of the individual; r is the number of input layer nodes; s_1 is the number of hidden layer nodes; s_2 is the number of output layer nodes; E is output error value; f is fitness function; u is the size of the population; f_i is the fitness value of the i^{th} individual, and P_i is the probability of being selected; $z_i(k)$ and $z_{i+1}(k)$ denote the k^{th} gene of the i^{th} and $i + 1^{th}$ individuals, respectively; α and β are random numbers between 0 and 1; and q is the threshold width corresponding to the $p + 1$st gene value. n represents the input layer nodes; l represents the hidden nodes; m represents the output layer nodes; Y_k is the predictive value; O_k is the actual value; e_k is the model error; ω_{ij} is the weight between the input layer and the hidden layer; ω_{jk} is the weight between the hidden layer and the output layer; H_j is the hidden layers value; a_j is the threshold between the input layer and the hidden layer; b_k is the threshold between the hidden layer and the output layer.

Model evaluation phase

The root mean square error (RMSE), mean absolute error (MAE), mean relative error (MRE), and coefficient of

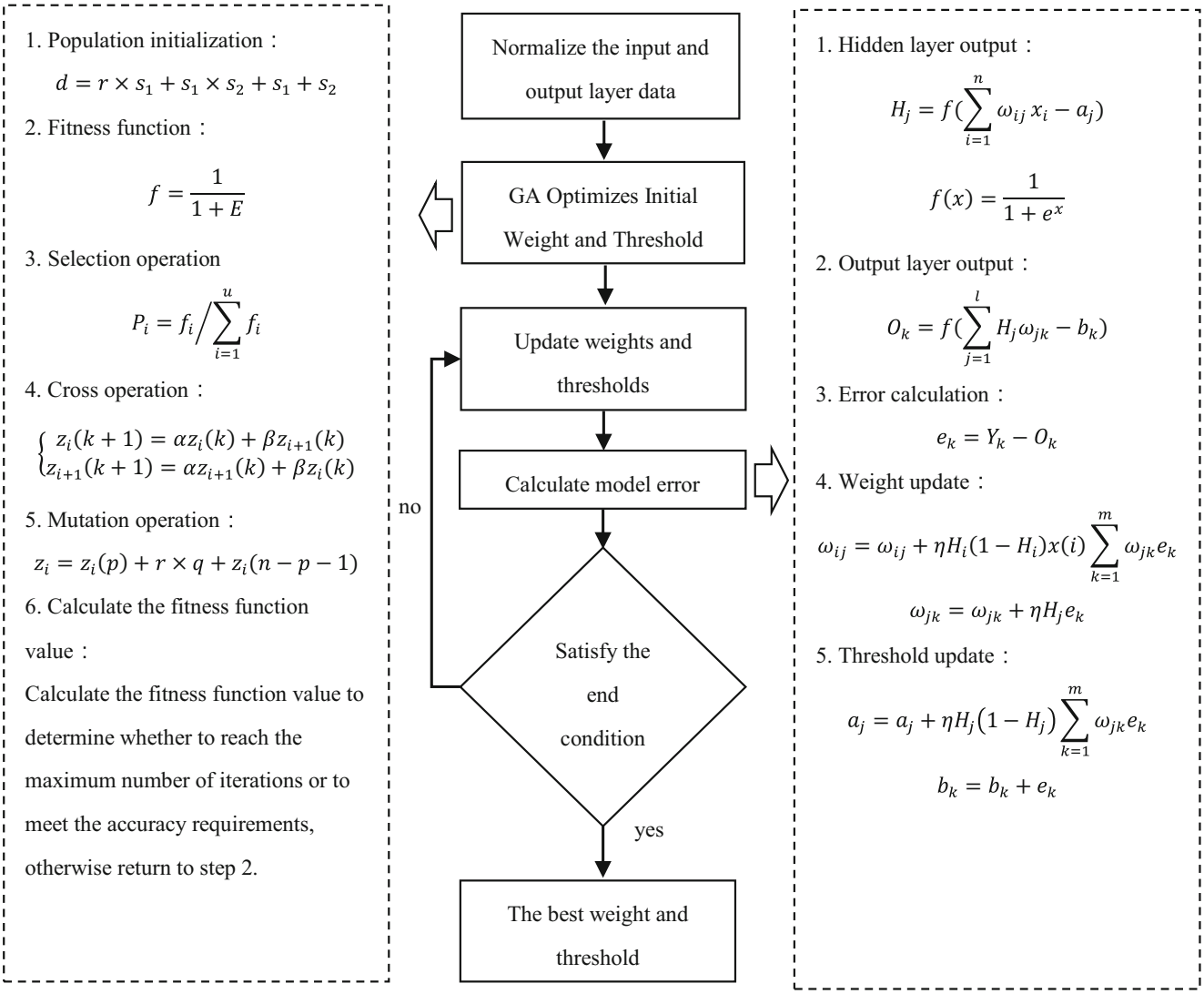

Fig. 2 Flowchart of the GA-BP neural network model

determination (R^2) are used as metrics to assess the performance of different models:

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(X_i - \overline{X_i}\right)\sum_{i=1}^{n}\left(Y_i - \overline{Y_i}\right)\right]^2}{\sum_{i=1}^{n}\left(X_i - \overline{X_i}\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y_i}\right)^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{n}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i|$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|X_i - Y_i|}{X_i} \times 100\%$$

where X_i and Y_i denote the measured and forecasted values of Cd concentration in grain and n is the number of groups. A larger R^2 is associated with a smaller RMSE, MAE, and MRE, representing higher accuracy of the model.

Results and discussion

Main factors influencing Cd concentration in rice grain

The key to establishing an accurate prediction model is selecting appropriate factors as input factors. The mechanism of Cd transfer and accumulation in soil-crop systems is very complex. Absorption of Cd by rice is influenced by the chemical form of Cd in the soil. Based on the results of previous studies, soil physicochemical properties such as pH, clay content, OM, and other heavy metals can change the chemical form of Cd in soil to affect the amount of Cd absorbed by rice,

and the Cd concentration in rice is affected by joint action of these factors (Wang et al. 2017). Therefore, the factors that are highly correlated with Cd concentration in rice grain and great impact on the process by which rice absorbs Cd from soil should be selected. Correlative analysis showed that the Cd concentration in rice was positively correlated with total Cd concentration in soil, negatively correlated with clay content > total Ni concentration in soil > CEC > OM, and uncorrelated with soil pH.

The total Cd concentration in soil has a large effect on Cd uptake (McBride 2002; Novotna et al. 2015). As shown in Table 1, the correlation coefficient between Cd concentration in soil and in rice grain was 0.651, indicating that this is an optimal factor to determine the Cd concentration in rice.

The OM is considered an important factor influencing the availability of heavy metals in soils (Bester et al. 2013; Chaudri et al. 2007; Chen et al. 2016). As shown in Table 1, OM was significantly correlated with Cd concentration in grain, with a correlation coefficient of − 0.301. Studies conducted to investigate the sorption mechanisms of OM onto Cd have shown that the main mechanism is binding of functional groups of OM with Cd ions via strong chemical bonds to form stable complexes that are difficult to desorb, thereby reducing the biological activity of Cd in soil and inhibiting the absorption of rice (Bradl 2004; Guo et al. 2006; Weng et al. 2002).

Clay content was significantly correlated with Cd concentration in grain, with a correlation coefficient of − 0.495. Clay is an important adsorbent for Cd in soils in addition to OM (Hooda and Alloway 1998). Studies have shown that clay minerals adsorb Cd ions in soil solution through their unique and negatively charged surfaces, then exchange their ions, thereby affecting the exchangeable content of Cd and reducing its bioavailability (Bradl 2004; Song et al. 1999).

As shown in Table 1, the correlation coefficient between CEC and Cd concentration in rice grain was found to be − 0.338, indicating a significant negative correlation. The CEC is a measure of the total amount of negative surface charges, which are primarily contributed by clay and organic fractions (Hooda and Alloway 1998). Studies have shown that as the CEC increases, the negative charge of soil becomes greater, providing more adsorption sites to fix Cd ions, thereby reducing the availability of Cd in soil and the absorption of Cd by rice (Bradl 2004; Hooda and Alloway 1998).

Soil pH is another important factor that determines the bioavailability of Cd in soil (Chen et al. 2016; Li et al. 2014b; Ye et al. 2014; Zhao et al. 2010). The Cd concentration in rice grain was expected to be greatly affected by soil pH (Chen et al. 2016; Romkens et al. 2009). However, there was no significant correlation between pH and Cd concentration in grain observed in our data. These findings indicate that soil properties other than pH, such as Cd concentration and OM, had a strong influence on the available Cd in soil and therefore the amount of Cd absorbed by rice grains. Similarly, Wiggenhauser et al. (2016) found that there was no significant correlation between soil pH and available Cd in a soil-wheat system (Wiggenhauser et al. 2016).

In addition to the above factors, other heavy metals in soil can also affect Cd absorption by rice. Specifically, we found that the total Ni concentration in soil was significantly correlated with Cd in rice grain, with a correlation coefficient of − 0.433. There is an antagonistic effect between Cd and Ni. Pot experiments showed that Ni treatment reduced Cd concentration in rice grain (Bingham et al. 1980), and that the uptake of Cd by seedlings growing on medium containing both Cd and Ni was significantly lower than that observed by seedlings grown on medium containing only Cd (Artiushenko et al. 2014). However, we did not find a significant correlation between Cd concentration in grain and other heavy metals in soil. This may have occurred because the concentrations of Zn, Pb, As, and Cu in soil were below the thresholds in accordance with the Chinese environmental quality standard for soils (GB 15618–1995).

Interaction among influence factors

The interaction detector of the GDM was used to determine if the soil factors had interactive effects on Cd concentration in rice grain. The symbol $q(X_1 \cap X_2)$ denotes the determinant power of interaction of new layers that comprised a combination of X1 and X2. As shown in Table 3, the interaction relationship is

defined in the coordinate axis, which has five intervals, and the interaction relationship is determined by the location of $q(X_1 \cap X_2)$ in the five intervals (Wang et al. 2010). The results show that all of the interactive values were higher than the q value of sole factors (Table 2). Moreover, most of those combinations consisted of nonlinear enhancement interactions, while few had bi-enhance interactions (Table 3). In particular, the interaction between Cd and Ni showed the highest power of determinant for Cd content of rice, reaching 0.889.

Table 1 Correlation between Cd concentration in rice grain and soil factors

Soil factor	Total Cd concentration in soil	Clay content	Total Ni concentration in soil	CEC	OM	pH
Correlation coefficient	0.651**	− 0.495**	− 0.433**	− 0.338*	− 0.301*	0.231

*$p < 0.05$; ** $p < 0.01$

Table 2 q-values for interactions between pairs of soil factors influencing the Cd concentration of grain

	Total Cd concentration in soil	Clay content	Total Ni concentration in soil	CEC	OM	pH
Total Cd concentration in soil	0.400	–	–	–	–	–
Clay content	0.639	0.326	–	–	–	–
Total Ni concentration in soil	0.889	0.746	0.222	–	–	–
CEC	0.671	0.587	0.594	0.234	–	–
OM	0.852	0.695	0.749	0.364	0.148	–
pH	0.575	0.586	0.753	0.513	0.395	0.180

GA-BP neural network prediction and comparison with other models

To establish and validate the neural network model, we used 36 groups of samples as a training set and nine groups of samples as a test set for the model. Matlab software was used to simulate and compute (Li et al. 2014a; Peng 2013). The GA was used to optimize and assign initial weights and thresholds of the BP neural network, while the BP neural network was used to search for the local optimization values. By trying different parameters, architectures, and training algorithms, satisfactory GA-BP neural network models were developed.

In the experiments, 50, 100, 0.5, and 0.09 were set as the population size, the maximum genetic algebra, the crossover rate, and the variation rate, respectively. Tansig functions were used as activation functions for hidden neurons and output neurons. The Levenberg-Marquardt back-propagation was used as a training function to ensure the rapid convergence of the network. The most suitable model structure was found to be 6-11-1 based on the experiments. Finally, the optimal model with the lowest RMSE, errors, and highest R^2 value was obtained.

To assess and compare the performance of GA-BP neural network models, a BP neural network model and a multiple regression model were established. The equation for multiple regression was $\lg Cd_{rice} = 0.768 \lg Cd_{soil} - 1.760 \lg SOM - 0.142 \lg pH + 0.663 \lg CEC - 1.049 \lg Clay - 2.073 \lg Ni + 3.890$.

The RMSE, MSE, and MAE values of the training and test sets were used to compare the credibility and stability of the models. The results suggested the application of GA-BP neural network in predicting Cd concentration in rice grain achieved relatively high prediction accuracy. As shown in Table 4, the accuracies of the GA-BP neural network and BP neural network models based on the six factors were better than those of the multiple regression model. Furthermore, the accuracy of the GA-BP neural network model was slightly better than that of the BP neural network model. The errors of train sets of the GA-BP neural network model and the BP neural network model had reached a relatively low level, and R^2 had reached a relatively high level. However, the RMSE and MAE of the test set of GA-BP neural network are 0.040 and 0.023, respectively, which are much lower than the 0.117 and 0.077 of the BP neural network. The MRE of the GA-BP neural network and the BP neural network are approximately equal. The R^2 of test set of GA-BP

Table 3 Types of interaction between two factors

Graphical representation	Description	Interaction type	Interaction factors
	q(X1 ∩ X2) > q(X1)+ q(X2)	Enhance, nonlinear	Cd concentration ∩ OM; Cd concentration ∩ CEC; Cd concentration ∩ Ni concentration; OM ∩ pH; OM ∩ clay; OM ∩ Ni concentration; pH ∩ CEC; pH ∩ clay; pH ∩ Ni concentration ; CEC ∩ Ni concentration ; clay ∩ Ni concentration;
	q(X1 ∩ X2) = q(X1)+ q(X2)	Independent	—
	q(X1 ∩ X2) > Max(q(X1), q(X2))	Enhance, bi-	Cd concentration ∩ pH; Cd concentration ∩ clay; OM ∩ CEC; CEC ∩ clay
	Min(q(X1),q(X 2))<q(X1 ∩ X2)<Max(q(X1)), q(X2))	Weaken, uni-	—
	q(X1 ∩ X2) < Min(q(X1), q(X2))	Weaken, nonlinear	—

Min (q(X1), q(X2)); Max (q(X1), q(X2)); q (X1)+q(X2); q (X1 ∩ X2)

Table 4 Results of evaluation indexes of training and test samples for models

Model	RMSE		MAE		MRE		R^2	
	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set
GA-BP neural network	0.041	0.040	0.024	0.023	0.337	0.346	0.991**	0.989**
BP neural network	0.045	0.117	0.028	0.077	0.302	0.342	0.991**	0.884**
Multiple regression	0.304	0.092	0.143	0.077	0.970	0.859	0.594**	0.808**

*$p < 0.05$; **$p < 0.01$

neural network is 0.989, which was much higher than 0.884 of the BP neural network. Therefore, the prediction effect of the GA-BP neural network was better than that of the BP neural network. Overall, the optimal model for estimating the Cd concentration of rice grain is the GA-BP neural network model.

Figure 3 shows the measured and predicted values of the training and test sets in different models. Figures 4 and 5 show the absolute errors and relative errors between the model-predicted and measured values, respectively. As shown in Figs. 3, 4, and 5, the predicted values of the GA-BP neural network model and the BP neural network model matched the measured values well. Moreover, the errors of the GA-BP neural network model were smaller than those of the BP neural network model. However, there was a relatively large difference between the measured and predicted values of the multiple regression model.

The results indicated that the accuracy of neural networks was better than that of the multiple regression model. This was primarily because the BP neural network model can be used as a black box model to predict a certain variable through a complex interaction factor and to process complex and fuzzy mappings relationships without knowing the relationship between the distribution form and variables. Therefore, when compared with multiple regression analysis, the BP neural network can reveal the nonlinear relationship between Cd concentration in rice grain and soil properties better, which overcomes the shortcomings of simulation by multiple regression model using complex factors. Moreover, the higher accuracy of the GA-BP neural network than the BP neural network at predicting heavy metal concentrations can be explained by the advantage of GA in global optimization.

Compared with neural network model, the mechanism models are in various forms, and the factors used include the root water uptake, the transpiration rate, the relative humidity, the change of total plant mass, the water flux in different parts of plants, the concentration of metals at particles in air, the dry and wet deposition velocity of particles and so on (Hough et al. 2005; Legind and Trapp 2010; Rein et al. 2011). These factors are relatively complex and difficult to obtain, and some of them can only be obtained in pot experiments. There are also some empirical parameters that were given by experts based on experience and theory, such as an empirical parameter that describes the distribution of root length density with depth, which may have some deviations from the actual (Hough et al. 2005). In addition, before applying the mechanism model, it is necessary to analyze which model is suitable for the area. And most mechanism models are only suitable for specific areas, crops, and heavy metals. When applied to other conditions, the models need to be calibrated and expanded (Hough et al. 2005). However, the neural network model does not have too many regulations and limitations on the choice of factors. The factors that have a great impact on the predicted objects should be selected. Data acquisition can be done by sampling in the field, rather than being limited to pot experiments.

The application of GA-BP neural networks can be used as a new method to complement the original methods. It can achieve as high as possible prediction accuracy with limited

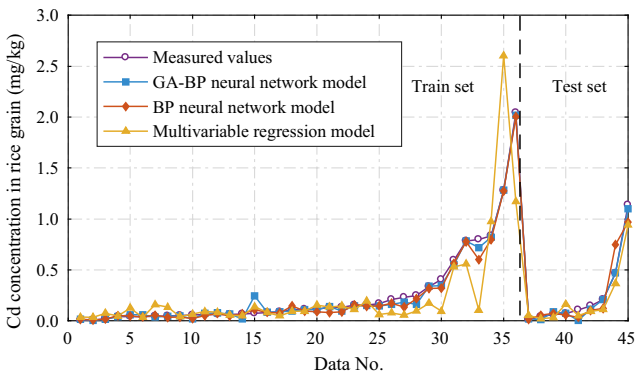

Fig. 3 Measured values and model prediction results of Cd concentration in rice grain

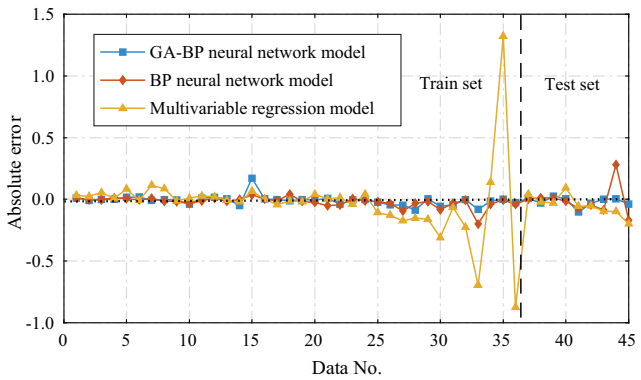

Fig. 4 Absolute errors of GA-BP neural network model, BP neural network model and multiple regression model

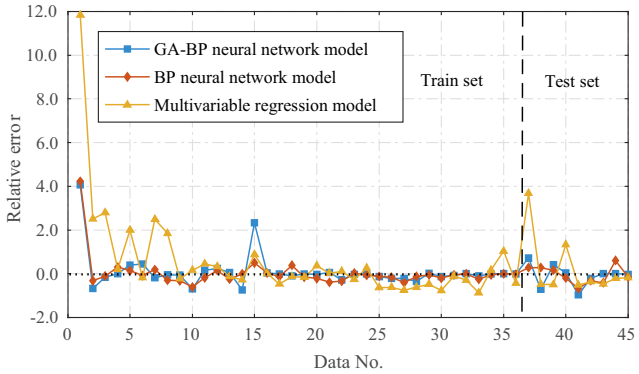

Fig. 5 Relative errors of GA-BP neural network model, BP neural network model and multiple regression model

data. We can choose the method based on the characteristics of the study area, data acquisition, dataset characteristics, and expected predictions effect that need to be achieved. Based on the neural network method, we can predict the Cd concentration in rice by measuring critical factors, without waiting for measuring the Cd concentration in the grain after the rice is matured. According to the prediction of Cd concentrations of rice grains in farmland, we can develop agricultural plans to plant rice in areas with low Cd concentration, which can reduce the flow of Cd from soil to the crops. In this study, we selected and measured factors that have a more significant impact on Cd concentration in rice grain such as Cd concentration in soil, SOM, pH, and so on. Although the prediction results of the neural network model based on the critical factors were relatively accurate, there were still some errors. Many factors such as the atmospheric environment, crop varieties, Fe concentration, S concentration, and microbes in the soil have not been considered for the data are not readily available. Whether the addition of these factors will further improve the prediction accuracy of the network can be found in further research.

Conclusion

In this study, we applied GA-BP neural network on predicting the Cd concentration in rice grain and achieved high prediction accuracy. Pearson's correlation analysis and Geodetector were used to identify the main factors influencing Cd concentration in rice grain for model construction. The total soil Cd concentration, clay content, Ni concentration, CEC, OM, and pH have important impacts and interactions on Cd concentration in rice grain. The BP neural network model optimized by GA was used to predict the concentration of Cd in rice grain based on the main influencing factors. The GA-BP neural network model showed higher prediction accuracy than the BP neural network model and multiple regression model. Although the movement of Cd in the soil-rice system is very complicated, the black box feature of the BP neural network enabled the concentration of Cd in rice grain to be predicted through learning and training by the neural network, and the purpose of accurate prediction was achieved. We also found that the limited explanatory power of this model for the concentration of Cd in rice grain based on soil properties can be improved by considering irrigation patterns, crop variety and atmospheric factors to reduce model errors in further studies. Overall, this study demonstrates the applicability of the GA-BP neural network for accurate prediction of Cd concentration of rice grain in large areas, which can contribute to human exposure and health risk assessment and the development of measures to reduce the transfer of Cd from soil to the food chain.

Acknowledgments We thank the International Science Editing (http://www.internationalscienceediting.com) for editing this manuscript.

Funding information This study was financially supported by the national key R&D program of China (No. 2017YFD0800305) and special funds for scientific research on public causes of ministry of land and resources of China (No. 201511082).

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

References

Adams ML, Zhao FJ, McGrath SP, Nicholson FA, Chambers BJ (2004) Predicting cadmium concentrations in wheat and barley grain using soil properties. J Environ Qual 33:532–541

Artiushenko T et al (2014) Metal uptake, antioxidant status and membrane potential in maize roots exposed to cadmium and nickel. Biologia 69:1142–1147

Barber S.A., Cushman J.H (1981) Nitrogen uptake model for agronomic crops, in modeling waste water renovation- land treatment

Bester PK, Lobnik F, Erzen I, Kastelec D, Zupan M (2013) Prediction of cadmium concentration in selected home-produced vegetables. Ecotoxicol Environ Saf 96:182–190. https://doi.org/10.1016/j.ecoenv.2013.06.011

Bingham FT, Page AL, Strong JE (1980) Yield and cadmium content of rice grain in relation to addition rates of cadmium, copper, nickel, and zinc with sewage sludge and liming. Soil Sci 130:32–38

Bradl HB (2004) Adsorption of heavy metal ions on soils and soils constituents. J Colloid Interface Sci 277:1–18. https://doi.org/10.1016/j.jcis.2004.04.005

Chaudri A, McGrath S, Gibbs P, Chambers B, Carlton-Smith C, Godley A, Bacon J, Campbell C, Aitken M (2007) Cadmium availability to wheat grain in soils treated with sewage sludge or metal salts. Chemosphere 66:1415–1423. https://doi.org/10.1016/j.chemosphere.2006.09.068

Chen J (2007) Rapid urbanization in China: a real challenge to soil protection and food security. Catena 69:1–15. https://doi.org/10.1016/j.catena.2006.04.019

Chen H, Yuan X, Li T, Hu S, Ji J, Wang C (2016) Characteristics of heavy metal transfer and their influencing factors in different soil-crop systems of the industrialization region, China. Ecotoxicol Environ Saf 126:193–201. https://doi.org/10.1016/j.ecoenv.2015.12.042

Dayton EA, Basta NT, Payton ME, Bradham KD, Schroder JL, Lanno RP (2006) Evaluating the contribution of soil properties to modifying lead phytoavailability and phytotoxicity. Environ Toxicol Chem 25:719–725

Guo X, Zhang S, Shan XQ, Luo L, Pei Z, Zhu YG, Liu T, Xie YN, Gault A (2006) Characterization of Pb, Cu, and Cd adsorption on particulate organic matter in soil. Environ Toxicol Chem 25:2366–2373

Hooda PS, Alloway BJ (1998) Cadmium and lead sorption behaviour of selected English and Indian soils. Geoderma 84:121–134. https://doi.org/10.1016/s0016-7061(97)00124-9

Hooyberghs J, Mensink C, Dumont G, Fierens F, Brasseur O (2005) A neural network forecast for daily average PM10 concentrations in Belgium. Atmos Environ 39:3279–3289

Hough RL, Young SD, Crout NMJ (2003) Modelling of Cd, Cu, Ni, Pb and Zn uptake, by winter wheat and forage maize, from a sewage disposal farm. Soil Use Manag 19:19–27. https://doi.org/10.1079/sum2002157

Hough RL, Tye AM, Crout NMJ, McGrath SP, Zhang H, Young SD (2005) Evaluating a 'free ion activity model' applied to metal uptake by *Lolium perenne* L. grown in contaminated soils. Plant Soil 270: 1–12. https://doi.org/10.1007/s11104-004-1658-5

Hu Y, Wang J, Li X, Ren D, Zhu J (2011) Geographical detector-based risk assessment of the under-five mortality in the 2008 Wenchuan earthquake, China. PLoS One 6:e21427. https://doi.org/10.1371/journal.pone.0021427

Huang JX, Wang JF, Bo YC, Xu CD, Hu MG, Huang DC (2014) Identification of health risks of hand, foot and mouth disease in China using the geographical detector technique. Int J Environ Res Public Heath 11:3407–3423

Ingwersen J, Streck T (2005) A regional-scale study on the crop uptake of cadmium from sandy soils: measurement and modeling. J Environ Qual 34:1026–1035. https://doi.org/10.2134/jeq2003.0238

Jiang J, Liu X, Xu Z, Jin M, Liu F (2016) An improved BP neural network model for estimating cd stress in rice using remote sensing data. In: International conference on fuzzy systems and knowledge discovery

Legind CN, Trapp S (2010) Comparison of prediction methods for the uptake of As, Cd and Pb in carrot and lettuce. SAR QSAR Environ Res 21:513–525. https://doi.org/10.1080/1062936X.2010.502296

Li C, Yang Z, Yan H, Wang T (2014a) The application and research of the GA-BP neural network algorithm in the MBR membrane fouling. Abstr Appl Anal 2014:8. https://doi.org/10.1155/2014/673156

Li W, Xu B, Song Q, Liu X, Xu J, Brookes PC (2014b) The identification of 'hotspots' of heavy metal pollution in soil-rice systems at a regional scale in eastern China. Sci Total Environ 472:407–420. https://doi.org/10.1016/j.scitotenv.2013.11.046

Liu M, Liu X, Wu M, Li L, Xiu L (2011) Integrating spectral indices with environmental parameters for estimating heavy metal concentrations in rice using a dynamic fuzzy neural-network model. Comput Geosci 37:1642–1652. https://doi.org/10.1016/j.cageo.2011.03.009

Lu J, Yang X, Meng X, Wang G, Lin Y, Wang Y, Zhao F (2017) Predicting cadmium safety thresholds in soils based on cadmium uptake by Chinese cabbage. Pedosphere 27:475–481. https://doi.org/10.1016/s1002-0160(17)60343-6

Martin I, Jeffries J (2008) Updated technical background to the CLEA model

McBride MB (2002) Cadmium uptake by crops estimated from soil total Cd and pH. Soil Sci 167:62–67. https://doi.org/10.1097/00010694-200201000-00006

Novotna M, Mikes O, Komprdova K (2015) Development and comparison of regression models for the uptake of metals into various field crops. Environ Pollut 207:357–364. https://doi.org/10.1016/j.envpol.2015.09.043

Pang NS, Shi YL (2008) Research on short-term load forecasting based on adaptive hybrid genetic optimization BP neural network algorithm. Int C Manage Sci Eng 1563–1568. https://doi.org/10.1109/Icmse.2008.4669113

Peng HY (2013) The BP neural network's GA optimization and its realization on MATLAB. In: Chinese Control and Decision Conference. IEEE, 2013:536–539. https://doi.org/10.1109/CCDC.2013.6560982

Qian YZ, Chen C, Zhang Q, Li Y, Chen ZJ, Li M (2010) Concentrations of cadmium, lead, mercury and arsenic in Chinese market milled rice and associated population health risk. Food Control 21:1757–1763. https://doi.org/10.1016/j.foodcont.2010.08.005

Ran J, Wang DJ, Wang C, Zhang G, Zhang HL (2016) Heavy metal contents, distribution, and prediction in a regional soil-wheat system. Sci Total Environ 544:422–431

Rein A, Legind CN, Trapp S (2011) New concepts for dynamic plant uptake models. SAR QSAR Environ Res 22:191–215. https://doi.org/10.1080/1062936X.2010.548829

Rizwan M, Ali S, Abbas T, Zia-ur-Rehman M, Hannan F, Keller C, al-Wabel MI, Ok YS (2016a) Cadmium minimization in wheat: a critical review. Ecotoxicol Environ Saf 130:43–53. https://doi.org/10.1016/j.ecoenv.2016.04.001

Rizwan M, Ali S, Adrees M, Rizvi H, Zia-ur-Rehman M, Hannan F, Qayyum MF, Hafeez F, Ok YS (2016b) Cadmium stress in rice: toxic effects, tolerance mechanisms, and management: a critical review. Environ Sci Pollut Res Int 23:17859–17879. https://doi.org/10.1007/s11356-016-6436-4

Romkens PF, Guo HY, Chu CL, Liu TS, Chiang CF, Koopmans GF (2009) Prediction of cadmium uptake by brown rice and derivation of soil-plant transfer models to improve soil protection guidelines. Environ Pollut 157:2435–2444. https://doi.org/10.1016/j.envpol.2009.03.009

Shen CY, Wang LX, Li Q (2007) Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method. J Mater Process Technol 183:412–418

Song Y, Wilson MJ, Moon HS, Bacon JR, Bain DC (1999) Chemical and mineralogical forms of lead, zinc and cadmium in particle size fractions of some wastes, sediments and soils in Korea. Appl Geochem 14:621–633

Sterckeman T, Perriguey J, Caël M, Schwartz C, Morel JL (2004) Applying a mechanistic model to cadmium uptake by Zea mays and Thlaspi caerulescens: consequences for the assessment of the soil quantity and capacity factors. Plant Soil 262:289–302. https://doi.org/10.1023/B:PLSO.0000037049.07963.ab

Tongle X, Yingbo W, Kang C (2016) Tailings saturation line prediction based on genetic algorithm and BP neural network. J Intell Fuzzy Syst 30:1947–1955. https://doi.org/10.3233/ifs-151905

Tudoreanu L, Phillips CJC (2004) Empirical models of cadmium accumulation in maize, rye grass and soya bean plants. J Sci Food Agric 84:845–852. https://doi.org/10.1002/jsfa.1730

Wang KR (2002) Tolerance of cultivated plants to cadmium and their utilization in polluted farmland soils. Acta Biotechnol 22:189–198

Wang JF, Li XH, Christakos G, Liao YL, Zhang T, Gu X, Zheng XY (2010) Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China. Int J Geogr Inf Sci 24:107–127. https://doi.org/10.1080/13658810802443457

Wang S, Zhang N, Wu L, Wang Y (2016) Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method. Renew Energy 94:629–636. https://doi.org/10.1016/j.renene.2016.03.103

Wang S, Wu W, Liu F, Liao R, Hu Y (2017) Accumulation of heavy metals in soil-crop systems: a review for wheat and corn. Environ Sci Pollut Res Int 24:15209–15225. https://doi.org/10.1007/s11356-017-8909-5

Weng L, Temminghoff EJ, Lofts S, Tipping E, Van Riemsdijk WH (2002) Complexation with dissolved organic matter and solubility control of heavy metals in a sandy soil. Environ Sci Technol 36:4804–4810

Wiggenhauser M, Bigalke M, Imseng M, Müller M, Keller A, Murphy K, Kreissig K, Rehkämper M, Wilcke W, Frossard E (2016) Cadmium isotope fractionation in soil-wheat systems. Environ Sci Technol 50: 9223–9231

Yan XD, Gao D, Zhang F, Zeng C, Xiang W, Zhang M (2013) Relationships between heavy metal concentrations in roadside topsoil and distance to road edge based on field observations in the Qinghai-Tibet plateau, China. Int J Environ Res Public Health 10:762–775

Ye XX, Li HY, Ma YB, Wu L, Sun B (2014) The bioaccumulation of Cd in rice grains in paddy soils as affected and predicted by soil properties. J Soils Sediments 14:1407–1416

Yu F, Xu X (2014) A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. Appl Energy 134:102–113. https://doi.org/10.1016/j.apenergy.2014.07.104

Zhang L, Wang F, Sun T, Xu B (2016) A constrained optimization method based on BP neural network. Neural Comput & Applic 29:413–421. https://doi.org/10.1007/s00521-016-2455-9

Zhang D, Liu J, Jiang C, Liu A, Xia B (2017) Quantitative detection of formaldehyde and ammonia gas via metal oxide-modified graphene-based sensor array combining with neural network model. Sensors Actuators B Chem 240:55–65. https://doi.org/10.1016/j.snb.2016.08.085

Zhao KL, Liu XM, Xu JM, Selim HM (2010) Heavy metal contaminations in a soil-rice system: identification of spatial dependence in relation to soil properties of paddy fields. J Hazard Mater 181:778–787

Zheng B-H (2017) Material procedure quality forecast based on genetic BP neural network. Mod Phys Lett B 31:1740080. https://doi.org/10.1142/s0217984917400802