DATA MODELING ANALYSIS ON REMOVAL EFFICIENCY OF HEXAVALENT CHROMIUM

RUNQIN HAO

Key Laboratory of Water Quality Science and Water Environment Recovery Engineering
Beijing University of Technology
Beijing 100124, China

GUANWEN ZHANG¹, DONG LI² and JIE ZHANG²,³

¹School of Electronics and Information, Northwestern Polytechnical University
Xi'an 710129, China
²Key Laboratory of Water Quality Science and Water Environment Recovery Engineering
Beijing University of Technology
Beijing 100124, China
³State Key Laboratory of Urban Water Resource and Environment
Harbin Institute of Technology
Harbin 150090, China

(Communicated by Xiaochun Cao)

Abstract. Chromium and its compounds are widely used in many industries in China and play a very important role in the national economy. At the same time, heavy metal chromium pollution poses a great threat to the ecological environment and human health. Therefore, it’s necessary to safely and effectively remove the chromium from pollutants. In practice, there are many factors which influence the removal efficiency of the chromium. However, few studies have investigated the relationship between multiple factors and the removal efficiency of the chromium till now. To this end, this paper uses the green synthetic iron nanoparticles to remove the chromium and investigates the impacts of multiple factors on the removal efficiency of the chromium. A novel model that maps multiple given factors to the removal efficiency of the chromium is proposed through the advanced machine learning methods, i.e., XGBoost and random forest (RF). Experiments demonstrate that the proposed method can predict the removal efficiency of the chromium precisely with given influencing factors, which is very helpful for finding the optimal conditions for removing the chromium from pollutants.

1. Introduction. Chromium (Cr) is in the fourth cycle VIB group of the periodic table, whose atomic number is 24, density is 7.20 g/cm³ and content in the earth’s crust is 0.01%. Chromium is a kind of silver-white-luster metal. It is hard and corrosion resistant, insoluble in water but soluble in strong alkali solution. Chromium is widely found in soil, atmosphere, water, and bodies of plants and animals. At normal temperature, Chromium keeps chemically stable to air and water. Chromium...
exists in a variety of oxidation states (from 0 to 6). In nature, it is mainly in the form of Cr(III) and Cr(VI). Cr(III) is an essential micro-element to human body, which is relatively stable and less toxic. However, it exhibits biotoxicity at higher concentrations [9]. The main existing forms of Cr(III) are chromium oxide and chromium hydroxide. By comparison, the existing forms of Cr(VI) are $HCrO_4^-$ and $CrO_4^{2-}$ etc.

Fig. 1 illustrates the relationship of pH-Eh of Cr(VI). Cr(VI) is usually not readily present in soil but can be deposited in water. Without oxygen, Cr(VI) can be transformed to Cr(III) by organic matter, the ions of $S^{2-}$, $Fe^{2+}$, etc. Cr(VI) is 100 times more toxic than Cr(III) and has carcinogenic, teratogenic and mutagenic effects on humans. The transformation between Cr(III) and Cr(VI) is:

$$Cr_2O_7^{2-} + 14H^+ + 6e^- \rightarrow 2Cr^{3+} + 7H_2O.$$ (1)

In addition, studies have shown that Cr(VI) can be easily absorbed by the human body. Cr(VI) may cause skin eczema, ulcers, dermatitis or allergic reactions. Inhalation of higher concentration of Cr(VI) can cause sneezing, runny nose, pruritus and nosebleed. About 1.5-1.6g of Cr(VI) compound leads to death, nausea and vomiting, dizziness and headache, restlessness, thickening of oral mucosa, severe abdominal pain and other conditions. Chromium compounds can also cause damage to the eyes, leading to foreign body sensation, painful tears, conjunctival congestion, vision loss and corneal epithelium loss.

With the rapid development of industry, the pollution of Cr(VI) becomes increasingly serious. The Cr(VI) pollution ranks second, being below the lead pollution in the heavy metal pollution. It has been shown that about 600,000 tons of chromium slag are produced in China every year, and the total amount of chromium slag is 6 million tons, but over 4 million tons have not been detoxified or comprehensively utilized [12] [18]. Therefore, it’s of great importance to remove Cr(VI), which arouses

![Figure 1. The relationship of pH-Eh of Cr(VI)](image-url)
much attention in recent years. A lot of removal methods have been proposed, which can be classified into physical, chemical and biological methods generally.

Firstly, the physical methods mainly contain adsorption methods, ion exchange methods and membrane separation methods. The adsorption methods refer to removing Cr(VI) by the high void structure or specific surface area absorbents, such as activated carbon, peanut skin, microbial cells, chestnut shell, chitosan, etc. [2] [14] [8]. The ion exchange method can make Cr(VI) ion exchange with ion exchanger to separate Cr(VI) from wastewater, which is suitable for the treatment of high concentration Cr(VI) wastewater. The common ion exchangers are exchange resin. Membrane separation method uses specific semi-permeable membrane osmosis to make the substances in solution out from the semi-permeable membrane osmosis, so as to separate different substances. After passing through the semi-permeable membrane, Cr(VI) will be enriched on the side of the deposits, and then the permeable materials will be either processed or discharged directly [17] [6].

Secondly, the chemical methods for removing Cr(VI) include chemical precipitation, ferrite and electrolysis methods. Chemical precipitation methods remove Cr(VI) by converting it to Cr(III) precipitation. The common chemical precipitation methods include Barium salt precipitation and sulfur dioxide precipitation. Ferrite method involves addition of excessive FeSO$_4$ into acidic (pH 2-3) Cr(VI) waste water to transform Cr(VI) to Cr(III) and oxidize Fe(II) to Fe(III). Electrolysis is a method in which Cr(VI) is precipitated by hydroxide through oxidation and reduction reactions between iron and Cr(VI). In this process, iron serves as anode.

Thirdly, biochemical methods convert soluble Cr(VI) ions into insoluble compounds and remove them from polluted water through catalytic transformations of microbial enzymes, reduction of metabolites, flocculation and precipitation [7] [25] [1].

The above mentioned methods of removing Cr(VI) suffer from similar drawbacks, such as high cost, high energy consumption and secondary pollution to the environment. Therefore, researchers in recent years investigated green synthesis of nano-iron for removing Cr(VI) [13] [11] [10] [20]. Compared with traditional methods, the green synthesis of nano-iron has the advantages of simple operation, low synthesis cost, fast reaction speed, avoiding the use of toxic and harmful reducing agents, and no burden of environmental pollution, which has been widely applied in removing the environmental pollutants [24] [16] [21]. There are many factors that can influence the removal efficiency of Cr(VI) [23] [22] [19]. However, few studies have investigated this issue. Therefore, in this paper, we deeply investigate the impacts of multiple factors on the removal efficiency of Cr(VI). Specifically, we resort to advanced machine learning methods, i.e. XGBoost and random forest (RF) to establish a mapping model to predict the removal efficiency of Cr(VI) under given conditions, which can predict the removal efficiency of Cr(VI) precisely.

The rest of this paper is organized as follows. In Section 2, we formulate the problem mathematically. In Section 3, we introduce the machine learning methods, XGBoost and RF in detail. In Section 4, we introduce the process and settings of the experiments to collect data. In Section 5, we analyze the obtained experimental results and evaluate the effectiveness of the proposed model. At last, we conclude this paper in Section 6.

2. Problem description. In real applications, multiple factors that influence the removal efficiency of Cr(VI) should be investigated. We use green tea synthesized
iron nanoparticles in this experiment and the factors contain the pH value of the reaction solution, the dosage of the green tea synthesized iron nanoparticles, the initial concentration of Cr(VI) solution, reaction temperature, different content of green tea extract, the ratio of green tea extract and $Fe^{2+}$, the preparation temperature of green tea extract, and the synthesizing temperature of green tea iron nanoparticles. We denote the above eight different factors as $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$. In experiments, it is impossible to sample or traverse each dimension of the high dimension data densely. Therefore, we introduce the statistical learning methods, such as random forest [3], XGBoost [4], etc. to investigate the relationship between the influencing vector $X$ and the removal efficiency of Cr(VI). Supposing the removal efficiency of Cr(VI) as $Y$, we learn the mapping model from $X$ to $Y$ through the statistical learning methods, written as:

$$Y = M(X)$$

where $M(\cdot)$ refers to the mapping function.

3. Random forest and XGBoost. In machine learning, the objective function of the regression problem is defined as:

$$\arg \min_\theta \{L(\theta) + \Omega(\theta)\}$$

where $\theta$ is the model parameter, $L(\theta)$ is the training error, representing the fitting accuracy of the model to the training set. In regression, the Mean Square Error (MSE) is used as the loss function, $\Omega(\theta)$ represents the regular term, indicating the complexity of the model, which is usually defined by $l_1$ or $l_2$ norm.

Random Forest (RF) was proposed by Leo Breiman to approximate multi-non-linear relation, which belongs to statistical learning methods [3]. The basic idea of RF is through constructing a number of regression trees and assembling the regression trees according to the predefined principle. Due to randomness, the constructed regression trees are different in each time. Fig. 2 illustrates the structure of an RF. Suppose RF contains $K$ regression trees, then the final prediction result is:

$$\hat{y}^i = \sum_{k=0}^{K} f_k(X_i), f_k \in F$$

where $F$ refers to the RF, $f_k(X_i)$ is the weight of the $i$-th sample on the leaf of the $k$-th tree. In the training of the RF, we only consider the training error, then the training objective function can be denoted as:

$$L(\theta) = \sum_{i=1}^{n} l(y^i, \hat{y}^i) = \sum_{i=1}^{n} l(y^i, \sum_{k=1}^{K} f_k(x_i))$$

On the basis of RF, we introduce the regular term $\Omega(\theta)$ to constrain the complexity of the system and construct the regression tree successively as follows:

$$\hat{y}^0 = 0$$

$$\hat{y}^1 = f_1(x_i) = \hat{y}^0 + f_1(x_i)$$

$$\hat{y}^2 = f_1(x_i) + f_2(x_i) = \hat{y}^1 + f_2(x_i)$$

$$\ldots$$
\[
\hat{y}^k = \sum_{k=0}^{K} f_k(x_i) = \hat{y}^{k-1} + f_k(x_i)
\]

where the complexity of the regression tree is further defined as the function of the number of leaves \( T \) and the weight of leaves \( \omega \):

\[
\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2
\]

where \( \gamma \) and \( \lambda \) represent parameters to adjust the significance of the two terms. \( f_t \) refers to the \( t \)-th tree. Then we transform the regression problem to the optimization problem and construct the regression tree by the gradient solving method. Specifically, constructing the regression trees successively can improve the availability of the training samples. The regular term of the regression trees can reduce the risk of over-fitting. The gradient regression method can guarantee the best performance of the learned regression model.

XGBoost is a machine learning function library focused on gradient lifting algorithm, which was proposed by Chen and Guestrin \[4\]. This function library has attracted wide attention due to its excellent learning effect and efficient training speed. In 2015, of the 29 algorithms that won the Kaggle contest, 17 used the XGBoost library, compared with 11 for the recent boom in deep neural network methods \[5\]. XGBoost is highly-efficient, flexible and highly-transplantable, which also supports distributed training on multiple work stations, such as Amazon Web Services (AWS), Google Compute Engine and AZURE etc. \[15\].

4. **Experimental setup.** The experimental setup is listed in Table 1. The simulation experiments were performed on the Anaconda based Python environment.
Table 1. Experimental Setup.

| Experimental Parameters               | Setup       | Units of measurement |
|---------------------------------------|-------------|----------------------|
| Green tea extract content             | 20, 30, 40, 50, 60 | mg/L                 |
| green tea extract / Fe^{2+}           | 1:3, 1:2, 1:1, 2:1, 3:1 | -                   |
| green tea extract preparation temperature | 40, 60, 80, 100 | °C                  |
| GT-Fe NPs synthesis temperature       | 25, 35, 45, 55 | °C                  |
| pH value                              | 3, 5, 7, 9, 11 | -                   |
| dosage of GT-Fe NPs                  | 0.01, 0.02, 0.04, 0.06, 0.12 | g/L                 |
| Cr(VI) initial concentration         | 40, 60, 80, 100, 160, 200 | mg/L               |
| reaction temperature                 | 15, 25, 35, 45, 55 | °C                  |

on Windows system. The version of Python is 3.6, the RF and XGBoost were implemented under two machine learning libraries, which are Scikit-Learn with 0.20.0 and XGBoost with 0.84.

After preparing the experimental data, we randomly divided the data into 3:2 training set and test set according to the machine learning method. We learned the prediction model on the training set and tested its prediction performance on the test set.

5. Experimental results and analysis. In XGBoost, we adjusted the number of regression trees, the depth of regression trees and the parameter of the regular l^{2}-norm on the training set. We started from 1 and added 10 regression trees every time, the experimental results are shown in Fig. 3, where the horizontal axis represents the number of regression trees, the vertical axis represents the regression coefficient R^{2}. The closer to 1 R^{2} is, the better the fitting performance is. From Fig. 3, we can observe that the regression performance tends to converge as the number of regression trees increase. The maximum value of R^{2} appears around the value of 321. Therefore, we refined the search around 321, the searching results are shown in Fig. 4. From Fig. 4, we determine the best number of regression trees is 306. Then we investigate the effect of the maximum depth of the regression trees on the regression accuracy. We also start from 1 and add the depth of the

![Figure 3](image-url)
regression trees, the results are shown in Fig. 5. As can be seen, when the depth is 5, the performance is best. Then we investigate the $l_2$-norm. We start from 0 and increase by the step of 0.1, the results are shown in Fig. 6. From this figure, we can see that the regular parameter $\lambda$ has little effect on the performance. The final prediction results by XGBoost are shown in Fig. 7.

**Figure 4.** The impact of the number of decision trees on XGBoost performance (fine search)

**Figure 5.** The impact of the max depth of decision trees on XGBoost performance

**Figure 6.** The impact of the regular lambda of decision trees on XGBoost performance
From Fig. 7, we can observe that the prediction performance is excellent on the training set. The prediction performance on the test set is also satisfied, and the regression coefficient on the test set is 0.963.

Similarly, we also employ RF for learning the prediction model. We investigate the impact of the number of regression trees, the depth of the regression trees on the fitting performance. We start from 0 and increase the number of regression trees by the step of 1. The experimental results are shown in Fig. 8. From this figure, we can find that the optimal number of the trees is 9, at which the regression coefficient $R^2$ attains the maximum value. Then we start from 1 and increase the maximum depth of the trees by the step of 1. The experimental results are shown in Fig. 9. From this figure, we observe that the optimal depth is 13. The prediction results of the removal efficiency of Cr(VI) are shown in Fig. 10.

From Fig.9, the prediction performance of RF is worse than that of XGBoost on the training set. However, the performance of RF is better than that of XGBoost on the test set. The final regression coefficient is 0.981.
From the above experimental results, we can have the following conclusions. First, we can predict the removal efficiency of Cr(VI) through machine learning methods. The Cr(VI) removal rates are 87.26% and 93.23% when the feature $X = \{60, 1/3, 80, 25, 3.8, 0.06, 80, 25\}$ and $X = \{60, 1/3, 80, 25, 3.8, 0.06, 80, 45\}$. These two conditions are all the same except the removal temperature or $x_8$. Therefore, we can find that higher removal temperature can improve the removal efficiency of Cr(VI). In addition, the Cr(VI) removal rates are 57.76% and 87.26% when the feature $X = \{60, 1/3, 80, 25, 3.8, 0.02, 80, 25\}$ and $X = \{60, 1/3, 80, 25, 3.8, 0.06, 80, 45\}$. These two conditions are all the same except the dosage of the green tea synthesized nanoparticles iron or $x_6$. Therefore, we can obtain that higher dosage of the green tea synthesized nanoparticles iron can improve the removal efficiency of Cr(VI).

In general, according to the mapping model learned from RF or XGBoost, we are able to analyze the influencing degree of different factors on the removal efficiency of Cr(VI). Accordingly, we can obtain the optimal condition, under which the removal
efficiency of Cr(VI) earns the maximum value, which is impossible to obtain through the experiments by exhaustive traverse.

In this experiment, XGBoost performs better than RF on the training set, while RF outperforms XGBoost on the test set. Such observations are mainly restricted by the experimental conditions. It can be expected that the results can be refined as increasing the number of samples. It’s also note that we present the two-dimension plot, while the prediction model contains eight-dimension data, which can’t be intuitively demonstrated due to the representation restriction.

6. Conclusion. The removal of Cr(VI) is of great importance to the ecological environment and human health. In this paper, we have focused our attention on the investigation of the influence of multiple factors on the removal efficiency of Cr(VI). Specifically, we collected the data by elaborately-designed experiments and constructed a dataset. Through modern machine learning methods, such as XGBoost and RF algorithms, we establish a model that can predict the removal efficiency of Cr(VI) accurately with given multiple factors. This work can help to find the best conditions for removing Cr(VI) so as to guide the Cr(VI) removal process effectively.

REFERENCES

[1] R. Benniceli, Z. Stpniewska, A. Banach, et al., The Ability of Azolla Caroliniana to Remove Heavy Metals (Hg(II),Cr(III),Cr(VI)) from Municipal Waste Water, Chemosphere, 55 (2004), 141–146.
[2] S. Bosinco, J. Roussy, E. Guibal, et al., Interaction Mechanisms between Hexavalent Chromium and Corn Cob, Environmental Technology, 17 (1996), 55–62.
[3] L. Breiman, Random forest, Machine Learning, 45 (2001), 5–32.
[4] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system[C]. Acm sigkdd international conference on knowledge discovery and data mining, Water Research, (2016), 785–794.
[5] M. Chen, Q. Liu, S. Chen, Y. Liu, C. Zhang and R. Liu, XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system, IEEE Access, 7 (2019), 13149–13158.
[6] J. E. Cruver, Reverse Osmosis—Where it Stands Today, Water Sewage Works, 1973.
[7] V. Dushenkov, P. B. A. N. Kumar, H. Motto, et al., Rhizofiltration: The use of plants to remove heavy metals from aqueous streams, Environmental Science & Technology, 29 (1995), 1239–1245.
[8] I. C. Eromosele and S. S. Bayero, Adsorption of chromium and zinc ions from aqueous solutions by cellulosic graft copolymers, Bioresource Technology, 71 (2000), 279–281.
[9] A. M. Farag, T. May, G. D. Marty, et al., The effect of chronic chromium exposure on the health of chinook salmon (Oncorhynchus Tshawytscha), Aquat Toxicol, 76 (2006), 246–257.
[10] J. Farrell, J. P. Wang, P. O’Day, et al., Electrochemical and Spectroscopic Study of Arsenate Removal from Water Using Zero-Valent Iron Media, Environmental Science & Technology, 35 (2001), 2026–2032.
[11] S. Gandhi, B. T. Oh, J. L. Schnoor, et al., Degradation of TCE, Cr(VI), Sulfate, and nitrate mixtures by granular iron in flow-through columns under different microbial conditions, Water Research, 36 (2002), 1973–1982.
[12] C. R. gibbs, Characterization and application of ferrozine iron reagent as a ferrous iron indicator, Anal.Chem., 48 (1976), 1197–1200.
[13] R. W. Gillham And S. F. O’Hannesin, Enhanced degradation of halogenated aliphatics by zero-valent iron, Ground Water, 32 (1994), 958.
[14] T. Karthikeyan, S. Rajgopal and L. R. Miranda, Chromium(VI) adsorption from aqueous solution by hevea brasiliensis sawdust activated carbon, Journal of Hazardous Materials, 124 (2005), 192–199.
[15] S. M. H. Mahmud, W. Chen, H. Jahan, Y. Liu, N. I. Sujan and S. Ahmed, IDTi–CSamoteB: Identification of Drug–Target Interaction Based on Drug Chemical Structure and Protein Sequence Using XGBoost With Over–Sampling Technique SMOTE, IEEE Access, 7 (2019), 48699–48714.
[16] D. O’Arroll, B. Sleep and M. Krol, et al., Nanoscale zero valent iron and bimetallic particles for contaminated site remediation, *Advances in Water Resources*, 51 (2013), 323–332.

[17] H. Ozakai, I. Sharma and W. Saktaywin, Performance of an ultral-low-pressure reverse osmosis membrane (UI PROM) for separating heavy metal: effects of interference parameters, *Desalination*, 144 (2002), 287–294.

[18] C. D. Palmer And P. R. Wittrodt, Processes affecting the remediation of chromium-contaminated sites, *Environ. Health Persp.*, 92 (1991), 25–40.

[19] M. Sherman, J. G. Ponder, et al., Surface Chemistry and Electrochemistry of Supported Zerovalent Iron Nanoparticles in the Remediation of Aqueous Metal Contaminants, *Chemistry of Materials*, 13 (2001), 479–486.

[20] I. B. Singh And D. R. Sing, Effects of pH on Cr-Fe interaction during cr(vi) removal by metallic iron, *Environment Technology*, 24 (2003), 1041–1047.

[21] T. Tosco, M. P. Papini And C. C. Viggi, et al., Nanoscale zero valent iron particles for groundwater remediation: A review, *Journal of Cleaner Production*, 77 (2014), 10–21.

[22] C. Uzum, T. Shahwan and A. E. Eroglu, Application of zero-valent iron nanoparticles for the removal of aqueous co(2+) ions under various experimental conditions, *Chemical Engineering Journal*, 144 (2008), 213–220.

[23] D. B. Vance, Nanoscale iron colloids: The maturation of the technology for field scale applications, *Pollution Engineering*, 37 (2005), 16–18.

[24] P. Westerhoff And J. James, Nitrate removal in zero-valent iron packed columns, *Water Research*, 37 (2003), 1818–1830.

[25] X. H. Zhang, J. Liu, H. T. Huang, et al., Chromium accumulation by the hyperaccumulator plant leersia hexandra swartz, *Chemosphere*, 67 (2007), 1138–1143.

Received April 2019; 1st revision May 2019; 2nd revision June 2019.

E-mail address: hrunqin@163.com
E-mail address: guanwen.zh@nwpu.edu.cn
E-mail address: lidong2006@bjut.edu.cn
E-mail address: 6282031@163.com