EKSPLOATACJA I NIEZAWODNOŚĆ MAINTENANCE AND RELIABILITY

Eksploatacja i Niezawodnosc – Maintenance and Reliability

Volume 24 (2022), Issue 2

journal homepage: http://www.ein.org.pl

Article citation info:
Lin L, Guo H, Lv Y, Liu J, Tong C, Yang S. A machine learning method for soil conditioning automated decision-making of EPBM: hybrid GBDT and Random Forest Algorithm. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2022; 24 (2): 237–247, http://doi.org/10.17531/ein.2022.2.5.

Indexed by: Web of Science Group

A machine learning method for soil conditioning automated decision-making of EPBM: hybrid GBDT and Random Forest Algorithm

Lin Lin[a,*], Hao Guo[a], Yancheng Lv[a], Jie Liu[a], Changsheng Tong[a], Shuqin Yang[b]

[a] Harbin Institute of Technology, School of Mechatronics Engineering, Harbin, 150001, China
[b] China Railway Construction Corporation Limited, Changsha, 410100, China

Highlights

- A method hybrid two machine learning algorithms to predict the dosage of foam is proposed.
- GBDT is used to select geological parameters with big impact on the dosage of foam.
- Considering drive parameters as decision-making factors improves the practicability.
- Results shows that the prediction model performs better than other algorithms in accuracy.
- The proposed method can realize real-time decision-making compared with experiment.

Abstract

There lacks an automated decision-making method for soil conditioning of EPBM with high accuracy and efficiency that is applicable to changeable geological conditions and takes drive parameters into consideration. A hybrid method of Gradient Boosting Decision Tree (GBDT) and random forest algorithm to make decisions on soil conditioning using foam is proposed in this paper to realize automated decision-making. Relevant parameters include decision parameters (geological parameters and drive parameters) and target parameters (dosage of foam). GBDT, an efficient algorithm based on decision tree, is used to determine the weights of geological parameters, forming 3 parameters sets. Then 3 decision-making models are established using random forest, an algorithm with high accuracy based on decision tree. The optimal model is obtained by Bayesian optimization. It proves that the model has obvious advantages in accuracy compared with other methods. The model can realize real-time decision-making with high accuracy under changeable geological conditions and reduce the experiment cost.

Keywords

soil conditioning, automated decision-making, hybrid algorithm, geological parameters, drive parameters, feature selection.

1. Introduction

Shield machine is one of the most important large-scale equipment in the field of construction machinery, widely used in tunneling such as subway, railway and highway. According to statistics, in developed countries, the amount of tunnel excavation by the shield method accounts for more than 90% of the total amount of tunnel excavation. And EPB shield, whose excavation quantity accounts for more than 80% of all shield methods, is the most frequently used method because of the high mechanization level, fast construction speed, strong environmental adaptability, as well as relatively low cost [10].

It is necessary to improve the properties of the excavated material to ensure that the EPBM runs at a low failure rate and reduce the wear of the cutterhead [6, 33]. Firstly, the excavated material is used to transfer thrust from the jacks to keep the excavation face pressure balanced and ground surface settlement stable. Secondly, the excavated material is also used to generate a face-supporting muck to prevent water inflows at the tunnel face [10]. So stable geotechnical properties are required, including good plastic deformation, small inner friction angle. Thirdly, material flowing out from the excavation chamber should be conveyed steadily and continuously by the screw conveyor to ensure the earth pressure balanced, so the geotechnical properties such as plasticity and liquidity need to be suitable [9]. Soil with poor geotechnical properties likely causes spewing, mud cake, etc., and increase the wear of the cutters and screw conveyors, then damage the shield equipment, finally lower construction quality [11, 25].

The most mature method for soil conditioning is to inject soil additive to improve poor soil properties. Generally, additives such as foam and polymers apply to coarse grained soils, while water and suspension apply to fine grained soils [17]. At present, foam is one of the most commonly used soil additives in engineering application. The advantages are that it is universal for a variety of soils, convenient to transport and use [27]. In recent years, EPBM is gradually used in areas with complex and changeable geological conditions in China, such as Yunnan and Tibet. In addition, the foam injection system in the EPBM is automated, and the geological prediction accuracy of a certain distance in front of the working face is high, which provides a prerequisite for the automation technology of soil conditioning. One of the key issues in the field of soil conditioning is how to achieve accurate and rapid automated decision-making on the dosage of foam

(*) Corresponding author.
E-mail addresses: L. Lin (ORCID: 0000-0001-9525-1168): waiwaiyl@163.com, H. Guo (ORCID: 0000-0001-8725-0560): 1710044017@qq.com, Y. Lv (ORCID: 0000-0003-0843-683X): xgzlyc@163.com, J. Liu (ORCID: 0000-0002-1874-1507): 624003414@qq.com, C. Tong (ORCID: 0000-0001-8095-5453): 924534883@qq.com, S. Yang (ORCID: 0000-0003-1411-487X): yangshuqin@crchi.com

under changeable geological conditions, to improve the efficiency of the EPBM.

Traditional decision-making methods such as experimental method perform poor when applied to uncovered geological conditions, and don't take drive parameters that may impact the dosage of foam into consideration, see Section 2.1. Besides, they are too low in efficiency to do on-site decision-making. In fact, there is a lack of decision-making methods for soil conditioning with high accuracy and efficiency that comprehensively consider geological conditions and construction requirements to improve the automation level of soil conditioning decision-making.

It is the purpose of the present paper to provide a decision-making model with high accuracy and efficiency to determine the dosage of foam for soil conditioning (especially for coarse grained soil) in order to achieve automated decision-making for EPBM. The model is supposed to comprehensively consider geological parameters and drive parameters to improve its accuracy and application in engineering. A data-driven method integrating GBDT with random forest algorithm is proposed to establish the model. Sections 2 determines foam as the additive for soil conditioning by analyzing the grain-size distribution of the excavated material, and obtains target parameters. Section 3 makes the feature selection of decision parameters based on GBDT and correlation analysis, and three geological parameters sets are obtained. A decision-making method is presented in Section 4 based on random forest algorithm, and three decision-making models are obtained by Bayesian optimization according to the geological parameters sets. Finally, in Section 5, the optimal model is selected using fitting accuracy analysis. And the trend how the decision parameters affect the target parameters in the optimal model is analyzed, revealing the mechanism of soil conditioning to some extent.

2. Background

2.1. Previous researches about soil conditioning

Recent theories and technologies about soil conditioning can be divided into two: Method based on experiment and method based on data driving. Method based on experiment refers to sampling the soil at regular intervals in the work area before construction, and testing the influence of the amount of additive injected on the soil properties, finally determining the optimal dosage of the additive on each type of soil, as a reference for shield construction. This kind of method shows high retrieval efficiency because of structured domain knowledge. So far, many soil conditioning plans based on specific geological conditions such as water-rich sandy soil, loose sandy-clay have been proposed [15, 22]. The most reasonable soil conditioning plan can be obtained by experiments. The disadvantage is that because the sample cannot cover all geological conditions, a reasonable plan cannot be obtained for uncovered geological conditions since the samples cannot cover all geological conditions. Meanwhile, the influence of drive parameters on the dosage of the additive cannot be obtained through experiments, thus reducing the practicality of the experimental results.

Methods based on data driving refer to mining the data correlation of relevant parameters for soil conditioning based on historical construction data, and establishing a reasonable decision-making model to predict the dosage of certain additive. Unlike methods based on experiment, the decision-making model is used to predict the dosage of the additive for new geological conditions, thereby providing a reasonable plan based on data analysis. And the efficiency is significantly better than experiment. Moreover, the model can cover the drive parameters of EPBM and is more adjustable. Some researchers established mapping models between drive parameters and the dosage of certain additive for specific geological conditions [20]. Others studied mapping models between specific geological parameters and the dosage of certain additive [9, 18]. However, there is a lack of comprehensive consideration of both drive parameters and geological parameters in the decision-making model. This leads to the weak generalization ability to new geological conditions and low practical value.

2.2. The importance of automated decision-making for soil conditioning

As mentioned in Section 1, EPBM is gradually used in areas with complex and changeable geological conditions such as plateaus and flood plains. The conditions of these areas are particularly complex, and the stratigraphic section changes greatly every less than 10 meters. If the experimental method is used to determine the soil conditioning plan, it is necessary to sample the soil on site at short intervals. However, with such high-density sampling, the time and resource costs for sampling and testing are too high, and in many cases cannot meet the construction schedule requirements and cost constraints. If a low sampling density is adopted, a plan with similar geological conditions can only be used when geological conditions that have not been sampled occur, with poor conditioning effect [2].

With the improvement of the automation level of the foam injection system in the EPBM and the advancement of the geological prediction technology for a certain distance in front of the working face, the automated decision-making for soil conditioning becomes possible [14, 24]. Depending on the historical data of soil condition, a data-driven method is used to establish a high-precision correlation model between related parameters such as geological conditions and the dosage of additive. Based on this model and geological prediction technology, the dosage of additive under different working conditions is predicted and fed back to the foam injection system to realize the automation of soil conditioning decision-making.

The significance of the above-mentioned automation process for soil conditioning decision-making is as follows. Firstly, the model realizes high-precision decision-making within the coverage of historical data. And it can provide a relatively reasonable plan for uncovered geological conditions since the model mines the internal correlation of the relevant parameters of soil conditioning. Besides, the model can be continuously updated based on new data to improve its scope of application to working conditions. Secondly, the efficiency of this method is extremely high, and the resource cost is very low compared with experiment. Thirdly, the automation of soil conditioning decision-making realizes real-time decision-making with high accuracy for the dosage of foam under changeable geological conditions, thereby expanding the application conditions of the EPBM and improving the efficiency of shield construction.

2.3. Background of the project

The data source in this research is the construction data of the EPBM in the section from Wanqingsha Station to Hengli Station (W-H Section) of Guangzhou Rail Transit Line 18 in China. The total length of this section is around 2.4km, and the buried depth of the tunnel is 24.32-44.17m. And in terms of geological structure, this section covers over 7 types of rock stratums mainly including mucky medium coarse sand, medium coarse sand and fine sand. The grain-size distribution curves of the key soil sampling points are shown in Fig. 1.

As can be seen from Fig. 1, the main component of the excavated material is sand. And most of the soil has the grain size greater than 0.075mm, which is classified as coarse-grained soil. Previous researches and experiences show that foam applies to coarse grained soils [1]. Besides, foam is cost-effective, and the construction side has experience in using it. For all these reasons, foam was chosen as the additive for soil conditioning in this project.

As for the equipment, the EPBM with two screw conveyors was chosen. The shield machine has 12 foaming lines, corresponding to 12 foaming agent injection ports on the cutter head, to increase the flow rate of additives. Fig. 2 shows the selected EPBM in this project.

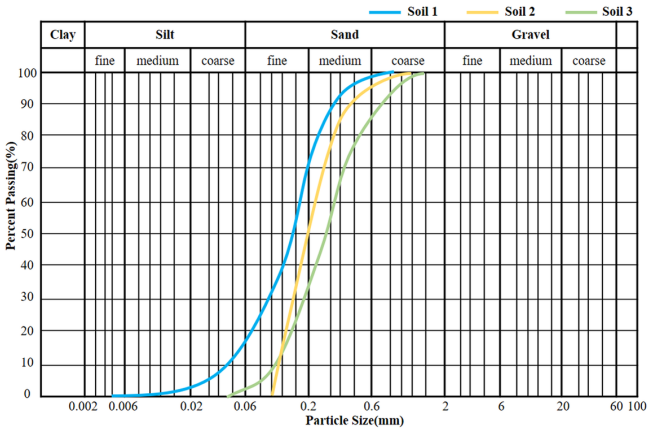


Fig. 1. The grain-size distribution curves of main soils

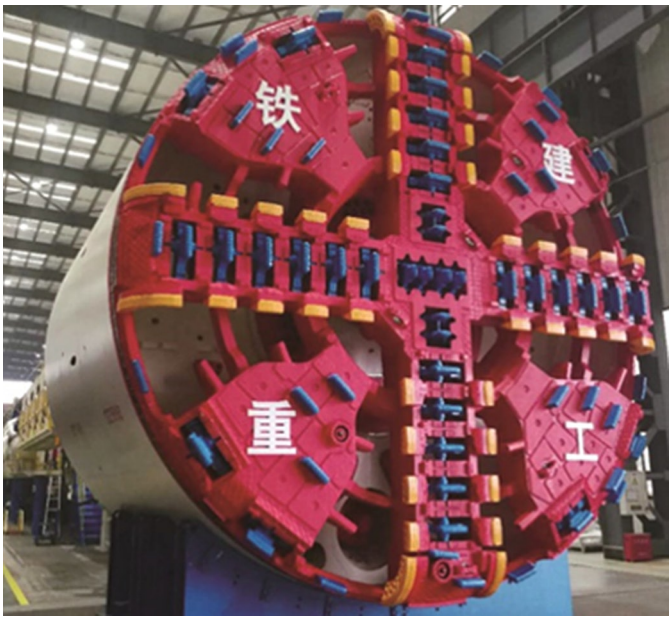


Fig. 2. The selected EPBM used in this project

2.4. Description of relevant parameters

Relevant parameters can be divided into two: target parameters and decisionrefer to the dosage parameters of foam, as the output of the decision-making model. Decision parameters have an important influence on the target parameters, as the input of the model.

Target parameters mainly depends on three indices: concentration of foaming agent within foaming liquid (c_f), foam expansion ratio (FER) and foam injection ratio (FIR) [28]. The equations of above indices are shown in Eq. (1), Eq. (2) and Eq. (3) [12]:

$$c_f = \frac{Q_f}{Q_L}, \tag{1}$$

$$\mathrm{FER} = \frac{Q_L + Q_A}{Q_L}, \tag{2}$$

$$\mathrm{FIR} = \frac{Q_L + Q_A}{Q_S}, \tag{3}$$

where Q_f is flow rate of foaming agent (m^3/min), Q_L is flow rate of foaming liquid (m^3/min), Q_A is flow rate of air (m^3/min), Q_S is flow rate of excavated soil (m^3/min).

According to above equations, the target parameters that determine the dosage of foaming agent are Q_f, Q_L, Q_A and Q_S. Methods for determining the decision parameters are presented in Section 3.

2.5. Data preparation

In this project, the data of drive parameters and target parameters can be obtained from the Database for EPBM Condition Monitoring System. And the source of the geological data is the geological report recorded by the operators. In fact, the operators collect soil every ring (Ring), then analyze the composition of soil and record it. A ring is completed when the segments are assembled in a circle for the EPBM. Before data analysis, the data from above database need to be processed in Rings to correspond to geological data. The data of poor effect for soil conditioning in each Ring, as well as the non-work data are eliminated to avoid the impact of bad data on the accuracy of the decision-making model. And the remaining data is averaged in Rings.

As for the geological data, according to the geological report, there are 12 key geological parameters, including plasticity index, silt particle content, void ratio, moisture content etc. When processing geological data in Rings, the matter is how to determine the value of above geological parameters since the cutterhead excavates several kinds of stratums. This paper provides a feasible plan as follows:

I Identify all strata, query the values of above geological parameters of each stratum, and sort them into the standard values matrix of geological parameters, denoted as M, as shown in Table 1.

II Calculate the proportion of each stratum in the n^{th} ring (Ring n), and sort them into the proportion matrix of each stratum, denoted as P_n.

III Calculate the actual values matrix of geological parameters in Ring n, denoted as D_n. The equation is as follows:

$$D_n = P_n \times M \ . \tag{4}$$

Specifically, the actual value of geological parameters is the weighted average of the standard values in each stratum, with the weight been P_n.

3. Methods for determining the decision parameters

The decision parameters can be divided into two based on physical interpretation: geological parameters and drive parameters. Feature selection is carried out to screen key factors.

3.1. Geological parameters

The geological parameters can be obtained in Section 2.5, including 12 parameters such as plasticity index, silt particle content, moisture content. Feature evaluation and selection is made next to eliminate irrelevant parameters.

(1) Methods of feature selection

The feature selection approach mainly includes filter and embedding method. The filtering method scores features according to indicators such as divergence and relevance, then sets thresholds to filter features. This method has a low computational cost, but it takes less consideration of target parameters and subsequent learners. As for embedding method, there are two steps. First, an efficient algorithm is chosen to train the model, and the weight coefficient of each feature is obtained for feature selection. Second, an algorithm with similar principle is selected to further optimize the model to get higher accuracy [31]. The reason is that theory suggests that the models trained by the algorithms with similar principle can be transferred mutually better [20]. Considering the computational cost and model accuracy, the embedding method is adopted for feature selection in this research.

Table 1. The standard values matrix of 12 key geological parameters in each stratum

Stratum	A	B	C	D	E	F	G	H	I	J	K	L
1	15	15.84	0	176.0	63.8	1.58	1.708	2.28	1.48	1.89	5.7	6.0
2	0	0	35.3	200	20	1.85	0.6	0	0.27	7	0	27.0
3	9	0	80	240.7	22	1.90	0.6	0	0.27	7	0	32.0
4	0	11.20	60	194.5	28.3	1.92	0.814	0.85	0.44	4.5	16.0	13.5
5	0	11.32	50.3	282.7	27.5	1.91	0.809	0.2	0.44	4.29	20.0	19.3
6	0	11.42	50.4	382.5	26.3	1.94	0.770	0.02	0.40	4.65	23.1	20.5
7	0	12.07	50.7	488.9	23.4	1.98	0.69	-0.09	0.38	4.73	24.3	20.2

A: clay particle content (%); B. plasticity index (%); C. silt particle content (%); D. shear wave velocity (m/s); E. moisture content (%); F. wet density (g/cm^3); G. void ratio; H. liquidity index (%); I. coefficient of compressibility; J. modulus of compressibility (MPa); K. cohesion (kPa); L. friction angle (^{o})

1: mucky medium coarse sand; 2: silty fine sand; 3: silt; 4: silty clay; 5: medium coarse sand;

6: coarse sand; 7: weathered granite

Algorithms based on decision tree are the most commonly used for feature selection. While the leaf nodes of the tree are divided, the importance of features is given through indicators such as information entropy, which naturally generates evaluation mechanism for feature selection. As an efficient algorithm based on decision tree, Gradient Boosting Decision Tree (GBDT) is employed for feature selection. The essence of GBDT is to generate Cart Tree and calculate the reduction of weighted impurity of all non-leaf nodes during splitting [8]. The larger the reduction, the more important the feature.

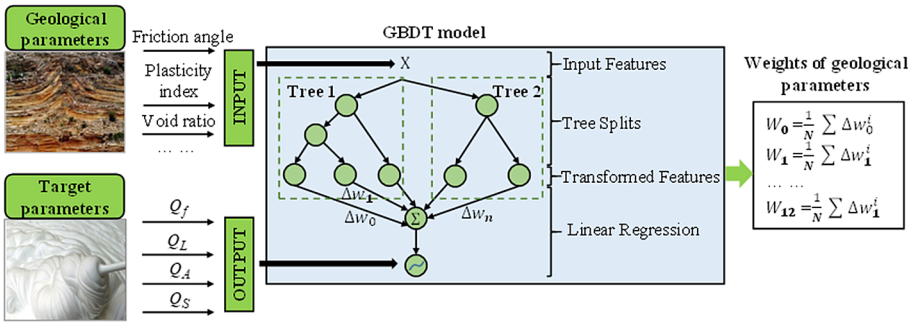


Fig. 3. The feature selection process for geological parameters based on GBDT

(2) Feature selection experiment based on GBDT

The feature selection process for geological parameters is shown in Fig. 3.

Step I. Take 12 geological parameters as input and 4 target parameters (Q_f, Q_L, Q_A and Q_S) as output.

Step II. Use GBDT algorithm to establish a regression model and adjust values of the model's key hyper-parameters to improve accuracy.

Step III. After determining a better model, output the weights of the geological parameters in the model, and then sort the weights to get key features.

A pre-experiment is carried out to determine the value of key parameters for the GBDT model in Step II. The result is that the accuracy is high when the number of estimators is 100 and the learning rate is 0.1 in the model.

(3) Analysis of experimental results

The GBDT model is repeatedly trained 100 times, and then the results are averaged to obtain the weights on 4 target parameters. The comprehensive weights of geological parameters are calculated as follows. Since the importance of all target parameters are equal, the above data is averaged with equal weights. The result is shown in Fig. 4. The conclusion is as below.

I The mean (M_e) of the comprehensive weights is greater than the median (M_o), which belongs to the right-skewed distribution with $P(X>M_o)>P(X>M_e)$. It means that more parameters are greater than M_o. For feature selection, it is necessary to cover as many effective features as possible. M_o is selected as the threshold for extracting effective features.

II The key parameters whose comprehensive weights are higher than M_o are determined: plasticity index, silt particle content, soil shear wave velocity, compressive modulus, cohesion, and friction angle. The sum of these parameters' weights is more than 0.8, indicating that the 80% of the influence of geological parameters on target parameters can be explained by them.

III There are 3 alternative parameters whose sum of weights are around 0.15.

According to the weights, three geological parameters sets are determined as the input geological parameters when modeling. The results are in Table 2.

3.2. Drive parameters

Some drive parameters have great influence on the dosage of foam. After literature analysis and interviews with experienced drivers, it was found that the soil conditioning process of the EPBM is mainly related to the cutterhead and screw conveyor. The cutterhead is responsible for excavation, and the screw conveyor is responsible for conveying the excavated mate-

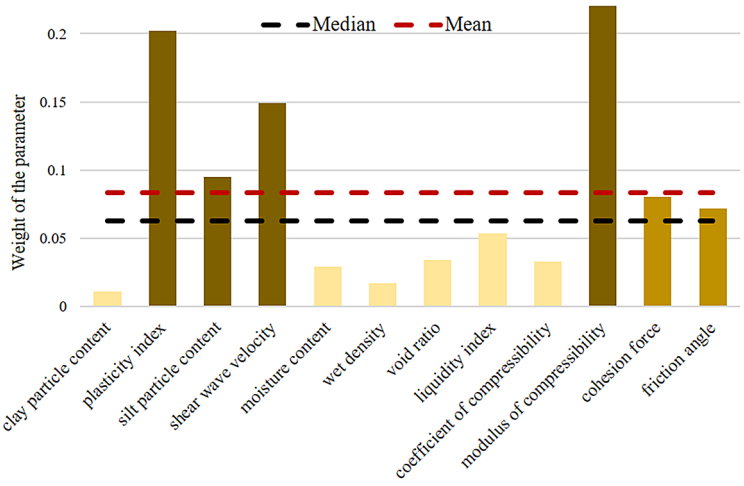


Fig. 4. Comprehensive weights of geological parameters on target parameters

Table 2. Three geological parameters sets according to the weights on the target parameters

Parameter Set	Parameter	Remarks	Sum of weights
1	shear wave velocity, friction angle, modulus of compressibility, plasticity index, cohesion force, silt particle content.	Significant for at least 2 target parameters	0.8
2	Set 1, void ratio, liquidity index, coefficient of compressibility.	Significant for at least 1 target parameters	0.95
3	Set 2, clay particle content, wet density, moisture content.	Without feature selection	1

rial. Specifically, there are 4 parameters: the cutterhead torque, the screw conveyor torque (screw torque), the screw conveyor pressure (screw pressure), and earth pressure around the screw conveyor (earth pressure) [26, 30].

(1) Data standardization

In order to avoid the influence of the magnitude difference of parameters on following analysis, the raw data is standardized by Z-score before modeling. The equation of certain parameter (x) is represented in Eq. (5):

$$Z_i = \frac{x_i - \mu}{\sigma}, \tag{5}$$

where x_i is the i^{th} data of x; μ is the mean of x; σ is the standard deviation of x; Z_i is the standardized value of x_i.

(2) Correlation analysis of drive parameters

The Pearson correlation coefficient analysis is used to exam the linear correlation of drive parameters. One of the parameters with strong correlation is eliminated to reduce input variables. The evaluation index is Pearson Correlation Coefficient (p). Table 3 presents the results in detail. It is clear that screw pressure and screw torque has a very strong linear relationship with $p = 0.966$. And this conclusion is significant at the error rate level $\alpha = 0.01$ based on hypothesis test.

(3) Regression analysis between the screw pressure and screw torque

Regression analysis is carried out to further test the linear correlation between the screw pressure and screw torque. A linear regression model is established with the screw torque as input and screw pressure as output. The equation is as follows:

$$P_{screw} = 0.9655 \times T_{screw}, \tag{6}$$

where P_{screw} is the screw pressure and in bar; T_{screw} is the screw torque and in kN·m. And Fig. 5 represents the fitting curve of this model.

The evaluation of the linear model is shown in Table 4. In terms of accuracy evaluation, the index is coefficient of determination (R^2), which represents the extent to which independent variables explain dependent variables in the model. It can be seen that the screw torque can explain 93% of screw pressure. And it proves that the model is good because the accuracy is considered high with R^2>0.85.

The validity evaluation of the model can be measured by F-test and t-test. F-test evaluates whether the model is statistically

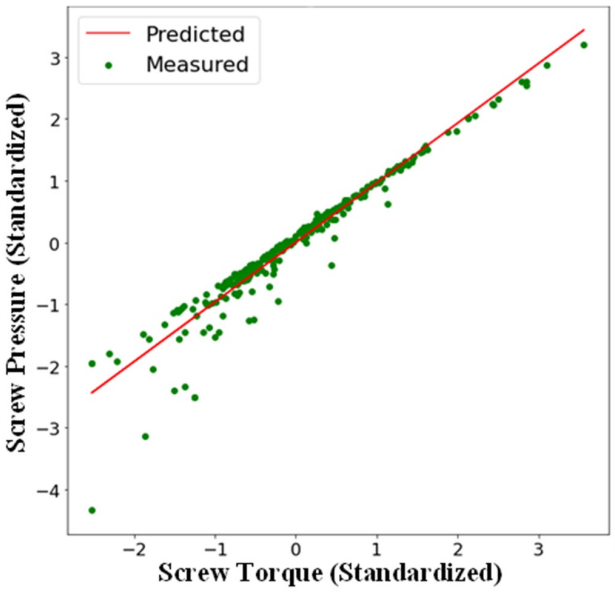


Fig. 5. The fitting curve of the linear regression model between screw pressure and screw torque

significant for all data. This model is statistically significant with a confidence of 99% since Sig. < 0.001. Moreover, t-test measures the significance of the independent variables on the model. It can be seen that the screw torque has a significant impact on the model with a confidence of 99%.

Table 4. The result of fitting accuracy evaluation and validity evaluation on the linear regression model between screw pressure and screw torque

Accuracy	F-test		t-test		
R^2	F	Sig.	B	t	Sig.
0.932	3742.836	.000	0.9655	61.179	.000
a. Independent variable: Screw torque, Dependent variable: Screw pressure					
b. Sig. is short for significance of the test; B is the coefficient of the linear model.					

Table 3. The result of Pearson correlation coefficient analysis for drive parameters

		Cutterhead Torque	Screw Pressure	Earth Pressure	Screw Torque
Cutterhead Torque	p	1	−.031	.047	−.040
	Significance	\	.613	.440	.514
Screw Pressure	p	−.031	1	−.422**	.966**
	Significance	.613	\	.000	.000
Earth Pressure	p	.047	−.422**	1	−.469**
	Significance	.440	.000	\	.000
Screw Torque	p	−.040	.966**	−.469**	1
	Significance	.514	.000	.000	\
***. At 0.01 level (two sided), it is significant.*					

To sum up, the screw pressure and screw torque has strong linear correlation, and the screw pressure is eliminated from drive parameters.

4. Methods for the establishment of soil conditioning decision-making model

GBDT algorithm is used to complete the feature selection of geological parameters in Section 3. According to the principle of embedding methods, an algorithm based on decision tree should be selected to establish the decision-making model in order to ensure the effectiveness of the selected features. The representative methods are Adaboost, random forest, GBDT, etc. [5, 21]. Random forest (RF) is employed to establish the decision-making model in present research.

Random Forest is an ensemble learning algorithm. This algorithm builds Bagging ensemble with the decision tree as the base learner, and introduces a random strategy for feature selection when training the tree [23]. The advantages are low computational cost, and good generalization ability. The main reason is that the diversity of base learners comes from sample disturbance and attribute disturbance, which improves the differences among individual learners [4]. The process of using random forest to establish a soil conditioning decision-making model is shown in Fig. 6.

4.1. Before modeling

(1) Data preprocessing

Although decision tree-based models are not sensitive to the magnitude of feature values, the raw data is standardized by Z-score before modeling according to Eq. (5). This process can exclude the influence of irrelevant factors on the performance comparison of the models in later section since the comparison models' accuracy and some accuracy indexes may be affected by the magnitude difference of parameters. 80% of total samples were randomly selected as training set and 20% as test set.

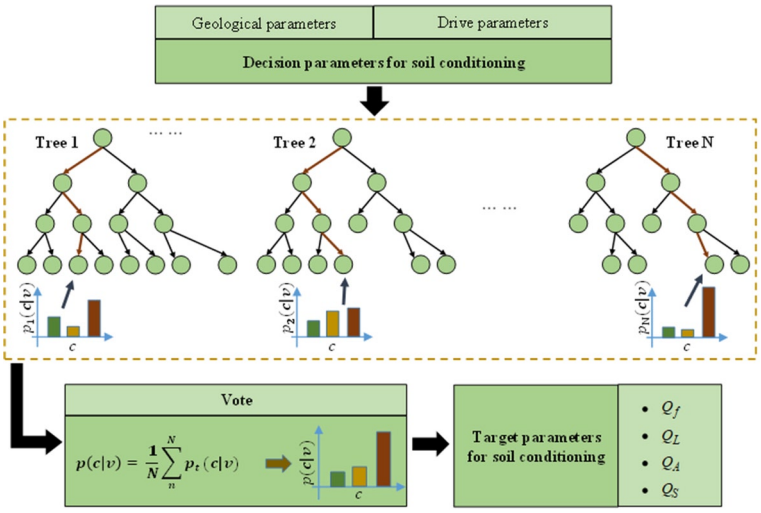


Fig. 6. The process of using random forest to establish a soil conditioning decision-making model

(2) Input and output parameters

The target parameters are taken as output of the model: Q_f, Q_L, Q_A and Q_S. The decision parameters are taken as input when modeling, including 3 drive parameters and some geological parameters. According to conclusions in Section 3.1, 3 geological parameters sets are considered when choosing the geological parameters.

Model 1: Take the geological parameters of top 6 on the weight as input parameters, assign weights when modeling, corresponding to Parameter Set 1 in Table 2.

Model 2: Take the geological parameters of top 9 on the weight as input parameters, assign weights and modeling, corresponding to Parameter Set 2 in Table 2.

Model 3: Take all the geological parameters without feature selection as input parameters, corresponding to Parameter Set 3 in Table 2.

4.2. Hyper-parameter optimization for the models

Hyper-parameters are the framework parameters of the model, and the reasonable values of hyper-parameters can greatly improve the fitting accuracy and generalization ability. Common methods for adjust hyper-parameters include grid search, random search, and Bayesian optimization. Grid search method is improper for this research due to the high calculation time and combinatorial explosion [19]. Besides, the optimization results of random search are unstable and unreliable [3]. By contrast, Bayesian optimization adjusts hyper-parameters of the model automatically. The principle is that an objective function is set first, then minimizes it using the probability model based on past evaluation result, hereby determine the optimal value of the hyper-parameters [29]. The superiority is that it will consult previous evaluation results when trying the next set of hyper-parameters, which improves the efficiency and has stable results. Hence Bayesian optimization is used to optimize the key hyper-parameters of the models which have great impact on the accuracy of the model. The key hyper-parameters are extracted based on the principle of RF including maximum number of features in decision tree (max features), the number of decision trees (tree number), the maximum depth of the decision tree (max depth), and the minimum number of samples for leaf nodes (min samples).

(1) Determining the objective function

The objective function is the evaluation index when adjusting hyper-parameters by Bayesian optimization, and it refers to the model's accuracy in this research. Specifically, the prediction error (RMSE) of the model with k-fold cross-validation is adopted as the objective function with k = 10.

(2) Setting the domain space

Domain space refers to the value range of each hyper-parameter. The method selects a set of hyper-parameters for the model from the domain space according to the probability distribution of each parameter and evaluate the accuracy while iterating [34]. Therefore, it is necessary to set the sampling probability distribution pattern for the hyper-parameters. The determination of the distribution pattern usually requires comprehensive consideration of the data type, value range, and empirical rules. The results are shown in Table 5. It is generally believed that when tree number is large enough, a small increment will not cause a large change in model performance. So logarithmic uniform distribution is used for sampling in order to reduce the calculation cost. And considering the fewer optional ranges, discrete uniform distribution is used for other hyper-parameters to cover as many values as possible. The kernel density estimation of the distributions for the hyper-parameters is shown in Fig. 7.

(3) Setting optimization algorithm

The common algorithms for Bayesian optimization include sequential model-based optimization (SMBO), Tree Parzen Estimator (TPE), etc. [7]. The TPE algorithm, which has been proven to perform well in accuracy and computational efficiency, is employed in this research [32].

(4) Results of hyper-parameter optimization

The hyper-parameter optimization of 3 models in Section 4.1 are performed following steps above. The optimization is repeated 100 times, and the results are averaged. The values of the hyper-parameters for the optimal models are shown in Table 6. The following conclusions can be drawn:

I The results of max features and max depth are the same for 3 optimal models. When max features are Auto (referring to selecting all features), and max depth is -1 (referring to full-grown), the prediction accuracy of 3 models is best.

II The tree number decreases, as the number of geological parameters increase. The reason might be that the more input parameters,

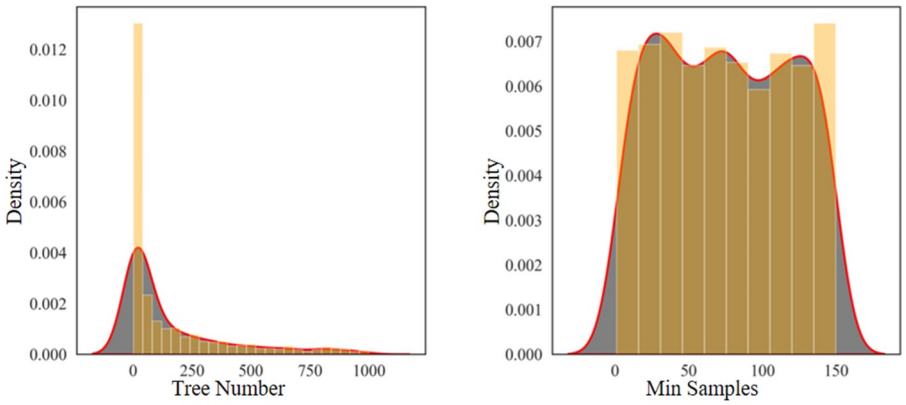


Fig. 7. The kernel density estimation of the distributions for the hyper-parameters

Table 5. Sampling probability distribution pattern for each hyper-parameter

Hyper-parameter	Description	Value Range	probability distribution pattern
Max Features	maximum number of features in decision tree when splitting	Auto, sqrt, log2	discrete uniform distribution
Tree Number	The number of trees in the forest	[1, 1000]	logarithmic uniform distribution
Max Depth	maximum depth of the decision tree	[-1, 20], where -1 means full-grown	discrete uniform distribution
Min Samples	minimum number of samples for leaf nodes	[1, 150]	discrete uniform distribution

Table 6. Results of hyper-parameter optimization for 3 models

Model	Max Features	Tree Number	Max Depth	Min Samples	Remarks (number of geological parameters)
Model 1	Auto	37	-1	3	6
Model 2	Auto	25	-1	4	9
Model 3	Auto	19	-1	2	12

the stronger the model's fitting ability to training set, causing fewer decision trees required to achieve the same accuracy.

5. Results and model interpretation

5.1. Performance measurement and evaluation of the models

There are two part for performance measurement of the models. On one hand, fitting accuracy analysis is performed on 3 models to select the optimal model. On the other hand, some artificial intelligence algorithms are used to establish decision-making models. Compare the prediction accuracy of above models to prove the advantages of the optimal model.

5.1.1. Fitting accuracy analysis of the models

The common evaluation indexes of the fitting accuracy are R^2 and RMSE, as mentioned above. The two indexes have different changing trends, and there are also differences between training set and test set. The TOPSIS method is used to integrate R^2 and RMSE to evaluate the fitting accuracy comprehensively [16]. Firstly, the original evaluation results are normalized based on unified standards to obtain an ideal solution. Then the distance between each evaluation object and the ideal solution (Distance) is calculated as a comprehensive index. The analysis results of the fitting accuracy are shown in Table 7. And following conclusions can be obtained.

I Model 2 is the best in fitting accuracy of all. Concretely, the Distance is the smallest, and R^2 on both training set and test set is high, and the RMSE is low, indicating robustness in the model.

II The fitting accuracy of Model 1 is lower than Model 3 on the training set, but better on the test set, showing stronger generalization ability but lower accuracy.

III Model 3 has the highest fitting accuracy on the training set, but the lowest on the test set, indicating that the model has learned the noise of the training set, which leads to decrease in generalization ability.

To visually analyze the prediction error, the models are applied to several Rings of the dataset to compare the predicted value and the measured value. Consecutive 100 Rings with continuous data are selected for testing. And Q_f is chosen as the target parameter to plot, as shown in Fig. 8. It shows that the fluctuation of error in Model 1 is stable, but there is always a large distance between the predicted value and the measured value, indicating that the mean of the error is high. Model 2 fits well with small and stable error. Model 3 fits well on some data, but poorly on the others with even completely opposite trends, showing very unstable predicted values.

With all the analysis, Model 2 is the optimal model, which has the best fitting accuracy.

5.1.2. Performance comparison with other models

A variety of common multi-output regression algorithms are employed to establish decision-making models to prove the advantages of Model 2. The models are obtained by 10-fold cross-validation on the training set, and the hyper-parameters are as follows. In the k-Nearest Neighbors

Table 7. The analysis results of the fitting accuracy for 3 models

Model	Training Set		Test Set		Distance
	R^2	RMSE	R^2	RMSE	
Model 1	0.8232	0.1832	0.7964	0.2052	0.1165
Model 2	0.8855	0.1487	0.8540	0.1588	0.0189
Model 3	0.9011	0.1380	0.7231	0.2841	0.1812

(KNN) model, the number of neighbors is 10. In the BPNN model, the number of hidden layer nodes is 20. In the SVR model, RBF is selected as the kernel function. In the remaining models which are based on decision tree, the number of trees is 25, the minimum number of samples for leaf nodes is 4, and the maximum depth of the tree is full-grown. These models are trained and tested 100 times on the dataset respectively, and the performances are evaluated by the mean of R^2 and *RMSE*. The results on the test set are shown in Table 8. Besides, 0 visually shows fitting effect between the predicted values and the measured values of the standardized Q_f. And the following conclusions can be drawn:

I Poor fitting accuracy. The R^2 of models (a), (b), (c) and (d) is less than 0.7, indicating that the models have poor accuracy on the test set. The reason may be that the learning ability of the model is too weak, such as the Multiple Linear model, or the learned noise leads to overfitting, such as the BPNN model.

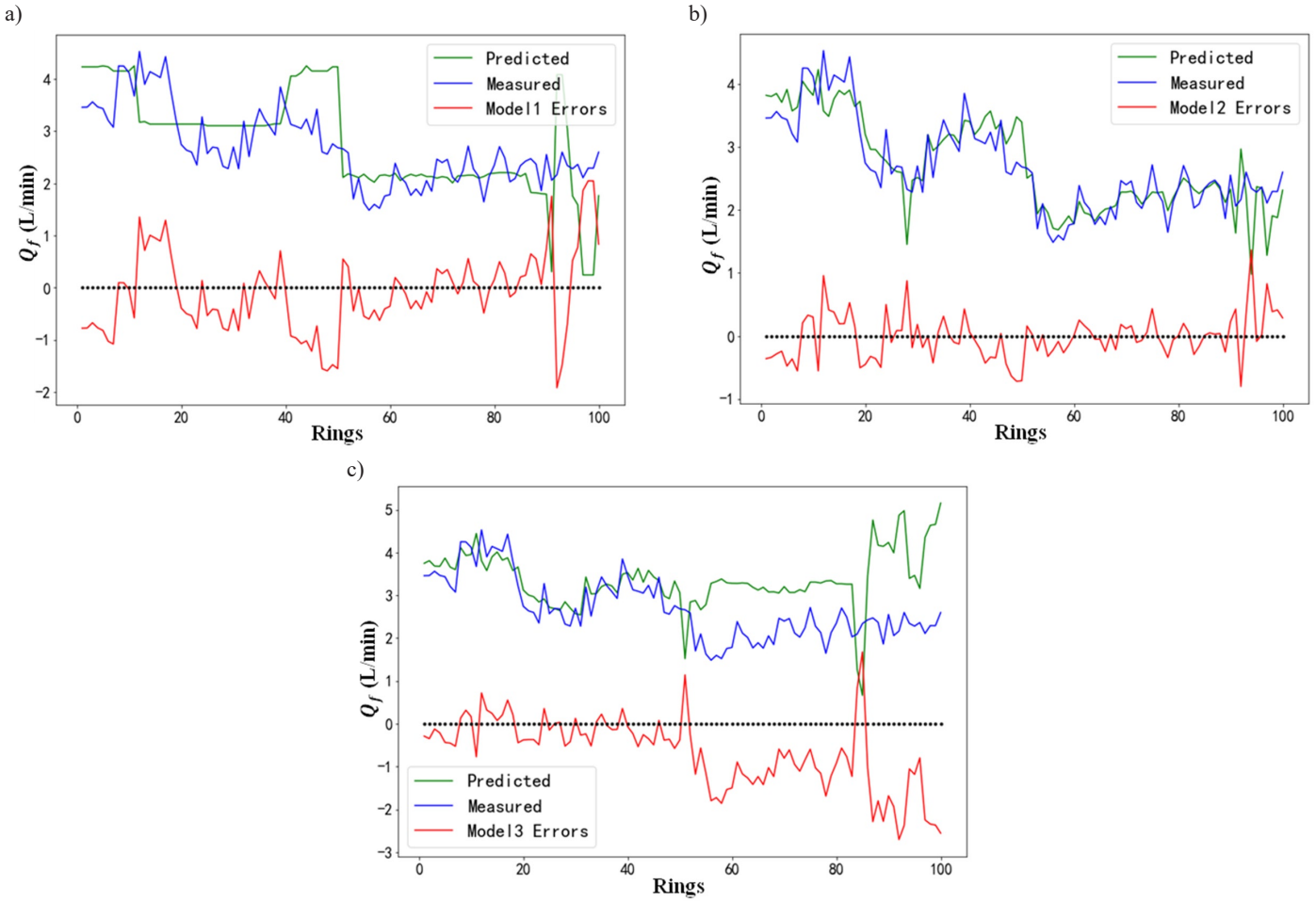


Fig. 8. Comparison between measured value and predicted value of three decision-making models: (a) Model 1: geological parameters of top 6 on the weight as input; (b) Model 2: geological parameters of top 9 on the weight as input; (c) Model 3: all geological parameters as input

Table 8. Performance comparison of soil conditioning decision-making models based on different regression algorithms

Algorithm	R^2	RMSE	Distance
Multiple Linear Regression	0.5539	0.4433	0.4992
k-Nearest Neighbors Regression	0.6898	0.3083	0.3070
BPNN Regression	0.6062	0.4027	0.4253
Support Vector Regression	0.6637	0.3343	0.3439
AdaBoost Regression	0.7874	0.2112	0.1690
LightGBM Regression	0.7407	0.2473	0.2350
XGBoost Regression	0.7623	0.2362	0.2045
Random Forest Regression (Model 2)	0.9069	0.1328	0

II Good fitting accuracy. The R^2 of the four models based on decision tree is basically greater than 0.75, and *RMSE* is small, indicating that the models fit the test data well. Among them, the Random Forest model (Model 2) is proved to be the most reasonable model with the Distance being 0. It also proves that, when using decision tree-based algorithms for feature selection, the accuracy of models based on the same principles is higher than that of models based on other algorithms.

5.2. Model interpretation

The change trend of output variables with input variables based on Model 2 is explored to further reveal the influence of the decision parameters on the target parameters. Control variable method is employed, and Q_f is chosen as the representative target parameter. The experimental results are shown in Table 9. And the following conclusions can be drawn.

I The trend how input variables affect foaming agent flow can be divided into 4 categories. Category 1 is a positive parabola, that is, with the increase of the input variable, Q_f decreases first and then increases. By contrast, Category 2 is a negative parabola, as the input variable increases, Q_f increases first and then decreases. Category 3 is a positive correlation curve, which means Q_f increases with the increase of the input variable. Finally, Category 4 is a negative correlation curve, Q_f decreases with the increase of the input variable.

II The slope of the curve represents the degree of influence of the decision parameter on the target parameter. The curve slope of drive parameters such as the cutterhead torque and screw torque

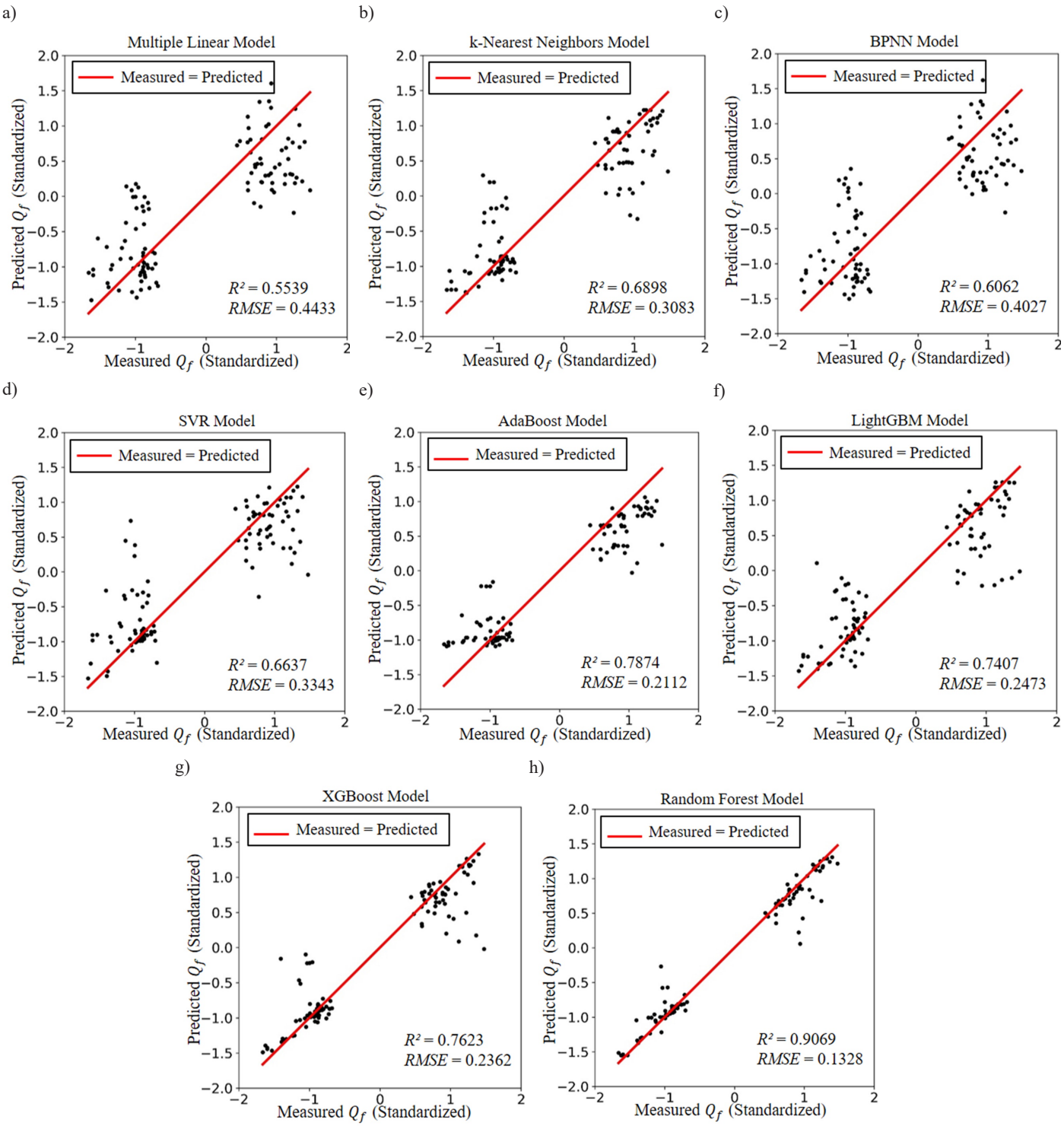


Fig. 9. Fitting results of decision-making models using a variety of multi-output regression algorithms

is obviously greater than the slope of geological parameters such as plasticity index. It indicates that the former has a greater effect on soil conditioning decision-making.

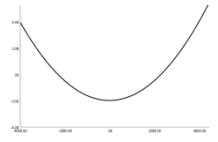
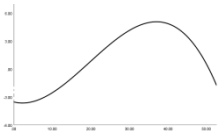
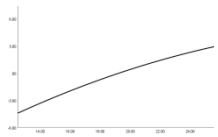
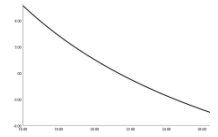
6. Conclusion

In order to solve the problems that traditional soil conditioning decision-making methods for the EPBM perform poor when applied to uncovered geological conditions, don't take drive parameters into consideration, and are too low in efficiency to do on-site decision-making, a method based on intelligent algorithms to predict the dosage of foam for automated decision-making for soil conditioning is studied. The following conclusions can be drawn.

I Based on the data in W-H Section, the target parameters and decision parameters for decision-making model are determined. And decision parameters include geological parameters and drive parameters. And it is the first time that drive parameters are considered in the decision-making for soil conditioning, which improves the practicability of the model in engineering.

II Taking GBDT, an efficient algorithm based on decision tree, as the feature selection method, three geological parameters sets are determined as the input of the decision-making model. The screw pressure as a drive parameter is filtered by correlation analysis since the results show it can be expressed by the screw torque.

Table 9. The trends how decision parameters affect Q_f

Category	Curve Type	Diagram	Trend	Decision Parameters
1	Positive Parabola		decreases first and then increases	cutterhead torque, earth pressure, shear wave velocity
2	Negative Parabola		increases first and then decreases	the screw torque
3	positive correlation curve		continuously increases	plasticity index, silt particle content, cohesion force
4	negative correlation curve		continuously decreases	modulus of compressibility, friction angle

III The random forest, also an algorithm based on decision tree, is used to establish three decision-making models differentiated by the geological parameters sets. And the optimal models are determined through Bayesian optimization to optimize the hyper-parameters. And Model 2 (with 9 geological parameters) is the best model based on the results of fitting accuracy analysis compared with the models based on five regression algorithms. And it also proves that, when using decision tree-based algorithms for feature selection, the accuracy of models based on the same principles is higher than that of models based on other algorithms.

IV The trends how input variables affect foaming agent flow can be divided into 4 categories: positive parabola, negative parabola, positive correlation curve and negative correlation curve. And the curve slope shows drive parameters have greater effect on soil conditioning decision-making than geological parameters.

To sum up, a hybrid method of GBDT and random forest algorithm for soil conditioning decision-making using foam is proposed, which presents a new idea for automated decision-making for soil conditioning. The model comprehensively considers the influence of geological parameters and drive parameters on the dosage of foam, thereby improving the adaptability to new geological conditions and engineering practicability compared with method based on experiment. The accuracy of the model is proved to be higher than traditional data-driven methods. Apparently, this method is more efficient than experiment. The method can realize real-time decision-making with high accuracy for the dosage of foam under changeable geological conditions, broaden the application conditions of the EPBM, improve the efficiency and reduce the experiment cost for soil conditioning. It must be noted that this model is based on the construction data of the EPBM, so the application area should be limited to similar engineering backgrounds.

Acknowledgement

This work was supported by National Key R&D Program of China (grant numbers 2019YFB1705203).

References

1. A CB, B MT. Application ranges of EPB shields in coarse ground based on laboratory research. Tunneling and Underground Space Technology 2015; 50: 296-304, https://doi.org/10.1016/j.tust.2015.08.006.
2. Ahmadi S, Moosazadeh S, Hajihassani M, Moomivand H, Rajaei MM. Reliability, availability and maintainability analysis of the conveyor system in mechanized tunneling. Measurement 2019; 145: 756-764, https://doi.org/10.1016/j.measurement.2019.06.009.
3. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 2012; 13: 281-305.
4. Biau G. Analysis of a Random Forests Model. Journal of Machine Learning Research 2012; 13: 1063-1095.
5. Chan J, Paelinckx D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment 2008; 112: 2999-3011, https://doi.org/10.1016/j.rse.2008.02.011.
6. Kozłowski E, Antosz K, Mazurkiewicz D, Sęp J, Żabiński T. Integrating advanced measurement and signal processing for reliability decision-making. Eksploatacja i Niezawodnosc - Maintenance and Reliability 2021; 23: 777-787, https://doi.org/10.17531/ein.2021.4.20.
7. Feurer M, Springenberg J, Hutter F. Initializing Bayesian Hyperparameter Optimization via Meta-Learning. Proceedings of the AAAI Conference on Artificial Intelligence 2015; 29(1), https://ojs.aaai.org/index.php/AAAI/article/view/9354.
8. Friedman J. Greedy function approximation: A gradient boosting machine. Annals of Statistics 2001; 29: 1189-1232, https://doi.org/10.1214/aos/1013203451.
9. Galli DIM, Thewes DIM. Investigations for the application of EPB shields in difficult grounds. Geomechanik und Tunnelbau 2014; 7: 31-44, https://doi.org/10.1002/geot.201310030.
10. Galli M, Thewes M. Rheological Characterization of Foam-Conditioned Sands in EPB Tunneling. International Journal of Civil Engineering

2018, https://doi.org/10.1007/s40999-018-0316-x.
11. Hollmann F, Thewes M. Assessment method for clay clogging and disintegration of fines in mechanised tunnelling. Tunneling and Underground Space Technology 2013; 37: 96-106, https://doi.org/10.1016/j.tust.2013.03.010.
12. Hu Q, Wang S, Qu T, et al. Effect of hydraulic gradient on the permeability characteristics of foam-conditioned sand for mechanized tunnelling. Tunneling and Underground Space Technology 2020; 99, https://doi.org/10.1016/j.tust.2020.103377.
13. Jerbi W, Ben Brahim A, Essoussi N. A Hybrid Embedded-Filter Method for Improving Feature Selection Stability of Random Forests. In. Cham 2017: 370-379, https://doi.org/10.1007/978-3-319-52941-7_37.
14. Jianjun Z, Diming C, Dongyuan W, Lu-Lu Z, Li-Min Z. Failure Probability of Transverse Surface Settlement Induced by EPB Shield Tunneling in Clayey Soils. Asce Asme Journal of Risk & Uncertainty in Engineering Systems Part a Civil Engineering 2018; 4: 4018030, https://doi.org/10.1061/AJRUA6.0000981.
15. Kim TH, Kim BK, Lee KH, Lee IM. Soil Conditioning of Weathered Granite Soil used for EPB Shield TBM: A Laboratory Scale Study. KSCE Journal of Civil Engineering 2019; 23: 1829-1838, https://doi.org/10.1007/s12205-019-1484-1.
16. Opricovic S, Tzeng G. Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. European Journal of Operational Research 2004; 156: 445-455, https://doi.org/10.1016/S0377-2217(03)00020-1.
17. Peila D. Soil conditioning for EPB shield tunnelling. KSCE Journal of Civil Engineering 2014; 18: 831-836, https://doi.org/10.1007/s12205-014-0023-3.
18. Pourmand S, Chakeri H, Sharghi M, Bonab M, Ozcelik Y. Laboratory Studies on Soil Conditioning of Sand in the Mechanized Tunneling. Journal of Testing and Evaluation 2020; 48: 3658-3672, https://doi.org/10.1520/JTE20170395.
19. Putatunda S, Rama K. A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. In. NEW YORK; 2018: 6-10, https://doi.org/10.1145/3297067.3297080.
20. Qu T, Wang S, Hu Q. Coupled Discrete Element-Finite Difference Method for Analysing Effects of Cohesionless Soil Conditioning on Tunneling Behaviour of EPB Shield. KSCE Journal of Civil Engineering 2019; 23: 4538-4552, https://doi.org/10.1007/s12205-019-0473-8.
21. Sakata R, Ohama I, Taniguchi T. An Extension of Gradient Boosted Decision Tree incorporating Statistical Tests. 2018 IEEE International Conference on Data Mining Workshops (ICDMW) 2018: https://doi.org/10.1109/ICDMW.2018.00139.
22. Selmi M, Kacem M, Jamei M, Dubujet P. Physical Foam Stability of Loose Sandy-Clay: a Porosity Role in the Conditioned Soil. Water and Soil Pollution 2020; 231, https://doi.org/10.1007/s11270-020-04598-8.
23. Scornet E. Random Forests and Kernel Methods. IEEE Transactions on Information Theory 2016; 62: 1485-1500, https://doi.org/10.1109/TIT.2016.2514489.
24. T GA, M HA. Reliability assessment of EPB tunnel-related settlement. Geomechanics and Engineering 2010; 2: 57-69, https://doi.org/10.12989/gae.2010.2.1.057.
25. Thewes M, Hollmann F. Assessment of clay soils and clay-rich rock for clogging of TBMs. Tunneling and Underground Space Technology 2016; 57: 122-128, https://doi.org/10.1016/j.tust.2016.01.010.
26. Vinai R, Oggeri C, Peila D. Soil conditioning of sand for EPB applications: A laboratory research. Tunneling and Underground Space Technology 2008; 23: 308-317, https://doi.org/10.1016/j.tust.2007.04.010.
27. Wang S, Hu Q, Wang H, Thewes M, Liu P. Permeability Characteristics of Poorly Graded Sand Conditioned with Foam in Different Conditioning States. Journal of Testing and Evaluation 2020, https://doi.org/10.1520/JTE20190539.
28. Wei Y, Yang Y, Tao M, Wang D, Jie Y. Earth pressure balance shield tunneling in sandy gravel deposits: a case study of application of soil conditioning. Bulletin of Engineering Geology and the Environment 2020; 79: 5013-5030, https://doi.org/10.1007/s10064-020-01856-1.
29. Xia Y, Liu C, Li Y, Liu N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications 2017; 78: 225-241, https://doi.org/10.1016/j.eswa.2017.02.017.
30. Xiao C, Tan L, Xia Y, et al. Excavation parameters characteristics of earth pressure balanced shield based on soil modification. Journal of Railway Science and Engineering 2017.
31. You M, Liu J, Li G, Chen Y. Embedded Feature Selection for Multi-Label Classification of Music Emotions. International Journal of Computational Intelligence Systems 2012; 5: 668-678, https://doi.org/10.1080/18756891.2012.718113.
32. Zhao M, Li J. Tuning the Hyper-parameters of CMA-ES with Tree-structured Parzen Estimators. 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI) 2018: 613-618, https://doi.org/10.1109/ICACI.2018.8377530.
33. Zheng G, Sun W, Zhang H, Zhou Y, Gao C. Tool wear condition monitoring in milling process based on data fusion enhanced long short-term memory network under different cutting conditions. Eksploatacja i Niezawodnosc - Maintenance and Reliability 2021; 23: 612-618, https://doi.org/10.17531/ein.2021.4.3.
34. Zou T, Dang W, Zhang G, Liu K, Li P. Prior distribution selection criterion in accelerated degradation. 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC) 2018: 694-698,