Civil and Environmental Engineering
Vol. 22, Issue 2, 764-783, DOI: 10.2478/cee-2026-0051

Rock mass properties prediction method in TBM tunnel based on principal component analysis (PCA)

Qingting MENG[1], Hai WANG[2], Liang BAI[1], Fei ZHAO[1], Jian BAI[2,*]

[1] Sinohydro Bureau 6 Co., Ltd., Shenyang, China.

[2] School of Water Conservancy and Environment, University of Jinan, Jinan, China.

* corresponding author: 252292645@qq.com

Date of Submission: 01 September 2025
Revision Date: 23 October 2025
Date of Acceptance: 23 October 2025

Civil and Environmental Engineering
Journal of the Faculty of Civil Engineering | University of Žilina

Abstract

Accurate prediction of the rock mass parameters is of great significance in the safety and efficiency of tunnel construction by tunnel boring machine (TBM). In this fields, machine learning method has become the main stream to build the mapping between rock mass parameters and TBM driving data and achieves good results. However, tunneling data have hundreds of features and high acquiring frequency, which results in the structural difference between tunneling data and rock mass data. Current solution aiming at these differences rely on human experience and may lose data information, or increase the model complexity and time costs. This paper proposed a pre-treated method suitable for TBM tunneling data based on principal component analysis (PCA), with can reduce the feature number of tunneling data, and simplify its structure to match the rock mass data, with low data information lost. In addition, the proposed method can match multiple machine learning method, such as back propagation neuron network (BPNN), support vector regression (SVR), and random forest (RF), the corresponding integrity prediction models of the uniaxial compressive strength and joint frequency achieve acceptable accuracy. The method is verified by totally 1272 filed samples from five different projects. The predicted mean absolute percentage errors (MAPE) of different targets and different projects are controlled within 15%, which is acceptable in actual projects. In addition, compared with model involved full tunneling variable as input, the proposed pre-treated method can reduce the time cost from 3-5 minutes to 3-6 seconds, which improves the calculation efficiency.

Keywords

Tunnel boring machine; Machine learning; Principal component analysis; Rock mass parameters.

1. Introduction

Given the high efficiency and safety of the tunnel boring machine (TBM), it has become an important and widely used equipment in tunnel excavation, especially for long and deep tunnels. To improve safety and efficiency, methods for determining reasonable operating parameters have become the object of much research. Many researchers, such as Sun et al. (2016), Xue et al. (2018), Zhang et al. (2018), and Liu et al. (2010), are proposed or developed method for choosing or optimizing the TBM tunnelling parameters. These researches have great reference value for improving the safety and efficiency of TBM tunneling. Actually, nearly all the optimizing method for TBM tunnelling parameters requires the

geological data as priori information. However, limited by the complex machine structure and narrow space of tunnels, in-field rock mass parameter testing methods can be used in tunnel environment are lack, which has become the hotpot in TBM tunnelling field. For example, Nacimipour et al. (2016) developed the rock strength boring probe (RSBP) to test the uniaxial compressive strength and tensile strength of the tunnel face. Wang et al. (2020) developed the true triaxial rock drilling test system (TRD) to test rock mass parameters under multiple rock type conditions. Goh et al. (2011) used share wave velocity by spectral analysis of surface waves (SASW) to measure the rock quality designation (RQD). Kong et al. (2018) and Liu et al. (2018) used the point-load testing results to evaluate the rock compressive strength.

In-field-testing is an effective kind of method for obtaining rock mass parameters in TBM tunnels. However, most of these methods spend time on testing or analysis process, which yields geological data with hysteresis, and it is difficult to acquire the data at the speed of TBM excavation (Wang and Zhang, 2023). To solve this problem, it is feasible to build a mapping between real-time TBM tunneling data and rock mass parameters by statistical methods, and a number of researchers made attempts on this topic. Currently, machine learning method has become the main stream of this kind of methods, with its powerful capability to solve regression problems with large quantities and strong nonlinearity of the rock mass and tunnelling data. For example, Armaghani et al. (2017), Zare et al. (2017, 2018), Mahdeveri et al. (2014), Liu et al.(2019), Yagiz et al.(2011), and Minh et al. (2017) applied artificial neural networks, particle swarm optimization, fuzzy logic, and gene expression programs to research the relationship between tunnelling data and rock mass parameters and obtained acceptable results in actual project. The mentioned researches prove that applying machine learning algorithms on establishing the mapping between tunneling data and rock mass data is feasible and effective.

Using machine learning method to solve the relationship between tunneling data and rock mass parameters are potential methods. They still have shortcomings. In actual project, the tunnelling data and rock mass data shows different volume and structure. Therefore, the data set composed of TBM tunneling data and rock mass data is always complex and difficultly used. Limited by space and testing time, rock samples are limited and each sample is always used to express the rock properties for an area with the length from tens of meters to hundreds of meters. Differently, TBM tunneling data have hundreds of features and high acquiring frequency, which results in more complex structures. Therefore, in the same tunneling area, a single combination of rock mass data corresponds to high volume tunneling data with a number of features and time-series samples (Wang et al. 2025). To simplify the structure of tunneling data, effective preprocessing and dimensionality reduction methods are necessary.

Aiming at the data structure problem, this paper proposed a pre-treated method for TBM tunnelling data based on the principal component analysis method (PCA). On this basis, a rock mass parameters prediction method based on artificial neural networks with pre-treated tunnelling data as input is proposed. Furthermore, totally 1272 rock samples come from five tunnel projects and the corresponding tunnelling data are collected to test and verify the proposed method. The results verified the proposed method has relatively high accuracy and calculation efficient, which also pointed out that the tunneling data pre-treated method is a valuable research point.

This paper is organized as follow: In the 2nd Section, the pre-treated method based on PCA is introduced. Further, data set with 1272 samples from five tunnel projects is collected, and the corresponding prediction models of rock mass parameters are established in the 3rd Section. The 4th Section introduced the pre-treated and prediction results. The 5th Section discussed the advantages and shortcomings of the method, and the 6th Section conclude this paper.

2. Methodology

2.1. Data preparation

To build the mapping for prediction rock properties, suitable data sets are necessary. According to the task of the method, the TBM tunnelling parameters are used as the input, and the rock mass parameters are the output. The acquirement and selection of these data is introduced in this section.

1. TBM tunnelling data

TBM is a huge and complex tunneling machine with many sub-systems, and each sub-system generates working parameters during operation. For instance, cutterhead generates the thrust, torque, and revolution speed, electrical system generates electric current and power, transport system generates the speed and load of the belt conveyor, and so on. Most of the current TBMs are equipped with data acquisition system, which can monitor and record about 200 working parameters as mentioned above with high frequency, always 1 record per second. Those tunneling parameters can be divided into two types, the first type is mainly human-controlled, such as the penetration rate and revolution speed, whose value is always closed to the preset value, and this type can be called as controlling parameters. The second type is affected by combined action of the controlling parameters and rock condition, such as the thrust and torque of the cutterhead, which are called the performance parameters. A single type of the tunneling data is insufficient to reflect changes in rock mass conditions and predict rock mass parameters. Therefore, the input needs to be including both controlling and performance parameters.

Selecting suitable tunnelling parameters as input is significant to the prediction accuracy of rock mass parameters. Generally, the number of input variables are less than 10 to reduce algorithm complexity and computational costs. A number of research used several parameters directly with rock-breaking process by TBM cutterhead as input, including thrust, torque, penetration rate and revolution speed et al. (Mahdeari et al. 2014, Gong and Zhao, 2009, Liu et al. 2019, Liu et al.2017). Recently, some researches used all tunnelling parameters after dimensionality reduction instead of directly selecting several variables as input to overcome the data information losses, and the prediction accuracy is improved (Wang and Zhang, 2023, Wang et al. 2025). This paper followed this idea and developed a dimensionality reduction method for tunnelling data based a principal component analysis (PCA).

2. Rock mass parameters

The strength and integrity are regarded as the most significant factors on TBM tunneling among the rock mass parameters. Rock strength includes multiple aspects, such as compressive, tensile and shear strength. There is a high correlation between the uniaxial compressive strength (UCS) to another strength parameters, and it also has a strong characterization effect on the properties of rock mass. Therefore, the UCS is widely regarded as a key factor on tunneling and involved in prediction model (Gong et al. 2020, Minh et al. 2017, Liu et al. 2018). Integrity represents the distribution of joints in the rock mass, multiple parameters such as distance between planes of weakness (DPW) (Armaghani et al. 2017), the joint orientation (α)(Gao et al. 2018), rock quality designation (RQD) (Liu et al. 2020), and joint frequency (J_f) (Wang and Zhang, 2023) are involved in researches. Among them, *DPW*, *RQD* and J_f express the frequency of joint in rock mass, and they can be converted to each other. In this paper, *UCS* and J_f are selected as the prediction target, also the output of the proposed model.

3. Dataset preparation

Generally, prepared dataset should follow “key-value” type to meet the requirement of most of machine learning algorithm. It means that the input tunneling data and output rock mass data should have same number of series and one-to-one correspondence. In actual project, rock samples are collected in several specific mileage, while the tunneling data are recorded continuously with high frequency, which results multiple series of input are corresponding to one series of output in the original dataset, and it cannot be used by machine learning algorithm directly. To solve this problem, the original “many-for-one” dataset is pre-treated in this research. In detail, the tunneling data is captured within the area of -0.5m to 0.5m at each mileage position of field test used to obtain rock mass parameters. On this basis, triple standard deviation method is used to remove abnormal samples in the multiple tunneling data samples corresponding to each rock mass sample, and the average value of each variable is calculated as the pre-treated tunneling data, as shown in Eq.1 and 2 (Wang et al. 2025, Gong et al. 2020).

$$x' \in [\mu - 3\sigma, \mu + 3\sigma] \quad (1)$$

$$X = \frac{1}{n}\sum_{i=1}^{n} x' \quad (2)$$

Where:

x' - a tunneling variable obtained at a certain moment, which is regarded as the normal data and reserved,

μ and σ - the average value and standard deviation of the variable corresponding to a series of rock mass data,

n - the number of the reserved samples,

X - the pre-treated tunneling data used to establish the prediction model for rock mass parameters.

2.2. PCA principle

According to Section 2.1, there are about 200 variables of tunneling data which requires dimensionality reduction to retain as much information as possible while reducing the algorithm complexity. In this paper, principal component analysis (PCA) is used conduct the dimensionality reduction of tunneling data.

The basic principle of PCA is to reorganize original variables to form new and independent variables through orthogonal transformation (Majdi and Beiki, 2019). Before conduct PCA, each tunneling parameter is normalized in the range from -1 to 1, and the covariance can be calculated and recorded as *C*, like shown in Eq.3.

$$\boldsymbol{C} = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \dots & cov(x_1, x_n) \\ cov(x_2, x_1) & cov(x_2, x_2) & \dots & cov(x_2, x_n) \\ \dots & \dots & \dots & \dots \\ cov(x_n, x_1) & cov(x_n, x_2) & \dots & cov(x_n, x_n) \end{bmatrix} \quad (3)$$

Where:

n - the number of variables,

$cov(x_i, x_j)$ - the covariance between the *i*th and *j*th variables.

The greater the absolute value of the covariance, the greater the influence of the *i*th and *j*th variable on each other. A positive covariance means the two features increase or decrease together. This covariance has *n* eigenvalues and corresponding eigenvectors λ_i and u_i, respectively, each one should require Eq.4.

$$Cu = \lambda u. \quad (4)$$

There is a high positive correlation between the absolute value of λ_i and the information amount of the *i*th eigenvectors. In other words, Eq.4 give the information order of the *n* eigenvectors. On this basis, any number of eigenvectors with high average absolute eigenvalues can be selected as input, and the number is recorded as *r* according to the actual requirement. The larger the value of *r*, the more data information is retained, and the complexity and computational burden of the algorithm also larger. Usually, the *r* value is chosen according to the data used, as discussed in section 4.

2.3. Prediction models of rock mass parameters

After data pre-treating and conducting PCA, the dataset involved rock mass parameter and tunneling data is transform to standard “key-value” dataset. On this basis, regression can be conducted with the input of tunneling data and output of rock mass parameter by multiple machine learning algorithms. In this paper, totally three algorithms are used to verify the positive effect of PCA, including BP neural networks (BPNN), support vector regression (SVR) and random forest

(RF), whose basic principles are introduced in this section. The mentioned algorithms are programed by Weka 3.8.6, a machine learning workbench, which provided an extensive collection of machine learning algorithms and data pre-processing methods. Accordingly, prediction models of rock mass parameters are established, which can calculate the *UCS* and J_f of target tunneling mileage with the known tunneling data.

1. Back Propagation Neural Network (BPNN)

Back Propagation Neural Network (BPNN) is one of the most widely used machine learning algorithm which is proposed in 1986 (Almasaeid et al. 2022, Truong 2023, Ali et al. 2024). The network with normal structure is composed of three layers, called input, hidden, and output layer, in which there are multiple neurons that character variables. Among them, the number of neurons in input layers are consistent with the input variables, while the number of neurons in output layers are consistent with the output variables. Variables in different layers are connected with each other by weights and activation functions. Based on known input and output, the BPNN model can be expressed by Eq.5 (Hecht-Nielsen 1989).

$$Y = f_o(\sum_{j=1}^{H} w_{kj} \left(f_h\left(\sum_{i=1}^{I} w_{ji} X_i + b_j\right) + b_k\right)) \quad (5)$$

Where:

Y and X_i - the output and the i^{th} input variables,

i, j, and *k* - the order of neurons in input, hidden, and output layers, respectively,

I and *H* - the number of neurons in input and hidden layers,

w_{ji}, b_j, w_{kj}, and b_k - the weight between input and hidden layers or between hidden and output layers, which needs to be optimized during training process.

Especially, the calculation of the variables in different layers is not linear because of the existence of f_h and f_o, which represent the activation functions, such as *sigmoid*, *tanh,* or *ReLU* functions as shown in Eq.6, 7 and 8 (Liu et al. 2020).

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (6)$$

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (7)$$

$$ReLu(x) = \max(0, x) \quad (8)$$

2. Support Vector Regression (SVR)

Support Vector Regression (SVR) is a generalized form of Support Vector Machine (SVM) used to solve regression problems. Aiming at the nonlinear problems, data is mapped to high-dimensional space through kernel functions, and any nonlinear continuous function is approximated with arbitrary accuracy in high-dimensional space by minimizing structural risk (Mahdevari et al. 2014). In terms of regression problems, given a training sample set with input $\boldsymbol{x_i}$, which is a vector with *m* dimensions, and output y_i, and the optimal regression hyperplane can be expressed by Eq.9, which is also the objective functions of SVR (Liu et al. 2019).

$$\min_{w,b}\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i^+ + \xi_i^-)\right]$$

$$s.t.\, f(\boldsymbol{x_i}) - y_i \le \epsilon + \xi_i^+$$

$$y_i - f(\boldsymbol{x_i}) \le \epsilon + \xi_i^- \quad (9)$$

Where:
i - the order of the samples,
ϵ, ξ_i^+ and ξ_i^- - given tolerance, which is non-negative number,
C - the penalty parameters.

During the training process, ϵ, ξ_i^+ , ξ_i^- and C needs to be selected to minimize the objective functions (Liu et al. 2019). There is a key task of SVR is to balance the regression error and structural complexity, which is also to balance the accuracy and generalization, which means the model adaptability on different data. Take the penalty parameter C as an example, the higher C, the higher the structural complexity of the training model, which may lead to relatively simple model with low accuracy on training set and relatively good generalization. On the contrary, too low C may obtain a relatively high accuracy on training set, but when the data changed, there might be a high difference between test and training accuracy.

3. Random Forest (RF)

Breiman proposed the Random Forest (RF) algorithm in 2001 (Hou et al. 2020), which is an ensemble learning algorithm based on decision trees and suitable for classification and regression problems. It uses bootstrap technology to extract randomized samples from the original samples to construct a single decision tree. At each node of the decision tree, a split point is selected through a random feature subspace for splitting. Finally, these decision trees are combined together and the final prediction result is obtained through majority voting. Compared with decision trees, random forests introduce the ideas of bagging and random subspaces. In each regression tree (RT), dichotomy is used to calculate the purity of the tree model by Eq. 10 (Yu et al. 2018).

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \tag{10}$$

Where:
$i(t)$ - the purity before division,
$i(t_L)$ and $i(t_R)$ - the purity of the left and right of the tree,
P_L and P_R - the probability of sample located on the left and right of the tree.

Based on Eq.10, Gini index is involved to calculated the impurity of each sample, as Eq.11 (Sun et al. 2018).

$$Gini(x_i) = \sum_{i=1}^{m} p_i(1 - p_i) \tag{11}$$

Where:
m - the number of possible values of each variable,
p_i - the probability of the i^{th} value.

In training process, the division of tree calculated by iteration, until the objective function (Eq.11) meets the requirement. In addition, the RF model is formed by multiple decision tree, and each decision tree trained by Eq.11 can give a predicted result. Take a RF model including M decision trees as an example, the prediction results of m^{th} decision tree is recorded as $h_m(x)$, and the finial prediction results by RF $f(x)$ can be calculated by Eq.12 (Sun et al. 2018).

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} h_m(x) \tag{12}$$

3. Case study and model establishment

3.1. Case study

The prerequisite of solving the structural difference between rock mass and tunneling data and establishing rock mass prediction model is collecting a number of field rock mass and tunneling data. In this research, totally 1272 series of

data from 5 different TBM projects are collected. Given the difference of the TBM type and rock mass condition, there are significant difference between data distribution of the different project. In addition, there is one investigated area in the project YH and QD, while there are two investigated areas in project YS, ZJ, and HZ. Therefore, the 1272 series of data are used as 8 different datasets according to their sources to establish the corresponding 8 different rock mass prediction models, instead of being mixed into a same dataset. The basic information of the datasets and the corresponding projects are listed in Table 1.

Table 1: Basic information of the datasets and projects (Liu et al. 2019,2020)

Data sources	YS		YH	ZJ		HZ		QD
	Area I	Area II		Area I	Area II	Area I	Area II	
Number of samples [-]	460	80	50	230	46	250	56	100
TBM type [-]	Open type		Double-shield	Double-shield		Double-shield		Double-shield
Number of tunneling variables [-]	178		202	276		276		202
TBM diameter [m]	7.9		7.0	6.2		6.0		7.0
Number of cutters [-]	56		48	42		40		48

3.2. Dataset division

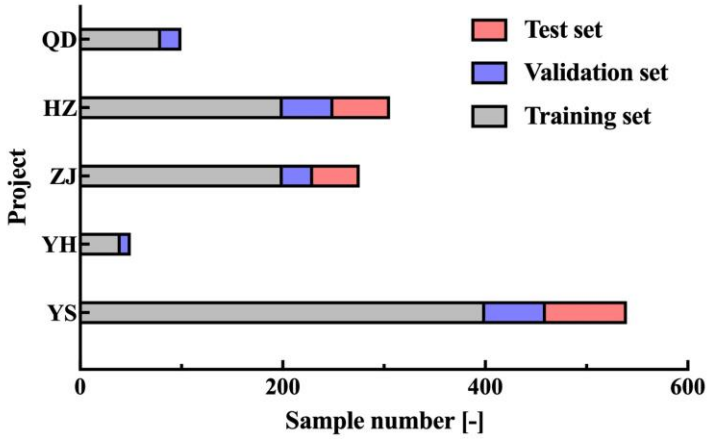


Figure 1: Division of the field datasets

Generally, in the machine learning tasks, the dataset should be divided into three parts, called training set, validation set, and test set. The three datasets typically have non overlapping samples and different function. The samples in training set are occupied the main parts in the total samples, always more than 60%. It is used for trial and error of existing algorithm architectures and parameter tuning. The weight combination that performs well on the training set is considered as the trained and usable model. The validation set is used to test the performance of the trained model, usually using the same data source as the training set. By comparing the accuracy of training and validation, the generalization of the model is verified to prevent overfitting; The test set is involved to the validated model to solve practical prediction problems, and its data sources are usually different from the training and validation sets, used to test the adaptability of the model to data with unknown and different distribution.

In this research, totally 1272 samples collected in 8 areas from 5 different project, and they are used to established 5 different models for different projects. According to the task, three principles for dataset division are proposed.

1. Model used for specific project should be established by samples collected in the project, because of the differences of the lithology and TBM type.

2. For each project, samples in the training and validation sets should be collected in the same study area.

3. For each project, samples in the test sets should be collected in the different study area. Especially, for projects with limited sample, test sets are not necessary.

Accordingly, the 1272 samples are divided as shown in Figure 1. Due to the small amount of the sample collected in YH and QD project, only 50 and 100 samples, respectively, there are no test set for the corresponding prediction model.

3.3. Data pre-treated by PCA

After field data preparation, the tunnelling data with a huge number of features should be simplified by PCA, as introduced in Section 2.2. Before pre-treating by PCA, the field data should be modified as key-value data. Each *UCS* and J_f sample matches an area with the length of 1m. During tunneling the area, usually more than 1000 series of tunnelling parameters are recorded and each one consists of 178 variables, because the tunneling data is sequential with a frequency of 1s-1, and tunneling an area with the length of 1m always takes more than 1000s. Part of those tunneling data is inefficient, because TBMs have interval stoppage and trial excavation (Wang et al. 2025). To remove the inefficient tunnelling data, a threshold of penetration rate is set as 10mm/min, and the series of tunneling parameters with lower penetration rate are removed. Subsequently, the time series of tunneling data corresponding to the same rock mass sample is eliminated using the average value, and the difference of ranges between the tunneling variables are unified by normalization (Wang et al. 2025). By conducting mentioned steps, the rock mass and tunneling data is modified as a standard key-value data, which will be pre-treated by PCA method.

Especially, the number of tunneling parameters varies among different projects because of the difference of the involved TBM manufacturers. In detail, tunnelling variables recorded in YS is 178, in YH and QD is 202, and in ZJ and HZ is 276. Although the number of tunneling variables are different, all dataset collected from the different project contain several common main variables. The common variables contained by all dataset can roughly divided into three groups. First group contains parameters directly related by tunneling and rock-breaking process, such as the cutterhead thrust, torque, and revolution speed. These parameters always have sharply fluctuations, because they are influenced by both geological condition and manual control. The second group contains the parameters which characterize the tunnelling direction and trajectory, such as the horizontal and vertical tunneling angle. These parameters always show both positive and negative value, which represent different direction, and they are also one of the most complex variable parameters. The third group contains the working status of the machine, such as motor current and hydraulic pressure. Compared with the above two groups, these variables are always stable and simple. However, due to multiple sub-system of TBM, the number of the variables in the third group is huge, because the motor current and hydraulic pressure of each sub-system is an independent variable. Thus, the variable number of different projects is different, and the features of variables are different, therefore, the PCA parameters and the number of retained variables are also different.

Take the YS project as an example. Totally 400 samples in training set are used by PCA, and the 178 features are reorganized to form new and independent variables through orthogonal transformation. The first step of PCA is to calculate the covariances among 178 variables and consist the covariance matrix. Figure 2 shows the part of the covariance matrix including some main tunneling variables, such as thrust (Th), torque (Tor), penetration rate (PR), penetration (P), cutterhead revolution speed (R), and cutterhead power (CP).

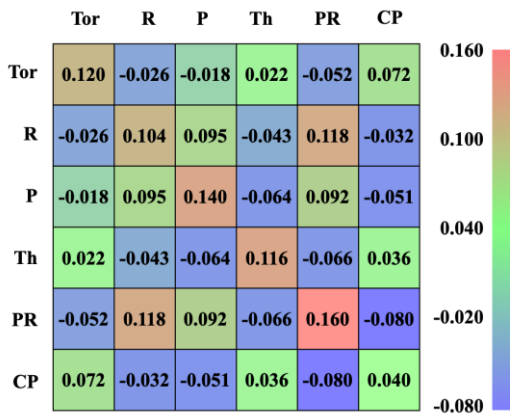


Figure 2: Part of the covariance matrix results

According to Eq.4, totally 178 eigenvectors and eigenvalues corresponding to the covariance matrix ***C*** are calculated. Each eigenvector and eigenvalue match a tunneling parameter. Then, the eigenvalues are sorted in descending order of their absolute value, and an index, called accumulated information, can be calculated by Eq. 13.

$$AI_r = \frac{\sum_{i=1}^{r} e_i}{\sum_{i=1}^{n} e_i}, \tag{13}$$

Where:

r - a changeable positive integer, and represents the number of retained tunneling variable,

AI_r - the accumulated information under *r* retained variable, it ranges from 0 to 1,

n - the total number of the tunneling parameters and it is 178 for YS project,

e_i - the absolute value of the i^{th} eigenvalues in order.

Generally, the accumulated information AI should be higher than 95%. On the basis of the AI higher than 95%, as few tunneling variable as possible should be retained. The absolute value of the eigenvalues and the AI of the top 10 eigenvectors in order is listed in Figure 3.

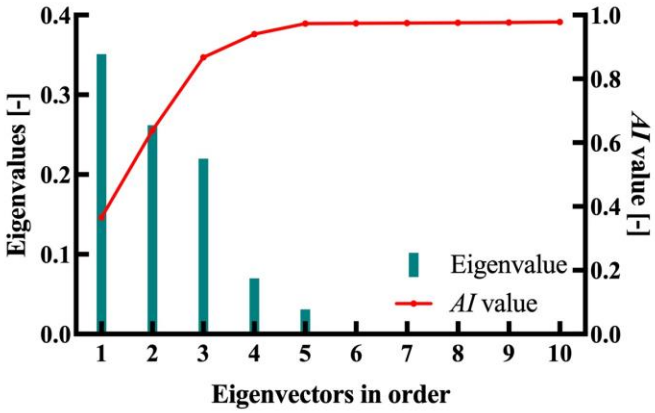


Figure 3: Part of the covariance matrix results

As Figure 3 shows, after orthogonal transformation, equal quantity (totally 178) variables are generated, and their eigenvalues are concentrated on the top 5 variables, and the eigenvalues of the variables after 6th place are lower than 0.005. The *AI* values of the top 4 variables are 0.941, while the one of the top 5 variables are 0.973. According to the principle of selecting variables by PCA, the top 5 tunneling variables are retained as the input, which will be used to train the prediction model of the rock mass.

3.4. Establishment of the prediction models of rock mass parameters

After pre-treated by PCA, the original field tunneling data form YS project with 178 variables are transformed to new data with 4 variables. Take the new variables as input and the corresponding rock mass parameters as output, and the mentioned machine learning method, including BPNN, SVR, and RF, are used to establish the mapping between the input and the output.

The 10-fold-corss-validation strategy is used in above mentioned training process. In detail, the 400 samples in training set are randomly and equally divided into 10 sub-sets, and named as sub-set 1 to 10. Among them, 9 sub-sets are used in training to establish a sub-model, and the other sub-set is used to test the sub-model and calculate the accuracy. By transforming the sub-set used for training and testing, totally 10 sub-models are established, and the corresponding accuracies are also calculated. The sub-model with the highest accuracy is chosen as the well-trained model. By conducting the 10-fold-corss-validation, three models are established by the mentioned three machine learning methods. The accuracies of the sub-models based on the training data collected from YS project are listed in Table 2.

Table 2: Mean absolute percentage error of the sub-models in the 10-fold-corss-validation

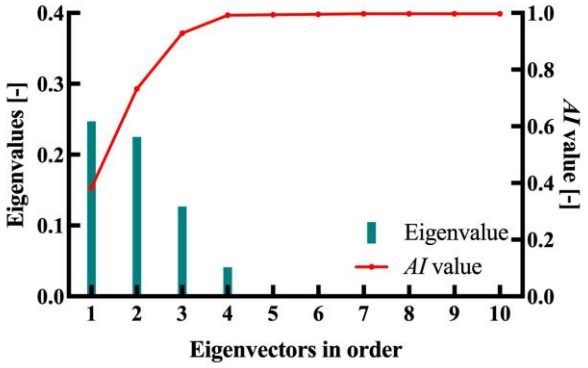
Order	Training sub-set	Test sub-set	BPNN		SVR		RF	
			UCS **[%]**	J_f **[%]**	***UCS*** **[%]**	J_f **[%]**	***UCS*** **[%]**	J_f **[%]**
1	2-10	1	9.7	8.4	9.2	9.2	9.0	8.9
2	1, 3-10	2	8.6	**8.2**	9.1	**8.3**	8.9	8.8
3	1-2, 4-10	3	9.0	9.4	**8.5**	8.6	**8.7**	9.0
4	1-3, 5-10	4	9.6	9.2	9.9	9.6	8.8	9.8
5	1-4, 6-10	5	10.0	9.7	8.8	9.2	10.1	9.1
6	1-5, 7-10	6	**8.4**	8.7	9.2	8.3	8.9	10.1
7	1-6, 8-10	7	9.7	8.9	9.5	9.6	9.4	**8.8**
8	1-7, 9-10	8	9.3	9.5	8.9	8.7	10.6	9.4
9	1-8, 10	9	9.7	9.4	9.2	9.8	9.1	8.9
10	1-9	10	9.4	8.9	9.2	9.2	9.3	8.8

In Table 2, sub-models with different prediction target and algorithms are selected separately. For example, for predicting UCS by BPNN, the 6th sub-model is selected as the well-trained model, since it has the lowest MAPE of 8.4%. According to the results of the 10-fold-cross-validation, the MAPE of all the well-trained models based on the pre-treated data by PCA are controlled within 9%, which proves that the proposed data-pre-treating method is suitable for machine learning, at least in the training stage.

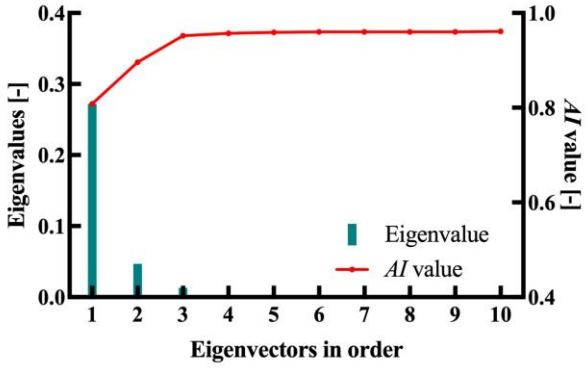
4. Results

4.1. PCA results of the other projects

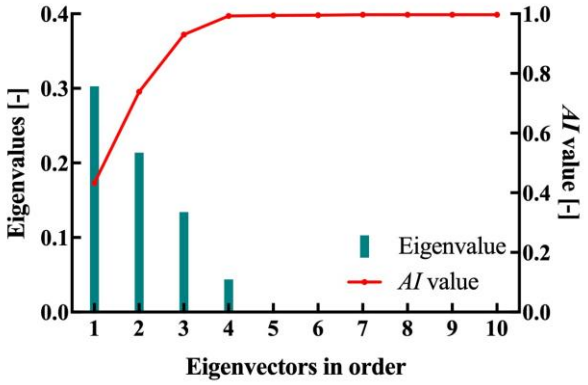
In Section 3.3 and 3.4, the data pre-treating by PCA and the prediction model establishing by machine learning are introduced, and they are also conducted in the field data sets collected from YH, ZJ, HZ, and QD project. The data sets from the four projects are transformed into key-value data, and this process are similar with the way introduced in Section 3.2. On this basis, PCA is applied to the four projects, and the eigenvalues and the accumulated information of the top 10 variables in the four projects are shown in Figure 4.



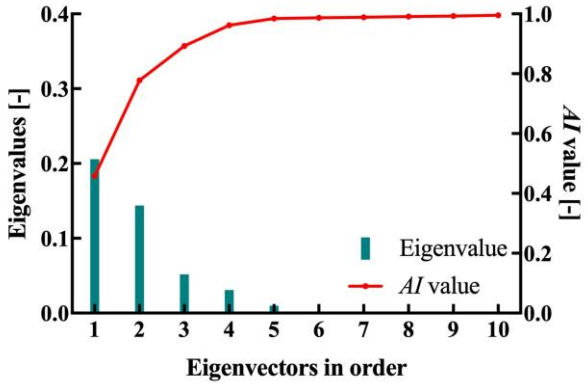
(a) YH project



(b) ZJ project



(c) HZ project



(d) QD project

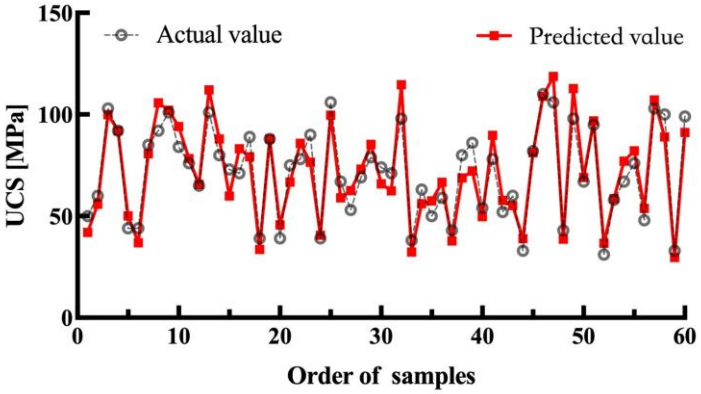
Figure 4: Eigenvalues and the accumulated information of the top 10 variable in the four projects

As Figure4 shows, the eigenvalues in the other four projects show similar rules with the one in YS projects. Only the top 3 to top 5 variables have relatively high eigenvalues, while the eigenvalues of the other variables are lower than 0.005. Take the 202 variables of YH project as an example, the *AI* of the top 3 variables is 0.929, while it of the top 4 variables are 0.992, and the 202 variables are reduced to 4 variables by PCA. In addition, 276 variables of ZJ and HZ project are reduced to 3 and 4 variables, respectively, while 202 variables of QD project are reduced to 4 variables. It shows that, PCA can significantly reduce the volume of tunneling data and the cost of calculation by machine learning, under prerequisite of retaining more than 95% input information.

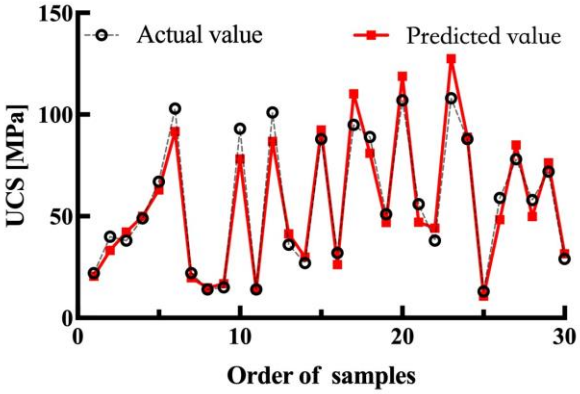
4.2. Prediction results of validation sets

On the basis of the rock mass parameters models trained by the training set, models are applied to the validation sets from corresponding projects. By comparing the prediction results of *UCS* and J_f, the prediction accuracy can be expressed by mean absolute percentage error (MAPE), which can be calculated according to Eq.13, where y and y' express the actual and prediction value of the target. In addition, Figure 5 and 6 shows the comparison between the predicted values by BPNN and the actual values of the *UCS* and J_f in each project.

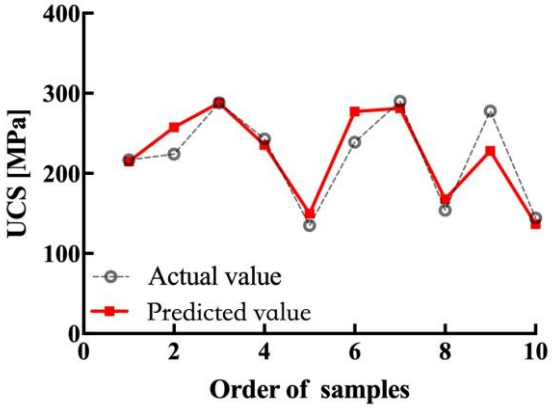
$$MAPE = \frac{|y'-y|}{y} \times 100\% \quad (13)$$



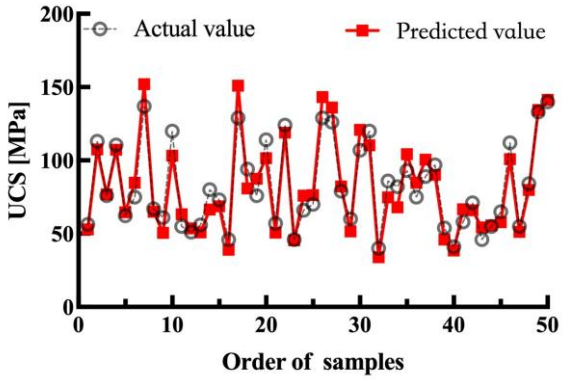
(a) Comparison results of YS project



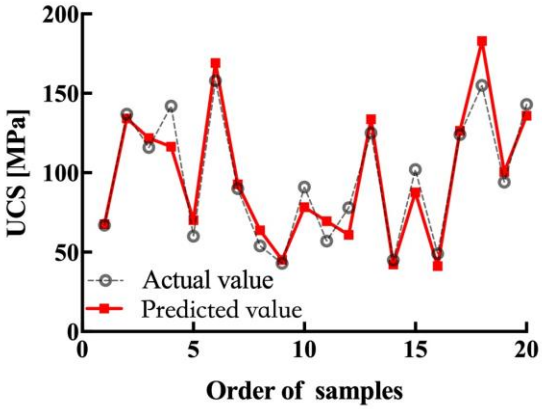
(b) Comparison results of ZJ project



(c) Comparison results of YH project

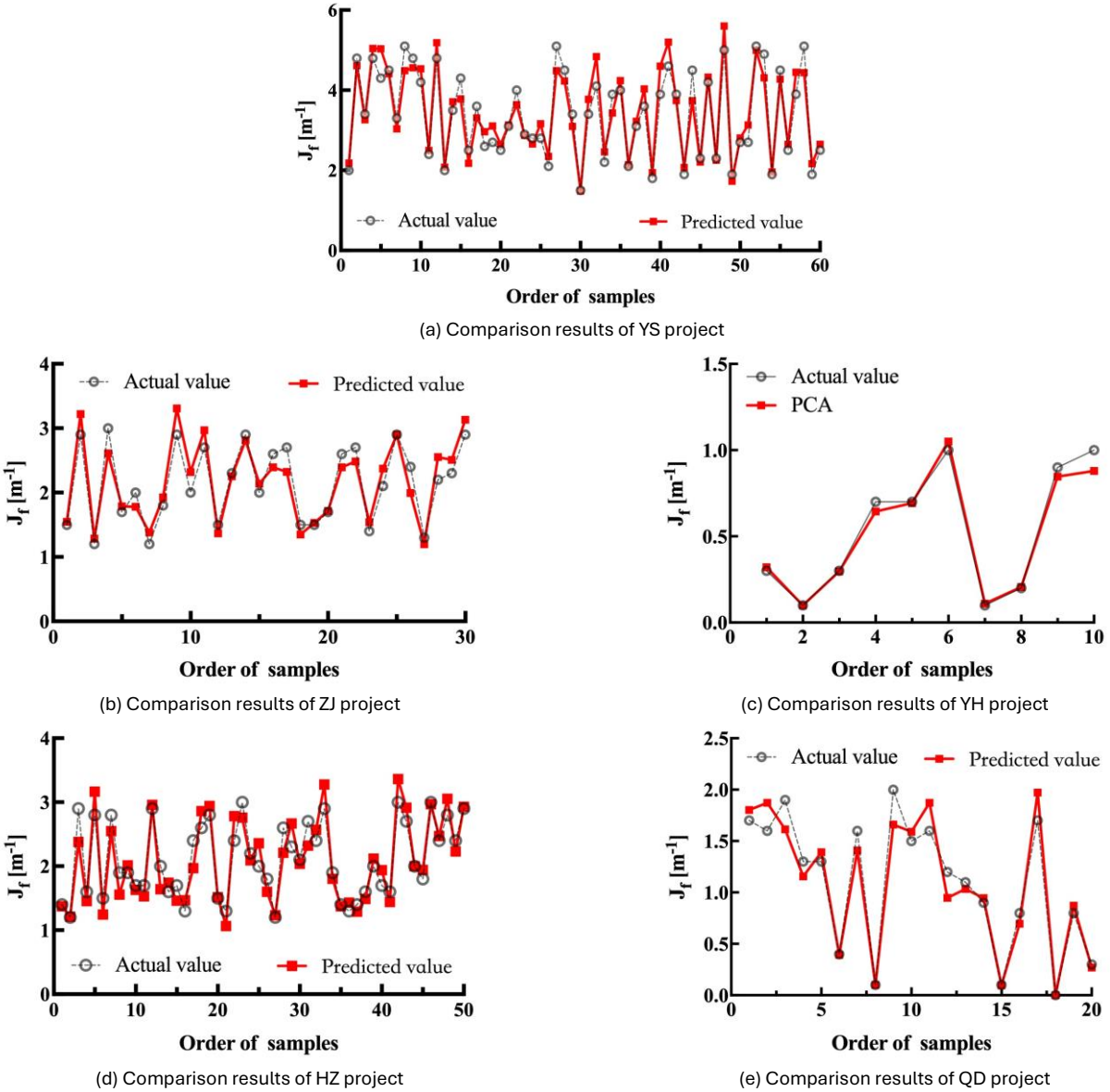


(d) Comparison results of HZ project



(e) Comparison results of QD project

Figure 5: Comparison between the predict *UCS* to the actual values in different project



(a) Comparison results of YS project

(b) Comparison results of ZJ project

(c) Comparison results of YH project

(d) Comparison results of HZ project

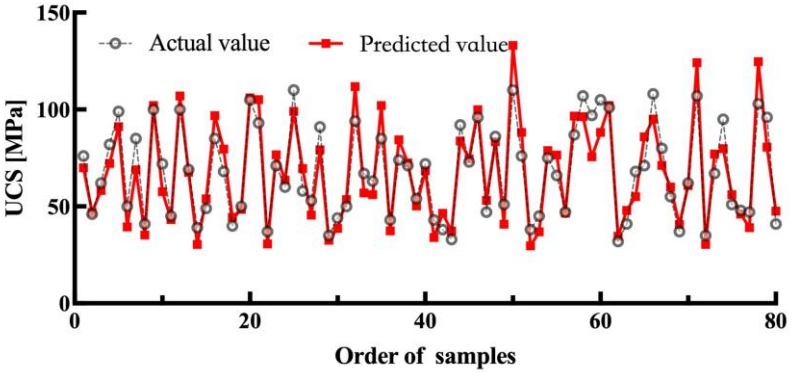
(e) Comparison results of QD project

Figure 6: Comparison between the predict J_f to the actual values in different project

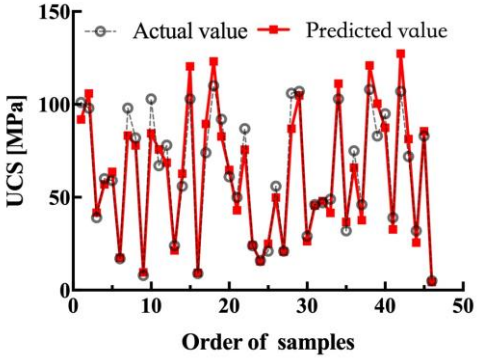
As Figure5 and Figure6 show, the trend of the predicted *UCS* and J_f are consistent with the actual values. In terms of the *UCS*, the average *MAPE* values of the 5 validation sets are 10.1%, 11.0%, 8.1%, 9.7%, and 10.5%, in order. The error between the predicted *UCS* and the actual values are lower than 20MPa, except part of sample from YH project, which is mainly located in the hard rock area and most of its *UCS* samples are higher than 200 MPa. Therefore, *UCS* errors of some samples is closed to 50MPa, such as the 6th and 9th samples. In terms of the J_f, the average *MAPE* are 8.4%, 8.8%, 5.3%, 9.1%, and 9.7%, which is better than the effect of *UCS*. Especially for YH project, whose rock is relatively integral and the most J_f is lower than 1 m^{-1}, which results in the low error. Errors of some samples are nearly to 0, such as the 1st, 2nd, 3rd, 7th, and 8th samples.

4.3. Prediction results of test sets

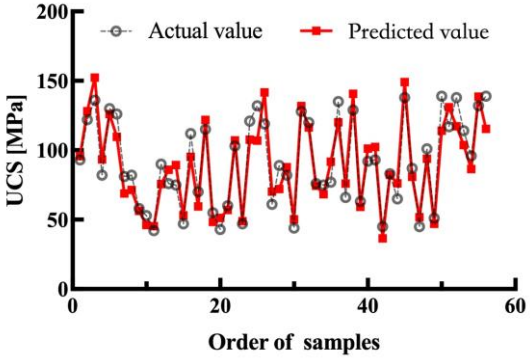
The proposed models based on BPNN are further used on the test set, and the comparison between the predicted *UCS* and J_f and their actual values are shown in Figure7 and 8. Especially, as explained in Section 3.2, there are no test sets in YH and QD project, because of the limited samples.



(a) Comparison results of YS project

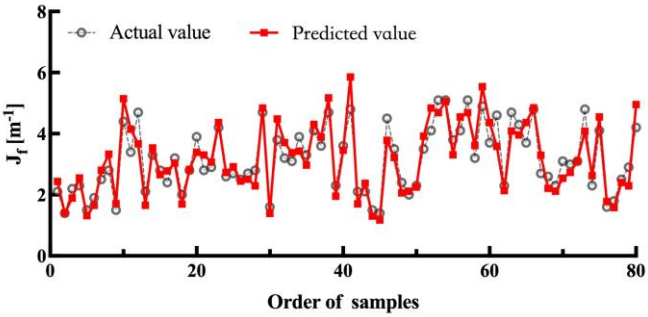


(b) Comparison results of ZJ project

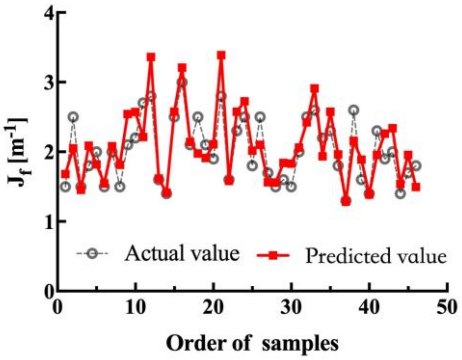


(c) Comparison results of HZ project

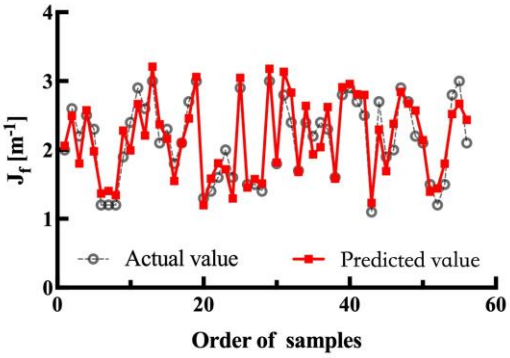
Figure 7: Predicted results of *UCS* in test sets



(a) Comparison results of YS project



(b) Comparison results of ZJ project



(c) Comparison results of HZ project

Figure 8: Predicted results of J_f in test sets

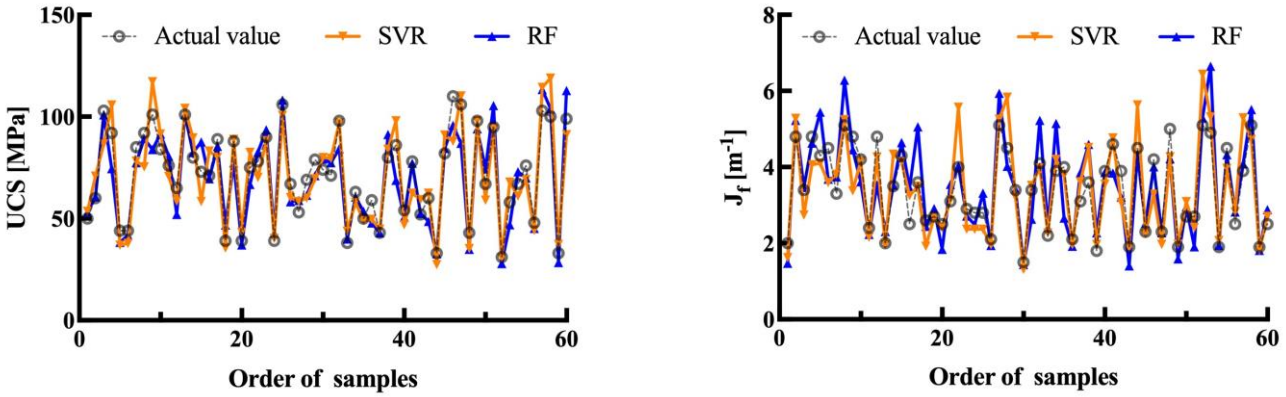
Figure7 and Figure8 also show the acceptable predicted results. The *MAPE* values of the *UCS* in test sets are 13.0%, 11.9%, and 11.9% in order, while the one of J_f are 12.7%, 12.3%, and 11.0%. In the aspects of *MAPE* values, the predicted errors in test sets are higher than the one in validation sets, because of the data distribution differences. Generally, errors lower than 13% are acceptable in actual projects, which proves the prediction models are accurate, and the tunneling data pre-treated by PCA is suitable for *BPNN* algorithm.

5. Discussion

By comparing the prediction results to the actual value of the validation set, it is proved that pre-treated tunneling data by PCA is useful for predicting rock mass parameters by BPNN. However, there are still two valuable issues for discussion. Firstly, only BPNN applied to tunneling parameters pre-treated by PCA is verified, and the universal adaptability of PCA still needs to discuss. Therefore, SVR and RF are further applied to the tunneling parameters pre-treated by PCA, which is discussed in Section 5.1. Then, the effect of PCA pre-treating on the accuracy have not been verified. For this purpose, tunneling data with all variables are used as input to build prediction models of rock mass parameters by same method, and the accuracy created by PCA pre-treated and all variables are compared with each other, which is discussed in Section 5.2.

5.1. Comparison between the prediction results of PCA pre-treated method and full variable

As introduced in Section 3.4, the expression of the tunneling data pre-treated by PCA on BPNN and the other two algorithm is relatively closed. The minimum *MAPE* of *UCS* and J_f by BPNN is 8.4% and 8.2%, while the one by SVR is 8.5% and 8.3%, by RF is 8.7% and 8.8%, which proves the universal adaptability of PCA on training sets. Further, to verify the universal adaptability of PCA on different machine learning algorithm, the validation and test sets used in Section 4 are also tested by two kinds of machine learning algorithms, SVR and RF, which is also programmed in Weka 3.8.6. Take the validation set from YS project as an example, and the comparison results are shown in Figure9. In addition, the *MAPE* values of the other four projects are listed in Table.3



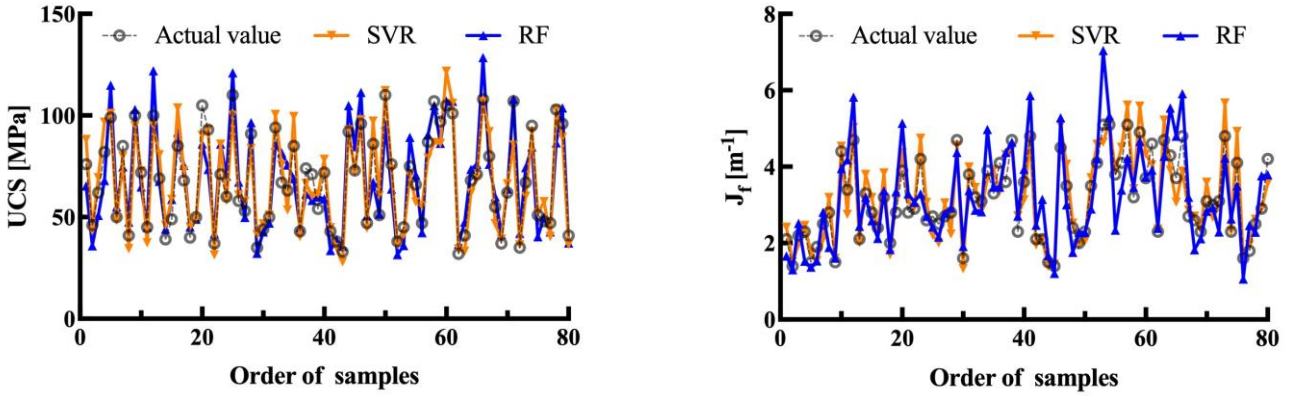
(a) Predicted results of *UCS* (b) Predicted results of J_f

Figure 9: Comparison between the predicted *UCS* and J_f by SVR, RF, and the actual values in validation sets

Table 3: Predicted *MAPE* values of the validation sets in other four projects

Projects	*UCS*			J_f		
	BPNN [%]	SVR [%]	RF [%]	BPNN [%]	SVR [%]	RF [%]
YS	10.1	10.3	10.7	8.4	8.5	8.8
ZJ	11.0	11.5	11.4	8.8	9.4	9.1
YH	8.1	8.3	9.0	5.3	6.0	6.2
HZ	9.7	8.9	9.6	9.1	8.8	9.9
QD	10.5	10.2	10.8	9.7	8.9	10.4
Average value	9.9	9.8	10.3	8.3	8.3	8.9

As shown in Table 3, in terms of the validation sets, the BPNN and SVR are shown similar error, 9.9% and 9.8% of *MAPE* for *UCS*, and 8.3% and 8.3% for J_f. In addition, the *MAPE* of *RF* on validation sets are slightly higher than the one of BPNN and SVR, which is 10.3% and 8.9% for *UCS* and J_f, respectively. The results prove that, at least to the data collected in the same area with the training set, the tunneling data pre-treated by PCA is suitable for not only BPNN, but also other algorithm, such as SVR and RF.



(a) Predicted results of *UCS* (b) Predicted results of J_f

Figure 10: Comparison between the predicted *UCS* and J_f by SVR, RF, and the actual values in test sets

Table 4: Predicted *MAPE* values of the test sets in other four projects

Projects	*UCS*			J_f		
	BPNN [%]	SVR [%]	RF [%]	BPNN [%]	SVR [%]	RF [%]
YS	13.0	12.8	14.4	12.7	13.2	14.8
ZJ	11.9	12.2	13.9	12.3	11.8	14.2
HZ	11.9	12.3	13.2	11.0	11.2	13.5
Average value	12.3	12.5	13.8	12.0	12.1	14.2

Table 4 shows the *MAPE* of the test sets predicted by the three algorithms. The *MAPE* of SVR keeps closed to BPNN, and increases by less than 3%, compared with the *MAPE* of the validation set. It proves that the SVR have similar effect with the BPNN, both on validation and test sets. Compared with SVR and BPNN, the generalization of RF is slightly weaker. The *MAPE* values are about 2% higher than the one of SVR and BPNN, and difference between the *MAPE* of validation and test sets reaches 3.5% (*UCS*) and 5.3% (J_f). However, the *MAPE* of lower than 15% is acceptable in actual projects, and although RF is not the best algorithm for predicting rock mass parameters, its results still have reference price.

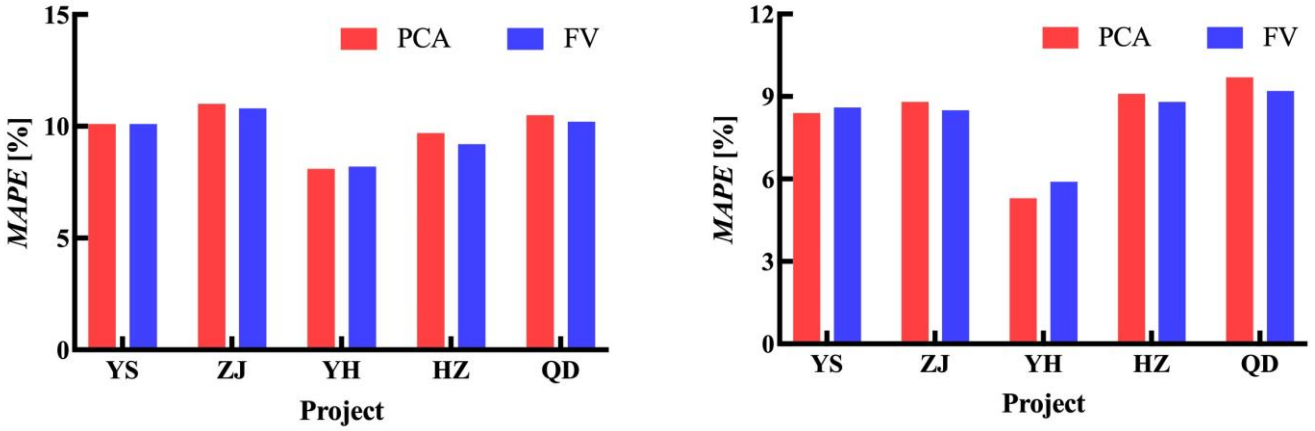
5.2. Comparison of the different machine learning method prediction results

This section mainly discusses the positive effect of PCA on the predicted accuracy by machine learning algorithm. To compare the predicted accuracy, another rock mass predicted models aiming at the mentioned five projects are added, respectively, with the tunneling data without pre-treated as input. The newly added models are established by BPNN algorithm, and the most of the hyper-parameters is consistent with models introduced in Section 3.4. Especially, the structure of the network changes. Since the input data are not pre-treated and have hundreds of variables, the neuron number of the input layers increases to be equal with the variables. In addition, the neuron number of the hidden layers also increases to the range from 50 to 80. Thus, the added models with full variables as input, named FV models, are established by networks with more complex structures than the network used in PCA models, which results in higher time costs during each calculation. The network structures and time costs of each project are listed in Table 5.

Table 5: The network structures and time costs of PCA and FV models

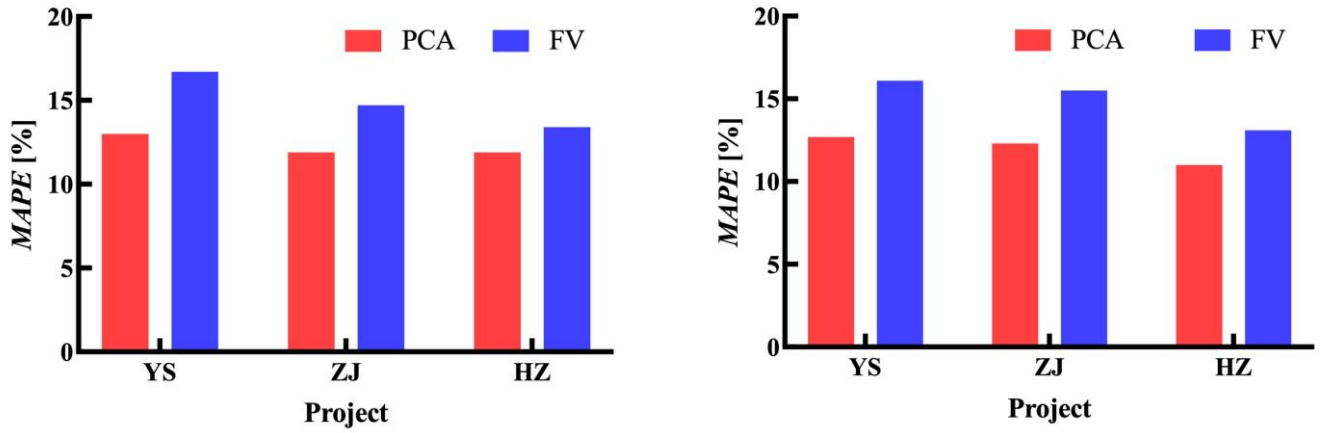
		PCA model					FV models				
		YS	YH	ZJ	HZ	QD	YS	YH	ZJ	HZ	QD
Number of input neurons [-]		5	4	3	4	4	178	202	276	276	202
Number of hidden neurons [-]		8	6	5	6	6	56	64	72	72	64
Time costs [s]	*UCS*	6	4	4	4	4	202	179	261	252	188
	J_f	6	4	3	5	3	196	165	249	266	193

As Table 5 shows, after pre-treated by PCA, the tunneling data are simplified, which results in the decreasing of the number of the input and hidden neurons. It makes a significant simplification of the network structure. In terms of the calculation efficiency, calculating the prediction results by PCA model needs about 3 to 6 seconds, while it needs about 3 to 4 minutes by FV model. It proves that, PCA pre-treating is helpful for simplifying the data and improving the calculating efficiency.



(a) *MAPE* values of *UCS* (b) *MAPE* values of J_f

Figure 11: Comparison between the calculating accuracy in validation sets of PCA model and FV models



(a) *MAPE* values of *UCS* (b) *MAPE* values of J_f

Figure 12: Comparison between the calculating accuracy in test sets of PCA model and FV models

In addition to the efficiency, the calculation accuracy of PCA and FV model is compared and listed in Figure11 and Figure12, which is also expressed by *MAPE* values. As Figure11 shows, the accuracies of FV models are closed to or lower than the accuracies of the corresponding PCA models. The *MAPE* of validation *UCS* collected in the five projects are 10.1%, 10.8%, 8.2%, 9.2%, and 10.2%, the average *MAPE* is 0.18% lower than the one of PCA models, while the *MAPE* of J_f is 8.6%, 8.5%, 5.9%, 8.8%, and 9.2%, the average *MAPE* is 0.06% lower than it of PCA models. It indicates that the

accuracy of PCA models in validation set is slightly weaker than FV model, the simplification of input data has almost no negative impact on the predicted effect.

Figure12 shows in predicted accuracy on test sets. The *MAPE* values of FV models are higher than them of PCA models. In terms of *UCS*, the *MAPE* values of FV models are 16.7%, 14.7%, and 13.4%, the average value is 2.7% higher than results of PCA models, while in terms of J_f, the *MAPE* values are 16.1%, 15.5%, and 13.1%, whose average value is 2.9% higher than the one of PCA models. The main reason of these phenomenon is that the too complexed tunneling data have been simplified, which is helpful for improving the generalization of the models.

In addition, it is found by compared with the prediction results of different projection that the validation results of the YH projects are much better than the other projects. In terms of the test set, prediction results of YS project have highest *MAPE*, especially predicted by FV model. The reason is inferred to be that there are the most samples collected from YS, while there are the least samples collected from YH. Therefore, the geological condition of the YH investigated area is relatively stable, while it of YS project is more complex. Poor performance in complex data is a common shortcoming of machine learning method, and it is also shown in this research. Further, more data with more complex feature should be involved in rock mass parameter prediction research by machine learning in the future research.

6. Conclusion

This research proposed a prediction method of rock mass parameters based on pre-treated tunneling data by PCA and multiple machine learning methods. The main conclusions of this research can be summarized as follows:

1. Given the structural difference between the tunneling data with a number of features and high acquirement frequency and rock mass data. The pre-treated method of the tunneling data by PCA is proposed. Totally 1272 samples collected from five different projects are collected and used to verify the method. The result shows that the PCA can reduce hundreds of tunneling variables to 3-5 variables with more than 95% data information reserved.

2. The field data pre-treated by PCA is used to train BPNN prediction models of *UCS* and J_f by 10-fold cross validation, and they are used to flied validation sets and test sets. The *MAPE* of *UCS* and J_f on validation sets are lower than 11.0% and 9.7%, while the one on test sets are lower than 13.0% and 12.7%, which are acceptable in actual project. The results prove that the BPNN prediction model is relatively accurate and has good generalization.

3. The tunneling data with full variables without pre-treated are used to build prediction model of rock mass parameters by BPNN, in order to verify the positive effect of the PCA method on the prediction accuracy. The results show that the network structure of PCA model is simpler than the model with full variables, and it leads to the time costs are reduced from 4-6 minutes to 3-6 seconds. In addition, the accuracies of PCA model on validation set is closed to them of model with full variables, while it on test set are higher than them of model with full variables. It proves that the PCA is helpful for the generalization of the prediction model of the rock mass parameters.

Acknowledgements

The research was supported by the Natural Science Foundation of Shandong Province (No. ZR2024QE150), and Postdoctoral Innovation Program Project of Shandong Province (No. SDCX-ZG-202503076).

Author Contributions

Q.M. proposed the idea and the method. **H.W.** conduct field investigation and acquired the field data. **L.B.** organized the original manuscript and finished the main calculation. **F.Z.** programmed the algorithm and conducted calculation. **J.B.** improved the method and written the manuscript. All authors critically reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Disclosure of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

Ali A M, Maha A S, Faten M R (2024) Constructing a Corrected Stations Network Using Some Geomatic Techniques. Civil and Environmental Engineering, 20(1): 78-85.

Almasaeid H H , Alkasassbeh A , Yasin B (2022) Prediction of Geopolymer Concrete Compressive Strength Utilizing Artificial Neural Network and Nondestructive Testing.Civil and Environmental Engineering, 18:655 - 665.

Armaghani DJ, Mohamad ET, Narayanasamy MS, Narita N, Yagiz S (2017) Development of hybrid intelligent models for predicting TBM penetration rate in hard rock condition. Tunneling and Underground Space Technology 63: 29-43

Gao F, Chen D, Zhou K (2018) A Fast Clustering Method for Identifying Rock Discontinuity Sets. KSCE Journal of Civil Engineering, 23(2):556-566.

Goh TL, Samsudin AR, Rafek AG (2011) Application of Spectral Analysis of Surface Waves (SASW) Method:Rock Mass Chara terization. Sains Malaysiana 40(5): 425-430

Gong QM, Lu J, Xu H (2020) A modified rock mass classification system for TBM tunnels and tunneling based on the HC method of China. International Journal of Rock Mechanics and Mining Sciences, 137(4): 104551.

Gong QM, Zhao J (2009) Development of a rock mass characteristics model for TBM penetration rate prediction. International Journal of Rock Mechanics and Mining Sciences, 46(1):8-18.

Hecht-Nielsen R (1989) Theory of the Backpropagation Neural Network. Neural Networks, 1989.

Hou SK, Liu YR, Li CY, (2020) Dynamic Prediction of Rock Mass Classification in the Tunnel Construction Process based on Random Forest Algorithm and TBM in situ Operation Parameters. IOP Conference Series: Earth and Environmental Science, 570(5):052056.

Kong F, Shang J (2018) A Validation Study for the Estimation of Uniaxial Compressive Strength Based on Index Tests. Rock Mechanics and Rock Engineering 51(7):2289-2297

Liu B, Wang RR, Guan ZD, Li JB, Xu ZH, Guo X, Wang YX (2019) Improved support vector regression models for predicting rock mass parameters using tunnel boring machine tunneling data. Tunneling and Underground Space Technology 91: 102958.1-102958.10

Liu B, Wang RR, Guan ZD, Li JB, Xu ZH, Guo X, Wang YX (2019) Improved support vector regression models for predicting rock mass parameters using tunnel boring machine driving data. Tunnelling and Underground Space Technology, 1: 102958.

Liu B, Wang RR, Zhao GZ, Guo X, Wang YX, Li JB, Wang SG (2020) Prediction of rock mass parameters in the TBM tunnel based on BP neural network integrated simulated annealing algorithm. Tunnelling and underground space technology, 95(6):103103

Liu DH, Zhou YQ, Jiao K (2010) TBM Construction Process Simulation and Performance Optimization. Transactions of Tianjin University 016(003): 194-202

Liu QS, Liu JP, Pan YC (2017) A case study of TBM performance prediction using a Chinese rock mass classification system – Hydropower Classification (HC) method. Tunnelling and Underground Space Technology, 65(5):140-154.

Liu QS, Zhao Y, Zhang X, Kong X (2018) Study and discussion on point load test for evaluating rock strength of TBM tunnel constructed in limestone. Rock and Soil Mechanics. 39(3): 977-984.

Liu QS, Zhao YF, Zhang XP, Kong XX (2018) Study and discussion on point load test for evaluating rock strength of TBM tunnel constructed in limestone. Rock and Soil Mechanics 39(3): 977-984

Mahdevari S, Shahriar K, Yagiz S, Shirazi MA. (2014) A support vector regression model for predicting tunnel boring machine penetration rates. International Journal of Rock Mechanics & Mining Sciences 72: 214-229

Majdi A, Beiki M (2019) Applying evolutionary optimization algorithms for improving fuzzy C-mean clustering performance to predict the deformation modulus of rock mass. International Journal of Rock Mechanics and Mining Sciences, 113: 172-182

Minh VT, Katushin D, Antonov M (2017) Regression Models and Fuzzy Logic Prediction of TBM Penetration Rate.Open Engineering, 7(1):60-68.

Naeimipour A, Rostami J, Buyuksagis IS (2016) Applications of rock strength borehole probe (RSBP) in underground openings. In: ISRM International Symposium-EUROCK 2016, International Society for Rock Mechanics and Rock Engineering

Sun W, Shi M, Zhang C (2018) Dynamic load prediction of tunnel boring machine (TBM) based on heterogeneous in-situ data. Automation in Construction, 92(8):23-34.

Sun W, Wang XB, Wang LT, Zhang J, Song XG (2016) Multidisciplinary design optimization of tunnel boring machine considering both structure and control parameters under complex geological conditions. Structural and Multidisciplinary Optimization 54(4): 1073-1092

Truong V (2023) Assessment of Slope Stability with the Assistance of Artificial Neural Network and Differential Evolution. Civil and Environmental Engineering, 19(1): 288-300.

Wang Q, Gao H, Jiang B, Li S, Zhang C (2020) Development and Application of a Multifunction True Triaxial Rock Drilling Test System. Journal of Testing and Evaluation 48(5): 3450-3467

Wang RR, Ni YD, Zhang LL, Gao BY (2025) Grouped machine learning methods for predicting rock mass parameters in a tunnel boring machine-driven tunnel based on fuzzy C-means clustering. Deep Underground Science & Engineering, 4(1): 12082.

Wang RR, Zhang LL (2023) K-means-based heterogeneous tunneling data analysis method for evaluating rock mass parameters along a TBM tunnel. Scientific Reports, 13(1): 21564.

Xue YD, Zhao F, Zhao HX, Li X, Diao ZX (2018) A new method for selecting hard rock TBM tunneling parameters using optimum energy: A case study. Tunneling and Underground Space Technology 78: 64-75

Yagiz S, Karahan H (2011) Prediction of hard rock TBM penetration rate using particle swarm optimization. International Journal of Rock Mechanics & Mining Sciences 48: 427-423

Yu B, Wang HZ, Shan WX. (2018) Prediction of bus travel time using random forests based on near neighbors. Computer-aided Civil and Infrastructure Engineering, 33:333-350.

Zare NM, Ramezanzadeh A (2017) Models for estimation of TBM performance in granitic and mica gneiss hard rocks in a hydropower tunnel. Bulletin of Engineering Geology and the Environment 76(4): 1627-1641

Zare NM, Samaei M, Ranjbarnia M, Nourani V (2018) State-of-the-art predictive modeling of TBM performance in changing geological conditions through gene expression programming. Measurement 126: 46-57

Zhang N, Li JB, Jing LJ, Li PY, Xu ST (2018) Study and Application of Intelligent Control System of TBM Tunneling Parameters. Tunnel construction 38(10): 1734-1740

How to Cite This Article

Meng, Q., Wang, H., Bai, L., Zhao, F., & Bai, J. (2026). Rock mass properties prediction method in TBM tunnel based on principal component analysis (PCA). Civil and Environmental Engineering, Vol. 22, Issue 2, 764-783. https://doi.org/10.2478/cee-2026-0051