Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.jzus.zju.edu.cn; www.springerlink.com
E-mail: jzus@zju.edu.cn

Machine learning-based prediction of soil compression modulus with application of 1D settlement*

Dong-ming ZHANG[1], Jin-zhang ZHANG[1], Hong-wei HUANG[†‡1], Chong-chong QI[†‡2], Chen-yu CHANG[3]

[1]*Department of Geotechnical Engineering, Tongji University, Shanghai 200092, China*

[2]*School of Resources and Safety Engineering, Central South University, Changsha 410083, China*

[3]*Bartlett Faculty of the Built Environment, University College London, London WC1E 7HB, UK*

[†]E-mail: huanghw@tongji.edu.cn; chongchong.qi@csu.edu.cn

Received Oct. 8, 2019; Revision accepted Mar. 9, 2020; Crosschecked May 23, 2020

Abstract: The compression modulus (E_s) is one of the most significant soil parameters that affects the compressive deformation of geotechnical systems, such as foundations. However, it is difficult and sometime costly to obtain this parameter in engineering practice. In this study, we aimed to develop a non-parametric ensemble artificial intelligence (AI) approach to calculate the E_s of soft clay in contrast to the traditional regression models proposed in previous studies. A gradient boosted regression tree (GBRT) algorithm was used to discern the non-linear pattern between input variables and the target response, while a genetic algorithm (GA) was adopted for tuning the GBRT model's hyper-parameters. The model was tested through 10-fold cross validation. A dataset of 221 samples from 65 engineering survey reports from Shanghai infrastructure projects was constructed to evaluate the accuracy of the new model's predictions. The mean squared error and correlation coefficient of the optimum GBRT model applied to the testing set were 0.13 and 0.91, respectively, indicating that the proposed machine learning (ML) model has great potential to improve the prediction of E_s for soft clay. A comparison of the performance of empirical formulas and the proposed ML method for predicting foundation settlement indicated the rationality of the proposed ML model and its applicability to the compressive deformation of geotechnical systems. This model, however, cannot be directly applied to the prediction of E_s in other sites due to its site specificity. This problem can be solved by retraining the model using local data. This study provides a useful reference for future multi-parameter prediction of soil behavior.

Key words: Compression modulus prediction; Machine learning (ML); Gradient boosted regression tree (GBRT); Genetic algorithm (GA); Foundation settlement

https://doi.org/10.1631/jzus.A1900515 **CLC number:** TU433

1 Introduction

With the rapid development of high-speed railways in China, the settlement of high-speed railway tracks is of great concern to engineers, as high-speed railways are extremely sensitive to foundation settlement which may cause serious accidents (Brabie and Andersson, 2008; Huang and Zhang, 2016). Therefore, there is a great practical need to effectively control ground surface and foundation settlement. The settlement of foundations in geotechnical engineering is determined mainly by the properties of the surrounding soil (Juang and Wang, 2013; Huang et al., 2017; Zhang et al., 2018). Among the soil properties, the compression modulus (E_s) and Poisson's ratio are two of the most significant parameters determining the deformation of soil and foundations (Fenton and Griffiths, 2008). Furthermore, the

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 51608380 and 51538009), the Key Innovation Team Program of the Innovation Talents Promotion Plan by Ministry of Science and Technology of China (No. 2016RA4059), and the Specific Consultant Research Project of Shanghai Tunnel Engineering Company Ltd. (No. STEC/KJB/XMGL/0130), China

ORCID: Hong-wei HUANG, https://orcid.org/0000-0001-6463-7869

compression modulus values are believed to show large spatial variation in deformation analysis (Fenton and Griffiths, 2008; Huang et al., 2017). Therefore, the ability to obtain accurate compression modulus values for soft clay precisely and efficiently is critical to the reliability of foundation design.

At present, the compression modulus is evaluated mainly in laboratory tests (Lee et al., 2010). However, disturbance and moisture loss during the process of collection, transportation, storage, and manual sample preparation are unavoidable (Sridharan and Nagaraj, 2000). This disturbance would affect the accuracy of the estimated value of the compression modulus. In addition, laboratory tests are relatively costly and time-consuming. The cone penetration test (CPT) is commonly used to estimate the E_s of soil because it can be carried out directly on the project site (Kulhawy and Mayne, 1990; Tong et al., 2013). Therefore, the geotechnical characteristics of the soil in the area of interest can be directly and accurately measured. The only requirement is a well-established translation from the cone tip resistance (P_s) from the CPT to the target compressive modulus E_s. In this regard, considerable effort has been devoted to establishing the correlation between CPT parameters and soil parameters (Kulhawy and Mayne, 1990; Lee et al., 2010). Much effort has also been devoted to establishing the relationship between the compressive modulus and other relatively readily available soil parameters, such as the plastic limit (w_P), liquid limit (w_L), plasticity index (I_P), and liquidity index (I_L) (Kulhawy and Mayne, 1990; Sridharan and Nagaraj, 2000). However, a generalized relationship between the empirical formula of one parameter and E_s is difficult to obtain. The parameters of different empirical formulas have great uncertainty. Therefore, empirical formulas are not very applicable to practical engineering. However, soil parameters are mutually influential (Ching and Phoon, 2014). Using multiple parameters to predict E_s will give better results than a single parameter, but this is difficult to achieve using traditional empirical formula methods.

To solve the above problem with geotechnical data analysis, machine learning (ML) algorithms recently developed in computer science have attracted substantial attention (Arditi and Pulket, 2005, 2010; Nejad et al., 2009). Through the application of ML algorithms, a system can become "intelligent" in self-understanding the relationship between input data and output data. ML models can learn the mapping correlation between inputs and outputs from the datasets. More specifically, these techniques have been proved to be practical for cases where the system's deterministic model is computationally expensive or there is no deterministic model to solve the problem. Lee et al. (2003) attempted to apply an ML algorithm to the prediction of unsaturated shear strength. Over the last decade, ML has been applied successfully to prediction problem in geotechnical engineering (Arditi and Pulket, 2010; Viswanathan and Samui, 2016; Tarawneh, 2017). Khanlari et al. (2012) implemented a method combining artificial neural networks and multivariate regression to predict the friction angle and cohesion of soils. Shahin (2016) provided a review of some selected artificial intelligence (AI) techniques and their application to pile foundations. Although there have been many attempts at ML approaches to deal with the prediction problem, the ML algorithms used were mostly single-base models. Ensemble learning algorithms provide a promising repertoire of tools in terms of their prediction accuracy. Instead of a single-base model, predictability can be strengthened by combining the outputs from multiple ML models. As a common and effective ensemble learning model, the gradient boosted regression tree (GBRT) model tends to be more stable and accurate than single-base models in prediction problems (Roe et al., 2005; Zhou et al., 2016). It has been demonstrated in several datasets that the prediction performance of GBRT is better than that of other ML algorithms (Roe et al., 2005; Qi and Tang, 2018). This method has been applied to various important engineering problems, including soil bulk density prediction (Jalabert et al., 2010), urban travel time (Gong et al., 2018), slope stability (Qi and Tang, 2018), and the strength of cemented paste backfill (Qi et al., 2018b).

Considering the limitations of existing models, in this study we develop a novel model by drawing upon the GBRT technique so that the prediction of compression modulus can consider the combined effects of P_s, w_P, w_L, I_P, I_L, and depth (H). The main contributions of this study can be summarized as follows: (1) a large dataset was prepared based on 65 site investigation reports in Shanghai; (2) a hybrid

method was used for relationship modelling that combines GBRT and genetic algorithm (GA); (3) hyper-parameter tuning was conducted and the predictive performance was validated; (4) the predictive performance of ML methods and empirical formulas was compared in calculating the 1D settlement of a shallow foundation.

2 Data acquisition and analysis

A practical prediction requires the easy input of parameters directly from in-situ or laboratory tests. Hence, the prediction of the parameter E_s in this study includes CPT data (i.e. P_s along with soil depth) from field investigation, and associated water content data from laboratory tests (i.e. w_P, w_L, I_P, and I_L).

2.1 Sampling and origin data acquisition

In this research, based on 65 site investigation reports in Shanghai, China (Fig. 1), E_s, w_P, w_L, I_P, I_L, and P_s values of soil were obtained. The P_s value was obtained from a CPT, and the E_s, w_P, w_L, I_P, and I_L values were acquired from borehole information. Two points can be considered strongly related when the distance between them is relatively small. This distance is expected to lie between 30 and 60 m (Huang et al., 2015, 2017). Therefore, the distance between the CPT and borehole locations was controlled within 30 m in this study. For inclusion, a sample required the six parameters to exist simultaneously. On this basis, a total of 211 samples were collected.

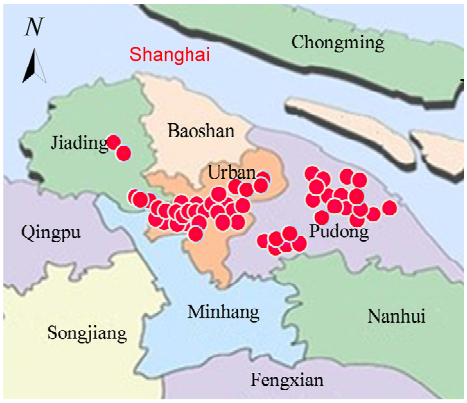

Fig. 1 Location of sampling sites in Shanghai, China

2.2 Data analysis

2.2.1 Model inputs and output

To obtain an accurate prediction, a thorough understanding of the factors influencing soil E_s is needed (Kulhawy and Mayne, 1990). Currently, most traditional prediction methods for E_s are empirical formulas with a single variable, such as w_P, w_L, I_P, or I_L (Kulhawy and Mayne, 1990). Previous studies (Lee et al., 2010; Tong et al., 2013) have shown that the E_s of soil can be estimated from P_s with reasonable accuracy. Prediction would be more effective using multiple measurement parameters than a single parameter (Ching and Phoon, 2014). Therefore, the input variables of the prediction model are P_s, H, w_P, w_L, I_P, and I_L. E_s is the single output variable in this study.

2.2.2 Original data analysis

The distributions of the six variables used in the prediction of E_s are illustrated in the diagonal line of Fig. 2. The vertical axis of the diagonal histogram represents the frequency. The upper triangle reveals the pairwise correlation of the model input variables. The correlation coefficients (R) are reported in the lower triangle. Most of the parameters are distributed in a concentrated manner. According to Koo and Li (2016), R values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively. There were relatively poor correlations between most input parameters (R<0.5). There was little correlation between depth and other model input variables.

3 ML and GA-GBRT modelling

In this study, the GBRT was employed as the ML algorithm to learn the non-linear pattern between compression modulus E_s and its influencing variables. The hyper-parameters need to be optimized to improve prediction performance. GA is considered an effective hyper-parameter optimization algorithm (Johari et al., 2011). Therefore, the GBRT hyper-parameters are tuned using GA. In this study, the GA-GBRT method was implemented by using Python Programming.

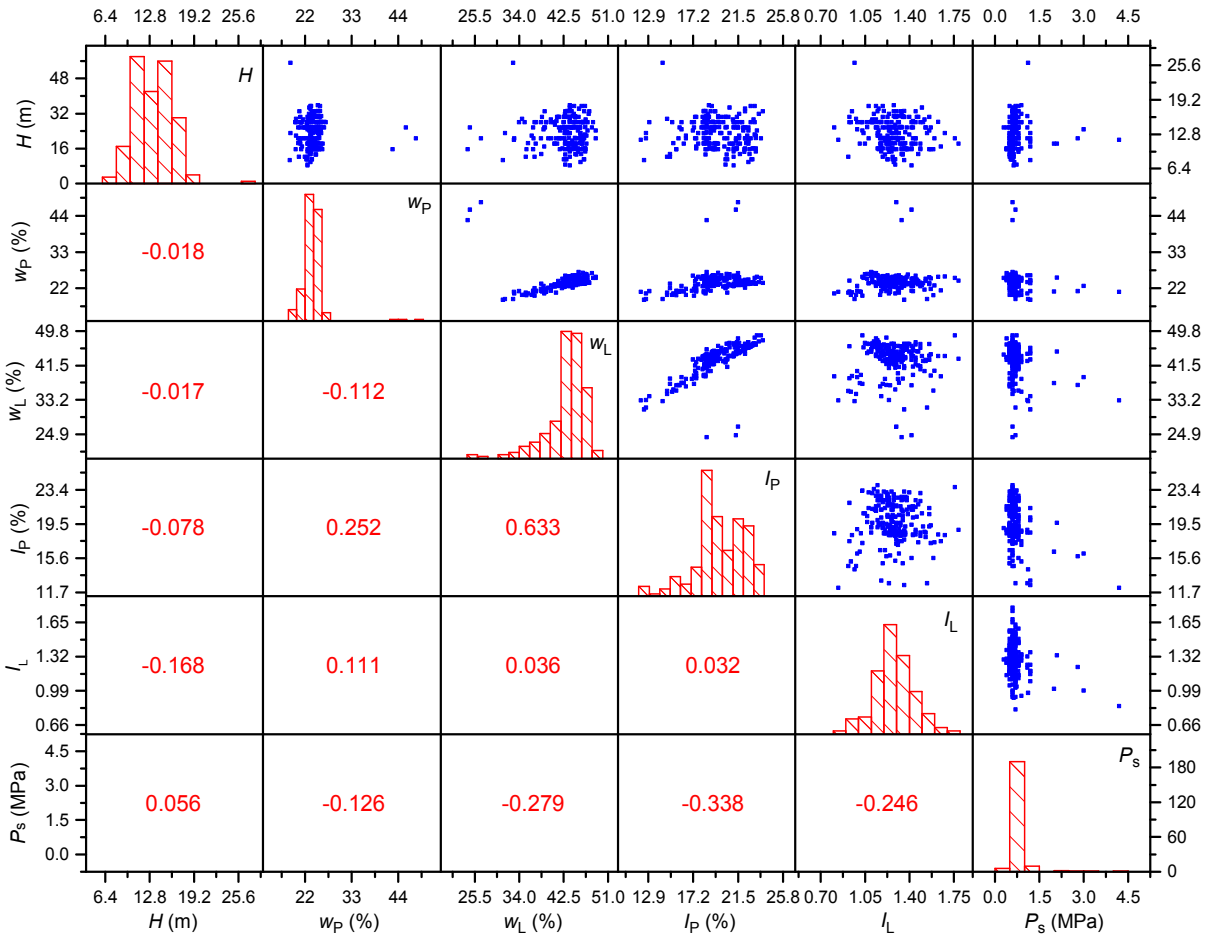

Fig. 2 Distributions and correlation coefficients of model input variables in the dataset

3.1 GBRT

As an ensemble learning model, the GBRT model tends to be more stable and accurate than single-base models (Zhou et al., 2016). The base learner of the GBRT model is a decision tree (DT). A DT is an algorithm that classifies and predicts new data by measuring historical data. The DT is also known as a regression tree (RT) for regression problems.

An RT consists of three main components: root decision nodes, intermediate nodes, and leaf nodes. These nodes are connected by branches. The structure of a typical RT is shown in Fig. 3. An RT can be used to make predictions by dividing the feature space into several regions and making predictions for each of them. Take Fig. 3 as an example with two influencing variables, namely X and Y. Based on these two variables, the dataset can be easily divided into four areas. Each of the four subsets has a central point or an average value: (1, 1), (1, 4), (4, 4), and (4, 1). In forecasting a new sample, the center point of the subset in which the sample falls is used as its predicted value. The training of an RT usually involves growth and pruning. The essence of the RT growth process lies in the process of repeated branching on the training set. Growth will stop when the data branching is no longer significant or the maximum tree depth is reached. Therefore, RT growth is chiefly controlled by branching criteria. Over-fitting can be avoided through a pruning process in which the weakest branches with little potential to improve the generalization capability are collapsed.

The core idea of the gradient boosting algorithm is that each tree can learn from the residuals of all previous trees. The negative gradient value of the loss function in the current model is used as an approximation of the residual in the boosting tree algorithm to fit an RT. Recently, this approach has gained increasingly popularity in various scientific and

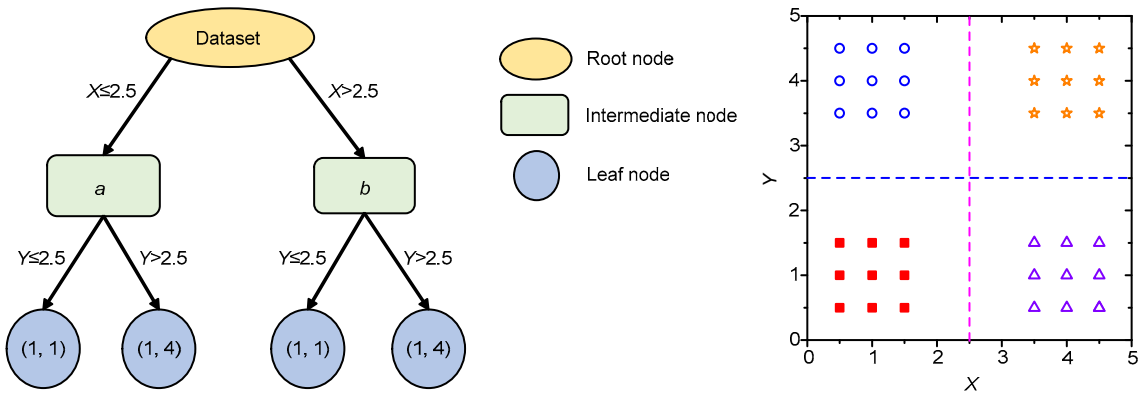

Fig. 3 An example of basic rules of the RT algorithm (*a* and *b* mean two intermediate categories)

engineering fields, such as energy consumption (Persson et al., 2017; Touzani et al., 2018), transportation (Gong et al., 2018), and civil and industrial engineering (Qi and Tang, 2018; Qi et al., 2018a, 2018b).

3.2 GA

A GA, also known as a genetic evolution algorithm, is a modern intelligent algorithm that simulates the survival of the fittest natural genetic mechanism in the biological universe. Local convergence can be overcome by using a GA method. In recent years, this method has been increasingly applied to various optimization problems (Johari et al., 2011; Juang and Wang, 2013; Tun et al., 2016; Yin et al., 2016).

A global and robust solution can be obtained by using the GA due to its global search strategy and optimization search method (Goldberg, 1989). As illustrated in the flowchart of Fig. 4, the GA procedure contains the following steps: (1) creation of initial individuals; (2) evaluation based on a fitness function; (3) creation of the next generation through selection, crossover, and mutation. The iterations are repeated until a specified stopping criterion is satisfied. A typical stopping criterion in a GA is a predefined maximum number of generations.

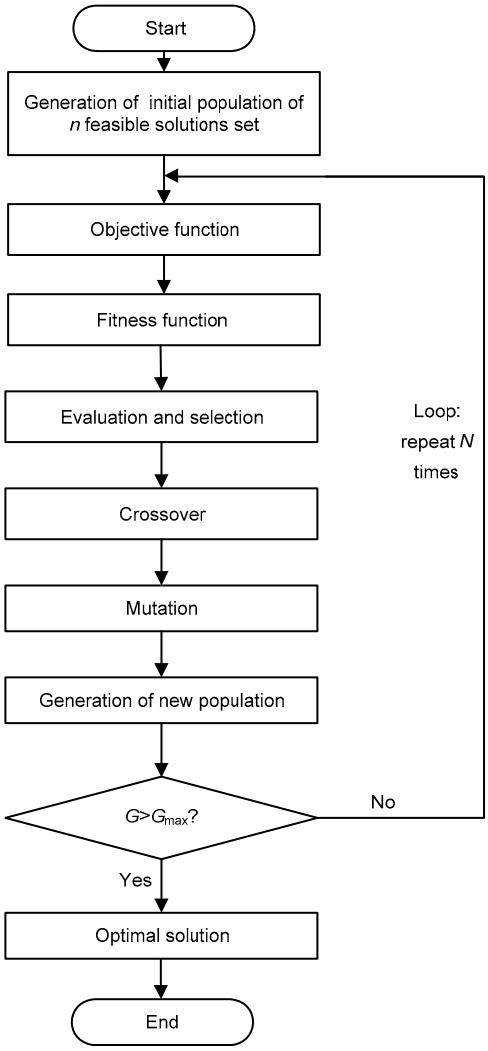

Fig. 4 Generic framework for a simple GA

3.3 GA-GBRT modelling

3.3.1 Data division

Before modelling, the whole dataset is divided into two subsets, namely the training and testing sets. The training set is used to train the model, and the testing set is applied to assess the generalization performance of the trained model. In practice, how to divide the data into the training and testing sets is often determined by an optimization analysis (Qi et

al., 2018a). In this study, a trial-and-error method was used to determine the percentage sizes of the training set and the testing set. More specifically, the training set size was increased from 30% to 90%, and the prediction performance was recorded. The optimum percentages of the training set and the testing set were determined to be 80% and 20%, respectively, after the trial-and-error analysis. The numbers of training and testing data points were 177 and 44, respectively. Generally, the testing set should be as mutually exclusive of the training set as possible. The training and testing subsets should have similar statistical characterizations since they are drawn randomly from the whole dataset (Shahin et al., 2004). A quick check can be done by examining important statistics of the input and output variables, including the mean, standard deviation, minimum, maximum, median, skewness, and kurtosis (Shahin et al., 2004; Nejad et al., 2009; Qi et al., 2018b). The skewness and kurtosis are defined by Eqs. (1) and (2), respectively. The statistics of the training and testing sets were generally very consistent (Table 1) and all the datasets could be considered to represent the same population.

$$\text{Skewness}=\frac{\overline{x}-x_{\text{median}}}{\text{STD}}, \tag{1}$$

$$\text{Kurtosis}=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\text{STD}}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}, \tag{2}$$

where x_i means the value of each point, $\overline{x}$ represents the mean value, x_{median} means the median value, n means the number of data points, and STD indicates the standard deviation of the data.

3.3.2 Performance measures

In this study, the predictive performance of the GBRT model was evaluated using the mean squared error (MSE) and the correlation coefficient R. The MSE is the mean value of the squares of error between the predicted data and the original data. The R value is an indicator of the degree of correlation between variables. A positive R value means that the dependent variable increases with the independent variable, and the fitted straight line rises from left to right. Conversely, a negative value of R shows a pattern declining from left to right. The prediction performance is better when the R value is close to +1. The MSE and R values can be calculated by:

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-x_i)^2, \tag{3}$$

$$R=\frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2}}, \tag{4}$$

where n is the number of samples, x_i and y_i are the experimental and predicted E_s values, respectively, and $\overline{x}$ and $\overline{y}$ represent the mean values of the experimental and predicted data, respectively.

Table 1 Statistical description of inputs and outputs

Variable	Type	Maximum	Minimum	Mean	Median	STD	Skewness	Kurtosis
P_s (MPa)	Input-training	3.0	0.3	0.681	0.600	0.282	0.289	32.875
	Input-test	4.20	0.40	0.798	0.600	0.641	0.310	21.236
H (m)	Input-training	26.0	7.1	12.864	12.850	2.748	0.005	1.816
	Input-test	17.70	7.00	13.021	13.100	2.799	−0.028	−0.958
w_P (%)	Input-training	48.2	18.6	23.983	23.500	3.327	0.145	32.497
	Input-test	26.60	18.40	23.344	23.700	1.871	−0.190	0.001
w_L (%)	Input-training	48.7	24.2	43.064	44.050	4.026	−0.245	6.267
	Input-test	47.30	30.90	42.330	42.900	3.772	−0.151	1.459
I_P (%)	Input-training	23.9	12.7	19.811	19.600	2.291	0.092	0.438
	Input-test	23.50	12.20	18.986	18.800	2.510	0.074	0.842
I_L	Input-training	1.79	0.81	1.278	1.270	0.161	0.051	0.843
	Input-test	1.57	0.84	1.279	1.290	0.156	−0.068	0.750
E_s (MPa)	Output-training	6.27	1.50	2.340	2.210	0.635	0.212	10.897
	Output-test	5.91	1.61	2.461	2.280	0.787	0.231	9.950

3.3.3 *k*-fold cross validation

To improve evaluation of the performance of the prediction model, it is essential to select appropriate validation methods carefully during hyper-parameter tuning. *k*-fold cross validation (CV) is the most popular method used to overcome a scarcity of data (Braga-Neto et al., 2004). The *k*-fold CV reduces the variance by averaging the results of *k* different folds of the training set. In this study, the *k* value was set to 10, based on the recommendation of Rodriguez et al. (2010). The 10-fold CV flow chart used in this study was as follows. First, the training set was randomly split into 10 equal-sized folds. Next, one-fold was selected as the validating fold and the remaining nine were used as the training folds. The training-validating process was then carried out 10 times so that each fold had an opportunity to serve as the validating fold. Performance evaluation indicators of the model (e.g. MSE and *R*) were calculated each time. Finally, the average performance of the 10 iterations was taken as the overall performance indicator for the model on the training set.

3.3.4 Hyper-parameter tuning

The hyper-parameters of the GBRT model must be pre-determined before its implementation. Performance may differ under a different combination of hyper-parameters. Therefore, hyper-parameter tuning is crucial for successful GBRT modelling. In this study, the GA was used to tune the hyper-parameters of the GBRT algorithms. The GA parameters used for hyper-parameter tuning of the GBRT are shown in Table 2.

Table 2 GA parameters used for hyper-parameter tuning

GA parameter	Description
Fitness function	Correlation coefficient
Selection method	Tournament (size=3)
Genetic possibility	Crossover (80%), mutation (5%)
Number of chromosomes	1000
Number of generations	200

The tuned hyper-parameters of the GBRT are shown in Table 3, together with their tuning ranges. An appropriate tuning range of hyper-parameters greatly improves the efficiency and accuracy of training. The range was determined in accordance with trial tests, modelling experience, and suggestions from previous studies (Qi et al., 2018a, 2018b). The whole procedure for the application of GA-GBRT approach to predict the compression modulus is summarized in Fig. 5.

Table 3 Explanation of hyper-parameters and tuning range

Hyper-parameter	Explanation	Type	Tuning range
Max_depth	The maximum depth of the RT	Integer	1–20
Min_samples_split	The minimum number of samples required to split an internal node	Integer	2–10
Min_samples_leaf	The minimum number of samples at the leaf node	Integer	1–20
Max_RT	The maximum number in the AdaBoost.R2 algorithm	Integer	50–1000
Learning rate	Learning rate shrinking the contribution of each RT model	Float	0.01–1

4 GA-GBRT results and discussion

4.1 Results of hyper-parameter tuning

To evaluate the capability of the GA in the tuning of GBRT hyper-parameters, the maximum and average *R* values of each generation were monitored during the evolution. Fig. 6 shows the maximum and average *R* values from the first sixteen generations of the GA evolution. There was no evident increase in the maximum *R* value after the sixth generation (Fig. 6a). Similarly, the average *R* value remained stable, though with small fluctuations, after the 12th generation (Fig. 6b).

The maximum and average *R* values eventually approached 0.816. The average *R* value increased from 0.700 to 0.816 during the hyper-parameter tuning. From the first iteration to the 12th iteration, the average *R* value increased by 0.116. In contrast, the improvement in the maximum *R* value (0.007) was not as significant. An increase of 16.6% in average *R*

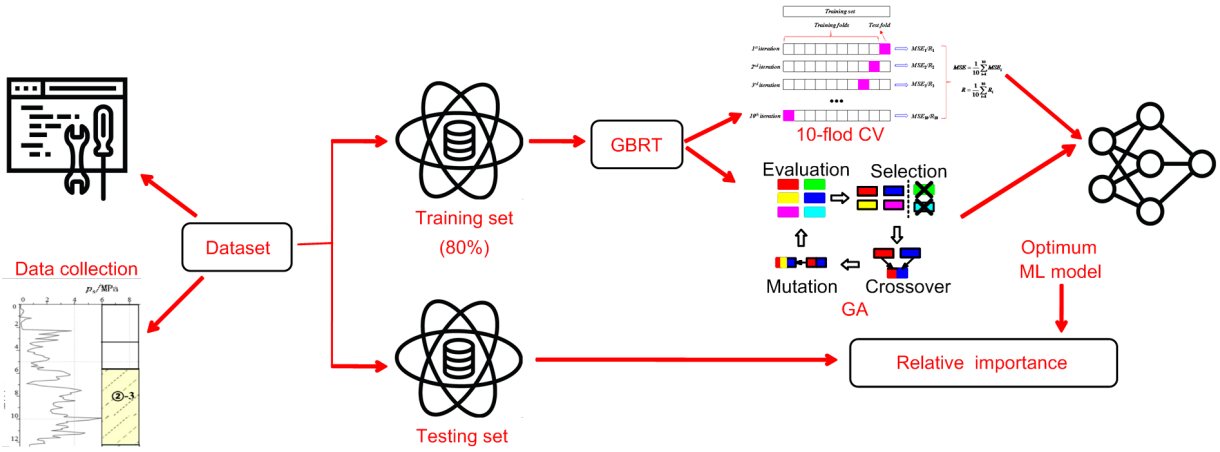

Fig. 5 Procedure for using GA-GBRT approach for compression module prediction

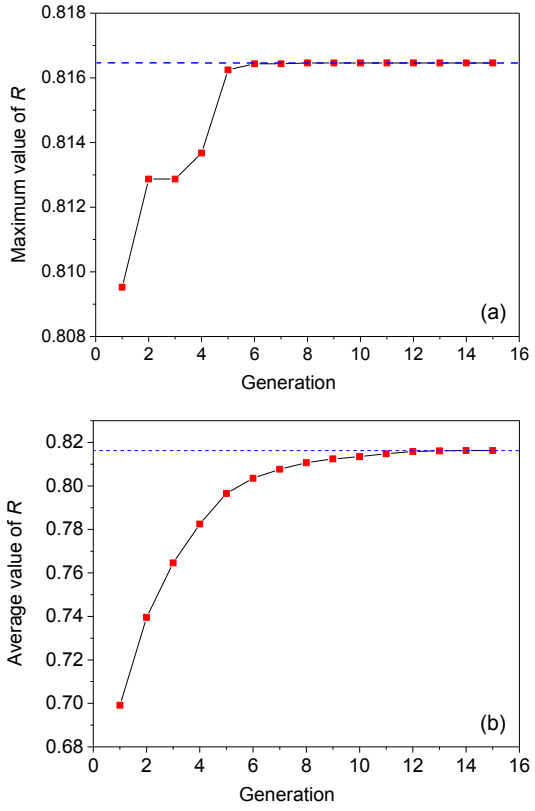

Fig. 6 Maximum (a) and average (b) of *R* values versus generation

value was achieved in the first six iterations. The largest *R* value was gained before the 12th generation and a rapid increase was achieved in the first six generations, indicating that the use of GA in GBRT hyper-parameter tuning was valid. The optimum hyper-parameters for the GBRT models are generalized in Table 4.

Table 4 Optimal hyper-parameter results

Max_ depth	Min_samples _split	Min_samples _leaf	Max_ RT	Learning rate
1	6	4	374	0.3967

4.2 Results of the optimum GBRT model

MSE and *R* were selected as evaluation indexes. The evaluation of the GBRT model was performed on both the training set and the testing set. The general predictive performance of the GBRT model with optimum hyper-parameters on the training set was quite acceptable (Fig. 7). Fig. 7a provides a visual comparison of predicted and experimental E_s values. Regardless of the E_s value, the predicted value was relatively consistent with the experimental value. Therefore, the successful modelling of the soil E_s using the optimum GBRT model indicated that GBRT modelling has great potential for making more reliable predictions.

The results of regression analysis of predicted and experimental E_s values are shown in Fig. 7b. Most points fell around the ideal fitting line, and the *R* value between the experimental and predicted E_s values was 0.82, which was consistent with the analysis result in Section 4.1. As suggested in previous studies, a GBRT model with *R* values larger than 0.8 can be regarded as acceptable (Roy and Roy, 2008; Qi et al., 2018b). The MSE of the GBRT model with optimal hyper-parameters was 0.15 MPa^2, also indicating that a relatively good GBRT model had been achieved on the training set.

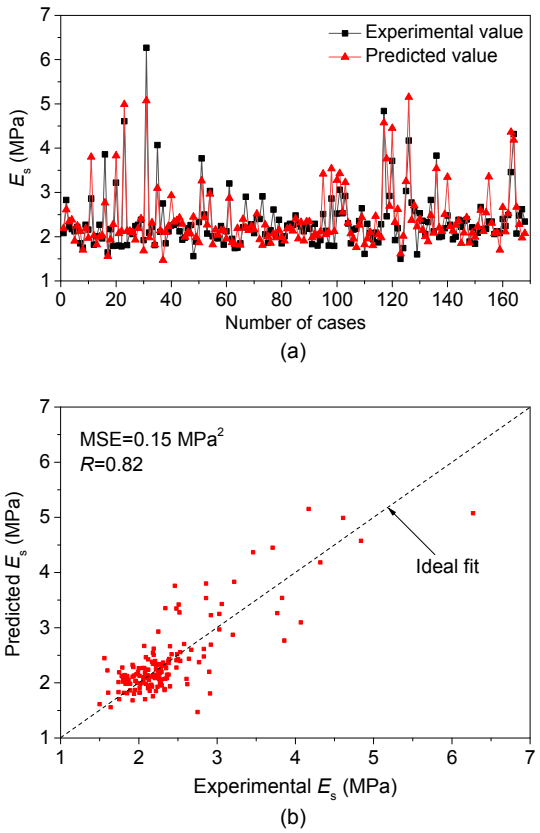

Fig. 7 Experimental versus predicted E_s for the GBRT model with the optimum hyper-parameters on the training set: (a) comparison of experimental and predicted E_s; (b) regression

Fig. 8 compares the experimental and predicted E_s values on the testing set. The prediction performance of the testing set was mostly excellent except for some extremely large or small samples (Fig. 8a). The MSE and R values of the optimum GBRT model on the testing set were 0.13 MPa^2 and 0.91, respectively (Fig. 8b), indicating excellent prediction performance of the optimum GBRT model (Koo and Li, 2016). The optimum GBRT model on the testing set had a higher prediction performance because more samples were observed during its training.

4.3 Relative importance of influencing variables

The ability to investigate the relative importance of the influencing variables is another crucial reason why the GBRT model is widely used in prediction problems. It can be used to rank the influencing variables according to their contribution to the performance of the prediction model. To reveal the effect of influencing variables on the predicted E_s values of soft clay soil, a sensitivity analysis was conducted for these influencing variables.

The importance of variables for E_s previously obtained from the optimum GBRT model is summarized in Fig. 9. The sum of all importance scores was scaled to 1 for ease of interpretation. I_L (liquidity index) was the most sensitive variable for predicting E_s with a relative importance score of 0.313. This score is defined as the ratio of the difference between the natural moisture content (w) and the plastic limit w_P to I_P (Eq. (5)). I_L is meant to capture the relative relationship between the natural moisture content and the boundary water content. The influence of moisture content on the E_s of soil has been extensively studied (Kulhawy and Mayne, 1990; Fan et al., 2006). I_L reflects the hard or soft natural state of clay soil. The greater the I_L, the softer the soil and the smaller the corresponding E_s. Therefore, I_L has a significant effect on the E_s of soil compared to other factors.

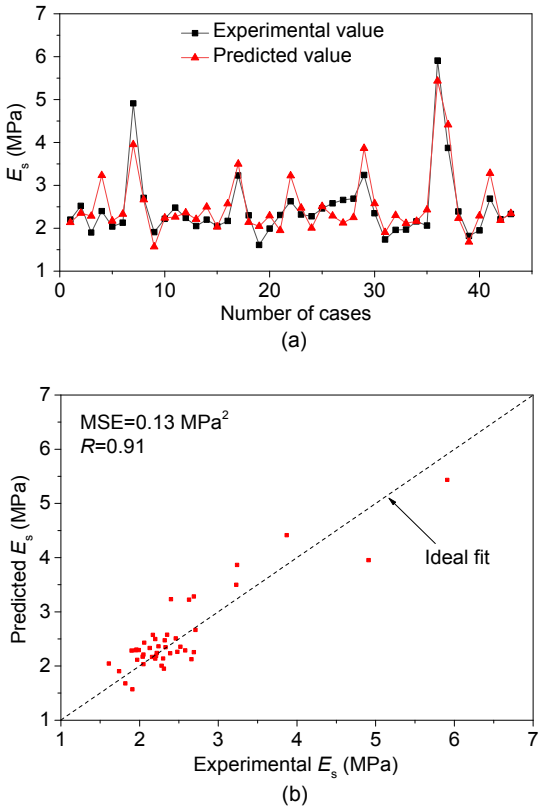

Fig. 8 Experimental versus predicted E_s for the GBRT model with the optimum hyper-parameters on the testing set: (a) comparison of experimental and predicted E_s; (b) regression

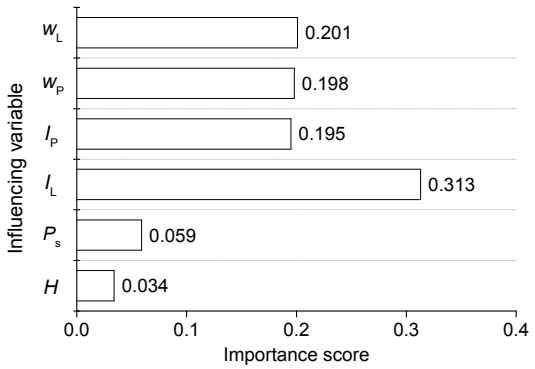

Fig. 9 Importance of variables influencing E_s predictions obtained by the optimum GBRT model

$$I_L = \frac{w - w_P}{I_P} = \frac{w - w_P}{w_L - w_P}. \tag{5}$$

The importance scores for w_L, w_P, and I_P were 0.201, 0.198, and 0.195, respectively. Therefore, these variables are also important predictors. w_L and w_P are moisture content indicators which can determine the physical state of the clay soil. I_P is an important index for measuring the plasticity of soil and can fully reflect its material composition. P_s and H had relatively low importance scores in this prediction model. Note that different importance scores may be obtained when using different datasets and ML models (Qi and Tang, 2018).

5 Application to a 1D settlement

To verify whether the performance of the GA-GBRT model is better than that of the existing empirical formulas, the predicted E_s was applied to the calculation of foundation settlement.

5.1 Comparison of empirical formulas and ML training effects

According to Clayton et al. (1995), the relationship between E_s and the compression index (C_c) is shown as

$$E_s = \frac{1 + e_1}{10 C_c \times \lg 2}, \tag{6}$$

where e_1 is the void ratio at a pressure of 0.1 MPa. The oedometer testing requires undisturbed samples and is quite time-consuming and expensive. For this reason, many previous studies correlated compressibility characteristics with other soil properties. Sridharan and Nagaraj (2000) correlated the compression index with various other individual soil parameters. For example, the relationships between C_c and w_L, and C_c and I_P are shown in the corresponding Eqs. (7) and (8). The empirical formula was obtained from 10 kinds of soil, including silty soil and clay. The dataset was mainly from Shanghai's clay soil, which is suitable for this formula.

$$C_c = 0.008 \times (w_L - 12), \tag{7}$$

$$C_c = 0.014 \times (I_P + 3.6). \tag{8}$$

Based on Eqs. (6)–(8), the relationships between E_s and w_L, and E_s and I_P are shown in the corresponding Eqs. (9) and (10). In this study, w_L and I_P were selected for comparing the results obtained by the optimum GA-GBRT in testing results.

$$E_s = \frac{1 + e_1}{0.08 \lg 2 \times (w_L - 12)}, \tag{9}$$

$$E_s = \frac{1 + e_1}{0.14 \lg 2 \times (I_p + 3.6)}. \tag{10}$$

The performance of the various prediction methods was compared using the testing set, and the results are shown in Fig. 10. The present ML method applied to the testing set was more accurate than the empirical formulas. For convenience of expression, the empirical formulas of E_s obtained by w_L and I_P are simply referred to as empirical formulas w_L and I_P. From Fig. 10a, the predicted E_s values from empirical formula w_L were generally greater than the experimental value. The predicted E_s values from empirical formula I_P were relatively good when the experimental values were small. However, the predicted results were relatively poor when the experimental values were large. Lower MSE and higher R values were achieved by the proposed ML model than the empirical formulas (Fig. 10b), showing that the proposed ML model was significantly better than the empirical formulas at predicting the soil E_s.

5.2 Comparison of foundation settlement

The method used in this study to calculate foundation settlement was a layer-wise summation

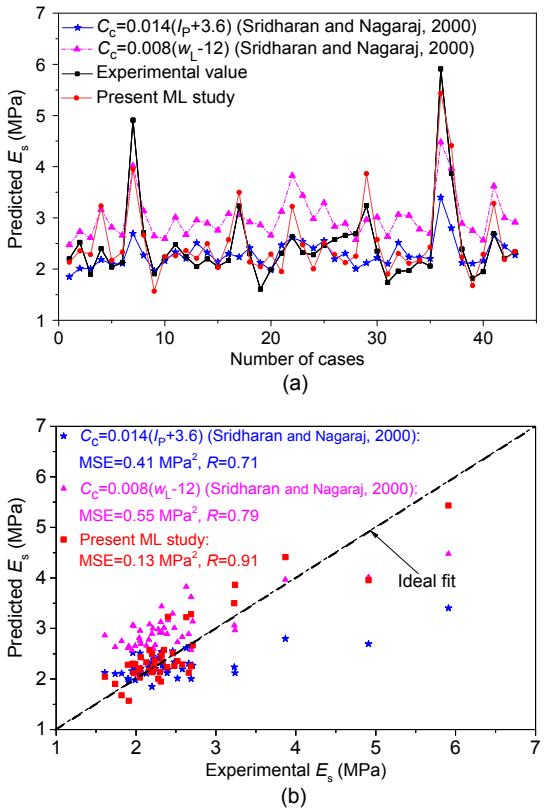

Fig. 10 Comparison of performance between empirical formulas and the proposed ML method for predicting E_s: (a) experimental and predicted E_s; (b) regression

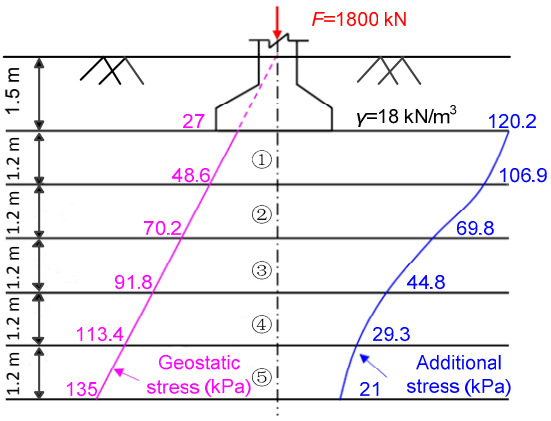

Fig. 11 Foundation soil stratification, geostatic stress, and additional stress

F is the central load; γ is the unit weight of soil

Table 5 Calculation of the foundation settlement by layer-wise summation method

Layer number	Depth H_i (m)	Geostatic stress (kPa)	Additional stress (kPa)	Average additional stress Δp_i (kPa)	Stratified compression Δs_i (mm)
–	0.0	27.0	120.2	–	–
1	1.2	48.6	106.9	113.6	$\Delta p_1 H_1/E_{s1}$
2	2.4	70.2	69.8	88.4	$\Delta p_2 H_2/E_{s2}$
3	3.6	91.8	44.8	57.3	$\Delta p_3 H_3/E_{s3}$
4	4.8	113.4	29.3	37.1	$\Delta p_4 H_4/E_{s4}$
5	6.0	135.0	21.0	25.2	$\Delta p_5 H_5/E_{s5}$

E_{si} indicates the compression modulus of the ith soil layer

method. The results of the calculation are shown in Fig. 11 and Table 5. The results from a comparison of the performance of the empirical formulas and the ML method in predicting foundation settlement are shown in Fig. 12. Six aspects of performance were compared: the maximum, minimum, mean, STD, R value, and the Mann-Whitney test. Fig. 12a shows the comparison between the calculated foundation settlement from the experimental E_s and predicted E_s using the empirical formula w_L. The predicted result distribution is more concentrated and the mean value of settlement much smaller. The R value was 0.69, indicating a relatively moderate prediction performance (Koo and Li, 2016). The null hypothesis that the two sets of samples came from the same probability distribution can be rejected at the customary 5% level of significance because the p-value was 0.00< 0.05. The poor consistency also implies a limitation of the empirical formula w_L for practical cases.

The prediction performance of the empirical formula I_P method is shown in Fig. 12b. Although the p-value of the Mann-Whitney test was 0.51>0.05, the distribution range was significantly narrower, especially without some small settlement. This means that in practice this method will overestimate the settlement of the foundation. The results obtained by the present ML method were more satisfactory than those from the empirical formulas (Fig. 12c). The distribution range, mean value, and STD value from the ML method were similar to the corresponding actual results. The R value was 0.81, indicating a relatively good prediction performance (Koo and Li, 2016). In addition, the p-value of the Mann-Whitney test was 0.77>0.05, which means that the possibility that the two sets of samples came from the same probability distribution cannot be rejected at the customary 5% level of significance. The high level of agreement also validates the rationality of the proposed ML model and its applicability to practical cases.

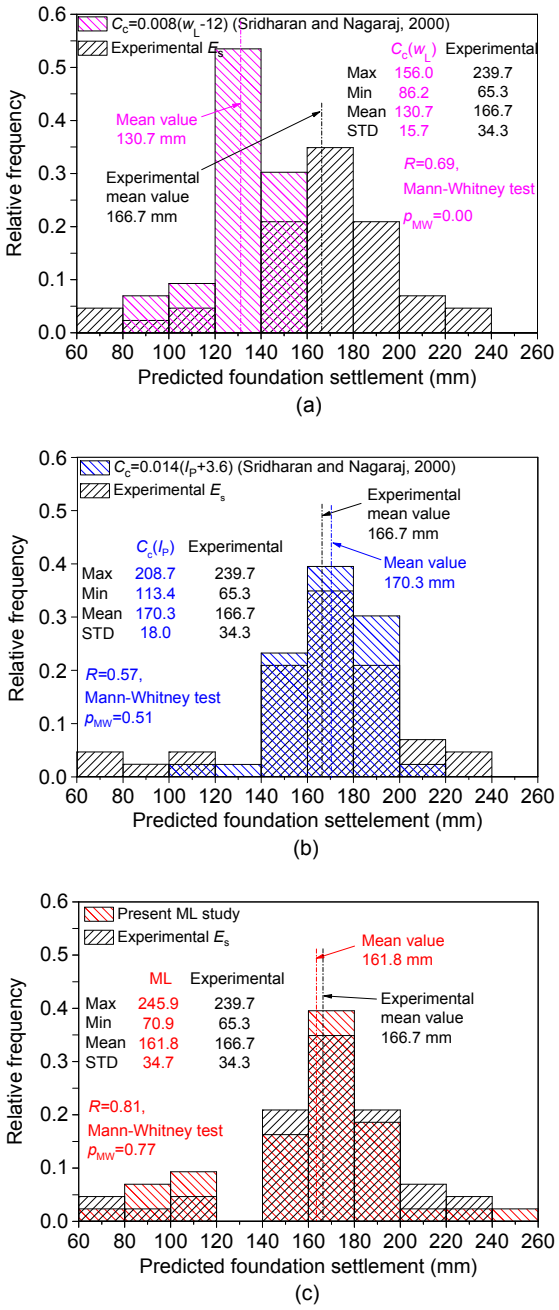

Fig. 12 Comparison of performance between empirical formulas and present ML method for predicting foundation settlement: (a) empirical formula w_L; (b) empirical formula I_P; (c) present ML method

6 Conclusions

In this study, an integrated GA-GBRT model was constructed to predict the E_s of soil and foundation settlement. A total of 221 soil samples from 65 engineering survey reports of Shanghai infrastructure projects were collected for preparation of the dataset. The input variables considered by the prediction model were P_s, H, w_P, w_L, I_P, and I_L, and the output variable was the E_s. A 10-fold CV was used as the validation method, and MSE and R values as the performance measures. The relative importance of influencing variables was investigated using the results from the optimum GA-GBRT model. Based on the results of the analysis, the following conclusions can be drawn:

1. A GA can effectively assist the hyperparameter tuning of ML algorithms, as the optimum R value was obtained within the first 12 iterations.

2. The optimum GA-GBRT model performed quite well on both the training and testing sets. The R values between the predicted and experimental E_s values were 0.82 and 0.91 on the training and testing sets, respectively, indicating that an accurate prediction was achieved by the optimum GA-GBRT model.

3. The relative importance of the influencing variables was studied, and the liquidity index was found to be the most important variable in this study, achieving an importance score of 0.313 out of 1.

4. For predicting the settlement of a foundation, the proposed ML method performed better than empirical formulas in terms of both the R value and Mann-Whitney test results. The results of this study can serve as a benchmark for further research, and the proposed GA-GBRT model can be used to obtain a more cost-effective and faster prediction of the E_s of soil.

Acknowledgement

The authors thank Li XIAO and Ye-lu ZHOU from Tongji University, China for their help in collecting the original data for this study.

Contributors

Hong-wei HUANG and Jin-zhang ZHANG designed the research. Dong-ming ZHANG and Chong-chong QI processed the corresponding data. Jin-zhang ZHANG wrote the first draft of the manuscript. Dong-ming ZHANG and Chong-chong QI helped to organize the manuscript. Hong-wei HUANG and Chen-yu CHANG revised and edited the final version.

Conflict of interest

Dong-ming ZHANG, Jin-zhang ZHANG, Hong-wei HUANG, Chong-chong QI, and Chen-yu CHANG declare that they have no conflict of interest.

References

Arditi D, Pulket T, 2005. Predicting the outcome of construction litigation using boosted decision trees. *Journal of Computing in Civil Engineering*, 19(4):387-393. https://doi.org/10.1061/(asce)0887-3801(2005)19:4(387)

Arditi D, Pulket T, 2010. Predicting the outcome of construction litigation using an integrated artificial intelligence model. *Journal of Computing in Civil Engineering*, 24(1): 73-80. https://doi.org/10.1061/(asce)0887-3801(2010)24:1(73)

Brabie D, Andersson E, 2008. An overview of some high-speed train derailments: means of minimizing consequences based on empirical observations. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 222(4):441-463. https://doi.org/10.1243/09544097jrrt149

Braga-Neto U, Hashimoto R, Dougherty ER, et al., 2004. Is cross-validation better than resubstitution for ranking genes? *Bioinformatics*, 20(2):253-258. https://doi.org/10.1093/bioinformatics/btg399

Ching J, Phoon KK, 2014. Correlations among some clay parameters—the multivariate distribution. *Canadian Geotechnical Journal*, 51(6):686-704. https://doi.org/10.1139/cgj-2013-0353

Clayton CRI, Steinhagen M, Steinhagen HM, et al., 1995. Terzaghi's theory of consolidation, and the discovery of effective stress (compiled from the work of K. Terzaghi and A.W. Skempton). *Proceedings of the Institution of Civil Engineers–Geotechnical Engineering*, 113(4):191-205. https://doi.org/10.1680/igeng.1995.28015

Fan HH, Wu PT, Gao JE, et al., 2006. Influence of density and water content on unconfined compression strength of solidified soil. *Science of Soil and Water Conservation*, 4(3):54-58 (in Chinese). https://doi.org/10.3969/j.issn.1672-3007.2006.03.011

Fenton GA, Griffiths DV, 2008. Risk Assessment in Geotechnical Engineering. Wiley, New York, USA, p.78-101. https://doi.org/10.1002/9780470284704.ch5

Goldberg DE, 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing, New York, USA.

Gong Y, Luo XQ, Wang DH, et al., 2018. Urban travel time prediction based on gradient boosting regression tress. *Journal of Zhejiang University (Engineering Science)*, 52(3):453-460 (in Chinese). https://doi.org/10.3785/j.issn.1008-973X.2018.03.006

Huang HW, Zhang DM, 2016. Resilience analysis of shield tunnel lining under extreme surcharge: characterization and field application. *Tunnelling and Underground Space Technology*, 51:301-312. https://doi.org/10.1016/j.tust.2015.10.044

Huang HW, Gong WP, Khoshnevisan S, et al., 2015. Simplified procedure for finite element analysis of the longitudinal performance of shield tunnels considering spatial soil variability in longitudinal direction. *Computers and Geotechnics*, 64:132-145. https://doi.org/10.1016/j.compgeo.2014.11.010

Huang HW, Xiao L, Zhang DM, et al., 2017. Influence of spatial variability of soil Young's modulus on tunnel convergence in soft soils. *Engineering Geology*, 228:357-370. https://doi.org/10.1016/j.enggeo.2017.09.011

Jalabert SSM, Martin MP, Renaud JP, et al., 2010. Estimating forest soil bulk density using boosted regression modelling. *Soil Use and Management*, 26(4):516-528. https://doi.org/10.1111/j.1475-2743.2010.00305.x

Johari A, Javadi AA, Habibagahi G, 2011. Modelling the mechanical behaviour of unsaturated soils using a genetic algorithm-based neural network. *Computers and Geotechnics*, 38(1):2-13. https://doi.org/10.1016/j.compgeo.2010.08.011

Juang CH, Wang L, 2013. Reliability-based robust geotechnical design of spread foundations using multi-objective genetic algorithm. *Computers and Geotechnics*, 48(4):96-106. https://doi.org/10.1016/j.compgeo.2012.10.003

Khanlari GR, Heidari M, Momeni AA, et al., 2012. Prediction of shear strength parameters of soils using artificial neural networks and multivariate regression methods. *Engineering Geology*, 137-138:11-18. https://doi.org/10.1016/j.enggeo.2011.12.006

Koo TK, Li MY, 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kulhawy FH, Mayne PW, 1990. Manual on Estimating Soil Properties for Foundation Design. Electric Power Research Institute, Palo Alto, USA.

Lee MJ, Hong SJ, Choi YM, et al., 2010. Evaluation of deformation modulus of cemented sand using CPT and DMT. *Engineering Geology*, 115(1-2):28-35. https://doi.org/10.1016/j.enggeo.2010.06.016

Lee SJ, Lee SR, Kim YS, 2003. An approach to estimate unsaturated shear strength using artificial neural network and hyperbolic formulation. *Computers and Geotechnics*, 30(6):489-503. https://doi.org/10.1016/s0266-352x(03)00058-2

Nejad FP, Jaksa MB, Kakhi M, et al., 2009. Prediction of pile settlement using artificial neural networks based on standard penetration test data. *Computers and Geotechnics*, 36(7):1125-1133. https://doi.org/10.1016/j.compgeo.2009.04.003

Persson C, Bacher P, Shiga T, et al., 2017. Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy*, 150:423-436. https://doi.org/10.1016/j.solener.2017.04.066

Qi CC, Tang XL, 2018. Slope stability prediction using integrated metaheuristic and machine learning approaches: a

comparative study. *Computers & Industrial Engineering*, 118:112-122. https://doi.org/10.1016/j.cie.2018.02.028

Qi CC, Fourie A, Ma GW, et al., 2018a. Comparative study of hybrid artificial intelligence approaches for predicting hangingwall stability. *Journal of Computing in Civil Engineering*, 32(2):04017086. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000737

Qi CC, Fourie A, Chen QS, et al., 2018b. A strength prediction model using artificial intelligence for recycling waste tailings as cemented paste backfill. *Journal of Cleaner Production*, 183:566-578, https://doi.org/10.1016/j.jclepro.2018.02.154

Rodriguez JD, Perez A, Lozano JA, 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):569-575. https://doi.org/10.1109/TPAMI.2009.187

Roe BP, Yang HJ, Ji Z, et al., 2005. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2-3):577-584. https://doi.org/10.1016/j.nima.2004.12.018

Roy PP, Roy K, 2008. On some aspects of variable selection for partial least squares regression models. *QSAR & Combinatorial Science*, 27(3):302-313. https://doi.org/10.1002/qsar.200710043

Shahin MA, 2016. State-of-the-art review of some artificial intelligence applications in pile foundations. *Geoscience Frontiers*, 7(1):33-44. https://doi.org/10.1016/j.gsf.2014.10.002

Shahin MA, Maier HR, Jaksa MB, 2004. Data division for developing neural networks applied to geotechnical engineering. *Journal of Computing in Civil Engineering*, 18(2):105-114. https://doi.org/10.1061/(asce)0887-3801(2004)18:2(105)

Sridharan A, Nagaraj HB, 2000. Compressibility behaviour of remoulded, fine-grained soils and correlation with index properties. *Canadian Geotechnical Journal*, 37(3):712-722. https://doi.org/10.1139/t99-128

Tarawneh B, 2017. Predicting standard penetration test *N*-value from cone penetration test data using artificial neural networks. *Geoscience Frontiers*, 8(1):199-204. https://doi.org/10.1016/j.gsf.2016.02.003

Tong LY, Tu QZ, Du GY, et al., 2013. Determination of confined compression modulus of soft clay using piezocone penetration tests. *Chinese Journal of Geotechnical Engineering*, 35(S2):569-572 (in Chinese).

Touzani S, Granderson J, Fernandes S, 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, 158:1533-1543. https://doi.org/10.1016/j.enbuild.2017.11.039

Tun YW, Pedroso DM, Scheuermann A, et al., 2016. Probabilistic reliability analysis of multiple slopes with genetic algorithms. *Computers and Geotechnics*, 77:68-76. https://doi.org/10.1016/j.compgeo.2016.04.006

Viswanathan R, Samui P, 2016. Determination of rock depth using artificial intelligence techniques. *Geoscience Frontiers*, 7(1):61-66. https://doi.org/10.1016/j.gsf.2015.04.002

Yin ZY, Jin YF, Huang HW, et al., 2016. Evolutionary polynomial regression based modelling of clay compressibility using an enhanced hybrid real-coded genetic algorithm. *Engineering Geology*, 210:158-167. https://doi.org/10.1016/j.enggeo.2016.06.016

Zhang DM, Hu QF, Huang HW, et al., 2018. Nonlinear subgrade reaction solution for circular tunnel lining design based on mobilized strength of undrained clay. *Canadian Geotechnical Journal*, 55(2):155-170. https://doi.org/10.1139/cgj-2017-0006

Zhou J, Li XB, Mitri HS, 2016. Classification of rockburst in underground projects: comparison of ten supervised learning methods. *Journal of Computing in Civil Engineering*, 30(5):04016003. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000553

中文概要

题　目：基于机器学习的土体压缩模量预测及一维基础沉降应用

目　的：土体压缩模量是影响岩土体结构变形的重要参数之一。本文旨在通过机器学习的方法实现对压缩模量的预测，并通过构建一个机器学习模型，得到塑限、液限、塑性指数、液性指数、比贯入阻力以及埋深这 6 个输入参数与压缩模量预测值之间的关系。

创新点：1. 构建一个机器学习算法框架以实现对土体压缩模量的预测；2. 此框架包括梯度提升回归树（GBRT）和遗传算法（GA），并采用 GA 对 GBRT 超参数进行获取。

方　法：1. 通过收集整理工程报告获取本次预测的数据集（样本 211 个）；输入参数有 6 个，分别为塑限、液限、塑性指数、液性指数、比贯入阻力以及埋深；输出参数为压缩模量。2. 采用 GBRT 算法识别输入变量与目标响应之间的非线性规律，并采用 GA 调整 GBRT 模型的超参数。3. 模型训练完成后，对压缩模量进行预测。4. 将测试集上的预测结果和传统方法进行对比分析并应用到一维基础沉降中。

结　论：1. 本文提出的 GA-GBRT 模型可以较好地实现对

土体压缩模量的预测；GA 可以对 GBRT 算法的超参数进行有效标定。2. 训练后的 GA-GBRT 模型在训练集和测试集上都表现良好；在训练集和测试集上的相关系数 *R* 值分别为 0.82 和 0.91，说明模型可以对压缩模量进行准确预测。3. 对输入变量相对重要性的研究发现，液性指标是本研究中最重要的变量，其重要性得分为 0.313（总数为 1）；其他指标的重要性排序依次为：液限、塑限、塑性指数、比贯入阻力和埋深。4. 对于地基沉降的预测，本文提出的模型在相关系数 *R* 值和 Mann-Whitney 检验结果上均优于经验公式。5. 本文提出的 GA-GBRT 模型可以更经济、更快速地预测土壤压缩模量。

关键词：压缩模量预测；机器学习；梯度提升回归算法；遗传算法（GA）；基础沉降