https://dx.doi.org/10.1016/j.enggeo.2019.105328

A Novel Hybrid Surrogate Intelligent Model for Creep Index Prediction based on Particle Swarm Optimization and Random Forest

Pin ZHANG[1, 2], Zhen-Yu YIN[2*], Yin-Fu JIN[2], Tommy H.T. Chan[1]

1 School of Civil Engineering & Built Environment, Science and Engineering Faculty, Queensland University of Technology (QUT), Brisbane, Qld 4001, Australia

2 Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China

* Corresponding author: Dr Zhen-Yu YIN, Tel: +852 3400 8470; Fax: +852 2334 6389; E-mail: zhenyu.yin@polyu.edu.hk; zhenyu.yin@gmail.com

Abstract: Long-term settlement issues in engineering practice are controlled by creep index C_α. But current empirical models of C_α are not enough reliable. Different from previous correlations, this study proposes a hybrid surrogate intelligent model for predicting C_α. The new combined model integrates the meta-heuristic particle optimization swarm (PSO) in the random forest (RF) to overcome the problem of user experience-dependent and local optimum. A total number of 151 datasets with four parameters (liquid limit w_L, plasticity index I_p, void ratio e and clay content CI) and one output variable C_α are collected from published works. 11 combinations of these four parameters (one combination with four parameters, four combinations with three parameters and six combinations with two parameters) are set as input variables in the RF algorithm for determining the optimum combination of variables. In this novel model, PSO is employed to determine the optimum hyper-parameters in RF algorithm, and the fitness function in the PSO is defined as the mean prediction error for ten cross-validation sets for enhancing the robustness of RF models. The performance of RF model is particularly compared with the existing empirical formulae. The results indicate that the combinations of I_P–e, CI–I_P–e, and CI–w_L–I_p–e are optimum RF models in respective group, and these models are recommended to predict C_α in engineering practice. Meanwhile, these three proposed models obviously outperform the empirical methods with lower prediction error. Parametric investigation indicates that the relationships between the C_α and four input variables in the proposed RF models harmonize with the physical explanation. Gini index generated in RF process indicates C_α is much more sensitive to e than the remaining three input variables, followed by CI, I_p and w_L, but the difference among later three variables can be negligible.

Keywords: Creep; Soft clay; Machine learning; Optimization; Physical properties; Correlation

1. Introduction

Natural soft clays exhibit significant creep under both laboratory and in situ conditions after primary consolidation, which significantly influences the long-term stability of slope and safety of infrastructures in various fields, such as tunneling (Meng et al. 2018; Shen et al. 2014), slope (Jin et al. 2003; Yang et al. 2019) and embankment (Karstunen and Yin 2010; Yao et al. 2018; Yin et al. 2015; Zhu et al. 2019), etc. Engineers have to calculate the long-term settlement before construction in order to control the post-construction settlement to a tolerated value. Actually, time-dependent behavior of soft clays has been studied for a long history, various methods in standard and advanced elastic viscoplastic (EVP) models have also been proposed to estimate the creep settlement (Tan et al. 2018a; Yin and Graha 1989; Yin et al. 2010a; Yin et al. 2010b; Zhou et al. 2018), for which the measurement of viscous parameters takes time, and thus it is not convenient for engineers and researchers.

Calculation of long-term settlement using methods recommended in standard or EVP models indicates the corresponding parameters have to be determined in advance. Creep index $C_\alpha = \Delta e / \Delta \log(t)$ generally determined by the one-dimensional oedometer test is a key parameter to calculate long-term settlement in engineering practice, and it is also applied in most standard and EVP models (Yin et al. 2014a; Yin et al. 2011). Although most clays in engineering practice are intact rather than reconstituted, the C_α of intact clays is not an intrinsic property because bonds in these natural soils are progressively destroyed under various loading or unloading formations, which causes the apparent C_α highly nonlinear (Karstunen and Yin 2010; Yin et al. 2017). The varying value of C_α depending on stress level for intact clay is thus not suitable for use in actual engineering problems, which may also result in wrong predictions. However, the C_α of reconstituted clay without the interruption of soil structures is an intrinsic property, which provides the base for understanding the creep behavior of soils and it is thus more adaptable in engineering design (Jin et al. 2019). Because of these factors, this research merely focuses on the C_α of reconstituted clays.

Currently, the value of C_α is primarily determined by the curve-fitting technique based on the experimental data. Researchers demonstrated that the creep property is affected by the microstructure of soft clays (Yin and Chang 2009; Yin et al. 2014b), and physical properties somehow represent the microstructure of clay (Jin et al. 2019). Nakase et al. (1998) proposed a linear formula to describe the relationship between C_α and plasticity index I_p, similar relationship was also formulated by (Yin 1999). Zeng et al. (2012) pointed out that the void ratio e and the void ratio at liquid limit e_L are the significant factors for the creep behavior of soft clays, and a C_α prediction model was thus proposed based on these two factors. Yin et al. (2015) formulated a linear expression of C_α with e in a double logarithm plane. More recently, Zhu et al. (2016) further developed a formulation of C_α considering both soil density and soil structure. Nevertheless, these empirical formulae are merely capable of describing few soft clays. Meanwhile, influential factors taken into consideration are limited, one or two in most formulae, although the value of C_α depends on more influential factors. Therefore, C_α calculated by these empirical methods is not enough reliable, and a model with wider adaptability between C_α and physical properties of soft clays needs to be determined.

Machine learning (ML) algorithms are characterized by the strong capability of capturing the non-linear relationships among high-dimension variables and worth to try. Various ML algorithms such as back-propagation neural network (BPNN) (Basheer 2000; Habibagahi and Bamdad 2003; He and Li 2009; Penumadu and Zhao 1999; Rashidian and Hassanlourad 2014; Turk et al. 2001), evolutionary neural network (ENN) (Johari et al. 2011), recurrent neural network (RNN) (Romo et al. 2001; Zhu et al. 1998), support vector machines (SVMs) (Kohestani and Hassanlourad 2016), evolutionary polynomial regression (EPR) (Faramarzi et al. 2012; Javadi and Rezania 2009; Nassr et al. 2018), genetic programming (GP) (Cabalar and Cevik 2011) and Bayesian-related methods (Gamse et al. 2018; Qi and Zhou 2017; Tan et al. 2016; Tan et al. 2018b; Zhou et al. 2012), have been extensively utilized in geotechnical engineering, e.g. the prediction of tunneling-induced settlement (Chen et al. 2019a; Chen et al. 2019b; Hasanipanah et al. 2016), slope displacement and stability (Qi and Tang 2018a; Xu and Niu 2018; Yang et al. 2019), pile behaviors (Pooya Nejad and Jaksa 2017), soil physical and mechanical characteristics (Feng et al. 2019; Kirts et al. 2018; Pham et al. 2018; Zhou et al. 2016b), etc. Recently, an advanced ensemble algorithm random forest (RF) has been applied in geotechnical engineering practice. The overfitting issue can be avoided and the importance of variables can be determined internally in the RF algorithm. The superiority of RF algorithm has been proved in other comprehensive studies such as prediction of slope stability (Qi and Tang 2018b), rockburst (Zhou et al. 2016a) and soil temperature (Feng et al. 2019). Furthermore, (Chen et al. 2019b) conducted a comprehensive comparison of different ML algorithms in predicting tunneling-induced settlement, and concluded that RF algorithm obviously outperforms other ML algorithms. However, there is no research to develop C_α prediction model based on ML algorithm. Meanwhile, the hyper-parameters such as the number of hidden layers and neurons in numerous ML prediction models are generally determined by trial and error method and the deterministic algorithms tend to optimize the general parameters (e.g., the weights and bias), which is time-consuming and tends to fall in local optimum resulting in a poor performance of the obtained model.

To resolve these deficiencies, this paper proposes a novel RF intelligent model with integrating the particle swarm optimization (PSO) algorithm to predict C_α of reconstituted clays. A database consisting of four input physical properties of remoulded clays and the corresponding C_α is formed first. Thereafter, 11 combinations of these four parameters (one combination with four parameters, four combinations with three parameters and six combinations with two parameters) are set as input variables for establishing RF models, thereby the optimum combination with lowest errors for predicting C_α can be determined. Herein, the hyper-parameters in each RF model are identified by PSO. To enhance the robustness of the proposed model, the average prediction error of 10-fold cross-validation sets is set as the fitness function in PSO algorithm. Finally, the relationships between C_α and input variables, and the sensitivity of input variables in the proposed model are investigated comprehensively.

2. Methodology

2.1 Random forest

Random forest (RF) is an ensemble algorithm consisting of a collection of tree-structured classifiers (Breiman 2001). Bagging (Breiman 1996) and random feature selection (Ho 1998) are integrated in the

RF algorithm. Each new bootstrap training set $\mathbf{N}_k$ with replacement samples is from the original training set $\mathbf{N}$. Then a tree that is a sub-predictor is built based on the new bootstrap training set $\mathbf{N}_k$. For each y, $\mathbf{x}$ in the training set, aggregate the votes only over those sub-predictors for which y, $\mathbf{x}$ does not exist in the $\mathbf{N}_k$. These sub-predictors are termed as out-of-bag (OOB) predictors and these datasets without the $\mathbf{N}_k$ are termed as OOB datasets (accounting for one-third of the original training datasets). OOB datasets are employed to evaluate the performance of sub-predictor developed upon the new bootstrap training set $\mathbf{N}_k$. The hyper-parameters in this algorithm are the number of trees and random features at each node, as shown in Table 1 with their ranges. Furthermore, random forest algorithm has been proved not to overfit with the increase in the number of trees (Breiman 2001). They are more robust with respect to noise, and the importance of input variables can be evaluated internally based on the value of Gini index (Breiman 2001). The detailed description of Gini index can refer to (Lovatti et al. 2019). Figure 1 presents the process of building a random forest, which is also showing as following:

1. Draw n bootstrap training sets from the original training data. Each bootstrap training set has about 2/3 of the original training datasets. The features in each bootstrap training set are selected at random.
2. Generate a decision tree based on each bootstrap training set. OOB datasets are used to evaluate the performance of a decision tree which ultimately selects the best features/split among the training set. All decision trees form a random forest.
3. Predict the new dataset by averaging the predictions of n decision trees. (i.e., average of regression).

The output of the RF can be expressed as:

$$y = \frac{1}{n}\sum_{i=1}^{n} y_i(\mathbf{x}) \tag{1}$$

where, y_i ($\mathbf{x}$) = individual prediction of a tree for an input $\mathbf{x}$; n = a total number of decision trees.

2.2 Particle swarm optimization

Particle swarm optimization (PSO) is a computational method that optimizes a problem by iteratively improving candidate solutions, here termed as particles (Kennedy and Eberhart 1995). Each particle has its position vector, X_i^k, velocity vector, V_i^k, and fitness value, where k is the current generation and i is the ith particle. In the search-space, these particles move toward the global best positions based on local best position and velocity. Herein, lower fitness value donates better position. The global best positions, that is, the optimum hyper-parameters of RF algorithm (see Table 1) can be determined when the fitness reaches the minimum value and keeps a constant. The velocity and position of each particle are updated using the following equations:

$$V_i^{k+1} = \omega V_i^k + c_1 r_1 \left(P_i^k - X_i^k \right) + c_2 r_2 \left(P_g^k - X_i^k \right) \tag{2}$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \tag{3}$$

where, c_1 and c_2 = acceleration coefficients; ω = a weight which is called "inertia weight", equal to 1 which is a typical value; r_1 and r_2 = random numbers, distributing in the interval [0, 1]; P_i = current best location of ith particles; P_g = the global best among all particles. In this study, both of c_1 and c_2 are equal to 1.49, which is the typical value used in PSO. Upper and lower bound of *Vk I* are 1 and –1, respectively. In the PSO, only three parameters (i.e., ω, c_1 and c_2) need to be set. These values mainly affect the convergence speed (El-Gallad et al. 2002; Zheng et al. 2003) and slightly affect the final optimization results as long as a large number of generations is performed. Therefore, the maximum generation is 100, which is large enough to find the best solution. To enhance the PSO performance through selecting proper parameters, many experiences on the selection of parameters of PSO can be found in previous studies (Shi and Eberhart 1998; Trelea 2003).

2.3 Evaluation indicators

Three common indicators root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are employed to evaluate the performance of RF models. RMSE and MAE can directly reflect the prediction error, but they are the scale-dependent indicators, whose scale is related to the scale of the data. In other words, low values of RMSE and MAE cannot indicate the great performance if the values of model output variables are small. Meanwhile, large values of RMSE and MAE cannot indicate the bad performance if the values of model output variables are large. RMSE has the same scale as the data, but it is more sensitive to outliers than MAE. MAPE is a scale-independent indicator, thereby it is not affected by the scale of data. Nevertheless, the MAPE value is infinite or undefined if the observed values are close to zero. A comprehensive comparison of these indicators can refer to (Hyndman and Koehler 2006). The combination of RMSE, MAE and MAPE can effectively evaluates model performance, and these three indicators are thus used in this study. The expression of these three indicators can be obtained by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(r_i - p_i\right)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|r_i - p_i\right| \tag{5}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{r_i - p_i}{r_i}\right| \times 100\% \tag{6}$$

where, *r* = measured output value; *p* = predicted output value; *n* = a total number of datasets. Low values of these three indicators indicate a model with great performance.

2.4 K-fold cross validation

The whole process of establishing a ML model includes three phases: training, validation and testing. *K*-fold cross-validation (CV) method has been extensively used to validate model (Stone 1974) for improving the robustness of ML models and avoiding overfitting problem. In this method, the original training set is randomly divided into *k* sub-datasets. Herein, *k*-1 sub-datasets are used to train model and a remaining dataset is used to validate model. Each sample thus has opportunity to train and test model. In open literatures, the *k* was recommended to be set as 10 (Kohavi 1995), therefore, 10-fold CV is applied in this study.

At each round, RF model with fixed hyper-parameters will be trained ten times by random nine sub-datasets, and the remaining one sub-dataset will be used to validate model. The performance of the RF model with fixed hyper-parameters will be evaluated by the mean prediction error for ten validation sets, that is, the fitness function in the PSO algorithm.

$$Fitness = \frac{1}{10}\sum_{i=1}^{10}\mathrm{MAE}_i \tag{7}$$

where, MAE_i = prediction error for *i*th validation set.

Note that the use of 10-fold cross validation can also reduce the effect of the selection of different amounts of data on the model performance.

2.5 Grey relational grade

Grey relational grade (GRG) has been extensively employed to evaluate uncertain correlations among variables (Jiang and He 2012; Li and Chen 2019). The geometric similarity of the time series of the two variables is taken into consideration in this method. Given a reference sequence $x_r = x_r$ ($x_r(1)$, $x_r(2)$, …, $x_r(n)$) and a compared sequence $x_i = x_i$ ($x_i(1)$, $x_i(2)$, …, $x_i(n)$). The grey relational coefficient between two sequences at jth ($j = 1, 2, \ldots, n$) criterion is defined as following

$$\gamma\left(x_r\left(j\right), x_i\left(j\right)\right) = \frac{\min_i \min_j \left|x_r\left(j\right) - x_i\left(j\right)\right| + \delta \max_i \max_j \left|x_r\left(j\right) - x_i\left(j\right)\right|}{\left|x_r\left(j\right) - x_i\left(j\right)\right| + \delta \max_i \max_j \left|x_r\left(j\right) - x_i\left(j\right)\right|} \tag{8}$$

where, δ = resolving coefficient, in the range of [0, 1], which is usually considered as 0.5. Thereafter, the GRG between sequences x_r and x_i can be obtained by

$$\gamma\left(x_r, x_i\right) = \frac{1}{n}\sum_{j=1}^{n} \gamma\left(x_r\left(j\right), x_i\left(j\right)\right) \tag{9}$$

where, large GRG value means the high correlations between sequences x_r and x_i.

3. Proposed intelligent model

3.1 Model framework

Figure 2 presents the process of establishing the proposed creep index C_α prediction model. The whole process can be categorized into three phases: data preprocessing, training and testing RF prediction models. At the first phase, main influential factors of C_α need to be determined and collected, a database consisting of input and output parameters are then established. Herein, 80% of data are randomly selected for training the model while the remaining are used to test the model. The selection of input variables are vitally important to the model performance. According to previous investigations, the liquid limit w_L, plasticity index I_p and void ratio e have been used to form an empirical equation to predict C_α (Nakase et al. 1998; Zeng et al. 2012; Zhu et al. 2016). In addition, the clay content (CI) also has a significant influence on predicting C_α (Jin et al. 2019). Therefore, these four parameters w_L, I_p, e and CI are preferably considered as the model input variables in this study. The correlation of selected parameters to C_α is examined by GRG method, as shown in Fig 4.

Feature selection method has been successfully conducted to process high-dimensional data and select the most relevant factors to the outputs of model (Gao et al. 2018; Lu et al. 2018). In order to determine the optimum combination of input variables for predicting C_α, a total number of 11 combinations of input variables are used to train the prediction model respectively, which can be divided into three groups: 6 combinations of two variables, 4 combinations of three variables, and 1 combination of four variables. The one variable as input parameter is not taken into consideration in this research, because the prediction model trained with only one variable suffers from underfitting, losing generalization capability. Therefore, a total number of 11 C_α prediction models with different combinations of input variables need to be trained. The model with the optimum performance will be recommended to predict C_α in practice engineering.

The objective at the process of training model is to identify the optimum hyper-parameters in 11 prediction models. PSO algorithm is employed to search for the optimum hyper-parameters in the RF algorithm. At each round, the proposed model starts from randomly assigning hyper-parameters to RF. Training set is then randomly divided into ten parts using 10-fold CV method. If the termination criterion is satisfied, that is, whether or not reaches maximum 100 generations and fitness value converges, the optimum hyper-parameters in one model can be determined. Otherwise, RF will be assigned a new set of hyper-parameters by the PSO algorithm. In this way, the hyper-parameters in 11 prediction models can be determined ultimately.

At the last phase, the 11 prediction models with optimum hyper-parameters will be evaluated by the test set. The model with the lowest error will be selected as the optimum C_α prediction model. Meanwhile, the optimum model in each group can also be determined. Therefore, engineers and researchers can select the most appropriate prediction model based on their existing experiment data.

3.2 Data source

The data used in this research were collected from published research works and consist of various remoulded clays in the world (Li et al. 2012; Yin 1999; Yin et al. 2015; Zeng et al. 2012; Zhu et al. 2016). A total number of 151 datasets were ultimately collected. The various remoulded clays in this database facilitate the development of a uniform C_α prediction model for all remoulded clays. The histogram of all variables in the database is presented in the diagonal line of Fig. 3. Scatter plots of pairwise variables are also plotted in this figure. It can be observed that all variables cover a wide range of values, which sufficiently extends the applicability of the proposed model.

To eliminate the effect of different magnitudes of input variables on the model's performance and also to reduce the computational cost, all datasets have been normalized into the range of (–1, 1) using the following expression:

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}\left(\bar{x}_{\max} - \bar{x}_{\min}\right) + \bar{x}_{\min} \tag{10}$$

where x = actual value of input variables, $x_{\min}$ = minimum value of input variables and $x_{\max}$ = maximum value of input variables. $\bar{x}_{\min} = -1$; $\bar{x}_{\max} = 1$.

Figure 4 presents the GRG values among input and output variables. The number in each panel represents the GRG value between pairwise variables. It can be observed that the largest GRG value reaches 0.81, showing high correlation between input e and output C_α, whereas the GRG values of remaining three input variables are roughly identical. Overall, GRG values of four input parameters exceed 0.65, suggesting the selected influential factors are appropriate to predict C_α.

4. Results

4.1 Determination of hyper-parameters

PSO algorithm is utilized for tuning two hyper-parameters in the RF model. Figure 5 shows the evolution of the fitness value within 100 generations in 11 C_α prediction models. In regard to the CV sets, the evolution of fitness is different and ultimately converges at various values among 11 models. RF model with the combination of I_p–e outperforms remaining five models trained by two input variables. The ultimate fitness values is only 0.00408 and it actually maintains steadily from the initial generation. The fitness value of this model is even less than fitness values produced by all RF models with three input variables, whereas the fitness values of remaining five RF models are larger than that. RF model with the combination of w_L–I_p–e yields the lowest fitness value of 0.0043, compared with remaining three models trained by three input variables. RF model with the combination of four variables shows the best performance with the fitness value of 0.00406. It is noteworthy that no further decrease in the fitness value is observed after the generation exceeds 16 in 11 RF prediction models, which indicates that 100 generations are large enough to determine the optimum hyper-parameters in RF models. Meanwhile, the fitness values of RF models with two input variables range from 0.00408 to 0.00591, whereas this value ranges from 0.0043 to 0.00504 in RF models with three input variables. It actually indicates that the performance of C_α prediction models is increasingly steady with the increase in the number of input variables. Overall, the fitness value in RF prediction model with two input variables is largest, and the lowest value appears in the RF prediction model with four input variables.

4.2 Prediction of C_α for the validation sets

In order to reveal the reason behind the difference in the converged fitness value in 11 RF prediction models, Figure 6 presents the values of three indicators in each CV set. It can be observed that the biggest difference of RMSE and MAE values appears in the second CV set, where the RF models with four combinations, that are, w_L–e, I_p–e, w_L–I_p–e, and CI–w_L–I_p–e, produces much less RMSE and MAE values than the remaining seven RF models. Meanwhile, the variation of RMSE and MAE values in these four models is slight, compared with the remaining RF models. Therefore, the fitness values produced by these four RF models are lowest in each group. Aside from the second CV set, the evolution of RMSE and MAE values is roughly identical in each RF model. It can be observed from Fig. 6(c) that the evolution of MAPE differs from RMSE and MAE due to its scale-independent characteristic. The performance of RF models in each CV set presents obvious difference, especially in the second, seventh and ninth CV sets, but the four RF models as mentioned above still show the lower MAPE value at the second CV set.

The distribution of RMSE, MAE and MAPE values in each RF model is presented by boxplot, as shown in Fig. 7. It is clear that the distribution of RMSE and MAE are roughly identical in each RF model. In each group, the RF models with better performance produce lower mean RMSE and MAE values. Meanwhile, the ranges of RMSE and MAE are much smaller. From the perspective of MAPE, RF model with the combination of I_p–e shows the best performance among all prediction models with the minimum mean prediction error of 20.8%. RF model with the combination of CI–w_L–e outperforms the remaining RF models trained by three parameters. These two characteristics are different from RMSE and MAE results. It will affect the performance of RF models for the test set, which will be revealed at the next section. Overall, for the CV sets, RF model with the combination of two parameters produces maximum prediction error as well as a wider range of errors.

4.3 Prediction of C_α for the test set

On the basis of hyper-parameters determined in the former section, 11 optimum RF models can be established. Figure 8 presents the scatter plot of predicted C_α for the training and test sets using optimum models. The predicted results of RF models with two input variables are illustrated in Figs. 8(a)-(f). The combinations of CI–e, w_L–e, and I_p–e clearly outperform the remaining three RF models. The predict C_α for the training set exist perfect agreement with the measured C_α, and the predict C_α for the test set are also close to the P = M line. In contrast, predicted results using the remaining three RF models widely scatter, and the disagreement of the predicted and measured C_α is frequently observed. Figures 8(g)-(j) present the predicted results using RF models with three input variables. The range of predicted C_α using RF model with the combination of CI–w_L–I_p is much smaller than the measured C_α, losing fidelity at most points. The predicted results using the remaining three models show great agreement with the measured results. The predicted C_α for both training and test sets using the RF model with four input variables are closer to the line with slope of 1 than remaining 10 models (see Fig. 8(k)).

Table 2 summarizes the values of indicators in 11 optimum RF models. The values of indicator are roughly consistent with the model performance presented in Fig. 8, that is, lower values of indicators harmonize with more reasonable distribution of the predicted C_α. Meanwhile, models with great performance in the CV sets also exhibit better performance in the test set. For instance, RF models with combination of I_P–e outperform the remaining models with two input variables in both CV and test sets, and RF model with combination of CI–w_L–I_p–e always presents the best performance among all models throughout the analysis. However, a special case is observed in the RF models with three input variables, where values of indicators of RF model with combination of CI–w_L–I_p are lowest for the test set, but Fig. 8(g) presents the distribution of predicted C_α using this model is not acceptable. In reality, the indicators values for the training set are larger than the remaining three RF models with three input variables. Figure 8 (g) also presents that the distribution of predicted C_α using this model is much more concentrated, the slight variation of predicted C_α thus leads to the lower values of indicators. Similar conditions also appear in the combinations of CI –w_L, CI–I_P, and w_L–I_P. However, the smaller range of predicted results indicates that the prediction applicability of these models is limited, that is, weak generalization capability. To sum up, the combinations of I_P–e, CI–I_P–e, and CI–w_L–I_p–e are optimum RF model in respective group, and these models are recommended to predict C_α in engineering practice.

5. Discussions

5.1 Comparison with empirical formula

In order to evaluate the predictive ability of the proposed model, five commonly used empirical methods are used for comparison (see Table 3). Several input parameters of the proposed model are also used in these empirical methods. Figure 9 presents scatterplot of the predicted C_α using five empirical methods. The predicted C_α values in the Figs. 9(a) and (b) are roughly identical, because I_p is the only parameter in these two methods. It leads to predicted C_α with the identical value and the small range. In Fig. 9(c), the values of predicted C_α vary dramatically, losing fidelity at most points. The predicted C_α using the empirical method proposed by (Zhu et al. 2016) exhibit great agreement with measured C_α (see Figs. 9(d) and (e)).

Figure 10 presents the comparison between the empirical and proposed models in predicting C_α, where the results of three optimum models in each group are plotted. It is clear that three proposed models obviously outperform the empirical methods with lower values of RMSE, MAE and MAPE. The mean RMSE and MAE values produced by the proposed models decrease by 0.0012 and 0.001, respectively, compared with the empirical methods. Further, the mean MAPE decreases from 29.33% to 18.09%.

5.2 Parametric investigation

A robust ML model exhibits smooth functions with respect to the input and output variables and exhibits physical explanation for the behaviors (Shahin et al. 2005). Therefore, the correlations between four input variables and the C_α at three typical points in the test set are investigated. Note that the C_α presented in this section are predicted by RF model with four input variables for comprehensively investigating the effects of all variables on the C_α. At each round, values of three input variables are fixed in the RF model, whereas the value of a remaining variable increases from 0.2v to 2v with an interval of 0.2v (v donates the original value of the investigated parameter).

Figure 11 presents the correlations between input and output variables in the RF model with four input parameters. The fixed values of three variables are also plotted in figures. Note that a perfect smooth correlation between input and output variables in RF model is hardly obtained, because RF model is developed based on measured data, and a smooth correlation is not observed in the measured data as shown in Fig. 3. Therefore, the correlations between input and output variables in the RF model merely reflect a general trend. It can be seen from Fig. 11(a) that the predicted C_α decreases initially with the increase in the CI, C_α then starts to increase after reaches minimum value. C_α ultimately holds steadily with the continuously increase in the CI. The correlation between w_L and C_α shows a similar condition. An opposite trend is observed between w_L and C_α. With an increase in the w_L, C_α increases initially, then starts to decreases after reaches maximum value, and C_α ultimately maintain a constant value. In regard to e, C_α increases monotonically with the increase in the value of e. After reaching maximum value, C_α is constant. The evolution of predicted C_α with the change in the four input variables is similar in three points, but the magnitude of C_α is different. Note that the predicted C_α using RF model always holds steadily when the input variables exceeds a certain range, which more or less violates laboratory test results. This is a limitation for all data-drive model, since the prediction capability of this kind of model will be useless when values of new input variables exceed the range of original database. Overall, the correlations presented in Fig. 11 are consistent with physical explanation, indicating robustness and reasonability of the proposed RF models.

To investigate the generalization ability of the proposed model, a database including a total number of 10000 random samples is established. Each sample has four input variables (CI, w_L, I_p, e), and it assumes that each variable complies with lognormal distributions (Cao and Wang 2014; Zhang et al. 2009; Zhang et al. 2018). Herein, the values of mean and standard deviation for each variable are consistent with the results presented in Fig. 3. Thereafter, the C_α is predicted by the RF model with four input variables.

Figure 12 shows the distribution of predicted C_α. It can be observed that the C_α roughly meet the lognormal distribution with the coefficient of determination of 0.91. The mean and standard deviation values of predicted C_α, 0.020 and 0.008, respectively, show great agreement with measured values, 0.019 and 0.011, respectively. Note that the range of predicted C_α are perfectly consistent with measured C_α, because the prediction ability of RF algorithm is useless when the datasets exceed the range of the original database. Overall, the performance of proposed RF model is absolutely reliable for the unseen datasets within the range of original database.

5.3 Sensitivity of variables

Variable importance measure (VIM) provides a basis for understanding the contributions of different input variables to the model output (Hapfelmeier et al. 2012). As mentioned in the Random Forest section, variable importance can be measured internally in RF algorithm, which is achieved by investigating the influence of the variation of input variables on the Gini index. Input variable that causes larger variation in Gini index is more significant to the model predictive capability (Breiman et al. 1984).

The mean decrease in Gini index caused by the change in each variable is shown in Fig. 13. It can be observed that model output C_α is much more sensitive to e than the remaining three variables, followed by CI, I_p and w_L, but the difference among these three variables can be negligible. It explains the reason that

e is a common input variable in three optimum models as mentioned above, meanwhile w_L as an input variable merely appears in the RF model with four input variables.

6. Conclusions

A new hybrid surrogate intelligent model based on particle swarm optimization (PSO) and random forest (RF) algorithms was proposed in this study for predicting C_α. A database with four input variables liquid limit w_L, plasticity index I_p, void ratio e, clay content CI and one output variable C_α was first established. 80% of data was used to train model, and the remaining 20% of data was used to test model. High values of grey relation grade (>0.65) between four input variables and one output variable indicated that the selected influential factors are appropriate to predict C_α.

To search the optimum combination of input variables with respect to predicting C_α, a total number of 11 combinations of input variables were used to train prediction model respectively, which can be divided into three groups: 6 combinations of two variables, 4 combinations of three variables, and 1 combination of four variables. Therefore, 11 RF models with different combinations of input variables were established.

To determine the optimum hyper-parameters in the RF algorithm, meta-heuristic algorithm PSO was integrated with RF algorithm. The fitness function in the PSO algorithm was defined as the mean prediction error for ten cross-validation sets, enhancing the robustness of RF models. The predicted results for the training and testing sets indicate that the combinations of I_P–e, CI–I_P–e, and CI–w_L–I_p–e are optimum RF models in respective group. Therefore, these models are recommended to predict C_α in engineering practice.

Compared with the empirical methods of predicting C_α, three proposed models obviously outperform the empirical methods with lower values of RMSE, MAE and MAPE. The mean RMSE and MAE values produced by the proposed models decrease by 0.0012 and 0.001, respectively, and the mean MAPE decreases from 29.33% in empirical methods to 18.09% in the proposed models.

Parametric investigation indicates that the relationships between the C_α and four input variables in the proposed RF models harmonize with the physical explanation, verifying the robustness and reasonability of the proposed models. Gini index in the RF algorithm was employed to evaluate the sensitivity of input variables. The results indicate that the model performance is much more sensitive to *e* than other three variables, followed by *CI*, I_p and w_L, but the difference among these three variables can be negligible.

As mentioned above, the performance of the RF model depends significantly on the datasets. Although the database used in this study includes numerous remoulded clays and the range of variables are large, it should be further expanded in the future with more data for facilitating the application of proposed models.

In order to allow readers to quickly perform the training and get results, the used datasets and the MATLAB source code for the proposed hybrid RF and PSO on predicting the C_α are provided and can be downloaded from the website of https://www.researchgate.net/publication/334450481_Matlab_code_for_predicting_creep_index_using_hybrid_Random_Forest_and_Particle_Swarm_Optimization_algorithms.

Acknowledgments

This research was financially supported by a RIF project (Grant No.: 15209119, PolyU R5037-18F) from Research Grants Council (RGC) of Hong Kong Special Administrative Region Government (HKSARG) of China, and the National Natural Science Foundation of China (Grant No. 51579179).

References

Basheer, I.A. 2000. Selection of methodology for neural network modeling of constitutive hystereses behavior of soils. Computer-Aided Civil and Infrastructure Engineering, **15**, 440–458.

Breiman, L. 1996. Bagging Predictors. Machine Learning, **24**, 123-140, doi: 10.1007/bf00058655.

Breiman, L. 2001. Random Forests. Machine Learning, **45**, 5–32.

Breiman, L.I., Friedman, J.H., Olshen, R.A. & Stone, C. 1984. Classification and Regression Trees (CART). Encyclopedia of Ecology, **57**, 582-588.

Cabalar, A.F. & Cevik, A. 2011. Triaxial behavior of sand–mica mixtures using genetic programming.

Expert Systems with Applications, **38**, 10358-10367.

Cao, Z.J. & Wang, Y. 2014. Bayesian model comparison and selection of spatial correlation functions for soil parameters. Structural Safety, **49**, 10-17, doi: 10.1016/j.strusafe.2013.06.003.

Chen, R.P., Zhang, P., Kang, X., Zhong, Z.Q., Liu, Y. & Wu, H.N. 2019a. Prediction of maximum surface settlement caused by EPB shield tunneling with ANN methods. Soils and Foundations, in press, doi: 10.1016/j.sandf.2018.11.005.

Chen, R.P., Zhang, P., Wu, H.N., Wang, Z.T. & Zhong, Z.Q. 2019b. Prediction of shield tunneling-induced ground settlement using machine learning techniques. Frontiers of Structural and Civil Engineering, in press.

El-Gallad, A., El-Hawary, M., Sallam, A. & Kalas, A. 2002. Enhancing the particle swarm optimizer via proper parameters selection. *IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No. 02CH37373)*. IEEE, 792-797.

Faramarzi, A., Javadi, A.A. & Alani, A.M. 2012. EPR-based material modelling of soils considering volume changes. Computers & Geosciences, **48**, 73-85, doi: 10.1016/j.cageo.2012.05.015.

Feng, Y., Cui, N., Hao, W., Gao, L. & Gong, D. 2019. Estimation of soil temperature from meteorological data using different machine learning models. Geoderma, **338**, 67-77, doi: 10.1016/j.geoderma.2018.11.044.

Gamse, S., Zhou, W.-H., Tan, F., Yuen, K.-V. & Oberguggenberger, M. 2018. Hydrostatic-season-time model updating using Bayesian model class selection. Reliability Engineering & System Safety, **169**, 40-50.

Gao, W., Hu, L., Zhang, P. & He, J. 2018. Feature selection considering the composition of feature relevancy. Pattern Recognition Letters, **112**, 70-74, doi: 10.1016/j.patrec.2018.06.005.

Habibagahi, G. & Bamdad, A. 2003. A neural network framework for mechanical behavior of unsaturated soils. Canadian Geotechnical Journal, **40**, 684-693, doi: 10.1139/t03-004.

Hapfelmeier, A., Hothorn, T., Ulm, K. & Strobl, C. 2012. A new variable importance measure for random forests with missing data. Statistics and Computing, **24**, 21-34, doi: 10.1007/s11222-012-9349-1.

Hasanipanah, M., Noorian-Bidgoli, M., Jahed Armaghani, D. & Khamesi, H. 2016. Feasibility of PSO-ANN model for predicting surface settlement caused by tunneling. Engineering with Computers, **32**, 705-715, doi: 10.1007/s00366-016-0447-0.

He, S. & Li, J. 2009. Modeling nonlinear elastic behavior of reinforced soil using artificial neural networks. Applied Soft Computing, **9**, 954-961, doi: 10.1016/j.asoc.2008.11.013.

Ho, T.K. 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis & Machine Intelligence, **20**, 832-844, doi: 10.1109/34.709601.

Hyndman, R.J. & Koehler, A.B. 2006. Another look at measures of forecast accuracy. International

Journal of Forecasting, **22**, 679-688, doi: 10.1016/j.ijforecast.2006.03.001.

Javadi, A.A. & Rezania, M. 2009. Applications of artificial intelligence and data mining techniques in soil modeling. Geomechanics and Engineering, **1**, 53-74, doi: 10.12989/gae.2009.1.1.053.

Jiang, H. & He, W. 2012. Grey relational grade in local support vector regression for financial time series prediction. Expert Systems with Applications, **39**, 2256-2262, doi: 10.1016/j.eswa.2011.07.100.

Jin, F., Zhang, C.H., Wang, G. & Wang, G.L. 2003. Creep modeling in excavation analysis of a high rock slope. Journal of Geotechnical and Geoenvironmental Engineering, **129**, 849-857, doi: 10.1061/(ASCE)1090-0241(2003)129:9(849).

Jin, Y.F., Yin, Z.Y., Zhou, W.H., Yin, J.H. & Shao, J.F. 2019. A single-objective EPR based model for creep index of soft clays considering L2 regularization. Engineering Geology, **248**, 242-255, doi: 10.1016/j.enggeo.2018.12.006.

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

Karstunen, M. & Yin, Z.Y. 2010. Modelling time-dependent behaviour of Murro test embankment. Géotechnique, **60**, 735-749, doi: 10.1680/geot.8.P.027.

Kennedy, J. & Eberhart, R. 1995. Particle swarm optimization. *1995 IEEE International Conference on Neural Networks*, Perth, Australia, 1942-1948.

Kirts, S., Panagopoulos, O.P., Xanthopoulos, P. & Nam, B.H. 2018. Soil-Compressibility Prediction Models Using Machine Learning. Journal of Computing in Civil Engineering, **32**, doi: 10.1061/(asce)cp.1943-5487.0000713.

Kohavi, R. 1995. A study of Cross-Validation and bootstrap for accuracy estimation and model selection. *International joint conference on artificial intelligence*. Morgan Kaufmann Publishers Inc., 1137-1143.

Kohestani, V.R. & Hassanlourad, M. 2016. Modeling the mechanical behavior of carbonate sands using artificial neural networks and support vector machines. International Journal of Geomechanics, **16**, 04015038, doi: 10.1061/(ASCE).

Li, Q., Ng, C.W.W. & Liu, G.B. 2012. Low secondary compressibility and shear strength of Shanghai Clay. Journal of Central South University, **19**, 2323-2332, doi: 10.1007/s11771-012-1278-9.

Li, Z. & Chen, L. 2019. A novel evidential FMEA method by integrating fuzzy belief structure and grey relational projection method. Engineering Applications of Artificial Intelligence, **77**, 136-147, doi: 10.1016/j.engappai.2018.10.005.

Lovatti, B.P.O., Nascimento, M.H.C., Neto, Á.C., Castro, E.V.R. & Filgueiras, P.R. 2019. Use of Random forest in the identification of important variables. Microchemical Journal, **145**, 1129-1134, doi: 10.1016/j.microc.2018.12.028.

Lu, Q., Li, X. & Dong, Y. 2018. Structure preserving unsupervised feature selection. Neurocomputing, **301**, 36-45, doi: 10.1016/j.neucom.2018.04.001.

Meng, F.Y., Chen, R.P. & Xin, K. 2018. Effects of tunneling-induced soil disturbance on the post-construction settlement in structured soft soils. Tunnelling and Underground Space Technology, **80**, 53–63, doi: 10.1016/j.tust.2018.06.007.

Nakase, A., Kamei, T. & Kusakabe, O. 1998. Constitutive parameters estimated by plasticity index Journal of Geotechnical Engineering, **114**, 844-858.

Nassr, A., Esmaeili-Falak, M., Katebi, H. & Javadi, A. 2018. A new approach to modeling the behavior of frozen soils. Engineering Geology, **246**, 82-90.

Penumadu, D. & Zhao, R.D. 1999. Triaxial compression behavior of sand and gravel using artificial neural networks (ANN). Computers and Geotechnics, **24**, 207-230.

Pham, B.T., Son, L.H., Hoang, T.-A., Nguyen, D.-M. & Tien Bui, D. 2018. Prediction of shear strength of soft soil using machine learning methods. Catena, **166**, 181-191, doi: 10.1016/j.catena.2018.04.004.

Pooya Nejad, F. & Jaksa, M.B. 2017. Load-settlement behavior modeling of single piles using artificial neural networks and CPT data. Computers and Geotechnics, **89**, 9-21, doi: 10.1016/j.compgeo.2017.04.003.

Qi, C.C. & Tang, X.L. 2018a. A hybrid ensemble method for improved prediction of slope stability. International Journal for Numerical and Analytical Methods in Geomechanics, **42**, 1823-1839, doi: 10.1002/nag.2834.

Qi, C.C. & Tang, X.L. 2018b. Slope stability prediction using integrated metaheuristic and machine learning approaches: A comparative study. Computers & Industrial Engineering, **118**, 112-122, doi: 10.1016/j.cie.2018.02.028.

Qi, X.-H. & Zhou, W.-H. 2017. An efficient probabilistic back-analysis method for braced excavations using wall deflection data at multiple points. Computers and Geotechnics, **85**, 186-198.

Rashidian, V. & Hassanlourad, M. 2014. Application of an artificial neural network for modeling the mechanical behavior of carbonate soils. International Journal of Geomechanics, **14**, 142-150, doi: 10.1061/(asce)gm.1943-5622.0000299.

Romo, M.P., Garcí a, S.R., Mendoza, M.J. & Taboada‐ Urtuzuá stegui, V. 2001. Recurrent and Constructive‐ Algorithm Networks For Sand Behavior Modeling. International Journal of Geomechanics, **1**, 371-387, doi: 10.1061/(asce)1532-3641(2001)1:4(371).

Shahin, M.A., Maier, H.R. & Jaksa, M.B. 2005. Investigation into the robustness of artificial neural networks for a case study in civil engineering. In: Zerger A, Argent RM, editors. MODSIM 2005

International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, pp.79-83.

Shen, S.L., Wu, H.N., Cui, Y.J. & Yin, Z.Y. 2014. Long-term settlement behaviour of metro tunnels in the soft deposits of Shanghai. Tunnelling and Underground Space Technology, **40**, 309-323, doi: 10.1016/j.tust.2013.10.013.

Shi, Y. & Eberhart, R.C. 1998. Parameter selection in particle swarm optimization. *International conference on evolutionary programming*. Springer, 591-600.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, **36**, 111-147, doi: 10.2307/2344741.

Tan, F., Zhou, W.-H. & Yuen, K.-V. 2016. Modeling the soil water retention properties of same-textured soils with different initial void ratios. Journal of Hydrology, **542**, 731-743.

Tan, F., Zhou, W.-H. & Yuen, K.-V. 2018a. Effect of loading duration on uncertainty in creep analysis of clay. International Journal for Numerical and Analytical Methods in Geomechanics, **42**, 1235-1254, doi: 10.1002/nag.2788.

Tan, F., Zhou, W.H. & Yuen, K.V. 2018b. Effect of loading duration on uncertainty in creep analysis of clay. International Journal for Numerical and Analytical Methods in Geomechanics, **42**, 1235-1254.

Trelea, I.C. 2003. The particle swarm optimization algorithm: convergence analysis and parameter selection. Information processing letters, **85**, 317-325.

Turk, G., Logar, J. & Majes, B. 2001. Modelling soil behaviour in uniaxial strain conditions by neural networks. Advances in Engineering Software, **32**, 805-812.

Xu, S.L. & Niu, R.Q. 2018. Displacement prediction of Baijiabao landslide based on empirical mode decomposition and long short-term memory neural network in Three Gorges area, China. Computers & Geosciences, **111**, 87-96, doi: 10.1016/j.cageo.2017.10.013.

Yang, B., Yin, K., Lacasse, S. & Liu, Z. 2019. Time series analysis and long short-term memory neural network to predict landslide displacement. Landslides, doi: 10.1007/s10346-018-01127-x.

Yao, Y.-P., Qi, S.-J., Che, L.-W., Chen, J., Han, L.-M. & Ma, X.-Y. 2018. Postconstruction Settlement Prediction of High Embankment of Silty Clay at Chengde Airport Based on One-Dimensional Creep Analytical Method: Case Study. International Journal of Geomechanics, **18**, doi: 10.1061/(asce)gm.1943-5622.0001191.

Yin, J.H. 1999. Properties and behaviour of Hong Kong marine deposits with different clay contents. Canadian Geotechnical Journal, **36**, 1085-1095.

Yin, J.H. & Graha, J. 1989. Viscous-elastic-plastic modelling of one-dimensional time-dependent behaviour. Canadian Geotechnical Journal, **26**, 199-209.

Yin, Z.-Y. & Chang, C.S. 2009. Microstructural modelling of stress-dependent behaviour of clay.

International Journal of Solids and Structures, **46**, 1373-1388, doi: 10.1016/j.ijsolstr.2008.11.006.

Yin, Z.-Y., Chang, C.S., Karstunen, M. & Hicher, P.-Y. 2010a. An anisotropic elastic–viscoplastic model for soft clays. International Journal of Solids and Structures, **47**, 665-677, doi: 10.1016/j.ijsolstr.2009.11.004.

Yin, Z.-Y., Yin, J.-H. & Huang, H.-W. 2014a. Rate-Dependent and Long-Term Yield Stress and Strength of Soft Wenzhou Marine Clay: Experiments and Modeling. Marine Georesources & Geotechnology, **33**, 79-91, doi: 10.1080/1064119x.2013.797060.

Yin, Z.-Y., Zhu, Q.-Y. & Zhang, D.-M. 2017. Comparison of two creep degradation modeling approaches for soft structured soils. Acta Geotechnica, **12**, 1395-1413, doi: 10.1007/s11440-017-0556-y.

Yin, Z.Y., Karstunen, M., Chang, C.S., Koskinen, M. & Lojander, M. 2011. Modeling Time-Dependent Behavior of Soft Sensitive Clay. Journal of Geotechnical and Geoenvironmental Engineering, **137**, 1103-1113, doi: 10.1061/(asce)gt.1943-5606.0000527.

Yin, Z.Y., Karstunen, M. & Hicher, P.Y. 2010b. Evaluation of the influence of elasto-viscoplastic scaling functions on modelling time-dependent behaviour of natural clays. Soils and Foundations, **50**, 203-214.

Yin, Z.Y., Xu, Q. & Yu, C. 2015. Elastic-Viscoplastic Modeling for Natural Soft Clays Considering Nonlinear Creep. International Journal of Geomechanics, **15**, doi: 10.1061/(asce)gm.1943-5622.0000284.

Yin, Z.Y., Zhu, Q.Y., Yin, J.H. & Ni, Q. 2014b. Stress relaxation coefficient and formulation for soft soils. Géotechnique Letters, **4**, 45-51, doi: 10.1680/geolett.13.00070.

Zeng, L.L., Hong, Z.S., Liu, S.Y. & Chen, F.Q. 2012. Variation law and quantitative evaluation of secondary consolidation behavior for remolded clays. Chinese Journal of Geotechnical Engineering, **34**, 1496-1500.

Zhang, J., Zhang, L.M. & Tang, W.H. 2009. Bayesian framework for characterizing geotechnical model uncertainty. Journal of Geotechnical and Geoenvironmental Engineering, **135**, 932-940, doi: 10.1061/共 ASCE 兲 GT.1943-5606.0000018.

Zhang, L., Li, D.-Q., Tang, X.-S., Cao, Z.-J. & Phoon, K.-K. 2018. Bayesian model comparison and characterization of bivariate distribution for shear strength parameters of soil. Computers and Geotechnics, **95**, 110-118, doi: 10.1016/j.compgeo.2017.10.003.

Zheng, Y.-L., Ma, L.-H., Zhang, L.-Y. & Qian, J.-X. 2003. On the convergence analysis and parameter selection in particle swarm optimization. *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693)*. IEEE, 1802-1807.

Zhou, J., Li, X. & Mitri, H.S. 2016a. Classification of Rockburst in Underground Projects: Comparison of Ten Supervised Learning Methods. Journal of Computing in Civil Engineering, **30**, 04016003, doi:

10.1061/(asce)cp.1943-5487.0000553.

Zhou, W.-H., Garg, A. & Garg, A. 2016b. Study of the volumetric water content based on density, suction and initial water content. Measurement, **94**, 531-537, doi: 10.1016/j.measurement.2016.08.034.

Zhou, W.-H., Yuen, K.-V. & Tan, F. 2012. Estimation of maximum pullout shear stress of grouted soil nails using Bayesian probabilistic approach. International Journal of Geomechanics, **13**, 659-664.

Zhou, W.H., Tan, F. & Yuen, K.V. 2018. Model updating and uncertainty analysis for creep behavior of soft soil. Computers and Geotechnics, **100**, 135-143, doi: 10.1016/j.compgeo.2018.04.006.

Zhu, J.H., Zaman, M.M. & Anderson, S.A. 1998. Modelling of shearing behaviour of a residual soil with Recurrent Neural Network. International Journal for Numerical and Analytical Methods in Geomechanics, **22**, 671-687.

Zhu, Q.Y., Jin, Y.F. & Yin, Z.Y. 2019. Modeling of embankment beneath marine deposited soft sensitive clays considering straightforward creep degradation. Marine Georesources & Geotechnology, in press, doi: 10.1080/1064119X.2019.1603254.

Zhu, Q.Y., Yin, Z.Y., Hicher, P.Y. & Shen, S.L. 2016. Nonlinearity of one-dimensional creep characteristics of soft clays. Acta Geotechnica, **11**, 887-900, doi: 10.1007/s11440-015-0411-y.

Captions of figures

Fig. 1 Flowchart of building a random forest

Fig. 2 Flow chart of the proposed C_α prediction model

Fig. 3 Distributions of all variables in the database

Fig. 4 GRG values among variables

Fig. 5 Evolution of fitness value in all prediction models

Fig. 6 Values of indicators in ten CV sets: (a) RMSE; (b) MAE; (c) MAPE

Fig. 7 Distribution of indicators values in ten CV sets: (a) RMSE; (b) MAE; (c) MAPE

Fig. 8 Predicted C_α for training and test sets by all prediction models

Fig. 9 Predicted C_α for the test set by all published methods

Fig. 10 Comparison of empirical methods and proposed models in predicting C_α

Fig. 11 Predicted C_α using RF model against (a) CI; (b) w_L; (c) I_p; (d) e

Fig. 11 Predicted C_α using RF model against (a) CI; (b) w_L; (c) I_p; (d) e

Fig. 12 Distribution of predicted C_α

Fig. 13 Mean decrease in Gini index of four input variables

Captions of tables

Table 1 Hyper-parameters in random forest

Table 2 Performance for all prediction models

Table 3 Performance for all prediction models

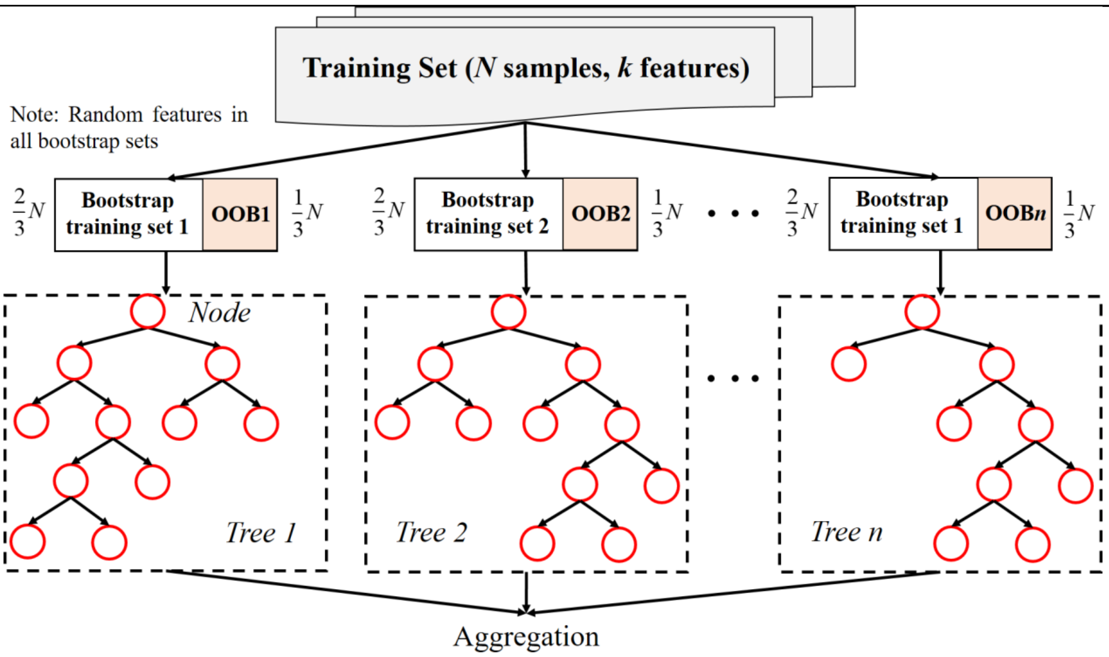

Fig. 1 Flowchart of building a random forest

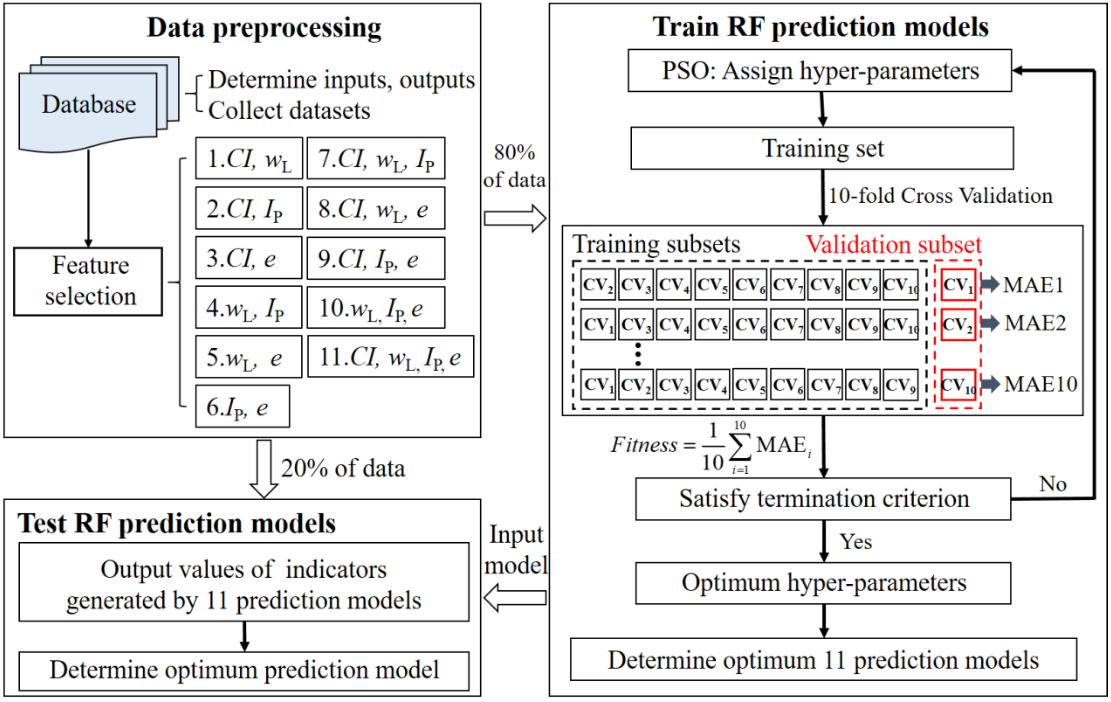

Fig. 2 Flow chart of the proposed C_α prediction model

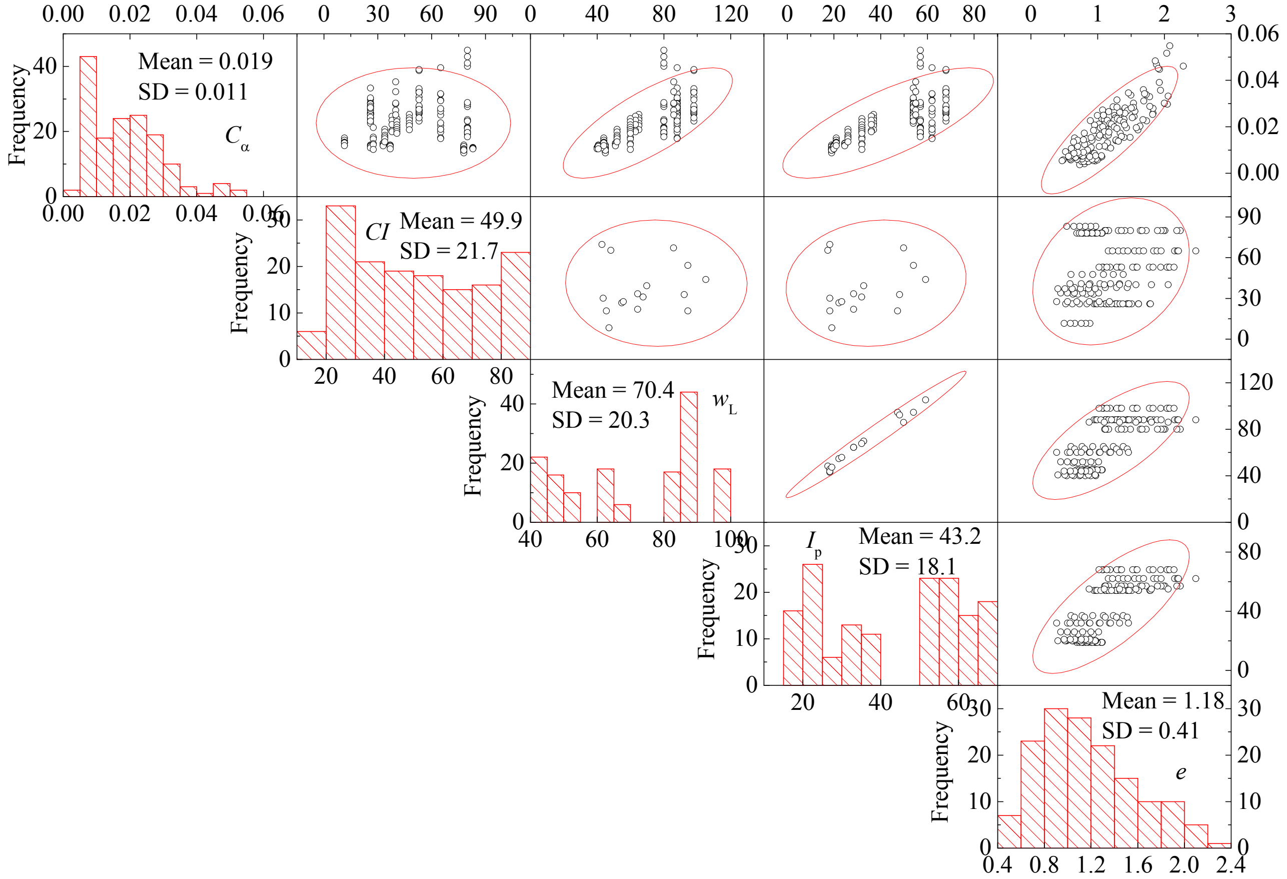

Fig. 3 Distributions of all variables in the database

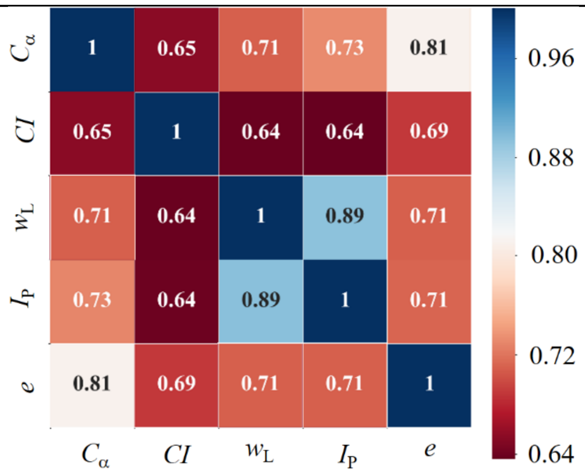

Fig. 4 GRG values among variables

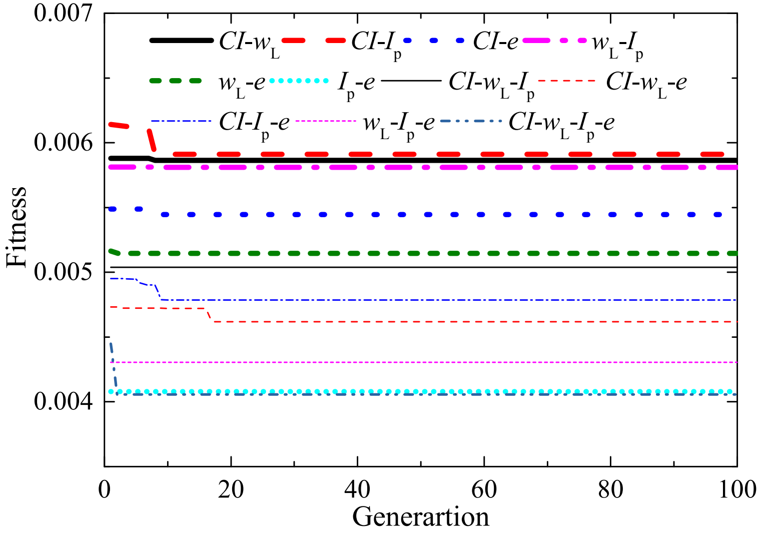

Fig. 5 Evolution of fitness value in all prediction models

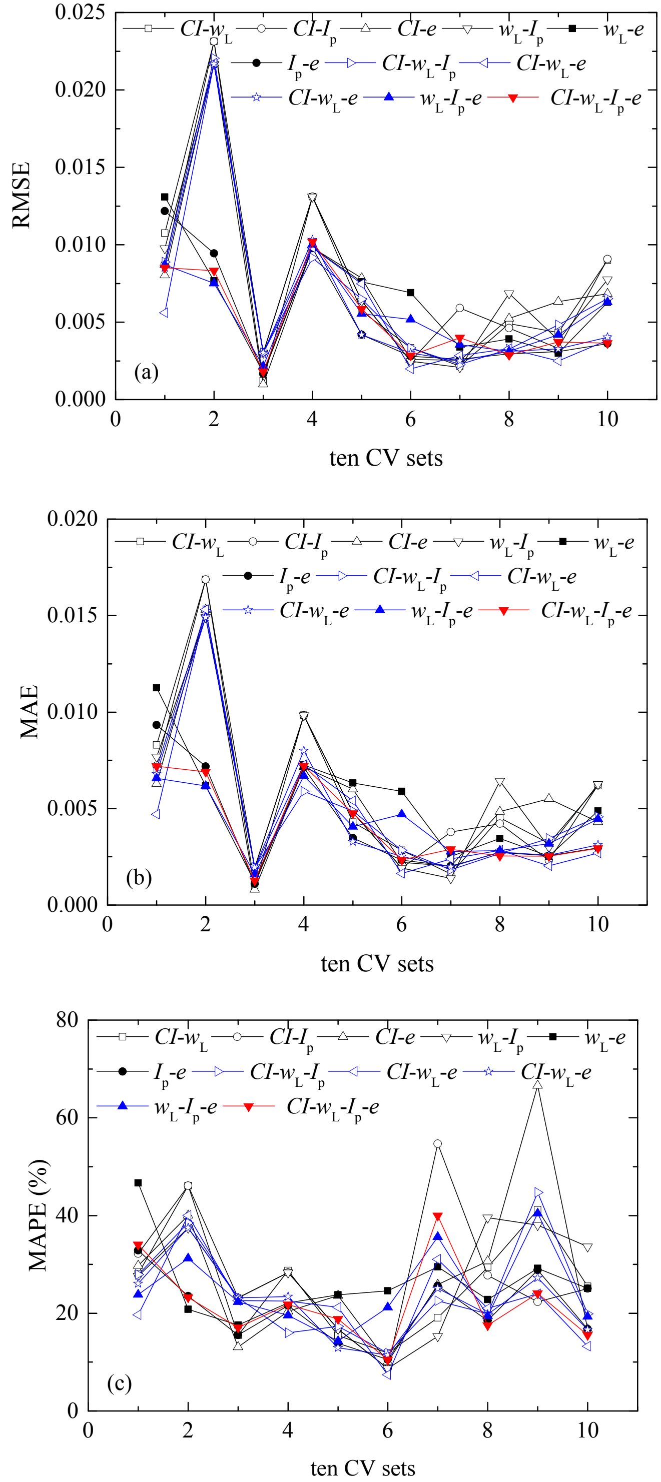

Fig. 6 Values of indicators in ten CV sets: (a) RMSE; (b) MAE; (c) MAPE

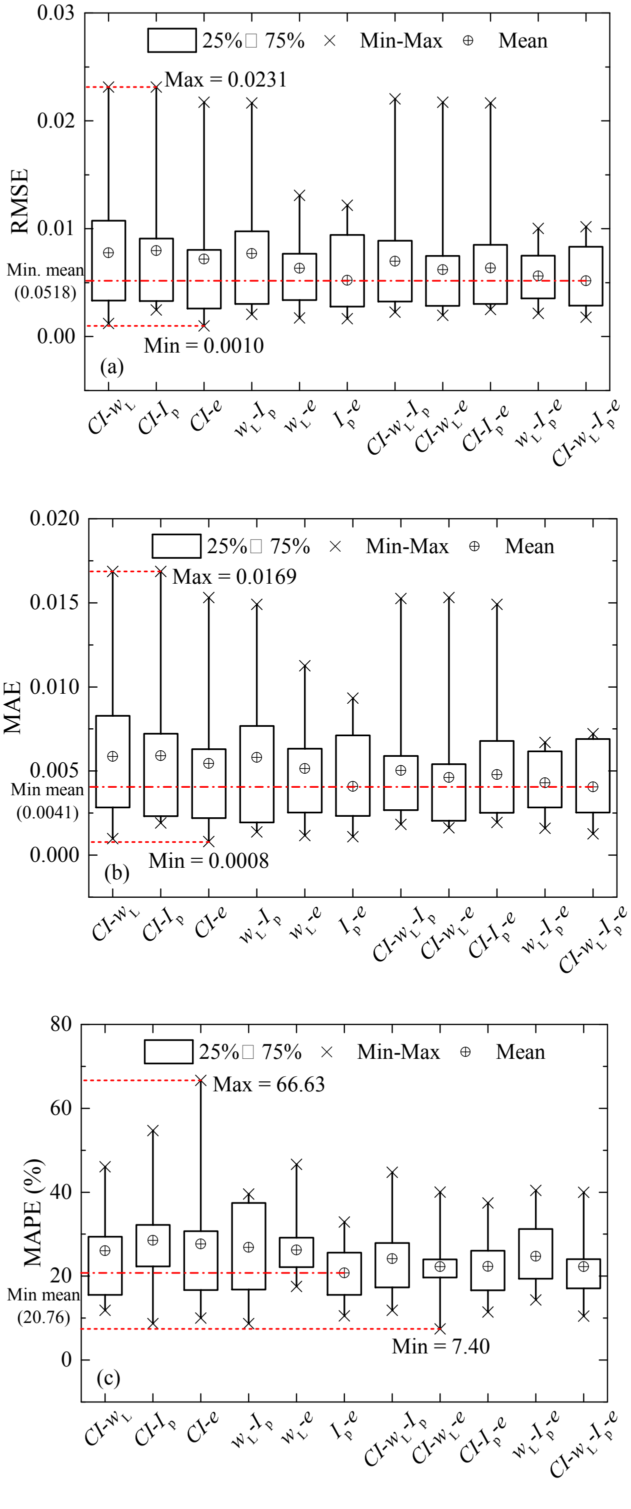

Fig. 7 Distribution of indicators values in ten CV sets: (a) RMSE; (b) MAE; (c) MAPE

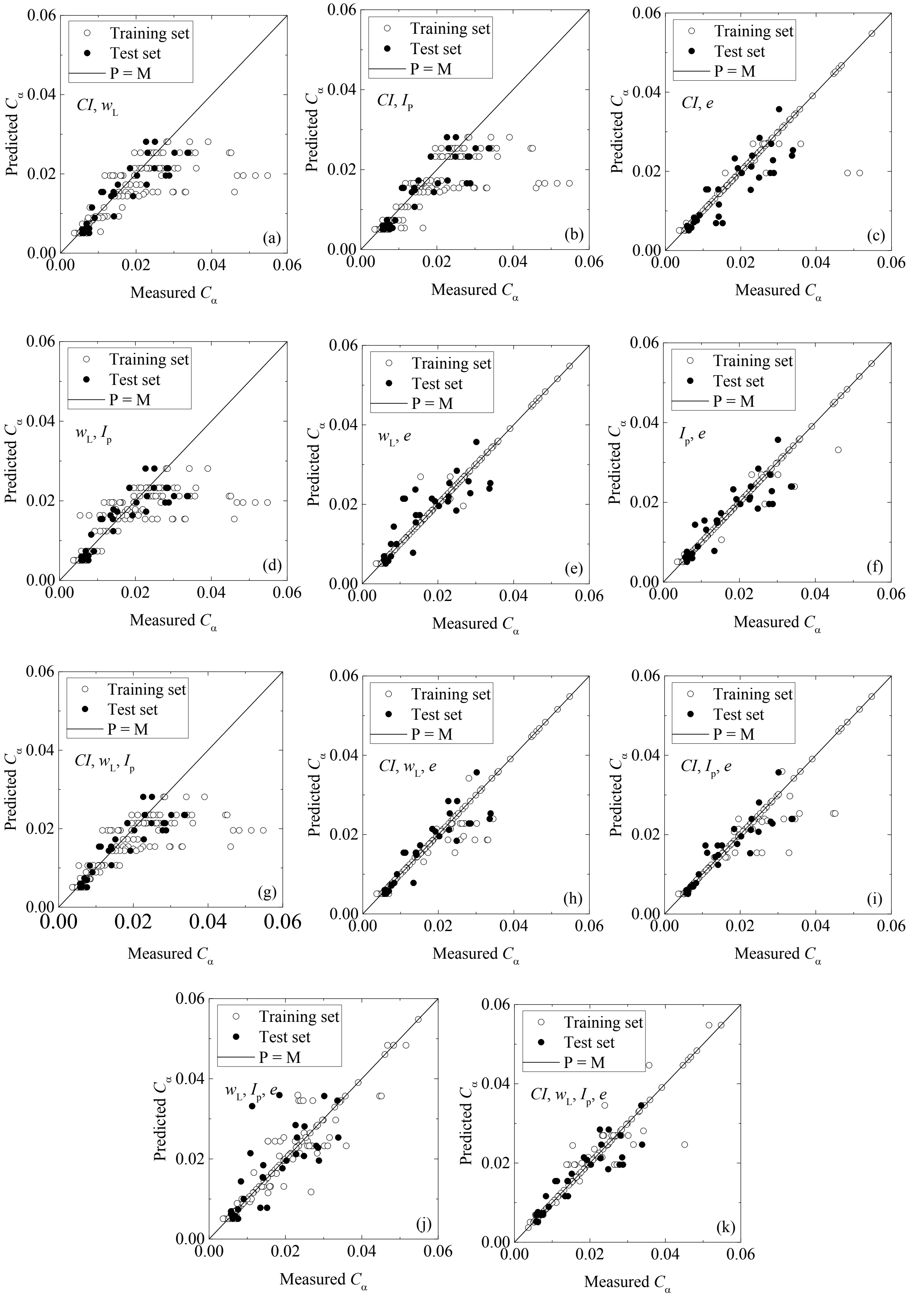

Fig. 8 Predicted C_α for training and test sets by all prediction models

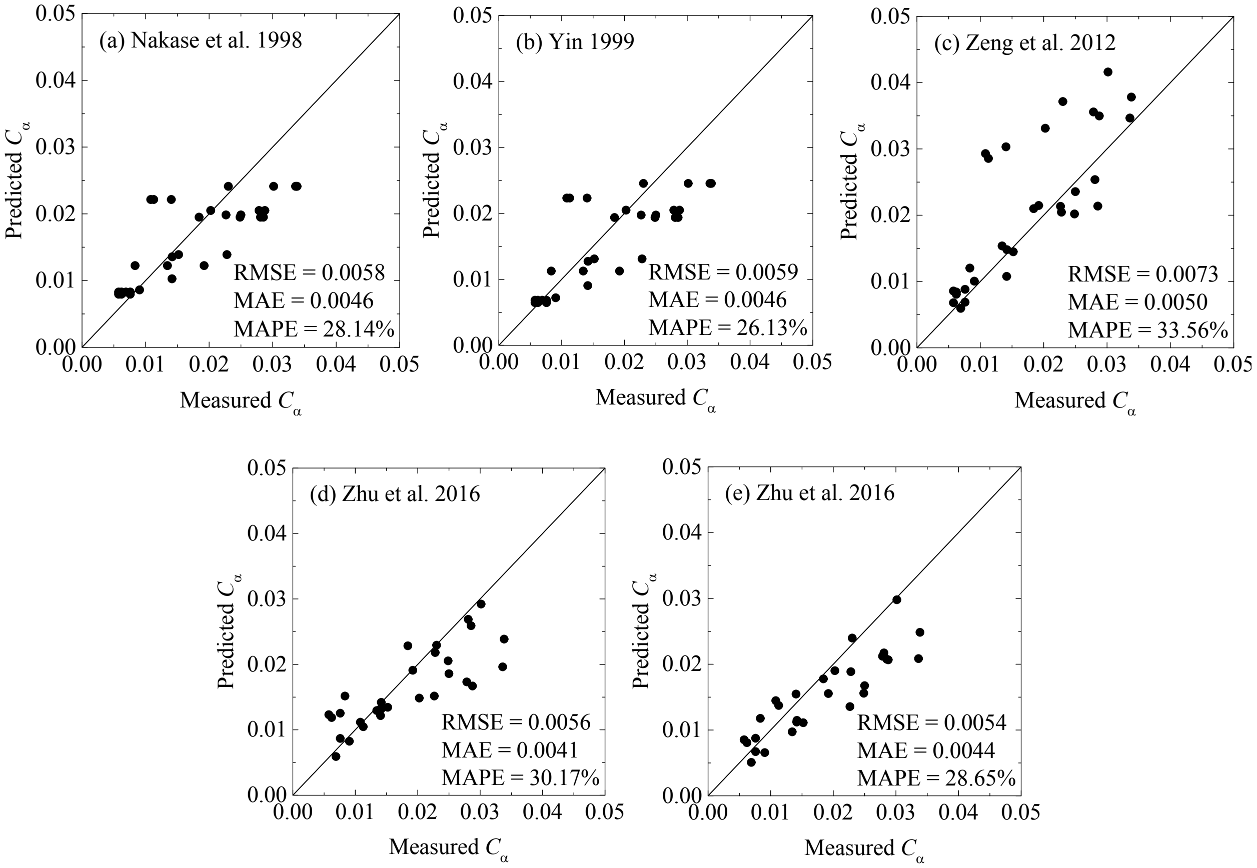

Fig. 9 Predicted C_α for the test set by all published methods

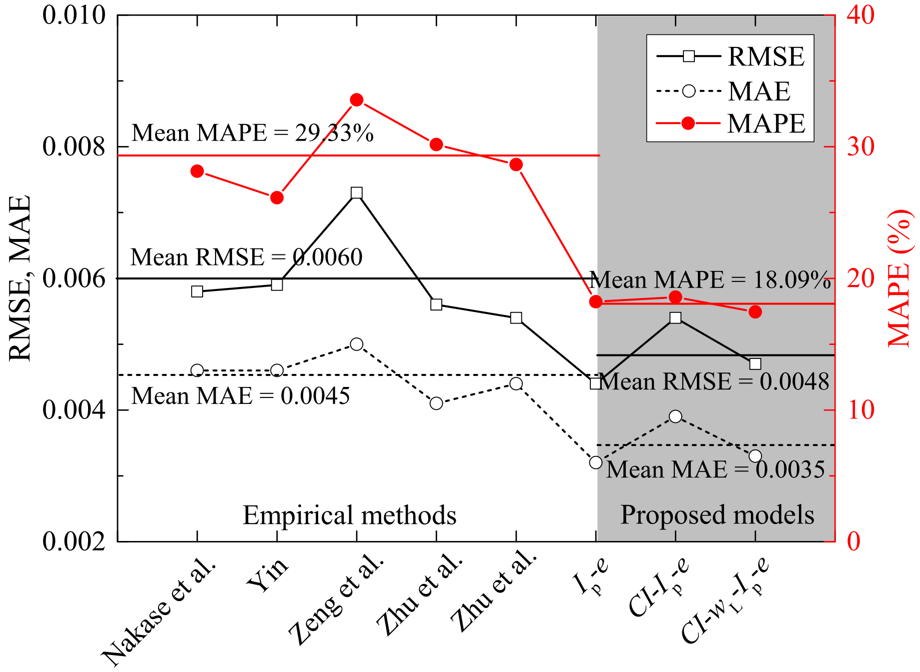

Fig. 10 Comparison of empirical methods and proposed models in predicting C_α

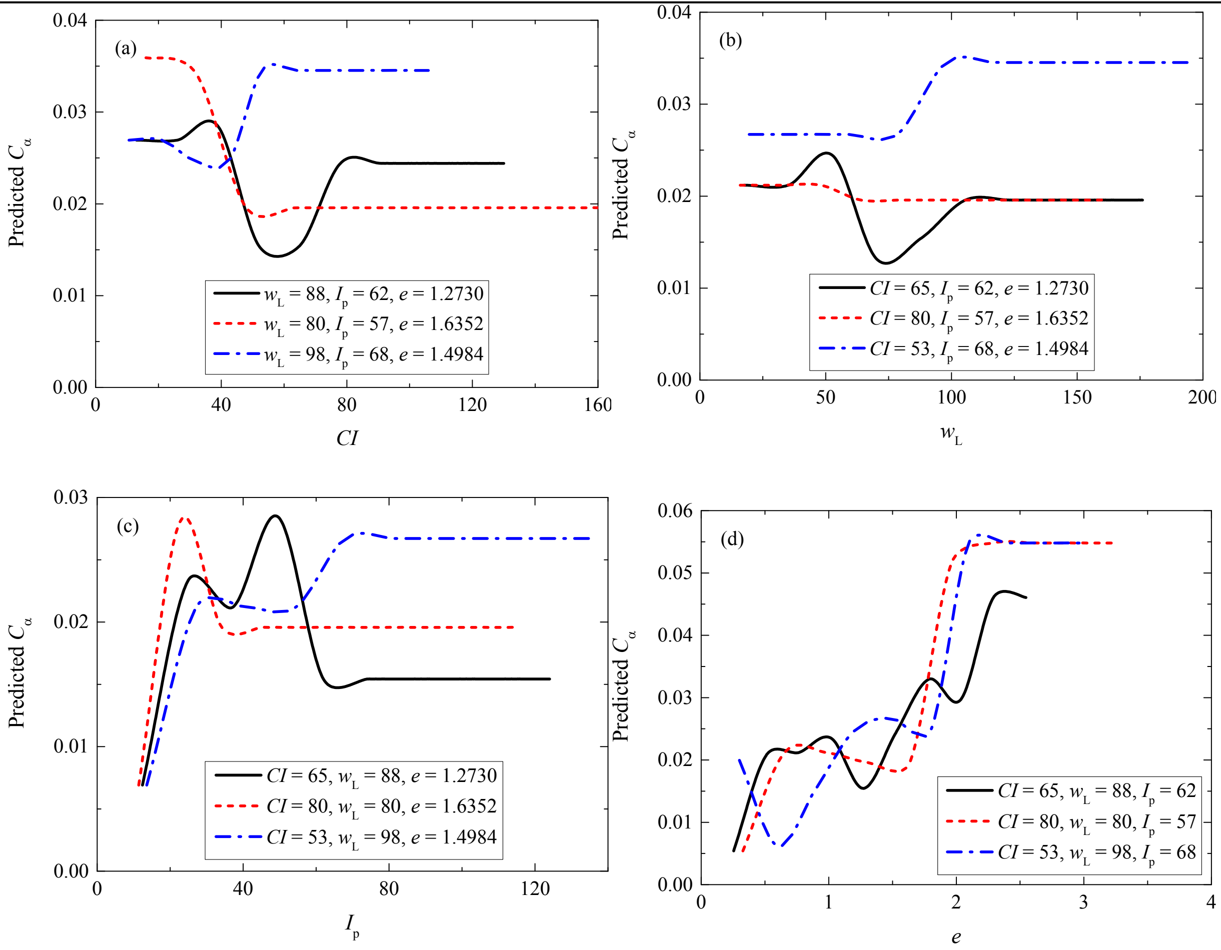

Fig. 11 Predicted C_α using RF model against (a) CI; (b) w_L; (c) I_p; (d) e

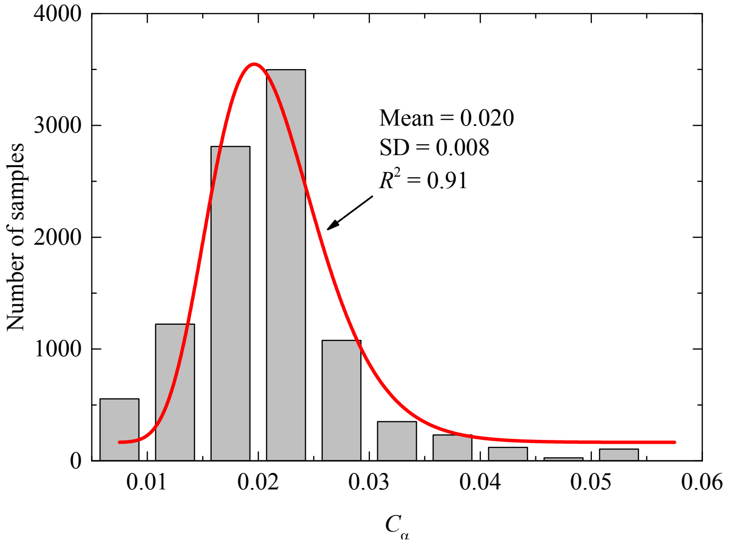

Fig. 12 Distribution of predicted C_α

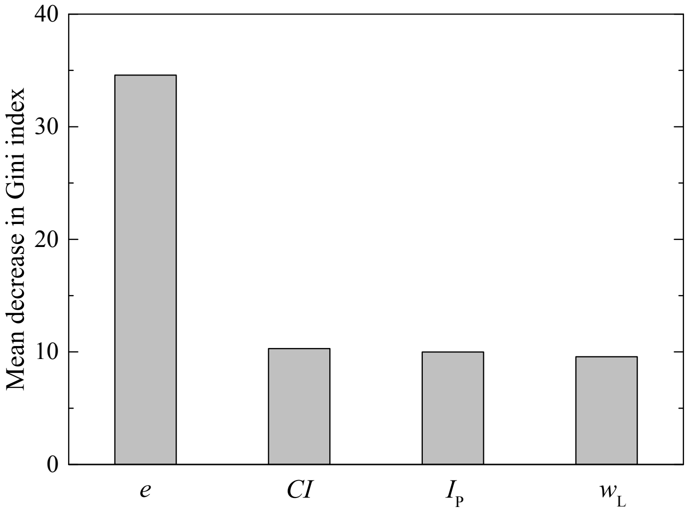

Fig. 13 Mean decrease in Gini index of four input variables

Table 1 Hyper-parameters in random forest

Hyper-parameters	Description	Range
ntree	Number of trees grown	0~300
mtry	Number of predictors sampled for spliting at each node	0~300

Table 2 Performance for all prediction models

Variables combination	Training set			Test set		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
CI, w_L	0.0084	0.0046	18.75%	0.0044	0.0035	19.05%
CI, I_P	0.0087	0.0046	18.36%	0.0045	0.0037	20.02%
CI, e	0.0041	0.0008	2.83%	0.0048	0.0037	19.57%
w_L, I_P	0.0087	0.0048	21.26%	0.0052	0.0039	20.56%
w_L, e	0.0013	0.0002	1.75%	0.0055	0.0048	27.77%
I_P, e	**0.0018**	**0.0004**	**2.35%**	**0.0044**	**0.0032**	**18.23%**
CI, w_L, I_P	0.0085	0.0047	19.01%	0.0047	0.0036	18.37%
CI, w_L, e	0.0030	0.0011	4.82%	0.0055	0.0040	19.42%
CI, I_P, e	**0.0037**	**0.0012**	**4.79%**	**0.0054**	**0.0039**	**18.58%**
w_L, I_P, e	0.0039	0.0020	9.84%	0.0055	0.0051	31.72%
CI, w_L, I_P, e	**0.0030**	**0.0011**	**5.05%**	**0.0047**	**0.0033**	**17.45%**

Note: bold format donates the optimum combination in each group

Table 3 Empirical correlations for C_α

References	Empirical formula
Nakase et al. (1998)	$C_\alpha = 0.00168 + 0.00033 I_p$
Yin (1999)	$C_\alpha = 0.000369 I_p - 0.00055$
Zeng et al. (2012)	$C_\alpha = \left(-0.0067 + 0.0115 e_L - 0.0016 \left(e_L\right)^2\right)\left(1+e\right)$
Zhu et al. (2016)	$C_\alpha = \left(-0.0274 + 0.0011 w_L - 0.00048 I_p\right)\left(\frac{w}{w_L}\right)^{0.7872-0.0369 w_L + 0.0619 I_p}$
Zhu et al. (2016)	$C_\alpha = \left(0.0007 w_L - 0.0223\right)\left(\frac{w}{w_L}\right)^{0.014978 w_L - 0.23031}$