https://dx.doi.org/10.1016/j.enggeo.2020.105911

Straightforward prediction for air-entry value of compacted soils using machine learning algorithms

Han-Lin Wang[1], Zhen-Yu Yin[1,*], Pin Zhang[1], and Yin-Fu Jin[1]

[1] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China.

*** Corresponding author:**

Dr. Zhen-Yu Yin, Tel: (+852) 3400 8470, Fax: (+852) 2334 6389, Email: zhenyu.yin@polyu.edu.hk; zhenyu.yin@gmail.com

Abstract: The straightforward prediction for the air-entry value of compacted soils is practically useful, but the investigation on this issue is scarce. This study presents three alternative straightforward prediction models for the air-entry value of compacted soils using the representative machine learning algorithms of multi expression programming (MEP), evolutionary polynomial regression (EPR) and random forest (RF). Five known soil properties (i.e. sand content, fines content, plasticity index, initial water content and initial void ratio) are used as input variables. All models are developed based on a large database, covering a wide range of soil classifications. The results show that all the three proposed models are appropriate to predict the air-entry values of different compacted soils, with high prediction accuracies for both the training and the testing data. The monotonicity, the sensitivity and the robustness of the three prediction models are evaluated, showing consistency among different models with a slight difference and providing a strong support for the model feasibility. On the whole, the MEP and the EPR models are recommended for more convenient applications with explicit expression, while higher prediction accuracy may require the RF model although no explicit expression can be derived.

Keywords: machine learning; air-entry value; multi expression programming; evolutionary polynomial regression; random forest; compacted soils

Introduction

Compacted soils are widely used in geotechnical or geoenvironmental engineering, such as the urban/landfill barrier system (Tinjum et al. 1997; Miller et al. 2002; Osinubi and Nwaiwu 2006; Birle et al. 2008; Krisdani et al. 2008), buffer for nuclear waste disposal (Delage et al. 1998; Cui et al. 2002; Sun et al. 2009; Ye et al. 2009; He et al. 2019), transportation substructure (Zhao et al. 2016; Zhou et al. 2016; Wang et al. 2017, 2018, 2019a, 2019b; Wang and Chen 2019; Chen et al. 2020b, 2020c, 2020d; de Freitas et al. 2020; Liu et al. 2020), etc. During their servicing period, the soils get saturated during the wetting conditions and the unsaturated state emerges when drying occurs. Through the wetting-drying cycles, the hydro-mechanical behaviours of these compacted soils are highly related to the soil-water characteristic curves (SWCC, Fredlund et al. 2012). Fig. 1 shows a conceptual SWCC for the drying process, plotted by the degree of saturation S_r (the percentage between the volume of liquid water and the volume of voids) against the matric suction ψ (the difference between the pore air pressure and the pore water pressure). In terms of the SWCC, two separation points can be easily identified: air-entry value (AEV) and residual degree of saturation. The air-entry value serves as the matric suction which is required to cause desaturation of the largest pores (i.e. beyond which suction value the air starts to enter the pores of the saturated soil). As shown in Fig. 1, the air-entry value can be obtained by extending the constant slope section of the SWCC to intersect the suction axis at $S_r = 100\%$. Before the air-entry value, water fills the pores of the saturated soil as the matric suction increases (I: boundary effect zone). When the suction increases beyond the air-entry value, water in the pores starts to lose with the continuous increase of the matric suction (II:

transition zone). After the matric suction reaches the point at the residual state, liquid water in the pores becomes discontinuous. In the residual zone (phase III), water is difficult to lose as the matric suction continues to increase.

Due to the importance of the air-entry value in determining the water holding capacity and further the hydro-mechanical behaviours of compacted soils, this parameter is considered in all typical SWCC models (e.g. Brooks and Corey 1964; van Genuchten 1980; Fredlund and Xing 1994). In the engineering practice, the straightforward prediction about the air-entry value of compacted soils is also very important for the design and the maintenance of geotechnical structures. For example, when a subgrade is planned to be constructed in a specific area with known humidity data throughout the year and the subgrade contains a barrier layer to hold the water under heavy precipitation, the prediction of the air-entry value of this layer is needed. During the raining conditions, the barrier layer stores rainwater. Under unfavourable conditions, the rainwater cannot go deeper or be lost if the humidity-induced suction is lower than the air-entry value. Thus, the threshold suction value can be used as a reference for the selection of the barrier materials and initial placement conditions for these materials.

In general, the air-entry value is determined through the SWCC, which can be obtained by laboratory tests or analytical models (Fredlund et al. 2012). The laboratory tests to measure the SWCC include pressure plate test, vacuum desiccator, chilled-mirror dew-point method, filter paper, unsaturated oedometer or unsaturated triaxial apparatus, etc. All these tests are costly and time-consuming (lasting for several months for a specific sample). Thus, several models were proposed to predict the SWCC of different soils from the soil properties

(Fredlund et al. 2002; Johari et al. 2006; Li et al. 2014; Zhou et al. 2014; Zhai et al. 2020). However, using these methods, the air-entry value still needs to be determined after fitting the SWCC. Thus, the accuracy of determining the air-entry value is highly dependent on the prediction of the SWCC. Furthermore, in these studies, the database used for predicting the SWCC was not large enough to cover a wide range of soils. To the authors' knowledge, the straightforward prediction for the air-entry value of the compacted soils remains scarce.

To solve nonlinear and complex problems for the prediction with a large database, several machine learning algorithms have been proven as effective approaches. In the field of geotechnical engineering, the machine learning algorithms have been successfully used to predict cyclic soil response (Shahnazari et al. 2010), creep index (Jin et al. 2019), bearing capacity of composite column (Sarir et al. 2019), spatiotemporal response of rooted soil (Cheng et al. 2020a), suction distribution close to tree (Cheng et al. 2020b), soil liquefaction (Njock et al. 2020), jet grouted diameter in soft soils (Shen et al. 2020), tunneling induced settlement (Zhang et al. 2020b), etc. In these studies, the algorithms cover the representative multi expression programming (MEP), evolutionary polynomial regression (EPR) and random forest (RF), etc. To date, the air-entry value of compacted soils has scarcely been predicted by the machine learning algorithms. To obtain the air-entry value of compacted soils in a fast and accurate manner, a comprehensive understanding of different machine learning algorithms on the prediction of the air-entry value is imperative and worth investigating.

In this study, alternative straightforward prediction models for the air-entry value of compacted soils are developed using three commonly used machine learning algorithms:

MEP, EPR and RF. A large database of soils with multi classifications is collected from the previous publications. Two-thirds of the data are chosen as the training data, while the remaining are used for testing. Each prediction model is developed with each algorithm using the respective optimum parameters. The prediction accuracy of the three models are verified by the training and the testing data. The feasibility is further examined by the monotonicity, sensitivity and robustness analysis for all three models, along with discussing their advantages and limitations.

Machine learning algorithms

Multi expression programming

Multi expression programming (MEP) is a representative approach to linear-based genetic programming (GP). In a chromosome of the MEP algorithm, multiple solutions (programs) can be encoded, starting with the creation of a random population of computer programs. The first gene of a chromosome must be a terminal randomly selected from the terminal set. In the following genes, a gene with a function has a pointer towards the function arguments. For a specific gene, the expression indices have lower values than the position of this gene in the chromosome. Through the calculation of the fitness of all expressions, the best encoding solution is determined to represent the chromosome by repeating the following steps, until the termination condition is reached (Oltean and Grosan 2003): (i) selecting two parents by a procedure of binary tournament and recombining them with a fixed crossover probability; (ii) obtaining two offspring by recombining two parents; (iii) mutating the offspring and replacing the worst individual in the current population with the best of them when the offspring is better than the worst one. After the identification of the best solution, the explicit expressions can be generated by reading the chromosome from top to bottom.

Evolutionary polynomial regression

Evolutionary polynomial regression (EPR) is another type of genetic programming (GP), to develop symbolic models following two steps: (i) structure identification, and (ii) parameter estimation (Giustolisi and Savic 2006). In the first step, the genetic algorithm (GA) is adopted to search for symbolic structures of EPR. During the second step, the values of parameters are estimated by solving the least squares (LS) linear problem. The advantage of EPR highlights that a simple explicit expression can be presented in the EPR algorithm, to describe the correlation between input and output variables. A general EPR expression is formulated as:

$$\boldsymbol{t} = \sum_{j=1}^{m} a_j \bullet \boldsymbol{z}_j + a_0 \tag{1}$$

where $\boldsymbol{t}$ is the predicted output; a_j is an adjustable parameter for the jth term; a_0 is an optional bias; m is the number of transformed terms; $\boldsymbol{z}_j$ is the jth transformed variable, which can be obtained by:

$$\boldsymbol{z}_j = \boldsymbol{x}_1^{\mathbf{ES}(j,1)} \bullet \ldots \bullet \boldsymbol{x}_i^{\mathbf{ES}(j,i)} \bullet \ldots \bullet \boldsymbol{x}_k^{\mathbf{ES}(j,k)} \tag{2}$$

where $\boldsymbol{x}_i$ is the ith input variable; k is a total number of input variables; $\mathbf{ES}_{m\times k}$ is the exponent matrix, determined by GA. The key objective of EPR is to identify the best form of the function: the number of transformed variables and a combination of vectors of independent input variables. Then, the adjustable parameters and an optional bias can be determined by the least squares regression. Finally, the optimum explicit expression can thus be deduced.

Random forest

Random forest (RF) is an ensemble learning algorithm, integrated with the methods of bootstrap aggregating (Breiman 1996) and random subspace (Ho 1998). Due to the integration of numerous decision trees, the prediction of RF shows a strong performance (Zhang et al. 2019, 2020a, 2020b, 2020c). In bagging, n_t bootstrap sets are built by sampling with the replacement of N training examples from the training database. The number of samples in the bootstrap training set is arbitrary, less than the original one. Then, each bootstrap set is used to develop a decision tree. Each node in a decision tree represents a classification criterion, with the leaves of the tree representing the output labels. Hence, a decision tree classifies a bootstrap training sample by testing random features at each node. As a result, a regression space can be determined. The ultimate predicted output t can be obtained by aggregating the outputs of all trees as (Liaw and Wiener 2002):

$$t = \frac{1}{n_t}\sum_{i=1}^{n_t} t_i(\boldsymbol{x}) \tag{3}$$

in which $t_i(\boldsymbol{x})$ is the predicted output for a tree with an input vector $\boldsymbol{x}$; n_t is the number of trees.

Model development

Database collection

To directly predict the air-entry value of compacted soils, 189 relevant samples are collected from the experimental data of previous publications (Han et al. 1995; Tinjum et al. 1997; Huang et al. 1998; Vanapalli et al. 1999; Ng and Pang 2000; Agus et al. 2001; Montanez 2002; Khalili et al. 2004; Yang et al. 2004; Indrawan et al. 2006; Puppala et al. 2006; Sun et al. 2006; Thu et al. 2006, 2007; Birle et al. 2008; Krisdani et al. 2008; Rahardjo et al. 2008; Li 2009; Zhang and Chen 2009; Gallage and Uchimura 2010; Zhou and Kong 2011; Mirzaii and Yasrobi 2012; Oh et al. 2012; Rahardjo et al. 2012; Lin and Cerato 2013; Salager et al. 2013; Sun et al. 2014; Sun and Gao 2015; Amadi and Osinubi 2016; Cuceoglu 2016; Han and Vanapalli 2016; Hashem and Houston 2016; Priono et al. 2016; Fattah et al. 2017; Jiang et al. 2017; Satyanaga et al. 2017; Chen et al. 2019, 2020a; de Freitas et al. 2020). According to these studies, the main influencing factors on the air-entry value of compacted soils include grain size distribution, plasticity and initial placement conditions. Thus, the soil properties of gravel content C_G, sand content C_S, fines content C_F, plasticity index PI, initial water content w_0, initial void ratio e_0 and air-entry value AEV are collected. The details of the soil properties and the relevant testing methods can be downloaded and referred to the supplementary database. Note that the gravel, sand and fines are separated by the grain size range of 75 mm to 4.75 mm, 4.75 mm to 0.075 mm and smaller than 0.075 mm, respectively (ASTM 2017). In this database, various soil classifications (ASTM 2017) are collected, including lean clay (CL), silty clay (CL-ML), fat clay (CH), silt (ML), elastic silt (MH), well-graded sand (SW), well-graded sand with clay (SW-SC), poorly graded sand (SP), poorly graded sand with clay (SP-SC), clayey sand (SC), silty clayey sand (SC-SM), silty sand (SM), well-graded gravel with silt (GW-GM), clayey gravel (GC) and poorly graded gravel (GP). In the supplementary database, these soils are ordered and numbered by the soil classification. For the same classification, the soils are ordered alphabetically by the name of the authors.

Table 1 lists the descriptive statistics of each variable in the database, with the values of minimum, maximum, mean and standard deviation. Fig. 2 shows the detailed frequency histogram of each variable, including the soil classification. As the gravel soil presents a relatively lower water holding capacity, the SWCC of this kind of soil was not widely investigated in the literature. Thus, the majority of the collected soils have the gravel content of less than 20% (Fig. 2a), leading to a mean gravel content of 3.8% (Table 1). Compared to the gravel content, the frequency distribution of the sand and the fines content is more uniform, with the highest value at 0% to 20% (Fig. 2b) and 80% to 100% (Fig. 2c), respectively. Regarding the plasticity index, around 120 soils are in the range of 0 to 20, whereas 19 samples show the plasticity index higher than 40 (Fig. 2d). The initial water content ranges from 0.5% to 48.6% (Table 1), showing the highest frequency at 10% to 20% (Fig. 2e). The majority of the initial void ratio concentrates in the range from 0.4 to 0.8 (Fig. 2f), which is also common for the compacted soils. In this database, a wide range of soil classifications from clay to gravel is introduced (Fig. 2g), showing the highest frequency for the samples of lean clay (48) and clayey sand (35). In terms of the air-entry value, most of the values are located in the range from 0 kPa to 20 kPa (Fig. 2h), while the highest air-entry value is 100 kPa (Table 1).

To have an overall understanding of the distribution of the air-entry value of compacted soils with different classifications, Fig. 3 is plotted with the air-entry value versus the specific and the general classification, respectively. Table 2 lists the minimum, the maximum, the mean and the median air-entry values for the general soil classifications. It is widely known that at the same initial placement conditions, the fine-grained soil should have a higher water holding capacity than the coarse-grained soil, leading to a higher air-entry value of the fine-grained soil. From Fig. 3 (a), it can be observed that the maximum air-entry value shows for the fat clay (CH, 100 kPa). However, the highest mean air-entry value (box symbol in the figure) locates for the well-graded sand with clay (SW-SC). This is because for this soil in the database, only 4 samples are collected and 3 samples have the fines with a very high plasticity index (88), resulting in high air-entry values. Regarding the general soil classification in Fig. 3 (b) and Table 2, the clay has the highest air-entry values of maximum, median (transverse line symbol) and minimum, and the silt has the highest mean air-entry value. In other words, the air-entry value of the fine-grained soil shows relatively higher values than the coarse-grained soil, supporting the feasibility of the supplementary database.

Before developing the prediction model, the basic linear fitting between the air-entry value and each input soil property is depicted to have an overall view of the monotonic variation trend, as shown in Fig. 4. The coefficient of determination R^2 is used to evaluate the fitting accuracy as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(h_i - t_i)^2}{\sum_{i=1}^{n}(h_i - \overline{h_i})^2} \tag{4}$$

where h_i and t_i are the actual and predicted output values for the i th output; n is the number of outputs; $\overline{h_i}$ is the average value of the actual outputs. The R^2 ranges from 0 to 1. The higher the R^2 value, the higher the fitting accuracy is obtained. It can be seen from Fig. 4 that the linear fitting between the air-entry value and the input soil property all shows a relatively low fitting accuracy ($R^2 \le 0.173$). For the grain size distribution, the air-entry value decreases as the gravel or the sand content increases (Figs. 4a and 4b), while the air-entry value increases with the increase of the fines content (Fig. 4c). When the plasticity index increases, the air-entry value increases accordingly (Fig. 4d). Regarding the initial placement conditions, the air-entry value increases as the initial water content increases (Fig. 4e) or as the initial void ratio decreases (Fig. 4f). On the whole, the linear fittings in Fig. 4 are not accurate enough to state the relationship between the air-entry value and a single soil property. Hence, a more comprehensive prediction model is needed to connect the air-entry value and the known input variables.

Model development

To develop the prediction model accurately and to check the validity of the model, about two-thirds of the samples (126 samples) in the supplementary database are chosen as the training data, and the rest (63 samples) are used for testing. As the samples are ordered by soil classification in the database, the samples with the line number as the multiple of three are picked out for testing. In this way, both the training and the testing data cover all kinds of soil classifications in the database. Note that the solid soil particles are constituted by gravel, sand and fines, with the summation of their contents as 100%. Hence, the gravel content is not considered in the training process. With the contents of sand and fines, the gravel content can be calculated accordingly. It is also worth mentioning that the slight discrepancy of the air-entry value induced by different testing methods (Lin and Cerato 2013; Sun et al. 2016) is not considered during the model development. More studies are needed to clarify this issue.

To evaluate the prediction precision, three indicators of mean absolute error MAE, root mean squared error $RMSE$ and coefficient of determination R^2 [see Eq. (4)] are introduced as:

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (h_i - t_i)^2}{n}} \tag{6}$$

The lower values of MAE or $RMSE$, or the higher value of R^2 indicate the higher precision of the model.

For the MEP, the source code of Oltean (2004) is used for training the data, with various pre-set parameters (Oltean and Dumitrescu 2002). The population size defines the number of programs in the population. The number of generations is the number of calculations before the run of a program terminates. The crossover and the mutation probability indicate the probability of an offspring that is imposed on the crossover and the mutation operator, respectively. When the crossover type is set as uniform, the offspring genes are randomly taken from one parent to another. The code length represents the number of genes in each chromosome. The replication number is the number of runs or the number of developed chromosomes. In this study, the determination of the optimum code parameters follows the trial and error method used by Wang and Yin (2020), keeping the prediction MAE at the minimum level. Table 3 lists the initial parameter setting, as suggested by several previous studies (Oltean and Dumitrescu 2002; Shahnazari et al. 2010; Wang and Yin 2020). Using the optimum parameter setting, explicit expressions are developed accordingly.

To obtain the best solution using EPR, the number of transformed terms in the explicit expression, the values and the interval of elements in the exponent matrix need to be optimised. Table 3 lists the initial setting of the EPR parameters. Following Zhang et al.

(2020d), the interval of elements is set as 0.5. In the EPR calculation, the number of terms needs to be pre-set at a fixed value. Then, the values of elements for each input variable is determined by GA, with the parameters shown in Table 3. These parameters are verified to be optimum for the EPR method (Zhang et al. 2020d). During each calculation, the values of elements are randomly assigned to machine learning algorithms by GA. The performance of the machine learning algorithms with these parameters is assessed by the fitness value until the terminal condition is satisfied. As the explicit EPR expression contains the input components with index and some of them serve as the denominators (Jin et al. 2019; Jin and Yin 2020), the input variable with the value of 0 is not applicable and may influence the prediction accuracy if directly used. To improve the accuracy of EPR and to guarantee the feasibility of the EPR expression, the input variables expressed as percentages are written as decimals, and then the input variables with the value of 0 are converted to a small value of 0.001. Note that the conversions take place for a minor proportion of the input variables of sand content, fines content and plasticity index, showing negligible conversion value compared to the variation range of these variables (see Table 1). After the calculation, the optimum combination of the values of elements is determined. By comparing the prediction MAE for the cases with a different number of terms, the optimum EPR explicit expression is deduced.

Regarding the RF, the numbers of trees and features at each node need to be optimised. In this study, the initial setting of these two parameters is listed in Table 3. GA is also used to search for the optimum setting of these two parameters, using the method by Zhang et al. (2020d). The parameter setting for GA is listed in Table 3, verified to be optimum for the RF method (Zhang et al. 2020d). By randomly assigning the parameters to machine learning algorithms and evaluating the fitness value, the optimum number of trees and features is determined. Accordingly, the air-entry values can be predicted using this optimum parameter setting.

Straightforward prediction of AEV

Following the respective procedure of each machine learning algorithm, the optimum parameters are determined. For the MEP, the optimum combination of code parameters shows population size of 3000, code length of 50, crossover probability of 0.9, crossover type of uniform, mutation probability of 0.01, number of generation of 3000, function set of +, -, ×, /, pow and replication number of 10. The optimum number of transformed terms for the EPR is 6. The optimum number of trees and features at each node for the RF is 116 and 3, respectively.

Using the optimum parameters, the MEP explicit expression for the air-entry value of compacted soils is derived as:

$$AEV = 3(A_4 + A_3 A_4^{A_3}) + A_6 + A_4^{A_3} + e_0 \tag{7}$$

in which A_1, A_2, A_3, A_4, A_5, A_6 and A_7 are parameters as:

$$A_1 = (2e_0)^{A_7} \tag{8}$$

$$A_2 = PI + w_0 + e_0^{e_0} \tag{9}$$

$$A_3 = \frac{2PI}{C_S + C_F - A_7} \tag{10}$$

$$A_4 = \frac{A_2 - 2e_0}{C_S + C_F} \tag{11}$$

$$A_5 = \frac{2e_0 A_7 (C_S + C_F - A_7)}{A_1 + A_2} \tag{12}$$

$$A_6 = \frac{A_5 + e_0 A_2}{e_0^{\,e_0} + 2C_S^{\,e_0} e_0} \tag{13}$$

$$A_7 = 2e_0 + C_S \tag{14}$$

The EPR explicit expression for the air-entry value of compacted soils is deduced as:

$$\begin{aligned} AEV = {} & 1.8116 - 0.0069\left(\frac{C_{S1}PI_1}{w_{01}}\right)^3\left(\frac{C_{F1}}{e_0}\right)^{2.5} + 29.2526\left(\frac{C_{F1}}{e_0}\right)^{2.5}\frac{C_{S1}^{\,3}}{w_{01}^{\,0.5}} + 27.4022\frac{C_{S1}^{\,0.5}C_{F1}^{\,3}}{e_0} \\ & + 470.6056\frac{(C_{S1}PI_1)^3 w_{01}^{\,1.5}}{C_{F1}^{\,0.5}e_0} + 1.4266\cdot 10^{-7}\frac{w_{01}^{\,2}}{(C_{S1}e_0)^3 PI_1^{\,0.5}} + 1.2272\frac{C_{F1}^{\,2.5}PI_1 e_0}{C_{S1}^{\,0.5}w_{01}} \end{aligned} \tag{15}$$

where C_{S1}, C_{F1}, PI_1, w_{01} and e_{01} are sand content, fines content, plasticity index, initial water content and initial void ratio all in decimals, respectively. Note also that the explicit expression cannot be derived from the RF model.

Fig. 5 presents the comparison between the predicted and the reference air-entry values for the training data by the three prediction models. The prediction indicators are also shown in each figure for evaluation. It can be observed from this figure that the predicted air-entry values have a good agreement with the reference data for all the three models, with the R^2 higher than 0.85. For the training data, the RF presents a higher prediction accuracy, while the prediction accuracies of MEP and EPR stay close to each other. Nevertheless, the high prediction accuracies of the training data indicate the appropriate development of all the prediction models.

Results and discussions

Model validation

The soil samples with the line number as the multiple of three are considered as the testing data in the supplementary database. Using the soil properties and the prediction models, the air-entry values of these testing samples are calculated. Fig. 6 plots the comparison between the predicted and the reference air-entry values of the testing data by each prediction model. Note that to stay consistent with the training data, when using the EPR model, the input variables with the percentage expressions are written as decimals and the input variables with the value of 0 are also converted to 0.001. From Fig. 6, it is observed that although the RF model shows a higher prediction accuracy for the training data than the other two models, similar testing accuracy can be identified for the three models, with the R^2 varying from 0.84 to 0.88. On the whole, satisfactory prediction accuracies are verified by all the three models for the air-entry values of compacted soils.

Monotonicity analysis

To check the availability of the prediction models, the monotonicity analysis is performed, using the method from Jin et al. (2019), Jin and Yin (2020), Wang and Yin (2020). To conduct the monotonicity analysis, the investigated soil property changes, while the other properties stay constant. With the input soil properties and the prediction model, the air-entry value of compacted soils can be calculated. Table 4 lists the basic setting of the soil properties for the monotonicity analysis. These soil properties are chosen as their respective mean values from the supplementary database (see Table 1). During the monotonicity analysis, the variation of the input soil property cannot exceed the threshold defined by their respective maximum and minimum values (see Table 1). Note that as the soils are constituted by gravel, sand and fines, the content of each material influences each other and the summation of their contents equals to 100%. Thus, for the monotonicity analysis of the sand and the fines content, the gravel content is fixed at 3.8%. For the monotonicity analysis of the other soil properties, the sand

and the fines content are set as the values shown in Table 4.

Fig. 7 depicts the results of the monotonicity analysis using the prediction models, showing the variation of the predicted air-entry value with each input soil property. In general, the monotonic variation trend predicted by the three models agrees with each other, also showing consistency with that by the basic linear fitting in Fig. 4. Despite the overall consistent trend, the smooth correlation between the RF predicted output variable and the input variable is difficult to obtain, because the RF model is developed strictly following the actual data from the database with no smooth relationships (Fig. 4). The predicted air-entry value increases with the decrease of the sand content or the increase of the fines content (Figs. 7a and 7b). As the plasticity index increases, the air-entry value increases accordingly (Fig. 7c; available for the RF prediction only when PI > 15). In terms of the initial placement conditions, the air-entry value increases when the initial water content increases or when the initial void ratio decreases (Figs. 7d and 7e). In spite of the general consistency of the monotonicity results for each model, some differences still exist, especially for the results regarding the initial placement conditions. With the increase of the initial water content, the increasing amplitude of the predicted air-entry value by EPR and RF is relatively small (Fig. 7d). Besides, the increasing trend only shows for the EPR model when the initial water content is higher than about 12%. With the initial void ratio lower than around 0.5, the monotonicity predicted by the RF shows the opposite trend (Fig. 7e). However, the general consistent monotonic variation trend of the predicted air-entry value with each input soil property between the prediction results and the original database directly verifies the validity of the developed models.

Sensitivity analysis

To have a better understanding of the contribution of the input soil property on the predicted air-entry value, the sensitivity analysis is conducted on the whole database. For a specific input variable x_i, the sensitivity R_{sen} is determined as (Wang and Yin 2020):

$$R_{sen} = \frac{\sum_{i=1}^{N}(x_i t_i)}{\sqrt{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} t_i^2}} \tag{16}$$

where t_i is the predicted output air-entry value using the proposed prediction models; N is the number of soil samples in the supplementary database ($N = 189$). The sensitivity R_{sen} ranges from 0 to 1, indicating the relevance between the predicted air-entry value and each input soil property. With the R_{sen} value closer to 1, the specific input soil property has a more remarkable influence on the predicted air-entry value. Note that to stay consistent with the prediction setup using the EPR algorithm, the input variables with the percentage expressions are written as decimals, also with the input variables of 0 value converted to 0.001 in this analysis.

Fig. 8 presents the distribution of the sensitivity value for each input soil property on the predicted air-entry value using the three prediction models. For a specific input soil property except for the plasticity index, the sensitivity shows a slightly higher value by the RF model, while the sensitivity value by the MEP model is relatively lower. However, this difference is not significant for a given input soil property. On the whole, the sand content has the least influence on the prediction of the air-entry value by all the three models. By contrast, the sensitivity value of the fines content and the initial water content rank the first- and the second-highest, respectively. Except for the sand content, the sensitivity value of the other four soil properties stays close to each other, showing the values between 0.6 and 0.8.

Robustness analysis

To validate the prediction model, the robustness analysis is another key aspect (Jin et al. 2019; Jin and Yin 2020), to guarantee that the output values are reasonable with the appropriate input variables. To assess the robustness of the prediction model, a robustness ratio r is defined as:

$$r = \frac{\text{Samples in the reasonable range}}{\text{Total testing samples}} \tag{17}$$

From the present database, the reasonable range for the air-entry value of compacted soils locate between 0.06 kPa and 100 kPa (Table 1). To generate the testing samples, the five input variables (C_S, C_F, PI, w_0, e_0) are first assumed to be independent to each other and to obey the lognormal distribution (Jin et al. 2019; Jin and Yin 2020). From the statistics in Table 1, 80,000 testing samples are randomly generated using the mean and the standard deviation of each input variable with the lognormal distribution. Then, the values exceeding the minimum and the maximum thresholds (Table 1) are deleted. Besides, the generated samples with the summation of sand and fines contents exceeding 100% are deleted. Finally, 10,000 samples are chosen as the testing samples, still showing close mean value and standard deviation for each input variable. Using the 10,000 samples, the robustness analysis is conducted for each prediction model.

Fig. 9 depicts the distribution of the predicted air-entry values for each prediction model, showing the robustness ratio in the legend. It can be seen that the majority of the predicted air-entry values locate in the range from 0 kPa to 20 kPa. For the models of MEP and EPR, some negative values exist. For the three models, some air-entry values exceed 100 kPa.

However, the robustness ratios are higher than 98% for all the three methods, suggesting the feasibility of the prediction models. By comparing the robustness ratio of each model, the RF and the MEP models are slightly more robust.

Limitations of the present algorithms

In this study, three representative machine learning algorithms are used to predict the air-entry values of compacted soils. All the prediction models by the present algorithms show relatively high accuracies, consistent monotonicity and strong robustness. However, there are still some limitations for each algorithm. The prediction model by MEP has a good performance for the complex and nonlinear problems. But the length of the expressions and the feasibility of the model needs to be balanced for practical applications. The EPR model is more convenient to use, with only one explicit expression. While due to the expression form of EPR, the limitation exists when the input variables include some values equalling to 0. The prediction accuracy of the training data and the robustness ratio of RF are the highest in the present study. Nevertheless, no explicit expressions can be generated using this algorithm, causing some inconvenience for the applications. In addition, the smooth monotonicity correlation between the predicted output and each input variable is difficult to obtain by RF. Hence, according to the specific scenario, a suitable machine learning algorithm needs to be selected to develop the prediction model with high accuracy, feasibility and convenience.

Conclusions

In this study, three alternative straightforward prediction models for the air-entry value of compacted soils have been developed using three representative machine learning algorithms:

multi expression programming (MEP), evolutionary polynomial regression (EPR) and random forest (RF). A large database with a wide range of soil classifications has been collected, covering clay, silt, sand and gravel.

The optimum parameter setting for each algorithm was determined firstly. Using their optimum parameter settings, three prediction models for the air-entry value of compacted soils were developed based on the training data, showing reasonable prediction accuracies. By comparison between the predicted air-entry values and the reference ones of the testing data, the prediction precisions were validated for all the three models.

The monotonicity, the sensitivity and the robustness analysis were conducted using the proposed models, showing consistent results among different models. From the monotonicity analysis, the predicted air-entry value increases monotonically as the fines content, the plasticity index or the initial water content increases, while it decreases with the increase of either the sand content or the initial void ratio. The variation trend of the monotonicity analysis shows a good agreement with the original database. The sensitivity analysis indicates that the sand content has a relatively lower relevance on the prediction of the air-entry value, whereas the influence of the other four soil properties stays close to each other at a high value. From the robustness analysis, the high robustness ratios of the predicted air-entry values strongly support the feasibility of the three prediction models.

Although three models have slight differences in terms of performance, the MEP model is first recommended due to its advantages of both explicit formulation and good monotonicity. For convenient engineering practice, the EPR model is also recommended because of the simple explicit formulation, despite of its deficiency to treat the input variable with 0 value. Without the explicit expression, the RF model is not as convenient as the other two models, but it can be used for the cases demanding a higher prediction accuracy.

Notations

a_j Adjustable parameter for *j*th item

a_0 Optional bias

A_1, A_2, A_3, A_4, A_5, A_6, A_7 Parameters for the prediction model

AEV Air-entry value

C_F Fines content

C_{F1} Fines content in decimal

C_G Gravel content

C_S Sand content

C_{S1} Sand content in decimal

e_0 Initial void ratio

e_{01} Initial void ratio in decimal

$\mathbf{ES}_{m\times k}$ Exponent matrix

h_i Actual output variable

$\overline{h_i}$ Average value of actual outputs

k Total number of input variables

m Number of transformed terms

MAE Mean absolute error

n Number of outputs

N Number of soil samples

n_t Number of trees

PI Plasticity index

PI_1 Plasticity index in decimal

r Robustness ratio

R^2 Coefficient of determination

R_{sen} Sensitivity

$RMSE$ Root mean squared error

S_r Degree of saturation

t_i Predicted output variable

$t_i(\boldsymbol{x})$ Predicted output for a tree with an input vector $\boldsymbol{x}$

$\boldsymbol{t}$, t Predicted output

w_0 Initial water content

w_{01} Initial water content in decimal

x_i Input variable

$\boldsymbol{x}_i$ ith input variable

$\boldsymbol{x}$ Input vector

$\boldsymbol{z}_j$ jth transformed variable

ψ Matric suction

Acknowledgement

The financial supports provided by the RIF project (Grant No. R5037-18F) from Research Grants Council (RGC) of Hong Kong are gratefully acknowledged.

References

Agus, S.S., Leong, E.C., Rahardjo, H., 2001. Soil-water characteristic curves of Singapore residual soils. Geotech. Geol. Eng. 19, 285-309.

Amadi, A.A., Osinubi, K.J., 2016. Soil-water characteristic curves for compacted lateritic soil-bentonite mixtures developed for landfill liner applications. In: Geo-Chicago 2016: Sustainability and Resiliency in Geotechnical Engineering, Chicago, Illinois, US, pp. 488-497.

ASTM., 2017. Standard practice for classification of soils for engineering purposes (unified soil classification system). ASTM D2487-17, West Conshohocken, PA.

Birle, E., Heyer, D., Vogt, N., 2008. Influence of the initial water content and dry density on the soil–water retention curve and the shrinkage behavior of a compacted clay. Acta Geotech. 3(3), 191.

Breiman, L., 1996. Bagging Predictors. Mach. Learn. 24 (2), 123-140.

Brooks, R., Corey, T., 1964. Hydraulic properties of porous media. Hydrology Papers, Colorado State University, 24, 37.

Chen, R., Tan, R., Chen, Z., Ping, Y., Mei, Z., 2020a. Influence of degree of compaction on unsaturated hydraulic properties of a compacted completely decomposed granite. Geofluids 2020, 7615361.

Chen, R.P., Qi, S., Wang, H.L., Cui, Y. J., 2019. Microstructure and hydraulic properties of coarse-grained subgrade soil used in high-speed railway at various compaction degrees. J. Mater. Civil Eng. 31 (12), 04019301.

Chen, W.B., Feng, W.Q., Yin, J.H., 2020b. Effects of water content on resilient modulus of a granular material with high fines content. Constr. Build. Mat. 236, 117542.

Chen, W.B., Feng, W.Q., Yin, J.H., Chen, J.M., Borana, L., Chen, R.P., 2020c. New model for predicting permanent strain of granular materials subjected to cyclic loadings. J. Geotech. Geoenviron. Eng. 146 (9), 04020084.

Chen, W.B., Liu, K., Feng, W.Q., Borana, L., Yin, J.H., 2020d. Influence of matric suction on nonlinear time-dependent compression behavior of a granular fill material. Acta Geotech. 15 (3), 615-633.

Cheng, Z.L., Zhou, W.H., Ding, Z., Guo, Y.X., 2020a. Estimation of spatiotemporal response of rooted soil using a machine learning approach. J. Zhejiang Uni.-SCI. A (Appl. Phys. Eng.), 21 (6), 462-477.

Cheng, Z.L., Zhou, W.H., Garg, A., 2020b. Genetic programming model for estimating soil suction in shallow soil layers in the vicinity of a tree. Eng. Geol. 268, 105506.

Cuceoglu, F., 2016. An experimental study on soil water characteristics and hydraulic conductivity of compacted soils. MSc dissertation, Virginia Tech.

Cui, Y.J., Yahia-Aissa, M., Delage, P., 2002. A model for the volume change behavior of heavily compacted swelling clays. Eng. Geol. 64 (2-3), 233-250.

de Freitas, J.B., de Rezende, L.R., de FN Gitirana Jr, G., 2020. Prediction of the resilient modulus of two tropical subgrade soils considering unsaturated conditions. Eng. Geol. 105580.

Delage, P., Howat, M.D., Cui, Y.J., 1998. The relationship between suction and swelling properties in a heavily compacted unsaturated clay. Eng. Geol. 50 (1-2), 31-48.

Fattah, M.Y., Salim, N.M., Irshayyid, E.J., 2017. Determination of the soil–water characteristic curve of unsaturated bentonite–sand mixtures. Environ. Earth Sci. 76 (5), 201.

Fredlund, D.G., Rahardjo, H., Fredlund, M.D., 2012. Unsaturated soil mechanics in engineering practice. John Wiley & Sons, Inc.

Fredlund, D.G., Xing, A., 1994. Equations for the soil-water characteristic curve, Can. Geotech. J. 31 (3), 521–532.

Fredlund, M.D., Wilson, G.W., Fredlund, D.G., 2002. Use of the grain-size distribution for estimation of the soil-water characteristic curve. Can. Geotech. J. 39 (5), 1103-1117.

Gallage, C.P.K., Uchimura, T., 2010. Effects of dry density and grain size distribution on soil-water characteristic curves of sandy soils. Soils. Found. 50 (1), 161–172.

Giustolisi, O., Savic, D.A., 2006. A symbolic data-driven technique based on evolutionary polynomial regression. J. Hydroinform. 8 (4), 235-237.

Han, K.K., Rahardjo, H., and Broms, B.B. 1995. Effect of hysteresis on the shear strength of a residual soil. In: Proceedings of the First International Conference on Unsaturated soil (UNSAT 95), Paris, France, pp. 6–8.

Han, Z., Vanapalli, S.K., 2016. Relationship between resilient modulus and suction for compacted subgrade soils. Eng. Geol. 211, 85-97.

Hashem, E.B., Houston, S.L., 2016. Volume change consideration in determining unsaturated soil properties for geotechnical applications. Int. J. Geomech. 16 (6), D4015003.

He, Y., Ye, W.M., Chen, Y.G., Cui, Y.J., 2019. Effects of K+ solutions on swelling behavior of compacted GMZ bentonite. Eng. Geol. 249, 241-248.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE T. Pattern Anal., 20 (8), 832-844.

Huang, S., Barbour, S.L., Fredlund, D.G., 1998. Development and verification of a coefficient of permeability function for a deformable unsaturated soil. Can. Geotech. J. 35 (3), 411-425.

Indrawan, I.G.B., Rahardjo, H., Leong, E.C., 2006. Effects of coarse-grained materials on properties of residual soil. Eng. Geol. 82 (3), 154-164.

Jiang, Y., Chen, W., Wang, G., Sun, G., Zhang, F., 2017. Influence of initial dry density and water content on the soil–water characteristic curve and suction stress of a reconstituted loess soil. B. Eng. Geol. Environ. 76 (3), 1085-1095.

Jin, Y.F., Yin, Z.Y., 2020. An intelligent multi-objective EPR technique with multi-step model selection for correlations of soil properties. Acta Geotech. 15, 2053–2073.

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. Eng. Geol. 248, 242-255.

Johari, A., Habibagahi, G., Ghahramani, A., 2006. Prediction of soil–water characteristic curve using genetic programming. J. Geotech. Geoenviron. Eng. 132 (5), 661-665.

Khalili, N., Geiser, F., Blight, G.E., 2004. Effective stress in unsaturated soils: Review with new evidence. Int. J. Geomech. 4 (2), 115-126.

Krisdani, H., Rahardjo, H., Leong, E.C., 2008. Effects of different drying rates on shrinkage characteristics of a residual soil and soil mixtures. Eng. Geol. 102 (1-2), 31-37.

Li, X., 2009. Dual-porosity structure and bimodal hydraulic property functions of coarse granular soils. Ph. D. Thesis, Hong Kong University of Science and Technology, Hong Kong.

Li, X., Li, J. H., Zhang, L.M., 2014. Predicting bimodal soil–water characteristic curves and permeability functions using physically based parameters. Comput. Geotech. 57, 85-96.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News, 23 (23), 18-21.

Lin, B., Cerato, A.B., 2013. Hysteretic soil water characteristics and cyclic swell–shrink paths of compacted expansive soils. B. Eng. Geol. Environ. 72 (1), 61-70.

Liu, K., Yin, J., Chen, W., Feng, W.Q., Zhou, C., 2020. The stress–strain behaviour and critical state parameters of an unsaturated granular fill material under different suctions. Acta Geotech. in press. doi: 10.1007/s11440-020-00973-1.

Miller, C.J., Yesiller, N., Yaldo, K., Merayyan, S., 2002. Impact of soil type and compaction conditions on soil water characteristic. J. Geotech. Geoenviron. Eng. 128 (9), 733-742.

Mirzaii, A., Yasrobi, S.S., 2012. Influence of initial dry density on soil-water characteristics of two compacted soils. Geotechnique Let. 2 (4), 193-198.

Montanez, J.E.C., 2002. Suction and volume changes of compacted sand-bentonite mixtures. PhD dissertation, University of London.

Njock, P.G.A., Shen, S.L., Zhou, A., Lyu, H.M., 2020. Evaluation of soil liquefaction using AI technology incorporating a coupled ENN/t-SNE model. Soil Dyn. Earthq. Eng. 130, 105988.

Ng, C.W.W., Pang, Y.W., 2000. Experimental investigations of the soil-water characteristics of a volcanic soil. Can. Geotech. J. 37 (6), 1252-1264.

Oh, S., Lu, N., Kim, Y.K., Lee, S.J., Lee, S.R., 2012. Relationship between the soil-water characteristic curve and the suction stress characteristic curve: Experimental evidence from residual soils. J. Geotech. Geoenviron. Eng. 138 (1), 47-57.

Oltean, M., 2004. Multi expression programming source code. Available at: https://www.mepx.org/source_code.html

Oltean, M., Dumitrescu, D., 2002. Multi expression programming. Technical Report, UBB-01-2002, Babes-Bolyai University, Cluj-Napoca.

Oltean, M., Grosan, C., 2003. A comparison of several linear genetic programming techniques. Complex Syst. 14 (4), 285-314.

Osinubi, K.J., Nwaiwu, C.M., 2006. Design of compacted lateritic soil liners and covers. J. Geotech. Geoenviron. Eng. 132 (2), 203-213.

Priono, Rahardjo, H., Chatterjea, K., Leong, E.C., Wang, J.Y., 2016. Effect of hydraulic anisotropy on soil–water characteristic curve. Soils Found. 56 (2), 228-239.

Puppala, A.J., Punthutaecha, K., Vanapalli, S.K., 2006. Soil-water characteristic curves of stabilized expansive soils. J. Geotech. Geoenviron. Eng. 132 (6), 736-751.

Rahardjo, H., Indrawan, I.G.B., Leong, E.C., Yong, W.K., 2008. Effects of coarse-grained material on hydraulic properties and shear strength of top soil. Eng. Geol. 101 (3-4), 165-173.

Rahardjo, H., Satyanaga, A., D'Amore, G.A., Leong, E.C., 2012. Soil–water characteristic curves of gap-graded soils. Eng. Geol. 125, 102-107.

Salager, S., Nuth, M., Ferrari, A., Laloui, L., 2013. Investigation into water retention behaviour of deformable soils. Can. Geotech. J. 50 (2), 200-208.

Sarir, P., Shen, S.L., Wang, Z.F., Chen, J., Horpibulsuk, S., Pham, B.T., 2019. Optimum model for bearing capacity of concrete-steel columns with AI technology via incorporating the algorithms of IWO and ABC. Eng. Comput. doi: 10.1007/s00366-019-00855-5

Satyanaga, A., Rahardjo, H., Zhai, Q., 2017. Estimation of unimodal water characteristic curve for gap-graded soil. Soils Found. 57 (5), 789-801.

Shahnazari, H., Dehnavi, Y., Alavi, A.H., 2010. Numerical modeling of stress–strain behavior of sand under cyclic loading. Eng. Geol. 116 (1-2), 53-72.

Shen, S., Njock, P.G.A., Zhou, A., Lyu, HM., 2020. Dynamic prediction of jet grouted column diameter in soft soil using Bi-LSTM deep learning. Acta Geotech. in press. doi: 10.1007/s11440-020-01005-8

Sun, D.A., Cui, H., Sun, W., 2009. Swelling of compacted sand–bentonite mixtures. Appl. Clay Sci. 43 (3-4), 485-492.

Sun, D.A., Gao, Y., 2015. Water retention behaviour of soils with different preparations. Chinese J. Geotech. Eng. 37 (1), 91-97 (in Chinese).

Sun, D.A., Gao, Y., Zhou, A.N., Sheng, D.C., 2016. Soil-water retention curves and microstructures of undisturbed and compacted Guilin lateritic clay. Bull. Eng. Geol. Environ. 75 (2), 781–791.

Sun, D.A., Liu, W.J., Lü, H.B., 2014. Soil-water characteristic curve of Guilin lateritic clay. Rock Soil Mech. 35 (12), 3345-3351 (in Chinese).

Sun, D.A., Sheng, D.C., Cui, H.B., Li, J., 2006. Effect of density on the soil-water-retention behaviour of compacted soil. In Proceedings of the Fourth International Conference on Unsaturated Soils, Carefree, Arizona, US, pp. 1338-1347.

Thu, T.M., Rahardjo, H., and Leong, E.C. 2006. Shear strength and pore-water pressure characteristics during constant water content triaxial tests. J. Geotech. Geoenviron. Eng. 132(3): 411–419.

Thu, T.M., Rahardjo, H., Leong, E.C., 2007. Soil-water characteristic curve and consolidation behavior for a compacted silt. Can. Geotech. J. 44 (3), 266-275.

Tinjum, J.M., Benson, C.H., Blotz, L.R., 1997. Soil-water characteristic curves for compacted clays. J. Geotech. Geoenviron. Eng. 123 (11), 1060-1069.

van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, J. Soil Sci. Soc. America 44, 892–898.

Vanapalli, S.K., Fredlund, D.G., Pufahl, D.E., 1999. The influence of soil structure and stress history on the soil–water characteristics of a compacted till. Geotechnique 49 (2), 143-159.

Wang, H.L., Cui, Y.J., Lamas-Lopez, F., Dupla, J.C., Canou, J., Calon, N., Saussine, G., Aimedieu, P., Chen, R.P., 2017. Effects of inclusion contents on resilient modulus and damping ratio of unsaturated track-bed materials. Can. Geotech. J. 54 (12), 1672-1681.

Wang, H.L., Cui, Y.J., Lamas-Lopez, F., Dupla, J.C., Canou, J., Calon, N., Saussine, G., Aimedieu, P., Chen, R. P., 2018. Permanent deformation of track-bed materials at various inclusion contents under large number of loading cycles. J. Geotech. Geoenviron. Eng. 144 (8), 04018044.

Wang, H.L., Chen, R.P., 2019. Estimating static and dynamic stresses in geosynthetic-reinforced pile-supported track-bed under train moving loads. J. Geotech. Geoenviron. Eng. 145 (7), 04019029.

Wang, H.L., Chen, R.P., Cheng, W., Qi, S., Cui, Y.J., 2019a. Full-scale model study on variations of soil stress in the geosynthetic-reinforced pile-supported track-bed with water level change and cyclic loading. Can. Geotech. J. 56 (1), 60–68.

Wang, H.L., Chen, R.P., Liu, Q.W., Kang, X., 2019b. Investigation on geogrid reinforcement and pile efficacy in geosynthetic-reinforced pile-supported track-bed. Geotext. Geomembranes 47 (6), 755-766.

Wang, H.L., Yin, Z.Y., 2020. High performance prediction of soil compaction parameters using multi expression programming. Eng. Geol. 276, 105758.

Yang, H., Rahardjo, H., Leong, E. C., Fredlund, D.G., 2004. Factors affecting drying and wetting soil-water characteristic curves of sandy soils. Can. Geotech. J. 41 (5), 908-920.

Ye, W.M., Cui, Y.J., Qian, L.X., Chen, B., 2009. An experimental study of the water transfer through confined compacted GMZ bentonite. Eng. Geol. 108 (3-4), 169-176.

Zhai, Q., Rahardjo, H., Satyanaga, A., Dai, G., 2020. Estimation of the soil-water characteristic curve from the grain size distribution of coarse-grained soils. Eng. Geol. 267, 105502.

Zhang, F., Chen, X., 2009. Experimental study on characteristics of deformation and strength of unsaturated clay. Chinese J. Rock Mech. Eng. 28 (2), 3808-3814 (in Chinese).

Zhang, P., Chen, R.P., Wu, H.N., 2019. Real-time analysis and regulation of EPB shield steering using Random Forest. Automat. Constr. 106, 102860.

Zhang, P., Jin, Y.-F., Yin, Z.-Y., Yang, Y., 2020a. Random forest based artificial intelligent model for predicting failure envelopes of caisson foundations in sand. Appl. Ocean Res. 101, 102223.

Zhang, P., Wu, H.-N., Chen, R.-P., Chan, T.H.T., 2020b. Hybrid meta-heuristic and machine learning algorithms for tunneling-induced settlement prediction: A comparative study. Tunnell. Undergr. Space Technol. 99, 103383.

Zhang, P., Yin, Z.-Y., Jin, Y.-F., Chan, T.H.T., 2020c. A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest. Eng. Geol. 265, 105328.

Zhang, P., Yin, Z.Y., Jin, Y.F., Chan, T., Gao, F.P., 2020d. Intelligent modelling of clay compressibility using hybrid meta-heuristic and machine learning algorithms. Geosci. Front. in press. doi: 10.1016/j.gsf.2020.1002.1014.

Zhao, L.S., Zhou, W.H., Fatahi, B., Li, X.B., Yuen, K.V., 2016. A dual beam model for geosynthetic-reinforced granular fill on an elastic foundation. Appl. Math. Model. 40 (21-22), 9254-9268.

Zhou, B.C., Kong, L.W., 2011. Effect of volume changes on soil-water characteristics of unsaturated expansive soil. J. Hydr. Eng. 42 (10), 1152-1160 (in Chinese).

Zhou, W.H., Xu, X., Garg, A. 2016. Measurement of unsaturated shear strength parameters of silty sand and its correlation with unconfined compressive strength. Measurement 93, 351-358.

Zhou W.H., Yuen K.V., Tan F., 2014. Estimation of soil-water characteristic curve for granular soils with different initial dry densities. Eng. Geol. 179, 1–9.

Table 1. Descriptive statistics of each variable

Variable	Minimum	Maximum	Mean	Standard deviation
C_G (%)	0	86.5	3.8	13.41
C_S (%)	0	100	40.3	29.93
C_F (%)	0	100	55.9	30.38
PI (%)	0	88	18.9	16.45
w_0 (%)	0.5	48.6	18.0	9.55
e_0	0.24	1.55	0.73	0.28
AEV (kPa)	0.06	100	15.37	17.82

Table 2. Air-entry value of different soils in the database

Soil	Minimum (kPa)	Maximum (kPa)	Mean (kPa)	Median (kPa)
Clay	2.5	100	18.33	14.63
Silt	0.43	73	22.62	12
Sand	0.2	71.16	10.44	4.07
Gravel	0.06	14.65	3.59	0.77

Table 3. Parameter setting for determination of the optimum combination

Algorithm	Parameter	Setting
All algorithms	Terminal set	C_S, C_F, PI, w_0, e_0
MEP	Population size	1000, 2000, 3000
	Number of generation	1000, 2000, 3000
	Crossover probability	0.1, 0.5, 0.9
	Crossover type	Uniform
	Mutation probability	0.01, 0.1, 0.9
	Code length	50, 100
	Function set	+, -, ×, /, pow
	Replication number	10
EPR	Number of terms	2-10
	Values of elements	[-3, 3]
	Interval of elements	0.5
RF	Number of trees	1-500
	Number of features	1-5
GA for EPR and RF	Population size	20
	Number of generation	500
	Crossover probability	0.7
	Mutation probability	0.1

Table 4. Basic setting of input soil properties for the monotonicity analysis

Soil parameter	Value
C_G (%)	3.8
C_S (%)	40.3
C_F (%)	55.9
PI (%)	18.9
w_0 (%)	18.0
e_0	0.73

List of Figures

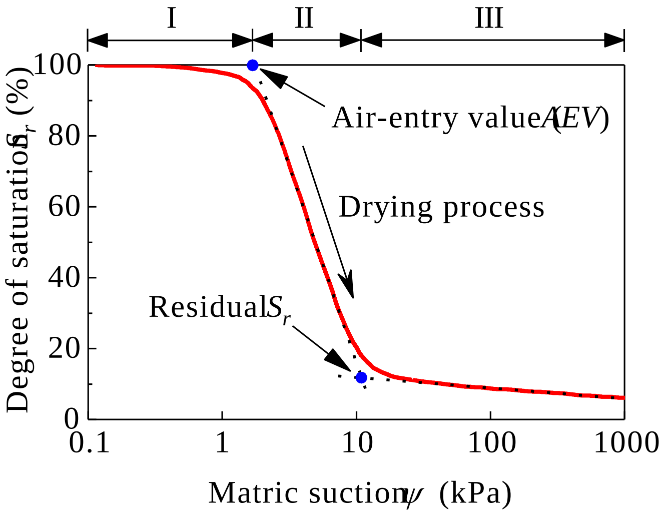


Fig. 1. Conceptual soil-water characteristic curve (I: boundary effect zone; II: transition zone; III: residual zone)

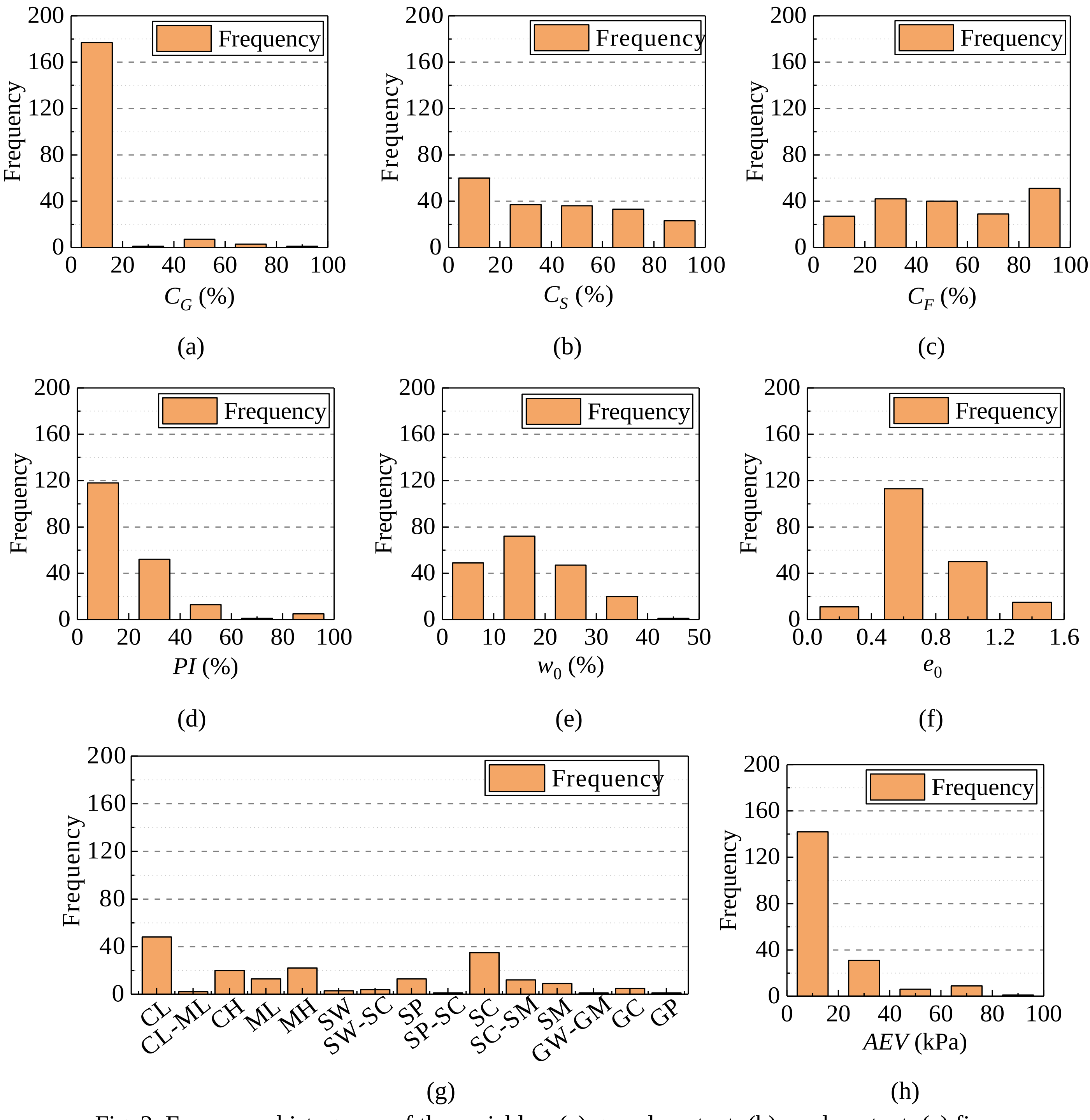


Fig. 2. Frequency histograms of the variables: (a) gravel content; (b) sand content; (c) fines content; (d) plasticity index; (e) initial water content; (f) initial void ratio; (g) soil classification; (h) air-entry value

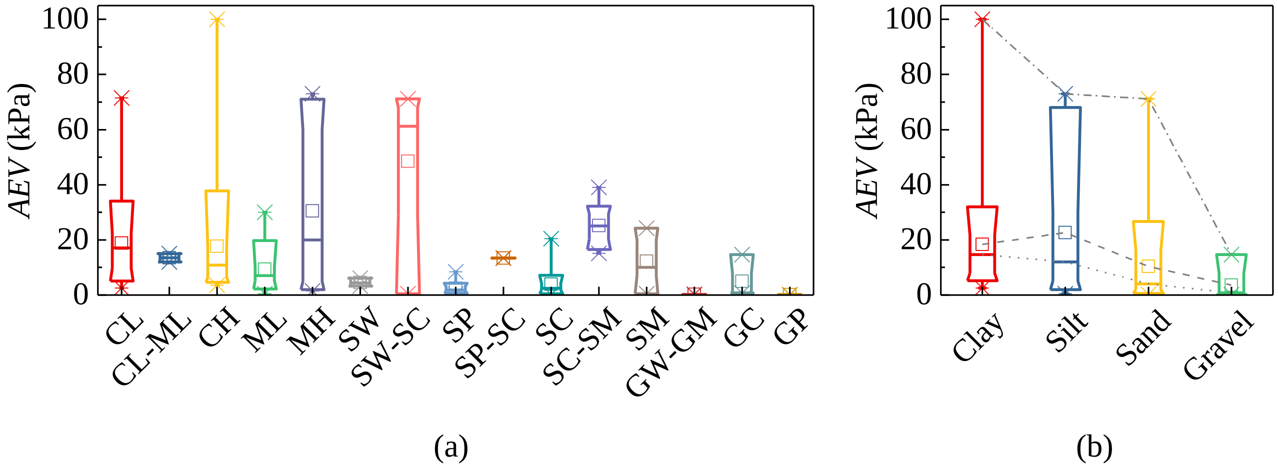


Fig. 3. Comparison about air-entry value of soils with different classifications in the database: (a) specific classification; (b) general classification

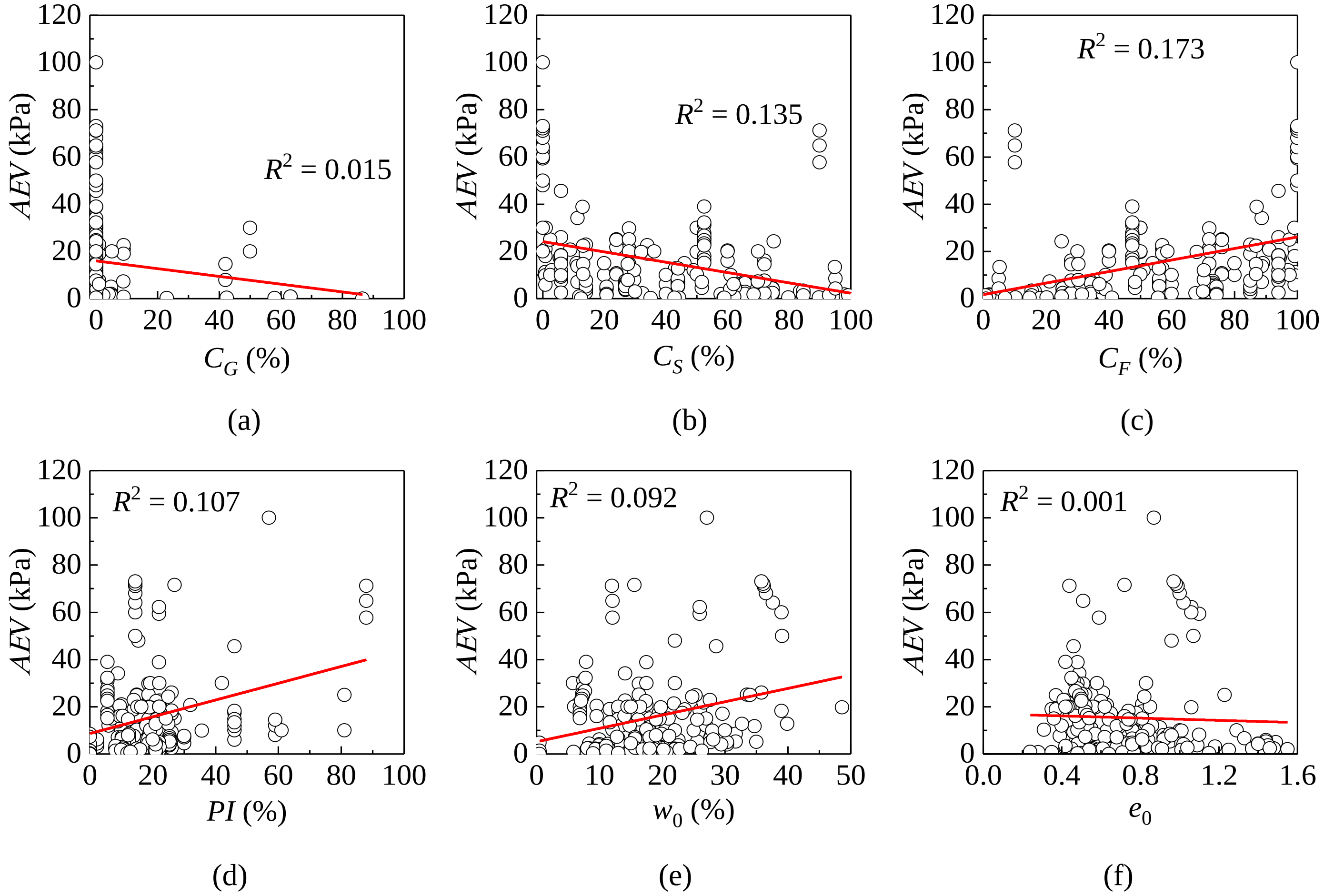
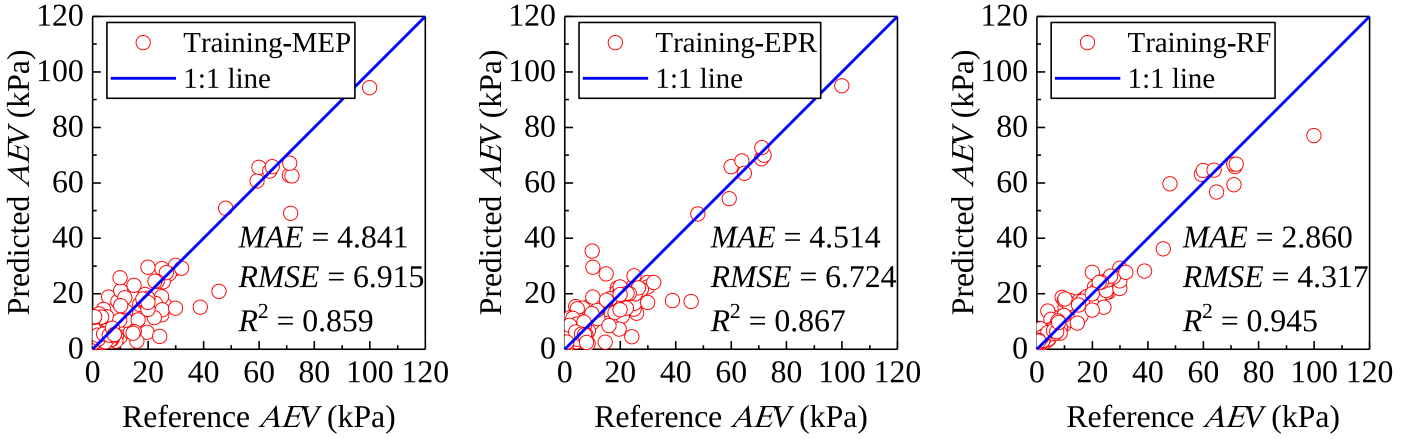
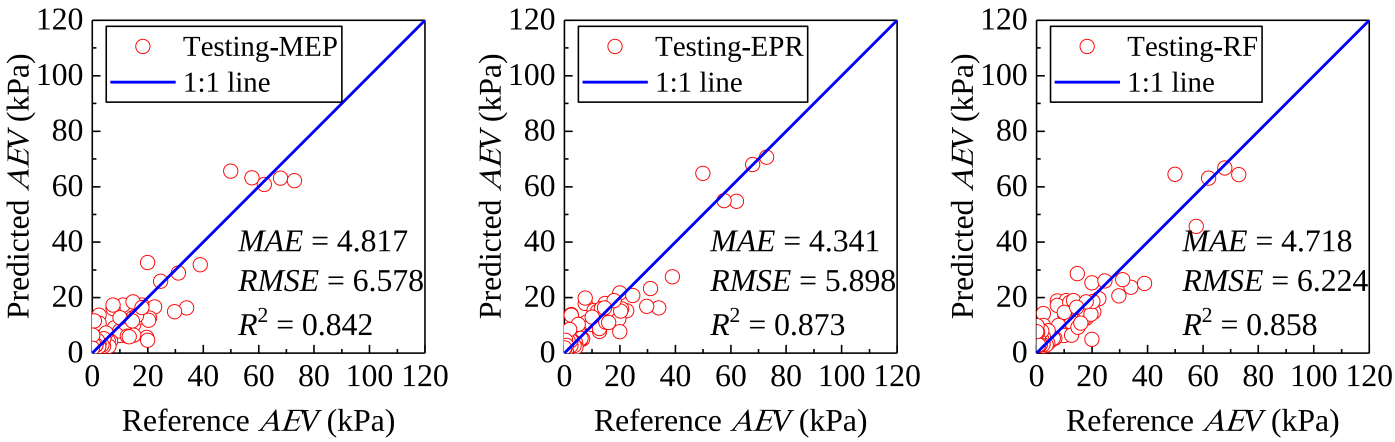


Fig. 4. Basic linear fittings between air-entry value and each input soil property: (a) gravel content; (b) sand content; (c) fines content; (d) plasticity index; (e) initial water content; (f) initial void ratio



(a) (b) (c)

Fig. 5. Comparison between predicted and reference air-entry values for the training data: (a) MEP; (b) EPR; (c) RF



(a) (b) (c)

Fig. 6. Comparison between predicted and reference air-entry values for the testing data: (a) MEP; (b) EPR; (c) RF

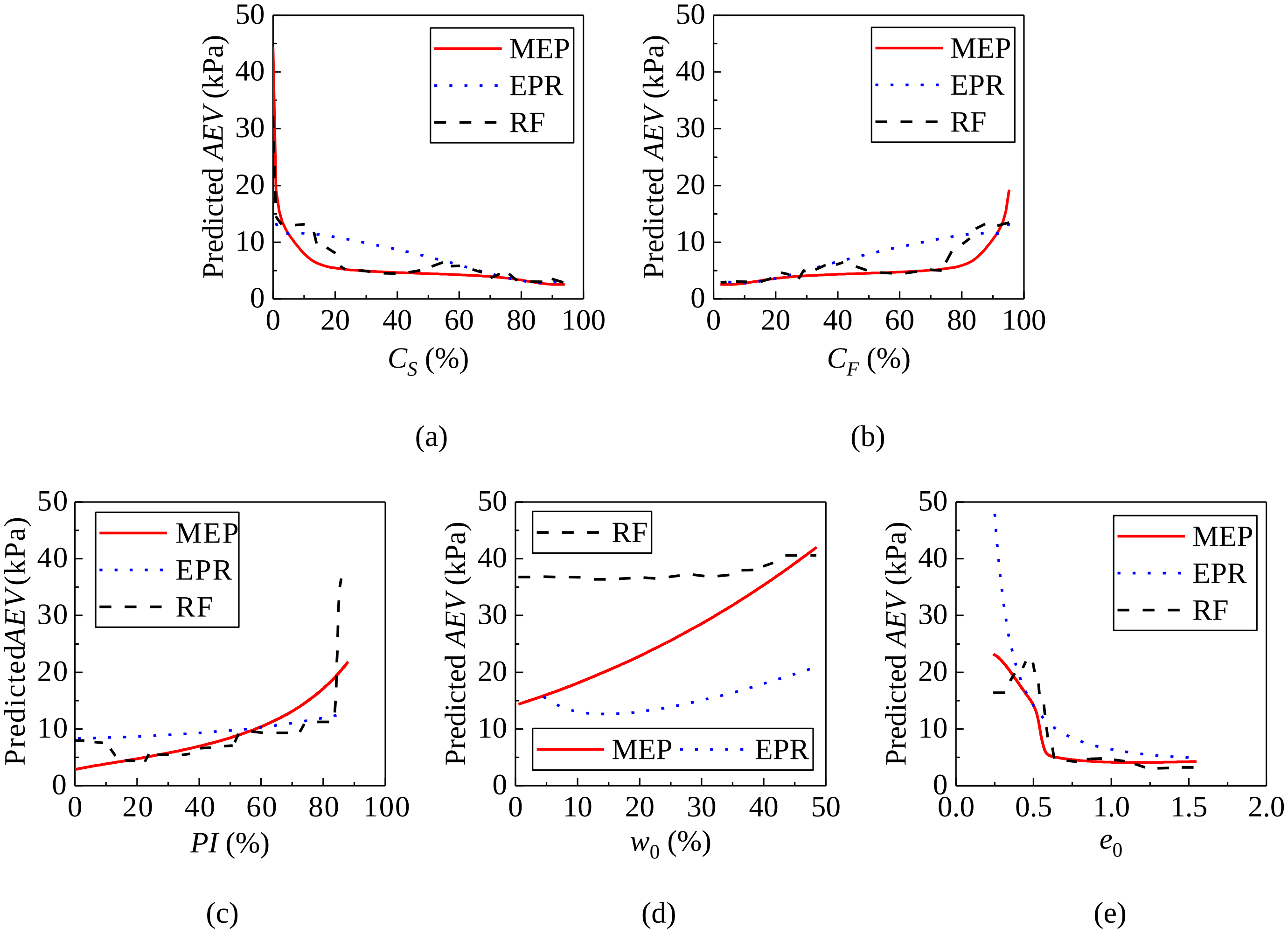


Fig. 7. Monotonicity analysis of the predicted air-entry value versus (a) sand content; (b) fines content; (c) plasticity index; (d) initial water content; (e) initial void ratio

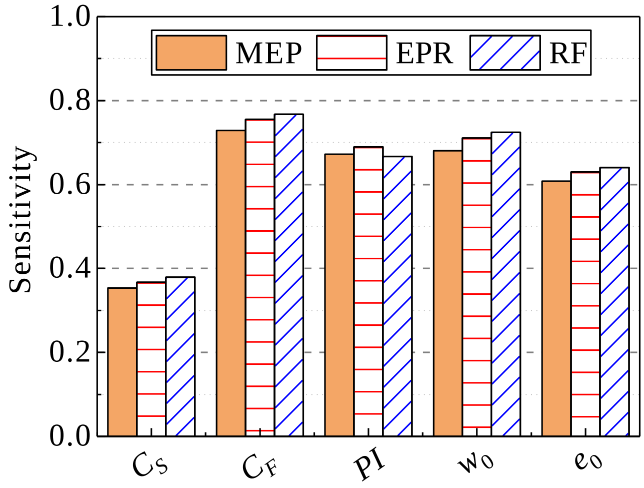


Fig. 8. Sensitivity analysis about the relevance of the input variables on the predicted air-entry value

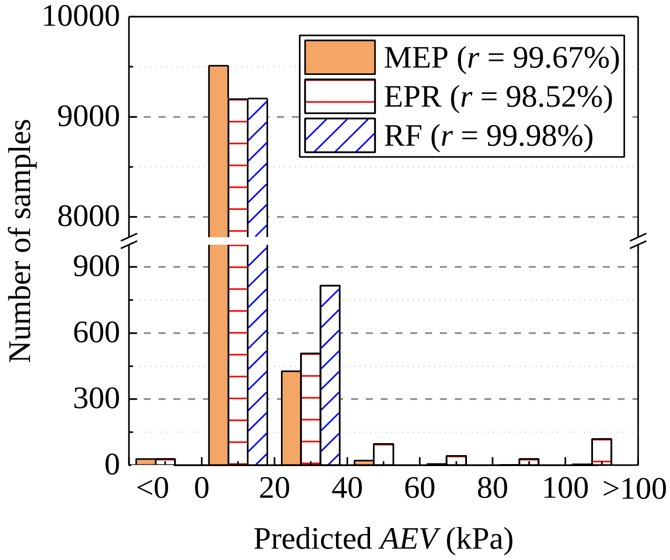


Fig. 9. Distribution of the predicted air-entry value in robustness analysis