Probabilistic analysis of soil-water characteristic curve based on machine learning algorithms

S Yang¹, P Q Zheng², Y T Yu³ and J Zhang¹*

1. Department of Geotechnical Engineering, Tongji University, Shanghai 200092, China.
2. Key Laboratory of Geotechnical and Underground Engineering of Ministry of Education and Department of Geotechnical Engineering, Tongji University, Shanghai 200092, China.
3. Senior Engineer, Shaanxi Key Laboratory of Behaviour and Treatment for Special Rocks and Soils, China JK Institute of Engineering Investigation and Design Co., Ltd., Xi’an, Shaanxi 710043, China

*E-mail: cezhangjie@tongji.edu.cn

Abstract: The soil-water characteristic curve (SWCC) is the constitutive relationship curve of soil suction and water content, which has important engineering significance for studying the soil strength, permeability coefficient and volume change of unsaturated soils. Considering the experimental measurement of the SWCC is time-consuming, many empirical methods have been suggested to estimate the SWCC. This paper proposed a machine learning algorithm to predict SWCC from limited sets of soil properties. By predicting the parameters of the Fredlund and Xing model (which is called FX model), the most probable SWCC can be obtained. Since SWCC can’t be accurately determined, the residual probability distribution functions of parameters of FX model are also derived in this paper. Finally, this paper predicted the 90% confidence interval of the SWCC. The result shows that the machine learning algorithm has high accuracy and generalization ability. The suggested method provides a practical means to estimate the SWCC and the variability of it, so that the error associated with the prediction model can be explicitly considered.

1. Introduction

The soil-water characteristic curve (SWCC) defines the relationship equation between water and soil suction in the soil. The water in the soil can be expressed as weight water content \( w \), volume water content \( \theta \) and saturation \( S \). The soil-water characteristic curve contains basic information for
estimating the properties of unsaturated soils, such as the estimation of shear stress and permeability coefficient[1-3]. Therefore, it is very necessary to estimate the water and soil characteristic curve correctly to predict the characteristics of unsaturated soils.

As the measurement of the SWCC is time-consuming and labour intensive, many empirical methods have been suggested to estimate the SWCC, such as such as the point estimate method [4], the parametric estimation method [5-6], and machine learning algorithms [7-9]. With the rapid development of big data, machine learning algorithms based on probability and statistics has received great attention from industry and academia in recent years[10]. Machine learning algorithms is a common research hotspot in the field of artificial intelligence and pattern recognition. Its theories and methods have been widely used to solve complex problems in engineering applications and science.

The purpose of this study is thus to suggest machine learning algorithms to estimate the SWCC considering both the soil particle size distribution and dry density of the soil. This paper is organized as follows. First, the soil database is constructed to extract a large enough amount of soil samples which are divided into training set and test set, and the soil-water characteristic curve is fitted based on the Fredlund and Xing model[1]. Then, the machine learning algorithms is used for estimating the SWCC using hybrid methods. Finally, the effectiveness of the suggested method is verified. The 90% confidence interval of SWCC is given by considering the uncertainty of parameters.

2. Establishment of soil database

2.1. SWCC model

SWCC is used to describe the relationship between the degree of saturation and the matric suction. In geotechnical engineering, the model suggested by Fredlund and Xing [1] is widely used, which is called the FX model in this paper. In this paper, the FX model will be selected as an example, in which the degree of saturation is related to the suction as follows:

\[
S = C(\psi)/\left\{ \ln \left[ e^c + \left( \frac{\psi}{a} \right)^c \right] \right\}^m \tag{1}
\]

\[
C(\psi) = 1 - \ln \left( 1 + \frac{\psi}{\psi_r} \right) / \ln \left( 1 + \frac{10^6}{\psi_r} \right) \tag{2}
\]

Where \(S\) is the degree of saturation, and \(\psi\) is the suction. \(\psi_r\) is the suction corresponding to the residual water content, which can be assumed to be 3000 kPa [1]. \(C(\psi)\) is the correction function to ensure that the degree of saturation is 0 when the suction reaches \(10^6\) kPa.

2.2. Parameters collection

According to the analysis in the previous paragraph, \(a, n, \) and \(m\) are parameters of the FX model to be calibrated. This article derived the original soil sample data from the SoilVision database [11] which contains various experimental data of soil. A total of 520 sets of soil samples were collected, of which 300 were used as training samples and 220 were used as verification samples. Figure 1 shows the distribution of the 520 soil samples in the USDA textural triangle[12].

According to the measured SWCC, the parameters of the FX model are fitted based on the least square method. Take sand, clay and coarse loam soil as examples, and then draw the soil-water
characteristic curve intervals of the three soil types, as shown in figure 2. It can be seen from figure 2 that the range of different soils is different, which is mainly due to the different micro-pore structure and particle composition. Compared with coarse soil, the fine grain soils have a stronger water-holding capacity. Therefore, under the same suction, the saturation of clay is higher than that of sand.

![Figure 1. Distribution of the 520 soil samples on the USDA textural triangle](image)

![Figure 2. Range of SWCC for different soil samples](image)

The morphology of SWCC is affected by many factors. Studies have found that the factors that affect SWCC include soil mineral composition, soil particle size and gradation, initial dry density, stress history, etc. These factors affect the matrix of unsaturated soils. The influence of suction is very complicated, and these factors also affect the model parameters in the SWCC empirical model. Many scholars such as Mohammadi[13] used the particle size curve to predict SWCC and achieved good prediction results. Arya[14] searched for the relationship between particle size distribution curve, soil bulk density and SWCC, and found that soil bulk density has a certain correction effect on SWCC prediction, which is more accurate than the result of single use of particle size distribution curve prediction. Li et al.[15] explored the relationship between SWCC morphology, saturation and dry density of weak expansive soil. The experiment showed that when the saturation is constant, the total suction and matrix suction both increase with the rise of dry density;

Based on the previous research results, this paper used soil physical parameters to predict SWCC, including particle size distribution parameters (such as $d_{10}$, $d_{50}$, $d_{60}$, the coefficient of uniformity $C_u$, the coefficient of curvature $C_c$), void ratio $e$, dry density $\rho$. The above parameters are also included in the established soil database.

3. Machine learning algorithms for estimation of SWCC

3.1. Model selection

Machine Learning algorithms usually solve three types of problems: regression, clustering, and classification. There are mainly two types of algorithms: single learning algorithm and integrated learning algorithm. There are three typical single learning algorithms: Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbour (KNN); Typical integrated learning algorithms mainly include: Bagging, AdaBoost, Gradient Boosting (GB), Random Forest (RF).
In the following calculations, the above machine learning algorithms are selected for comparison. Accuracy of machine learning algorithms was evaluated by calculating the coefficient of determination \( R^2 \), the root mean square error \( \text{RMSE} \) and mean absolute error \( \text{MAE} \) according to the following equations:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_m - y_p)^2}{\sum_{i=1}^{n} (y_m - \overline{y_m})^2}
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_m - y_p)^2}{n}}
\]

\[
\text{MAE} = \frac{\sum_{i=1}^{n} |y_m - y_p|}{n}
\]

The correlation coefficients among different variables are improved after the logarithm transformation. Therefore, in this article, we will take soil sample particle size parameters (\( \ln d_{10}, \ln d_{30}, \ln d_{60}, \ln C_u, \ln C_c \)) and soil physical parameters (\( \ln e, \ln \rho \)) as input, and parameter of FX model (\( \ln a, \ln n \) and \( \ln m \)) as output. For the input data, there is a correlation between particle size distribution parameters, such as \( d_{10}, d_{30}, d_{60} \), they reflect the particle size corresponding to different percentages, and there is a correlation between them. For example, the coefficient of correlation (COC) between \( \ln d_{10} \) and \( \ln d_{30} \) is 0.77. In addition, the coefficient of uniformity \( C_u \), the coefficient of curvature \( C_c \) is calculated by \( d_{10}, d_{30} \) and \( d_{60} \). Therefore, there is also a correlation between each other. The COC between \( \ln C_u \) and \( \ln C_c \) is 0.37. In addition, the dry density and void ratio have a weak correlation with the above indicators. In short, they have different meanings in the soil, which represent different characteristics of the soil. They are related to the soil-water characteristic curve. Therefore, many scholars used these parameters to predict SWCC predict SWCC and achieved good prediction results [13-14].

To identify the most effective parameters on SWCC, the coefficient of correlation between \( \ln a, \ln n, \ln m \) and \( \ln d_{10}, \ln d_{30}, \ln d_{60}, \ln C_u, \ln C_c, \ln e, \ln \rho \) are calculated (Table 1). COC is also called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The COC between \( \ln a \) and \( \ln d_{10} \) indicates that there is a stronger correlation between the two parameters between the two variables, which means \( \ln d_{10} \) is the most effective parameters on the parameters (\( \ln a \)) of SWCC. If we want to establish a simple empirical relationship, we can choose more relevant parameters to predict the parameters of SWCC. But for machine learning algorithms, in order to expand the amount of information, all relevant parameters are used as input data will get better prediction results, which will be verified later. For example, compared to selecting only \( \ln d_{10}, \ln d_{30}, \ln d_{60}, \ln C_u \) as the input data, the \( R^2 \) of \( \ln n \) that selects all soil physical parameters as input data is increased from 0.32 to 0.35.

Table 1. Coefficients of correlation between characteristic parameters of SWCC and soil physical parameters.

| \( \ln d_{10} \) | \( \ln d_{30} \) | \( \ln d_{60} \) | \( \ln C_u \) | \( \ln C_c \) | \( \ln e \) | \( \ln \rho \) |
|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|                  |                  |                  |                  |                  |                  |                  |

\[ \text{Table 1. Coefficients of correlation between characteristic parameters of SWCC and soil physical parameters.} \]
Previously, Schaap et al. [16] developed prediction equations for the parameters of the van Genuchten model [17] through linear regression. The $R^2$ between the predictions and actual values for the two parameters of van Genuchten model are 0.20 and 0.45, respectively. The correlation coefficient reported in this paper is similar to Schaap et al. [16]. The results presented in this study and those reported in Schaap et al. [16] imply that the SWCC may also be affected by many other factors and cannot be accurately predicted. Therefore, it is necessary to consider the uncertainty of SWCC forecasts. On the other hand, there are $R^2$ less than zero, which means that the prediction error of the fitting function is greater than the prediction error of the function that the predicted value is equal to the average value. This is mainly due to the over-fitting of the machine learning algorithm, this is mainly due to the over-fitting of the machine learning algorithm. For example, in the prediction of $\ln a$ by the DT method, the $R^2$ in the training set is 0.89, while the $R^2$ in the test set is -10.

Seven machine learning algorithms are used to predict $\ln a$, $\ln n$ and $\ln m$. And then the prediction effect is calibration on the test data. In order to select the optimal machine learning algorithm, a 10-fold cross-validation is used to preliminarily evaluate the effect of the algorithm on the test data before prediction. The average absolute error (MAE) of 10 times of 10-fold cross-validation are used. The average value of the coefficient of determination ($R^2$) is used as the evaluation criterion. According to principle of exhaustive method, the GridSearchCV function in Python is used to find the optimal model hyperparameter combinations on the training set, with $R^2$ as the evaluation index. In order to obtain better prediction accuracy, $\ln a$, $\ln n$, and $\ln m$ have different hyperparameter combinations.

Table 2 shows the cross-validation scores of 7 machine algorithms. The coefficient of determination $R^2$ for DTR, AdaBoost and Bagging less than 0 can indicate that the regression model residual sum of squares is greater than the average model residual sum of squares, and the fitting effect is very poor. The MAE and MSE of RF are almost the smallest and $R^2$ is the largest, but the $R^2$ of RF for $\ln a$ is only 0.24 which shows that the predicted results and the actual results still have a large discreteness. In that case, the parameter combination of RF are: {criterion: mae, n_estimator:100} for $\ln a$, {criterion: mse, n_estimator:120} for $\ln n$, {criterion: mse, n_estimator:55} for $\ln m$.

|        | DT  | SVM | KNN | AdaBoost | RF   | GB   | Bagging |
|--------|-----|-----|-----|----------|------|------|---------|
| MAE    | 2.24| 1.04| 1.06| 1.16     | 1.05 | 1.10 | 1.03    |
| $\ln a$ RMSE | 2.74| 1.55| 1.55| 1.60     | 1.54 | 1.58 | 1.52    |
| $R^2$  | -10.00| 0.21| 0.22| 0.16     | 0.24 | 0.19 | 0.24    |
| MAE    | 0.73| 0.57| 0.55| 0.59     | 0.53 | 0.56 | 0.55    |
| $\ln n$ RMSE | 1.08| 0.79| 0.79| 0.81     | 0.77 | 0.78 | 0.78    |
| $R^2$  | -0.27| 0.31| 0.32| 0.27     | 0.35 | 0.33 | 0.33    |
| MAE    | 0.66| 0.51| 0.55| 0.61     | 0.55 | 0.59 | 0.55    |
| $\ln m$ RMSE | 1.06| 0.72| 0.82| 1.00     | 0.85 | 0.80 | 0.85    |
| $R^2$  | -0.64| 0.25| 0.01| -0.46    | 0.25 | 0.07 | -0.05   |
It can be seen from the table 2 that the prediction effect of RF is the best. The predicted scatter plot of ln $a$, ln $n$ and ln $m$ are shown in figure 3. We can see that ln $a$ cannot be accurately predicted, so we will consider the uncertainty of the prediction and estimate the probability distribution of SWCC.

**Figure 3.** Comparison between actual and predicted values: (a) ln $a$; (b) ln $n$; (c) ln $m$

### 3.2. Probability distribution prediction of SWCC

For ease of presentation, the parameters of FX model to be estimated in this model are represented by $\mathbf{\theta} = \{\ln a, \ln n, \ln m\}$ and the input parameters of RF represented by $\mathbf{x} = \{\ln d_{10}, \ln d_{30}, \ln d_{60}, \ln C_u, \ln C_c, \ln e, \ln \rho\}$. Let $\varepsilon$ present the error in parameter prediction, $G(\mathbf{x})$ represent the parameters of FX model predicted by RF. The real parameters of FX model can be related to the model prediction as follows:

$$
\mathbf{\theta}_i = G(\mathbf{x}) + \varepsilon_i
$$

where $\varepsilon$ is a normal random variable with a mean of 0 and a standard deviation of $\sigma_{\varepsilon}$ which considers the effect of model uncertainty. According to the results of RF method, Figure 4 shows the predicted residual histograms of $\mathbf{\theta}$ for 220 groups of test soil samples. The figure also shows the normal distribution curve obtained according to the histogram fitting. Based on the data in these figures, the mean values of $\varepsilon_1$, $\varepsilon_2$ and $\varepsilon_3$ are zero, and the standard deviations of $\varepsilon_1$, $\varepsilon_2$ and $\varepsilon_3$ are 1.54, 0.77 and 0.85 respectively, i.e., $\sigma_{\varepsilon_1} = 1.54$, $\sigma_{\varepsilon_2} = 0.77$ and $\sigma_{\varepsilon_3} = 0.85$. For simplicity, $\varepsilon_1$, $\varepsilon_2$ and $\varepsilon_3$ are assumed to be normally distributed.

**Figure 4.** Histogram of samples of different variables : (a) ln $a$, (b) ln $n$, (c) ln $m$.

Let $\mathbf{\mu}_\mathbf{\theta}$ represent the mean value of $\mathbf{\theta}$, that is, $\mathbf{C}_\theta$ represent the covariance of $\mathbf{\theta}$. $\mathbf{C}_\theta$ can be calculated as follows:
\[
C_0 = \begin{bmatrix}
\sigma_{e1}^2 & \rho_{12}\sigma_{e1}\sigma_{e2} & \rho_{13}\sigma_{e1}\sigma_{e3} \\
\rho_{21}\sigma_{e2}\sigma_{e1} & \sigma_{e2}^2 & \rho_{23}\sigma_{e2}\sigma_{e3} \\
\rho_{31}\sigma_{e3}\sigma_{e1} & \rho_{32}\sigma_{e3}\sigma_{e2} & \sigma_{e3}^2
\end{bmatrix} = \begin{bmatrix}
2.38 & -0.13 & 0.61 \\
-0.13 & 0.60 & -0.33 \\
0.61 & -0.33 & 0.71
\end{bmatrix}
\]

(7)

Where \( \rho_{ij} \) denotes the correlation coefficient between \( e_i \) and \( e_j \). Since \( \ln a, \ln n \) and \( \ln m \) obey the normal distribution respectively, \( \theta = \{\ln a, \ln n, \ln m\} \) obeys the multivariate normal distribution. Based on the characteristics of the multivariate normal distribution, the joint probability density function of \( \theta \) is:

\[
f(\theta) = (2\pi)^{-3/2} |C_0|^{-1/2} \exp\left[ -\frac{1}{2} (\theta - \mu_\theta)^T C_0^{-1} (\theta - \mu_\theta) \right]
\]

(8)

According to equation 8, the probability distribution of the FX model parameters can be predicted by the machine learning algorithms. Then we can predict the soil-water characteristic curve based on the probability distribution of the FX parameters.

4. Validation of the proposed method

In order to further verify the reliability of the method, the proposed method is used to predict the soil-water characteristic curve of the test set. The soil samples 10754, 10785, 10886 and 11211 from the SoilVision database[11] were selected, and the input data were collected, as shown in Table 3. The RF method is used to obtain the predicted value of the FX parameter, and based on equation 8, the probability distribution of the FX parameter is obtained.

**Table 3. Input data of machine learning algorithms for soil samples**

| Soil code | \( \ln d_{10} \) | \( \ln d_{50} \) | \( \ln d_{60} \) | \( \ln C_u \) | \( \ln C_c \) | \( \ln e \) | \( \ln \rho \) |
|-----------|----------------|----------------|----------------|-------------|-------------|-------|-------|
| 10754     | -11.15         | -6.32          | -1.82          | -9.33       | 0.33        | -0.19    | 7.37   |
| 10785     | -11.24         | -9.63          | -1.71          | -9.53       | -6.31       | -0.03    | 7.24   |
| 10886     | -4.52          | -2.24          | -1.52          | -3.00       | 1.56        | -0.40    | 7.38   |
| 11211     | -7.88          | -5.40          | -4.30          | -3.58       | 1.37        | -0.05    | 7.26   |
Figure 5. Comparison between measured data and predicted SWCCs: (a) soil 10754; (b) soil 10785; (c) soil 10886; (d) soil 11211

Figure 5 shows the actual measured soil-water characteristic curve of the corresponding soil sample. The most probable prediction curve and the 90% confidence interval are given in this paper. As can be seen from the figure, for different soil samples, the data points of SWCC are fitted well and all fall within the 90% confidence interval, indicating the good applicability and effectiveness of the machine learning algorithms.

5. Conclusion

This paper proposed a method to predict the probability of soil-water characteristic curves, compares and selected seven types of machine learning algorithms, and used RMSE, MAE and $R^2$ to select the RF method with the highest prediction accuracy. Through the prediction of the FX model parameters, considering the uncertainty in the prediction process, the joint probability density equation of the FX model parameters is established, and finally the most probable curve of the SWCC and the 90% confidence interval are obtained. The conclusions of this paper are summarized as follows:

- Among most of machine learning algorithms, the RF’s prediction effects on SWCC is the best, and the predicted values of the FX model parameters are obtained.
- Considering the uncertainty of prediction, the probability prediction density function of the FX model is established.
- The most probable curve and the 90% confidence interval of the soil-water characteristic curve are obtained, and the validation shows that the effectiveness of the proposed method.

References

[1] Fredlund D G and Xing A Q 1994 Equations for the soil-water characteristic curve Can. Geotech. J. 31 521-532.

[2] Vanapalli S K, Fredlund D G, Pufahl D E and Clifton A W 1996 Model for the prediction of
shear strength with respect to soil suction Can. Geotech. J. 33 379-392.

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

[4] Saxton K E, Rawls W J, Romberger J S and Papendick R I 1986 Estimating generalized soil-water characteristics from texture Soil Sci. Soc. Am. J. 50 1031-1036.

[5] Zapata C E, 1999: Uncertainty in soil-water-characteristic curve and impacts on unsaturated shear strength predictions, Arizona State University.

[6] Perera Y Y, 2003: Moisture equilibria beneath paved areas, Arizona State University.

[7] Pachepsky Y A, Timlin D and Varallyay G 1996 Artificial neural networks to estimate soil water retention from easily measurable data Soil Sci. Soc. Am. J. 60 727-733.

[8] Koekkoek E J W and Booltink H 1999 Neural network models to predict soil water retention Eur. J. Soil Sci. 50 489-495.

[9] Johari A, Habibagahi G and Ghahramani A 2006 Prediction of soil-water characteristic curve using genetic programming J. Geotech. Geoenviron. Eng. 132 661-665.

[10] Zhu J and Hu W 2015 Recent Advances in Bayesian Machine Learning J. Comput. Res. Dev. 52 16-26.

[11] SoilVision 2002 SoilVision. Version 3 [computer program].

[12] Soil Survey Division Staff (SSDS) 1993 Soil survey manual U.S. Department of Agriculture(USDA) Handbook No. 18 ( U.S. Government Printing Office, Washington, D.C.).

[13] Mohammadi M H and Vanclooster M 2011 Predicting the Soil Moisture Characteristic Curve from Particle Size Distribution with a Simple Conceptual Model Vadose Zone J. 10 594-602.

[14] Arya L M and Paris J F 1981 A physic-empirical model to predict the soil moisture characteristics from particle-size distribution and bulk density data Soil Sci. Soc. Am. J. 45(6) 1023-1030.

[15] Li J, Ma L, Zhang R, Zhang T and Wang B 2020 Analysis of influence of saturation and dry density on soil characteristic curve of weak expansive soil Journal of Southwest University. Natural Science Edition. 42 142-148.

[16] Schaap M G, Leij F J and van Genuchten M T 2001 ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions J. Hydrol. 251(3-4) 163-176.

[17] van Genuchten M T 1980 A closed-form equation for predicting the hydraulic conductivity of unsaturated soils Soil Sci Soc Am J. 44(5) 892-898.