Determination of soil-water retention curve: an artificial intelligence-based approach

Ehsan Nikoosaei^{1*}, Rasoul Mirghafari¹, Ghassem Habibagahi¹, Alireza Ghadamgahi Khorassani² and Amir Mohammad Nouri¹

¹Shiraz University, Department of Civil and Environmental Engineering, Shiraz, Iran

²K. N. Toosi University of Technology, Faculty of Civil Engineering, Tehran, Iran

Abstract. Soil Water Retention Curve (SWRC) is a fundamental relationship in unsaturated soil mechanics, knowledge of which is essential for determining major mechanical and hydraulic properties of unsaturated soils. There are several empirical, semi-empirical and physically-based models which have been proposed to date for estimating SWRC. While the physically-based models which employ the basic soil characteristics such as grain-size and pore-size distributions are regarded superior to purely empirical models, their Achilles' heel is the several simplifying assumptions based on which these models are constructed, thereby, restricting their applications and influencing their accuracy. Given the complexity of the soil porous structure, one may resort to the new inference techniques rather than mechanistic modelling to find the relationship among soil physical characteristics and the retention properties. Therefore, an alternative approach to purely empirical relationships as well as physically-based and conceptual models for determining SWRC is the use of Artificial Intelligence (AI) based techniques to acquire a relationship for SWRC based on the soil basic properties, especially grain size distribution and porosity. Among AI-based methods, Multi-Gene Genetic Programming (MGGP), often used to establish a close form equation for a complex physical system, offers a suitable alternative to the current approaches. In this study, a database comprising of 437 soils (containing various soil types, namely, sand, clay, silt, loam, silt loam, clay loam, sandy loam, sandy clay loam, silty clay loam, silty clay, and loamy sand soils) was used along with MGGP to establish a relationship among suction, saturation, porosity and grain size distribution. The proposed equation had a reasonable agreement with the experimental data compared to the other grain-based and physically-based models.

1 Introduction

Often used to establish equations for the major mechanical and hydraulic characteristics of unsaturated soil such as permeability, shear strength, and volume change, soil water retention curve (SWRC) plays an indispensable role in the mechanics of unsaturated soil, knowledge of which is required, as a core element, for any hydromechanical model.

Soil water retention is determined experimentally through various methods such as pressure plate, Buchner funnel, tensiometers, and pressure membranes. Laboratory measurements of the soil water retention curve are, however, time-consuming, costly and laborious, particularly for fine-grained soils. That is why in the last couple of decades, several studies have been performed to present various techniques to simulate, estimate and predict soil water retention curve. The numerical methods used either to estimate or predict SWRC can be grouped into the following categories [1]:

1. Empirical equations whose parameters are found by fitting to the experimentally measured values, e.g., [2,3].
2. Correlation of the parameters of the equations of the first category to the soil physical characteristics such as dry density, and soil particle size distribution, e.g., [4, 5].
3. Physically-based techniques which utilize a conceptual model such as the bundle of capillary tube to determine pore size distribution based on the grain size distribution, analytically and then formulate the soil water retention curve.
4. Simulation techniques such as discrete element method (DEM) and pore network modelling (PNM) [6-8].
5. Water content at different suction values is correlated to various soil characteristics such as porosity, and an index representing grain size distribution (e.g., effective diameter, D_{10}); the correlation is obtained, often, by regression techniques [9, 10].

* Corresponding author: enikoosaei@shirazu.ac.ir and ehsan_nikoosaei@yahoo.com

6. Various Artificial Intelligence methods (AI) such as neural networks, evolutionary polynomial regression (EPR), and genetic programming have been resorted to, to find the equations of soil water retention curve [11-12].

All of the techniques suffer from certain limitations which have raised the need to perform the current study. In fact, the first category of the techniques requires the experimental datapoints. In the second and fifth category of the studies, mostly a certain mathematical form for the equation of soil water retention has been assumed whose parameters are, then, found through the regression techniques. One major factor is the number of soils used for the regression which can limit the generalization of the obtained equations for different soils. Moreover, one has to assume the mathematical form a priori.

In the third category of studies, the conceptual model employed to find the relationships often includes very simplifying assumptions. While the problems pertinent to the use of bundle of capillary conceptualization (used in the 3rd category of studies) such as lack of connectivity between the tubes have been dealt with and alleviated by using simulation techniques such as discrete element and pore network modelling (fourth category of the methods), models in the fourth category of methods are commonly computationally demanding and are not straight forward as the use of a closed form equation. Furthermore, the direct use of pore network models needs assumptions about pore structure such as pore size distribution and throat size distribution which are often obtained from imaging techniques, that are not always available. The use of discrete element for soil water retention determination requires a pore network extraction step in which tessellation techniques are often employed. Tessellation techniques, however, in some cases result in dividing large pores into many small pores creating unrealistic throats.

While various artificial intelligence-based techniques have, therefore, been used to address the aforementioned shortcomings, the methods such as neural network often do not provide the user with an explicit equation. In recent years, genetic programming, particularly, an advanced form of it, the so-called Multi-gene genetic programming (MGGP) has been emerged in the computer science which can be considered as a powerful tool to extract the relationship among various variables in a complex system. Given the complexity and the number of factors affecting the soil water retention curve, one option could be the use of MGGP where no mathematical expression is assumed beforehand to obtain the relationship between the soil water retention and soil physical characteristics such as porosity and grain size distribution.

Therefore, the main purpose of this paper is to establish an AI-based explicit pedo-transfer function for SWRC prediction by using the Multi-Gene Genetic Programming algorithm (MGGP) utilizing a relatively comprehensive database. In order to make the spectrum of the soils to which the obtained equation can be applied as wide as possible, several soils whose hydraulic data as well as grain size distribution were

present in SoilVision (2002) database [13] were used (437 soils). Eventually, an equation relating water content to suction based on the soil grain size distribution and porosity was obtained. The equation was then compared to the other typical physico-empirical equations [14].

2 Materials and methods

In order to use MGGP, a database consisting various soil types was compiled. The next sections describe the database, the MGGP approach, and then the selected physico-empirical techniques used for comparison, respectively and finally the model performance criteria are presented.

2.1. Database collection and variable selection

It is now well established that SWRC is dependent on various soil physical characteristics, such as grain-size distribution, porosity, and water content. Hence, suitable variables representing these characteristics had to be selected for MGGP. Furthermore, such data should be readily available from laboratory works reported the literature or from present databases. In order to represent grain size distribution, five grain size diameters (D_{10} , D_{20} , D_{30} , D_{50} , and D_{60}) were considered. The reason to select these diameters was their presence for all 437 soils selected from the Soil Vision Database [13]. These effective diameters along with porosity, as well as suction were input variables for the MGGP, and the volumetric water content reported at each suction was considered as the output. The compiled database comprised of a wide spectrum of soil types (i.e., sand, clay, silt, loam, silt loam, clay loam, sandy loam, sandy clay loam, silty clay loam, silty clay, and loamy sand).

In order to prepare data for AI-based model, 75% of the database is chosen as the training dataset, and 25% is the testing data that is selected randomly from this database. The training samples are used for formulating the model, and the test data samples are used for authenticating the extrapolation ability of the model. The descriptive statistics of the input and output variables are presented in Table 1.

2.2 Multi-Gene Genetic Programming and its implementation

MGGP is an artificial intelligence-based technique which is tailored to obtain explicit formula through solving symbolic regression problems and is based on the same principles of Genetic Programming (GP). GP is known as one of the evolutionary programming algorithms, inspired by the Darwin's theory of "survival of the fittest"; it employs the elements of the evolution process in nature in order to find the optimal solution [15]. GP functions the same as Genetic Algorithm (GA); however, the apparent difference is that GA is utilized for parameter optimization, while the GP is developing a tree-based structure in various lengths denoting a certain mathematical equation on which the structural

optimization approach is applied to obtain the best explicit symbolic form. At first, GP algorithm generates the initial population of solutions at random. This process is performed by randomly combining the elements from function and terminal set. The function set elements can consist of arithmetic operators, nonlinear functions, and Boolean operators. The terminal set is composed of two input process variables and random constants. Composing the function and terminal sets leads to a tree-shape exhaustive set of solutions [16].

In order to find the best individuals of the generated population, the performance of the GP model must be evaluated. It is measured using the fitness function which can be various error measures such as RMSE, in order to compare the predicted GP model values to that of the actual values. The fitness function should be minimized to attain the best performance, considering the train and test results. This process has a termination

criterion that controls the procedure of performance evaluation. The termination criterion is defined by the user, and it can be any of the maximum number of generations, a limitation on run time, or an error threshold for the fitness function. Until the termination criterion is met, the genetic operators are implemented on the selected individuals of the initial population with the aim of evolving a new population. The genetic operators include processes such as crossover, mutation, and reproduction. They aim to increase the diversity of the new population and consequently increasing the chance of capturing the global optimum and the best possible structure. The tournament selection method is employed to select the individuals for applying the genetic operators. It chooses the individuals with the lowest fitness values and different size to reproduce or copy to the next generation.

Table 1. Descriptive statistics of the input and output variables.

Parameters	Porosity (x_1)	D_{10} (x_2)*	D_{20} (x_3)	D_{30} (x_4)	D_{50} (x_5)	D_{60} (x_6)	Suction (x_7)**	Water Content (y)
Mean	0.4672	0.0279	0.0529	0.0739	0.1115	0.1317	583.1314	0.2837
Median	0.4520	0.0013	0.0096	0.0214	0.0548	0.0737	8.731	0.2940
Standard deviation	0.0927	0.0484	0.0801	0.1040	0.1348	0.1484	9366.419	0.1472
Variance	0.0086	0.0023	0.0064	0.0108	0.0182	0.0220	8772.979	0.0217
Kurtosis	2.886	6.1087	5.7709	7.8597	7.8923	7.0409	1168.093	-0.4187
Skewness	1.3042	2.3284	2.2486	2.4854	2.4467	2.3295	31.650	0.2107
Minimum	0.2640	1.17E-05	1.44E-05	1.89E-05	0.0001	0.0014	0.001	0.005
Maximum	0.9150	0.3035	0.4471	0.6880	0.8769	0.9235	416728.81	0.8240

* D_{10} , D_{20} , D_{30} , D_{50} , D_{60} are in 'mm'

** Suction unit is 'kPa'

In the crossover operation, two individuals are selected randomly, and a subtree branch of each one is swapping to produce the new children. With this approach, the mutation operation is replacing the subtree branch of an individual with the new randomly generated GP model to create a new child. The intensity of the genetic operations depends on their probability values, reported in Table 2.

GP and MGGP have the same procedure in finding an explicit model; however, the prominent feature of MGGP against GP is that the MGGP model participating in the evolution process is composed of several trees/genes where genes are randomly selected and combined whose coefficient of participation in the combination (weights) are found via the least square method. In this research, the open source MGGP toolbox, called, GPTIPS 2.0 [17] was employed. The best MGGP model was selected based on training data. The maximum number of depths, genes, population size, generations, and other MGGP input parameters are reported in Table 2.

Table 2. Parameter setting for MGGP.

Parameters	Values assigned
Runs	1
Population size	15000
Number of generations	700
Tournament size	8
Tournament pareto	0.6
Elite fraction	0.15
Max depth of tree	10
Max genes	10
Max nodes per tree	Infinitive
Functional set (F)	times, plus, minus, rdivide, square, cube, power, mult3, add3, sin, cos, abs, exp, negexp, log, plog*
Terminal set (T)	(d_1 , d_2) [-10, 10]
Crossover probability rate	0.75
Reproduction probability rate	0.05
Mutation probability rate	0.2

* plog is protected natural logarithm

2.3 Grain-size based physical models: Arya and Paris Model

The Arya and Paris (1981) model conceptualizes each particle size fraction i with a single capillary tube of radius r_i . For each capillary tube, its radius, is obtained from the following [14, 18]:

$$r_i = R_i \left[\frac{2\varepsilon}{3(1-\varepsilon)} \left(\frac{3M_i}{4\pi R_i^3 \rho_s} \right)^{1-\alpha} \right]^{1/2} \quad (1)$$

where, R_i is the representative particle size in i^{th} particle size class, M_i is mass fraction of the i^{th} particle size class, ε is porosity, and ρ_s is the density of the solid particles. The exponent α is a semi empirical coefficient [14]. The constant values of this coefficient for different soils introduced in [19], and implemented in soil vision software were used in this study.

The matric potential, ψ_i , is then obtained based on the capillary rise (Young-Laplace) equation as follows:

$$\psi_i = - \left(\frac{2\sigma \cos \beta}{\rho_w g r_i} \right) \quad (2)$$

where, σ is the surface tension, β is wetting angle of water menisci, ρ_w is water density, and g is the gravitational acceleration. The volumetric water content, θ_i , for the matric potential ψ_i is obtained from summing up the mass fractions:

$$\theta_i = \theta_s \sum_{j=1}^{j=i} M_j \quad (3)$$

where, θ_s is the saturated water content. It should be noted that the index j denotes the particle size fractions associated with the water-filled pores and i corresponds to all the particle size fractions.

2.4 Measures of the model performance

The measures of performance utilized for the model selection were mean absolute percentage error (MAPE), root mean squared error (RMSE), square of correlation coefficient (R^2), multi-objective error (MO), sum squared error (SSE), mean squared error (MSE) as described hereafter:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - M_i}{A_i} \right| \times 100 \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n |M_i - A_i|^2}{n}} \quad (5)$$

$$R^2 = \left(\frac{\sum_{i=1}^n (A_i - \bar{A}_i)(M_i - \bar{M}_i)}{\sqrt{\sum_{i=1}^n (A_i - \bar{A}_i)^2 \sum_{i=1}^n (M_i - \bar{M}_i)^2}} \right)^2 \quad (6)$$

$$\text{Multi-objective error} = \frac{MAPE + RMSE}{R} \quad (7)$$

where, M_i is predicted and A_i is actual values, $\bar{M}_i$ and $\bar{A}_i$ are the average predicted and actual values respectively and n is number of data samples.

3 Results and discussion

The formula obtained from MGGP is as follows:

$$y = \sum_{i=1}^{10} G_i + b \quad (8)$$

where, b is the bias, G_i is the i^{th} weighted gene consisting the input variables and y is the predicted volumetric water content (see and recall Table 1 for a description of input variables). The weighted genes, G_i , and bias are as follows:

Gene1 = $0.033 * \exp(-1.0 * x_1 * (\log(x_1 - x_5) + \exp(\log(x_1^4 * x_7)) * (\sin(x_1 * \log(x_2)) + \exp(x_1^2 * x_6 * x_7^2))))$

Gene2 = $+ 0.016 * \log(-\log(2 * x_1) * (x_2 * x_6 * \exp(-x_1 - x_4) * \exp(2 * x_1^2 * (x_5^3 + 4.8)) - x_2 * x_6 * (x_1 + x_6) * \exp(\exp(-\exp(2 * x_1)))) * (\cos(\log(x_3)) + \log(\exp(x_5) + \log(x_6) + x_1 * x_6)) * \exp(\exp(-x_5) + 5.1))$

Gene3 = $- 0.012 * \log(-\log(\log(x_7 + 6.5) * (x_2 * x_6 * (x_1 - 188.0) + x_2 * x_6 * \sin(x_6) * \sin(6.5/x_6)) * (x_6 * x_1^2 * x_7^2 + x_6 + 155.0)) * (0.68 * x_6 * (x_2 * \sin(\log(x_2) * (x_1 + x_4 + x_6)) * (x_6 + \cos(x_4)) * \exp(\exp(-x_5)) - x_1 * x_2 * x_5 * x_6 * (-x_1^3 + x_6)) + x_1 * \exp(x_5) * x_2 * x_4) * (\log(\cos(x_5))/x_4 + x_1 * \exp(x_6)))$

Gene4 = $+ 7.5e-3 * \log(-\log((x_4 * x_7 * \log(\cos(x_5))) * \sin(6.5/x_6^2)/x_7 + x_1^2) * (x_2 * x_6 * (-x_1^3 + x_1) - x_2 * x_6 * (x_4 + x_5 + \sin(x_4)) * \sin(6.5/x_6)) * (x_5 * \exp(-x_1 - x_4) + x_3 * \sin(155.0 * \log(x_2)) * (x_1 + \cos(x_1)) * \exp(\exp(-x_5)) - x_1 * x_2 * x_4 * x_5 * x_7)))$

Gene5 = $- 0.019 * \log(\log(2 * x_1) * (x_2 * x_6 * (x_1 + x_6) * \exp(\exp(-\exp(-x_2))) - x_2 * x_6 * (x_1 - x_5)) * (x_6 + x_5 * \sin(x_7) * (2 * x_7 * x_1^4 + x_5 + x_6) + x_5 * \exp(x_1) * (x_7 - x_6 + \log(x_1 - x_5 * x_7))))$

Gene6 = $+ 0.041 * \exp(2.0 * x_1)$

Gene7 = $+ 0.041 * \exp(-1.0 * x_6) * (1.0 * \exp(-1.0 * \exp(-1.0 * \exp(x_1^4 * x_4)) * (x_7 + \sin((x_5 - 1.0 * x_6) / \cos(x_1)))) - 4.7 * \exp(-1.0 * x_5 * x_7))$

Gene8 = $- 1.2e-6 * ((x_6 + x_7 * \log(\cos(x_5)) * (x_2 - 1.0 * x_3)) / x_4) * \sin(6.5/x_6) / x_2 * (x_1 + x_6)$

$$\text{Gene9} = + 11.0 * \exp(- (1.0 * x_1) / ((\exp(2.0 * x_5) * \exp(- 1.0 * x_5 * x_7))^{\wedge} x_5 + (x_2 * x_5^{\wedge} 4) / x_6^{\wedge} 6) - 4.9)^{\wedge} \exp(x_1 * x_2 * x_3 * x_7^{\wedge} 2)$$

$$\text{Gene10} = + 0.49 * x_1 * \exp(- 1.0 * x_5 * x_7) - (0.22 * x_5^{\wedge} 3 * \text{plog}(\cos(x_5))^{\wedge} 3) / \sin(x_4)^{\wedge} 3$$

$$b = - 0.058$$

Figures 1 and 2 present the performance of the formula obtained by MGGP for the train and test datasets, respectively. As these figures reveal the soil water retention values obtained by the formula resulting from MGGP agree reasonably well with the experimental datapoints. Table 3 presents the values of the performance measures for the MGGP, for both test and train datasets.

Table 3. Calculated performance measures for MGGP model.

Parameter	Train data	Test data
R ² (coefficient of determination)	0.92	0.90
R (correlation coefficient)	0.96	0.95
RMSE (root mean squared error)	0.044	0.047
MSE (mean squared error)	0.002	0.002
SSE (sum squared error)	6.90	2.59
MAPE (mean absolute percentage error)	3.30	3.50
MO (multi-objective error)	0.080	0.086

Fig. 1. Performance of MGGP model on training dataset.

Figures 3 to 5 depict the experimental, and predicted soil water retention curves for a sandy loam, and two silt loams. While generally the performance of the model for coarse-grained soils is reasonably well, it is not as good as those predicted for the clayey soils. The reason is the fact that the soil mineralogy has not been reflected in the considered variables. Furthermore, as it can be seen the physically-based formula (Arya and Paris 1981) performs relatively poor as compared to the model obtained from artificial intelligence (MGGP). It is due to

the several simplifying assumptions pertinent to this model, one of which is the assumption and conceptualization of the bundle of capillary tube used to formulate the relationships in Arya and Paris (1981) model as well as other available physically-based methods.

Fig. 2. Performance of MGGP model on testing dataset.

Fig. 3. Experimental versus predicted soil water retention curves for soil #11274 (sandy loam).

Fig. 4. Experimental versus predicted soil water retention curves for soil #11278 (silt loam).

Fig. 5. Experimental versus predicted soil water retention curves for soil #11298 (silt loam).

4 Summary and the concluding remarks

In this study, an advanced tool from artificial intelligence-based techniques was utilized to obtain a new pedo-transfer function for soil water retention. For this purpose, a database consisting of 437 soils was compiled and used. The new pedo-transfer function had a reasonably well agreement with the experimental datapoints and had a better performance as compared to Arya and Paris (1981), a typical model from the physically-based techniques. Although the obtained equation seems to be long and complex, it is a closed form equation which can easily be implemented in any excel worksheet or software to obtain the soil water retention curve. The sole input parameters are the five grain size diameters obtained from the grain size distribution curve and porosity. Ongoing study of the authors is devoted to enhance the presented equation for the fine-grained soils by considering additional variables.

References

1. A. Ahangar-Asr, A. Johari, and A. A. Javadi. An evolutionary approach to modelling the soil–water characteristic curve in unsaturated soils. *COMPUT. GEOSCI.* **43** (2012).
2. R. Brooks, and T. Corey. Hydraulics properties of porous media. *Hydrology Papers*, Colorado State University **24** (1964).
3. M. Th. Van Genuchten, A closed-form equation for predicting the hydraulic conductivity of unsaturated soils 1. *SOIL SCI SOC AM J.* **44**, 5 (1980).
4. H. P. Cresswell, and Z. Paydar. Water retention in Australian soils. I. Description and prediction using parametric functions. *SOIL RES.* **34**, 2 (1996).
5. J. Tomasella, and M. G. Hodnett. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. *SOIL SCI.* **163**, 3 (1998).
6. E. Nikooee, T. Sweijen, and S. Majid Hassanizadeh. Determination of the relationship among capillary pressure, saturation and interfacial area: a pore unit assembly approach. *E3S Web of Conferences*. Vol. 9. EDP Sciences, 2016.A
7. H. Khaksar, G. Habibagahi, and E. Nikooee. SWRC modeling in unsaturated soils: a pore network approach. *Poromechanics V: Proceedings of the Fifth Biot Conference on Poromechanics*. 2013.
8. A. Rostami, G. Habibagahi, M. Ajdari, and E. Nikooee, Pore network investigation on hysteresis phenomena and influence of stress state on the SWRC. *INT. J. GEOMECH.* **15.5** (2013): 04014072.A
9. J. L. Hutson, and A. Cass. A retentivity function for use in soil–water simulation models. *J. SOIL SCI.* **38**,1 (1987).
10. M. Aubertin, J. F. Ricard, and R. P. Chapuis. A predictive model for the water retention curve: application to tailings from hard-rock mines. *CAN. GEOTECH. J.* **35**,1 (1998).
11. A. Johari, and A. A. Javadi. Prediction of a soil–water characteristic curve using neural network. *5th International Conference on Unsaturated Soils* (2011).
12. A. Johari, G. Habibagahi, and A. Ghahramani. Prediction of soil–water characteristic curve using genetic programming. *J. GEOTECH GEOENVIRON* **132**, 5 (2006).
13. SoilVision, S. L. *SoilVision: a Knowledge-based database system for saturated/unsaturated soil properties, version 4.14*. SoilVision Systems Ltd, Saskatchewan (2005).
14. L. M., Arya, and J. F. Paris. A physicoempirical model to predict the soil moisture characteristic from particle-size distribution and bulk density data 1. *SOIL SCI SOC AM J.* **45**, 6 (1981).
15. J.R. Koza, D.E. Goldberg, D.B. Fogel, and R.L. Riolo, eds.. *Genetic Programming: Proceedings of the First Annual Conference, 1996*. Cambridge Mass: Mit Press, 1996.
16. A. Garg, A. Garg, K. Tai, S. Barontini, and A. Stokes, A computational intelligence-based genetic programming approach for the simulation of soil water retention curves. *TRANSPORT POROUS MED.* **103**, 3 (2014).
17. D. P. Searson, D. E. Leahy, and M. J. Willis. GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. *Proceedings of the International multiconference of engineers and computer scientists*. Vol. 1. Hong Kong: IMECS, 2010.
18. U. Buczeko, and H. H. Gerke. Evaluation of the Arya-Paris model for estimating water retention characteristics of lignitic mine soils. *SOIL SCI.* **170**,7 (2005).
19. L. M. Arya, F. J. Leij, M. Th. van Genuchten, and P. J. Shouse. Scaling parameter to predict the soil water characteristic from particle-size distribution data. *SOIL SCI SOC AM J.* **63**, 3 (1999).