Accepted manuscript

As a service to our authors and readers, we are putting peer-reviewed accepted manuscripts (AM) online, in the Ahead of Print section of each journal web page, shortly after acceptance.

Disclaimer

The AM is yet to be copyedited and formatted in journal house style but can still be read and referenced by quoting its unique reference number, the digital object identifier (DOI). Once the AM has been typeset, an 'uncorrected proof' PDF will replace the 'accepted manuscript' PDF. These formatted articles may still be corrected by the authors. During the Production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal relate to these versions also.

Version of record

The final edited article will be published in PDF and HTML and will contain all author corrections and is considered the version of record. Authors wishing to reference an article published Ahead of Print should quote its DOI. When an issue becomes available, queuing Ahead of Print articles will move to that issue's Table of Contents. When the article is published in a journal issue, the full reference should be cited in addition to the DOI.

Submitted: 06 July 2021

Published online in 'accepted manuscript' format: 27 July 2021

Manuscript title: Effective Stress Parameter in Unsaturated Soils; an Evolutionary-Based Prediction Model

Authors: Alireza Ahangar Asr[1], Akbar A. Javadi[2]

Affiliations: [1]School of Science Engineering, and Environment, Directorate of Civil Engineering, University of Salford, Manchester, United Kingdom. [2]Department of Civil Engineering, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Devon, United Kingdom.

Corresponding author: Alireza Ahangar Asr, School of Science Engineering, and Environment, Directorate of Civil Engineering, University of Salford, Manchester, United Kingdom.

E-mail: a.ahangarasr@salford.ac.uk

Accepted manuscript
doi: 10.1680/jsmic.21.00012

Abstract

Deformations and failures in unsaturated soils are influenced directly by the effective stress calculated using the stress equation affected by the effective stress parameter. A data mining-based approach, the Evolutionary Polynomial Regression (EPR), is implemented in this research to develop a prediction model for the effective stress parameter in unsaturated soils. The proposed modelling approach takes an evolutionary computing technique to for finding polynomial models that are structured and explicit. A combination of the well-established genetic algorithm method and the least square approach are implemented to search for the most suitable polynomial structures and their corresponding parameters for all terms in the developed polynomial structure. A set of unsaturated soil experimental results (triaxial tests) from literature were used in this study to develop the prediction model. Once the model completed it was evaluated based on its performance for making predictions using input parameters that were previously kept unseen to validate generalization capabilities (making predictions of the output for new input data). The predictions made by the model, were compared to actual measured data from the lab tests as well as an Artificial Neural Network based model. A sensitivity analysis was also done to assess the level and form of contributions that input parameters had to the developed model. The results showed that the developed model could successfully and to a high level of accuracy capture and redevelop the intrinsic connections between the input parameters involved in the model to help produce accurate the effective stress parameter predictions that can not only compete with the artificial neural network model in terms of accuracy of the model predictions and generalisation capabilities; but also outperform the artificial neural network model with regards to the structure, simplicity and transparency.

Notation

(u_a-u_w)	suction
F	a function constructed by the modellingprocess
X	matrix of input variables
f	a function defined by the user
m	number of terms of the target expression/polynomial model
y	estimated vector of output
a_j	a constant
y_a	experimental/actual parmaeter value
y_p	model prediction
COD	Coefficient of Determination
N	number of data points/lines on which the COD was calculated
θ_r	Volumetric water content at residual condition
θ_s	Volumetric water content in saturated condition
λ	Soil-water characteristic curve slope
$\sigma_3 - u_a$	Net confining stress
h_b	Air entry value
χ	Effective stress parameter

1. Introduction

Understanding the mechanical behaviour and analysing the stability of foundations, natural slopes and geotechnical structures at the first instance is reliant on evaluating the shear strength of soils involved. An important requirement for an economically feasible design is an accurate shear strength prediction of geo-materials in both saturated and unsaturated conditions.

Changes in soil properties inevitably affect the effective stress parameter and this in turn changes the shear strength in un-saturated soils. Öberg and Sällfors considered that considering the effective stress parameter to be same as (as equivalent to) the degree of saturation can make predicting the shear strength of unsaturated soils easier (Öberg & Sällfors, 1997); however, Loret and Khalili emphasise that, in addition to the suggestion by Öberg and Sällfors, the effective stress parameter and shear strength of unsaturated soils are dependent on the properties of considered soil and its structure too (Loret & Khalili, 2002). Properties of unsaturated soils can also be determined using the soil water characteristic curve (SWCC). Empirical methods are suggested in the literature to predict the shear strength in unsaturated soils with the soil water characteristic curve (Fredlund, Xing, Fredlund, & Barbour, 1996; Vanapalli, Fredlund, Pufahl, & Clifton, 1996). A relationship between the plasticity index of the soil and κ was presented by Garven and Vanapalli (Garven & Vanapalli, 2006). Experimental works outcomes from some researchers showed that net confining pressure noticeably affects the soil water characteristic curve and parameters of this curve change by variations in stress (I.-M. Lee, Sung, & Cho, 2005). Khalili and Khabbaz suggested that the effective stress parameter can be considered as 1 where suction values drop below the bubbling

pressure. They also established that the relationship between the matric suction logarithm and the effective stress parameter can be considered as linear (Khalili & Khabbaz, 1998).

A similar picture for the effective stress parameter was depicted by Xu by using fractal theory (Xu, 2004). In Xu's work the soil-water characteristic curve was implemented to make reasonable estimations of the surface fractal dimension in soil. Equation 1 (below) was suggested by Russell and Khalili to calculate the effective stress parameter in sands (A. Russell & Khalili, 2006):

$$\chi = \begin{cases} 1 & \text{for } \dfrac{(u_a - u_w)}{(u_a - u_w)_b} < 1 \\ \left(\dfrac{(u_a - u_w)}{(u_a - u_w)_b} \right)^{-0.55} & \text{for } 1 < \dfrac{(u_a - u_w)}{(u_a - u_w)_b} < 25 \\ 25^{0.45} \left(\dfrac{(u_a - u_w)}{(u_a - u_w)_b} \right)^{-1} & \text{for } \dfrac{(u_a - u_w)}{(u_a - u_w)_b} > 25 \end{cases} \quad (1)$$

In Equation 1, $(u_a\text{-}u_w)_b$ is the air entry value in the drying process that represents the air expulsion value in wetting conditions. Zargarbashi and Khalili highlighted the stress state as an influencing factor on the bubbling pressure and emphasised that the corrected value of this parameter needs to be used for the purpose of estimating the effective stress parameter to a high level of accuracy (Zargarbashi & Khalili, 2011). A number of empirical relationships are available in the literature to predict shear strength in unsaturated soils but, there is no comprehensive equation/formula that could apply to all unsaturated soil types in predicting shear strength (Garven & Vanapalli, 2006; Fazeli, Habibagahi, & Ghahramani, 2009).

A percolation theory based numerical method was proposed by Arvin et al. which estimated the effective stress parameter (Arvin, Veiskarami, Ajdari, & Habibagahi, 2007). In

the research work by Arvin et al the soil-water characteristic curve was used to determine the distribution of pore sizes in soil and a conceptual model was created using the percolation theory. The proposed model was used to directly determine the effective stress parameter, but this method would be considerably expensive from computation point of view, if it was considered to be extended to make predictions in high suction ranges.

In the past decade some researchers used the Artificial Neural Network (ANN) to capture relations between the shear strength of unsaturated soils and the contributing physical properties (Kayadelen, 2008; S. Lee, Lee, & Kim, 2003). These models, however, did not take the effects of sample preparation methods as well as the stress states during these processes into consideration. Results from processing the triaxial unsaturated shear tests revealed that the effective stress parameter varies significantly with changes in net mean stress values under constant suction (A. Russell & Khalili, 2006; A. R. Russell & Khalili, 2004).

Artificial Neural Networks (ANN) model to make estimations of the effective stress parameter was developed by Ajdari et al which is a requirement if the shear strength in unsaturated soils is intended to be estimated. The input variables used in the research by Ajdari et al were the matric suction, net mean stress and the soil-water characteristic curve parameters in unsaturated soils that were obtained from various triaxial test results from literature. Ajdari et al also investigated the effect that net stress value can have on the effective stress parameter (Ajdari, Habibagahi, & Ghahramani, 2012).

This paper applies the evolutionary polynomial regression modelling approach to create an innovative and comprehensive model for estimating effective stress parameter in saturated

soils. Presented model has a simple polynomial structure and is chosen as the optimal model based on satisfying accuracy (coefficient of determination parameter), simplicity and capability in representing the true effective stress variations based on the s intended intrinsic relations between contributing input scientific understanding that is available from literature about the complex problem of effective stress parameter estimation in geotechnical engineering. The proposed model reflects expected sensitivity to contributing/input parameters in line with the expectations based on the available knowledge on the matter.

2. Evolutionary-based polynomial modelling approach

Evolutionary polynomial regression (EPR) is a data-mining based modelling technique combining numerical and symbolic regression methods. As a result of this modelling methodology polynomial models are developed. Polynomial structures have many mathematical specifications that makes them very attractive for the users. An important specification of this method is that EPR uses genetic algorithm (GA) to conduct evolutionary searches to find suitable exponent values for the terms in the polynomial expressions. This feature facilitates computational implementation of the algorithm as well as efficiently searching for an explicit expression, and results in the better capacity to control the complexity of the generated polynomial expression (Giustolisi & Savic, 2006).

Evolutionary polynomial regression is a method based on machine learning and is data-driven, and relies on evolutionary computing, aiming to find polynomial structures to represent a system/parameter. Assume a physical system with an output value (y) which is dependent on a group of inputs (X) and parameters (θ). This system can the be formulated as

below (Equation 2) from mathematical point of view:

$$y = F(\mathbf{X}, \boldsymbol{\theta}) \tag{2}$$

In this equation F is defined as a function in an m-dimensional space with m being considered as the number of inputs. In EPR modelling process to control the complexity of the developed polynomial models, in other words, to avoid the problem of mathematical expressions that grow fast and continues to become longer and longer in length (increasing number of terms) with time, the evolutionary procedure is arranged to be conducted in such a way that it searches for the exponents of a polynomial function in which the maximum number of terms is fixed. As a result of a single run of the programme, several expressions with increasing numbers of terms up to a limit set by the user to allow the optimum number of terms to be selected, will be developed. The form of polynomial equations/expression of EPR-based models can in general be presented as (Giustolisi & Savic, 2006):

$$y = \sum_{j=1}^{m} F(\mathbf{X}, f(\mathbf{X}), a_j) + a_0 \tag{3}$$

In this equation form, y is the estimated vector of output of the process; a_j is a constant; F is a function constructed by the process; X is the matrix of input variables; f is a function defined by the user; and m is the number of terms of the target expression/polynomial model.

In general, EPR use a technique for constructing symbolic models that works in two stages. First, EPR implements standard genetic algorithm (GA) to look for the best structure for the function. This means vectors corresponding to independent inputs, X_s=1: k, are combined , and at the second satge a least squares regression is conducted with the aim of finding the θ (parameters that are adjustable) for each and every combination of inputs

constructed. This strategy ensurs a global and comprehensive search algorithm is used not only for finding the best set of input combinations but also the to search for the most suitable related exponents - all at the same time – considering the cost function that is defined by the model developer/user (Giustolisi & Savic, 2006). To evaluate the adjustable parameters, a_j, the linear least squares (LS) method is adopted which operates based on minimization of the sum of squared errors (SSE) as its cost function. The sum of squared errors function is implenented to lead the search process towards the best fit model and can be expressed as:

$$\mathbf{SSE} = \frac{\sum_{i=1}^{N} (y_a - y_p)^2}{N} \tag{4}$$

In this function y_a is the experimental/actual parmaeter value and y_p is the model prediction aiming to be closet to the actual value to reduce the error margin.

The global search aiming to find the best form for the EPR equation/model is conducted using a standard genetic algorithm (GA) over the values in the user defined “exponents vector”. Gentic Algorithm (GA) operates based on the principals of Darwinian evolution beginning by randomly creating an initial “solutions population”. Each parameter set in the population is considered to stand for chromosomes of the individuals. Based on how well/poor each individual performs in its relevant environment, a fitness is assigned to that individual. The next generation is then created via crossover and mutation operations, with the probabilities Pc (probability of crossover) and Pm (probability of mutation) respectively. Fit individuals are then chosen for the “mating” process whilst the weak ones (individuals) are removed/dead. A child (offspring) is the created by “mated parents” carrying a set of chromosomes which is a

mix of chromosomes of parents. The evolutionary polynomial regression modelling process uses integer genetic algorithm coding with single point crossover is to find where (on which term) the candidate exponents will be located in the polynomial model (Giustolisi & Savic, 2006).

A set of termination critera including the maximum number of generations, the maximum number of terms in the target mathematical expression or a particular allowable error are used to eventually bring the evolutionary polynomial regression modelling process to a stop. The modeling process comes to a halt when either one of the criteria is met. EPR has been used to develop various models so far to represent complicated behavior of various geotechnical and civil engineering parameters, systems and materials including unsaturated soils (Ahangar Asr, Faramarzi & Javadi, 2018; Ahangar Asr, Javadi, & Khalili, 2015; Ahangar Asr & Javadi, 2016; Cuisinier, Javadi, Ahangar Asr, & Masrouri, 2013; Hussain, Javadi, Ahangar Asr, & Farmani, 2015; Javadi & Rezania 2009). Figure 1 presents a flow diagram for the EPR modelling procedure.

3. Processing and preparation of data used in model development and validation

A comprehensive set of unsaturated triaxial test data from literature (Bishop & Blight, 1963; Khalili, Geiser, & Blight, 2004; I.-M. Lee et al., 2005; Rahardjo, Heng, & Choon, 2004; Rassam & Williams, 1999; A. Russell & Khalili, 2006; A. R. Russell & Khalili, 2004; Ajdari, Habibagahi, & Ghahramani, 2012) was used to develop the evolutionary polynomila regression baed model in this research to predict the effective stress parameter in unsaturated soils.

The developed model assumed to be affected by some parameters (inputs) proven to be

influential on the effective stress parameter in unsaturated soils from literature; therefore, these parameter (the air entry value, volumetric water content at residual condition, volumetric water content in saturated condition, the soil-water characteristic curve slope, net confining stress and suction, shown in Table 1) were considered as input parameters in the model development process with the single output parameter set to the effective stress parameter in unsaturated soil. Statistical analysis was performed on various combinations of training and testing data to make sure that the best most representative and closest combination (from the statistical point of view) in terms of key statistical parameter values (mean and standard deviation) was chosen to help develop the most powerful and representative EPR model.

For the model to be representative there needs to be enough data that is used and seen by the software to be able to adequately learn the intended intrinsic relations between contributing input parameters. On the other hand for the machine-learning based models the crucial capability expected of models is their “generalisation” potential meaning that the developed models have to be able to produce accurate predictions in case of any data being presented as input to them is unseen to them during the model development stage. For this reason, the available data for model development by EPR approach in this research was divided into training and testing sets and whilst the training data was used to develop the model the testing data was kept unseen to the software during the whole model development stage and was only used after the model was finalised to verify the generalisation capabilities of the developed model.

In the literature, where machine learning techniques used, division of the data has

traditionally been around 80% (used for training) to 20% (used for testing/verification/examining generalisation capabilities) to develop prediction models (Javadi and Rezania, 2009; Johari et al, 2006 a & b); therefore, a similar approach is taken in this research work.

Just over 120 lines of data (the total number of cases in the database used) were divided into training and testing datasets. Using the created training and testing data sets (finalised after completing a statistical analysis – explained below) the training data were used to train EPR to develop the evolutionary polynomial regression-based model whilst the remaining cases (test data) were kept unseen to EPR during the model development process and were used later on to validate the developed model.

The adopted strategy for testing the generalisation capabilities of the developed model could also be translated as the "Hold-out" approach, as 21 out of 121 lines of data (about 18% of the whole data available) were "held-out" during the training stage of the model development process. Once the model development completed, the "left-out" data lines were used to verify the generalisation capabilities of the model in making predictions based on the "left-out" (previously unseen to EPR during the model creation process) data.

The data used to develop and verify the model was statistically analysed for the most statistically consistent training and testing sets to be selected and used in the development of the presented models. The aim was to ensure utmost consistency between the training and testing data sets to optimise the learning process to help develop best possible models. In other words, the data analysis process intended to make sure that the statistical properties of the data

in the training and testing data sub-sets were closest possible to one another to reflect the fact that they represent the same population from the statistical point of view. Many combinations for testing/verification and training data bunches were considered to start with. The standard deviation as well as the mean values were calculated for all parameters that were considered to be contributing to the final model. Statistical parameter values for training and testing data for the final combination used in the model development and verification processes are shown in Table 1.

Around 18% of data was kept unseen during the model development stage and was used to examine generalisation capabilities of the developed model; however, whilst the developed model was created based on a wide range of data and presented very good capabilities in extrapolating when making predictions of the effective stress parameter, due to the fact that extrapolation does not/cannot have any certain limits and in order to ensure that the developed model is entirely reliable and safe to be implemented using various ranges of unseen data, efforts has been put in place in this research to avoid extrapolation in the developed model predictions (Javadi and Rezania, 2009). To do this, the training and testing data ranges chosen (before doing the comprehensive statistical analysis detailed above) were carefully checked to ensure that all parameter values in the testing data sets were within the ranges of data chosen to be used for training EPR and for developing the evolutionary polynomial regression model to avoid extrapolation in making predictions to ensure that the predictions were completely reliable from statistical and engineering point of view. Although (as highlighted above) the model predictions seem to be very good for any range of unseen data, to ensure the reliability

factor applies, if the ranges of input data fall outside of the ranges used in the training stage of the model development process (in industry applications of the developed model for instance), retraining of the model using new data which includes the new wider ranges is strongly recommended.

4. Model development

A computer programme running on MATLAB as the main platform, was developed over the years in collaboration between the University of Exeter in the United Kingdom and the Technical University of Bari, Italy to be used to implement the Evolutionary Polynomial Regression technique (Giustolisi & Savic, 2006; Giustolisi et al, 2008; Rezania et al, 2008).

In order to be able to control the length, complexity, type of functions used in developing the model, number of terms in the polynomial model, range of the exponents considered in the model structure, and also the number of generations to be used by the programme to complete the evolutionary modelling processes some constraints were considered. Coefficient of determination (COD) parameter (fitness equation - Equation 5) was used to check model fitness/accuracy level as the model development processes were progressing .

$$\mathbf{COD} = 1 - \frac{\sum_{N} (\mathbf{Y}_a - \mathbf{Y}_p)^2}{\sum_{N} \left(\mathbf{Y}_a - \frac{1}{N} \sum_{N} \mathbf{Y}_a \right)^2} \tag{5}$$

In this equation Y_a is the actual output parameter value; Y_p is the EPR model predicted value for the output parameter, and N shows the number of data points/lines on which the COD was calculated.

Coefficient of Determination, COD is a simple parameter in determining the accuracy of prediction models and has been used in the past years by researchers as an effective parameter to evaluate the level of accuracy of developed models in making predictions of output parameters particularly in data-mining and machine learning based modelling techniques/approaches (Ahangar Asr et al, 2015 & 2018).

If the fitness level of the model (accuracy) was not considered acceptable (according to the calculated COD value), or if any other criteria set to terminate the modelling process (user defined maximum number of generations and/or maximum number of terms) were not met at the end of each model development cycle, the current model will be subject to going through another evolution for a new model to be developed. This process continues until a final model is reached and the modelling process is terminated by meeting modelling process termination criteria.

When using modelling techniques that adopt regression as part of their modelling processes the term ‘fitness’ usually is used as a means of showing how closely the outcomes of equation developed by regression match with the actual data points. Also, there is a fact that is accepted by the majority of researchers and it is that the best model is the one that is the simplest possible satisfying the requirements (i.e. reflects all known contributing parameters). In other words, if there are models developed to represent a system/parameter; and all those models could be considered equivalent without considering the simplicity factor; therefore, the simplest model must be picked to represent the intended system/parameter (Giustolisi & Savic, 2006). In the EPR approach in this work to achieve the best choice of model from amongst all

the ones that were developed in the process, the simplest model that included all the parameters that are proved in the literature that affect the effective stress parameter in unsaturated soils, and representing the highest possible level of accuracy (measured by the coefficient of determination parameter) was chosen.

33 models were developed in the model development process in this research using the evolutionary polynomial regression approach with the number of terms ranging between 1 to 15. Some of these models were however ruled out on the basis that the in each one of them at least one of the input parameters reported in the literature to be affecting the effective stress parameter in unsaturated soils was excluded. The remaining models were showing various performance levels in terms of accuracy (COD value) and complexity (number of terms).

Equations 6, 7 and 8 are three example models that included all input parameters in them with equation 8 being chosen as the final model after completing further analysis stages (details of further model analysis will follow).

$$\begin{aligned}\chi = {} & 3.488\times10^{-11}(\sigma_3-u_a)^{0.5}+1.775h_b^{-2}-0.479h_b^{-0.5}.\theta_r^{0.5}-23.334\theta_s^{-1}.\lambda^{-0.5}+180.694\theta_s^{-1} \\ & +1.405\theta_s^{0.5}-0.016(u_a-u_w)^{0.5}.\theta_s^{0.5}+0.002(u_a-u_w) \\ & +9.302\times10^{-6}(u_a-u_w)\theta_r^{0.5}.\theta_s.\lambda^{-0.5}-1.975\times10^{-7}(u_a-u_w)^2h_b^{0.5} \\ & +6.657\times10^{-4}(\sigma_3-u_a)^{0.5}\lambda^{-2}+0.039(\sigma_3-u_a)^{0.5} \\ & -2.224\times10^{-6}(\sigma_3-u_a)^2\lambda^{-0.5}-11.477\end{aligned} \quad (6)$$

$$\begin{aligned}\chi = {} & 2.198\times10^{-10}(\sigma_3-u_a)^{0.5}.\theta_s^{0.5}+1.867.h_b^{-2}+0.496h_b^{-0.5}.\theta_r^{0.5}-23.176\theta_s^{-1}.\lambda^{-0.5} \\ & +167.482\theta_s^{-1}+1.332\theta_s^{0.5}-0.016(u_a-u_w)^{0.5}.\theta_s^{0.5}+0.002(u_a-u_w) \\ & +9.215\times10^{-6}(u_a-u_w)\theta_r^{0.5}.\theta_s.\lambda^{-0.5}-1.949\times10^{-7}(u_a-u_w)^2h_b^{0.5} \\ & +1.61\times10^{-4}(\sigma_3-u_a)^{0.5}\lambda^{-3}+0.04(\sigma_3-u_a)^{0.5}-2.291\times10^{-6}(\sigma_3-u_a)^2\lambda^{-0.5} \\ & -10.681\end{aligned} \quad (7)$$

$$\chi = 1.083\times10^{-3}.\lambda^{-4} + 0.197\theta_s^{0.5} - 7.351\times10^{-3}.h_b^{0.5}\lambda^{-2} + 0.065h_b^{0.5} - 0.077(u_a - u_w)^{0.5} + 1.204\times10^{-3}(u_a - u_w) + 7.014\times10^{-13}(u_a - u_w)^2\theta_r^4\lambda^{-3} + 1.419\times10^{-3}(\sigma_3 - u_a) - 1.142\times10^{-6}(\sigma_3 - u_a)^2\theta_r^{0.5} - 0.29 \quad (8)$$

Table 2 shows, for equations 6, 7 and 8, their corresponding coefficient of determination values calculated using Equation 5 as an indication of the level of accuracy of the models in making predictions of the output parameter (effective stress parameter in unsaturated soils) for both training and testing data as well as some other criteria applied to choose a final model. It presents the selected model with the optimum combination of COD values for testing and training data along with two other models included as examples of either highly complex models (with too many terms in the polynomial structure) or the ones with low accuracy levels and/or generalisation capabilities (poor performance in terms of accuracy of predictions for unseen data – low testing data COD values).

At the end of the model development process using the training data a validation process began which was completed by implementing the testing data set which was kept completely unseen by EPR during the model development process with the aim of creating the potential for a robust testing for the model to examine its generalisation capabilities to unseen cases of data. This was done to test the models developed to see to what extent exactly these models are capable of generalising the training data to the cases not experiences (seen) by EPR before. Two COD values were calculated for every one of the 33 developed models (values for 3 example models are shown in Table 2).

As it can be easily seen in Table 2, equations 6 and 7 are more complex than the selected final model (equation 8) and at the same time also show much lower coefficient of

determination values particularly for the testing/verification data. It can be seen that also equation 6 for instance is showing higher coefficient of determination value for the training data compared to the selected model; however, on the other hand, is more complex than other models and more importantly shows a very law COD for the training data suggesting low level of generalisation capabilities. Equation 8 however, shows a very high COD value for training data (although slightly lower than equations 6 and 7) and is much less complex (showing fewer terms) than equations 6 and 7, and more importantly presents an excellent generalisation capability by outperforming the other two models (equations 6 and 7) in this by a very high margin. In other words, although models presented in the form of Equations 6 and 7 are showing relatively higher training COD values; however, their generalisation capabilities to unseen cases of data (represented by COD [testing] values) are significantly lower compared to equation 8 which shows a reasonable balance in accuracy levels for predictions made based on both training and testing data as well as the number of terms (relatively lower level of complexity due to smaller number terms in the polynomial structure).

At the end of the model development process a combination of all criteria considered leads equation 8 to be chosen as the final model through EPR model development and analysis process to represent the effective stress parameter in unsaturated soils. This model can be considered as the strongest and most balanced one amongst the 33 models developed in the modelling process in terms of its prediction generalisation capabilities to unseen data, which is considered a crucial characteristic for a model as an indication of being able to be applied to the cases that data connected to them are not previously included in the training phase of the

model in any shape or form, level of complexity and also capability in engaging all know contributing parameters in the developed model.

Figures 2 and 3 present comparisons made between the EPR model (Equation 8) predictions for the effective stress parameter in unsaturated soils and the actual results obtained from the experiments (based on literature data - detailed in an earlier section of the paper dedicated to data preparation) for training and testing datasets , respectively.

Coefficient of determination parameter is a simple, accurate and powerful means of measuring the accuracy levels of models developed using machine learning approaches and have been used in various research works including recent publications (Ahangar Asr et al, 2018 & 2015; Ahangar Asr and Javadi, 2016; Javadi and Rezania 2009; Rezania et al, 2008). Table 2 shows the performance of the selected EPR model in this research in comparison to two other example models which were ruled out due to complexity as well as law testing/verification/generalisation coefficient of determination (COD) values. The COD values for the training and testing data for the selected model were 86 and 74 percent respectively. Ajdari et al (2012) developed a black/grey box adaptive learning Artificial Neural Network (ANN) based model for predicting the effective stress parameter in unsaturated soils. The proposed network was a multilayer perceptron network with six neurons in the input layer representing the air entry value, the volumetric water content at residual and saturated conditions, the slope of soil water characteristic curve, the net confining stress and suction, similar to the set of input parameters introduced in this research work. They used R-Squared (R^2) parameter to determine the accuracy level for their model. R^2 value for the ANN training

and testing data were presented as 0.96 and 0.75 respectively.

Figure 4 presents a comparison between the performances of the ANN and EPR models for the testing data that were unseen for EPR during the model development process with the aim of measuring the generalisation capabilities of the model. It can be clearly seen that the models perform similarly with the EPR model performing slightly better in some of the data points which is in line with the similar performance measurement parameter (R2 and COD) values of 0.75 and 74% for ANN and EPR models respectively. This could be considered as a clear testimony to the high capability of the COD parameter as a measure to reflect the accuracy level of models including the EPR model developed in this research work.

The comparison of the performances of ANN and EPR models (Figure 4) also reveals that the proposed EPR model is not only capable of making accurate predictions of the effective stress parameter in unsaturated soils but also has a clear advantage over the ANN model and it is the fact that EPR - unlike ANN that provides a black/grey box model - presents an explicit polynomial model that is accessible to the user to be able to scrutinise the model in ensuring that all known contributing parameters are playing a part and form and level of contribution of every contributing parameter could easily be examined by conducting the sensitivity analysis on the model.

5. Sensitivity analysis

To be able to confirm the capabilities of evolutionary polynomial regression model developed in this research with the aim of predicting the effective stress parameter in unsaturated soils, an analysis was conducted to measure the sensitivity of the developed model to contributing

parameters with known contribution level and patterns to the output. The aim of this analysis was to ensure that the model has been able to adequately and correctly learn the relationship between the contributing parameters and the output (effective stress parameter for unsaturated soils). To provide a better understanding on how important parameters affect the model, well-recognised variations of the predictions made by the model in relation to key contributing parameters (with previously recognised behavioural patterns) were considered. The proposed approach not only has the advantage of analysing sensitivity to two parameters on a single graph, but also helps provide clearer interpretations of conducted sensitivity analyses (Ajdari et al, 2012). Figures 5 and 6 show the variations in the effective stress parameter in relation with changing net mean stress and bubbling pressure respectively considering the effect of suction (Figure 5) and net mean stress (Figure 6) whilst remaining parameters in both analyses were set to an average/constant value to ensure representativeness, credibility and clarity.

For the sensitivity analysis to have understandable meaning a comparison between Soil Water Characteristic Curves presented by Bishop and Blight (1963) and Khalili et al (2004) was used. The comparison revealed that significant changes in the bubbling pressures of the soils happen in relation with the varying effective stress parameter whilst other Soil Water Characteristic Curve (SWCC) parameters - which were subjects of the comparisons too – do not vary noticeably (Ajdari et al, 2012). Therefore, the sensitivity analysis of the effective stress parameter to the changes in the bubbling pressure was conducted and presented in Figure 5 which shows variations in the effective stress parameter obtained from the selected EPR model (Equation 8) for various net mean (confining) stress values as the bubbling pressure

changes. It is clear that effective stress parameter increases with the air entry value which is a confirmation that the developed model is able to reproduce the expected trend correctly (Ajdari et al, 2012). Figure 6 shows changes in the effective stress parameter against the net confining pressure (mean stress) calculated using the developed selected EPR model whilst other soil-water characteristic curve parameters were kept unchanged for different suction values. The presented results demonstrate that the expected trends has been correctly captured by the developed/chosen final evolutionary polynomial regression model for the effective stress parameter in unsaturated soils and the model behaviour is consistent with the expected form of variations in the predictions form literature as the analysed parameters change (Ajdari et al, 2012; Johari, 2006).

6. Conclusions

Evolutionary Polynomial Regression was used to develop models to predict the effective stress parameter in unsaturated soils considering six contributing input parameters. Experimental triaxial test data from literature was used to develop and validate the models in this study. From amongst the developed models one was selected based on robustness and complexity factors (the most robust and least complex model with highest possible generalisation capabilities was chosen). After training, the generalization capabilities of the selected model were evaluated by verification of its performance using a set of data which was kept unseen to EPR during the model development process. The results revealed that the proposed model was efficient and robust in successfully capturing the complicated underlying relations between the contributing parameters and predicting the effective stress parameter in unsaturated soils directly from a set

of raw experimental measurements to a very good level of accuracy.

A sensitivity analysis was also conducted to investigate the capability of the model in capturing the expected behavioural trends in relations between the unsaturated soil effective stress parameter and the key contributing parameters. The sensitivity analysis results confirmed that the expected behaviour (approved by previous knowledge from literature) have been successfully picked up by the developed model and reflected in the predictions made.

A comparison was made between the performance of the EPR model developed in this research and an Artificial Neural Network based model from literature. The results confirmed strong capabilities of the proposed model in making accurate predictions to similar levels done by the ANN model with the exceptional advantage of being presented in the form of an explicit and easy to interpret polynomial model providing a clear insight into the connections/intrinsic relations between input parameters and the output, the unsaturated soil effective stress parameter.

Another interesting feature of EPR approach is that as more data becomes available, the quality of the model predictions can be improved by retraining EPR with the newly available more comprehensive set of data. This feature highlights the flexibility and strength of the methodology in being able to be stretched to include newly generated data in developing stronger more accurate models.

References

A. Ahangar Asr, A. Faramarzi & A. A. Javadi. (2018). An evolutionary modelling approach to predicting stress-strain behaviour of saturated granular soils. Engineering Computations: International Journal for Computer-Aided Engineering, DOI (10.1108/EC-01-2018-0025)

Ahangar Asr, A., Javadi, A. A., & Khalili, N. (2015). An evolutionary approach to modelling the thermomechanical behaviour of unsaturated soils. International Journal for Numerical and Analytical Methods in Geomechanics, 39(5), 539-557.

Ahangar Asr, A., & Javadi, A. (2016). Air losses in compressed air tunneling: a prediction model. Proceedings of the ICE-Engineering and Computational Mechanics, 169(3), 140-147.

Ajdari, M., Habibagahi, G., & Ghahramani, A. (2012). Predicting effective stress parameter of unsaturated soils using neural networks. Computers and Geotechnics, 40, 89-96.

Arvin, M., Veiskarami, M., Ajdari, M., & Habibagahi, G. (2007). A percolation approach to determination of effective stress parameter in unsaturated soils. Paper presented at the International Conference on Unsaturated Soils. In Yin ZZ, Yuan JP, ACF Chiu (eds) Proceeding 3rd Asian Conference on unsaturated soils, Nanjing, China, Science Press, Beijing.

Bishop, A. W., & Blight, G. (1963). Some aspects of effective stress in saturated and partly saturated soils. Geotechnique, 13(3), 177-197.

Cuisinier, O., Javadi, A. A., Ahangar-Asr, A., & Masrouri, F. (2013). Identification of coupling parameters between shear strength behaviour of compacted soils and chemical's effects

with an evolutionary-based data mining technique. Computers and Geotechnics, 48, 107-116.

Fazeli, A., Habibagahi, G., & Ghahramani, A. (2009). Shear strength characteristics of Shiraz unsaturated silty clay. Iranian Journal of Science and Technology, 33(B4), 327.

Fredlund, D. G., Xing, A., Fredlund, M. D., & Barbour, S. (1996). The relationship of the unsaturated soil shear to the soil-water characteristic curve. Canadian Geotechnical Journal, 33(3), 440-448.

Garven, E., & Vanapalli, S. (2006). Evaluation of empirical procedures for predicting the shear strength of unsaturated soils Unsaturated Soils 2006 (pp. 2570-2592).

Giustolisi, O., Doglioni, A., Savic, D.A., di Pierro, & F. (2008). An evolutionary multi-objective strategy for the effective management of groundwater resources. Water Resources Research Journal 44. doi:10.1029/2006WR005359

Giustolisi, O., & Savic, D. A. (2006). A symbolic data-driven technique based on evolutionary polynomial regression. Journal of Hydroinformatics, 8(3), 207-222.

Hussain, M. S., Javadi, A. A., Ahangar-Asr, A., & Farmani, R. (2015). A surrogate model for simulation–optimization of aquifer systems subjected to seawater intrusion. Journal of Hydrology, 523, 542-554.

Javadi, A.A. & Rezania, M. (2009). Intelligent finite element method; an evolutionary approach to constitutive modelling. Journal of Advanced Engineering Informatics 23(4),442–451.

Johari, A., Habibagahi, G. & Ghahramani, A. (2006a). Prediction of soil–water characteristic curve using genetic programming. Journal of Geotechnical and Geo-environmental

Engineering (ASCE) 5,661–665.

Johari, A., Habibagahi, G. & Ghahramani, A. (2006b). Prediction of soil–water characteristic curve using a genetic based neural network. Scientia Iranica 13 (3), 284–294.

Johari A. (2006). Predicting soil water characteristic curve using artificial intelligence. PhD. Thesis, Shiraz University, Shiraz, Iran

Kayadelen, C. (2008). Estimation of effective stress parameter of unsaturated soils by using artificial neural networks. International Journal for Numerical and Analytical Methods in Geomechanics, 32(9), 1087-1106.

Khalili, N., Geiser, F., & Blight, G. (2004). Effective stress in unsaturated soils: review with new evidence. International journal of Geomechanics, 4(2), 115-126.

Khalili, N., & Khabbaz, M. (1998). A unique relationship of chi for the determination of the shear strength of unsaturated soils. Geotechnique, 48(5).

Lee, I.-M., Sung, S.-G., & Cho, G.-C. (2005). Effect of stress state on the unsaturated shear strength of a weathered granite. Canadian Geotechnical Journal, 42(2), 624-631.

Lee, S., Lee, S., & Kim, Y. (2003). An approach to estimate unsaturated shear strength using artificial neural network and hyperbolic formulation. Computers and Geotechnics, 30(6), 489-503.

Loret, B., & Khalili, N. (2002). An effective stress elastic–plastic model for unsaturated porous media. Mechanics of Materials, 34(2), 97-116.

Öberg, A., & Sällfors, G. (1997). Determination of shear strength parameters of unsaturated silts and sands based on the water retention curve.

Rahardjo, H., Heng, O. B., & Choon, L. E. (2004). Shear strength of a compacted residual soil from consolidated drained and constant water content triaxial tests. Canadian Geotechnical Journal, 41(3), 421-436.

Rassam, D. W., & Williams, D. J. (1999). A relationship describing the shear strength of unsaturated soils. Canadian Geotechnical Journal, 36(2), 363-368.

Rezania, M., Javadi, A.A. & Giustolisi, O. (2008). An evolutionary-based data mining technique for assessment of civil engineering systems. Journal of Engineering Computations 25(6),500–517.

Russell, A., & Khalili, N. (2006). A unified bounding surface plasticity model for unsaturated soils. International Journal for Numerical and Analytical Methods in Geomechanics, 30(3), 181-212.

Russell, A. R., & Khalili, N. (2004). A bounding surface plasticity model for sands exhibiting particle crushing. Canadian Geotechnical Journal, 41(6), 1179-1192.

Vanapalli, S., Fredlund, D., Pufahl, D., & Clifton, A. (1996). Model for the prediction of shear strength with respect to soil suction. Canadian Geotechnical Journal, 33(3), 379-392.

Xu, Y. (2004). Fractal approach to unsaturated shear strength. Journal of Geotechnical and Geoenvironmental Engineering, 130(3), 264-273.

Zargarbashi, S., & Khalili, N. (2011). Discussion of “Shear Strength Equations for Unsaturated Soil under Drying and Wetting” by Goh Shin Guan, Harianto Rahardjo, and Leong Eng Choon. Journal of Geotechnical and Geoenvironmental Engineering, 137(12), 1310-1313.

Table 1. Parameters involved in developing EPR models and ranges of values of the parameters as well as statistical analysis parameter values for training and testing datasets

			Training data ranges	Testing data range	Mean – Training data	Mean – Testing data	Standard deviation – Training data	Standard deviation – Testing data
Input Parameters	$u_a - u_w$	Suction	0 – 612 kPa	76 -496 kPa	138	122	131	92
	θ_r	Volumetric water content at residual condition	0 – 28.35	5.502 – 21.54	10.74	8.76	10.38	11.08
	θ_s	Volumetric water content in saturated condition	23.82 -55.95	25.43 - 52	37.19	43.18	9.26	8.42
	λ	Soil-water characteristic curve slope	0.19 – 11.82	0.94 – 8.33	1.37	2.41	2.65	2.72
	$\sigma_3 - u_a$	Net confining stress	0 - 400 kPa	50 - 300 kPa	113	145	105	98
	h_b	Air entry value	1 – 200 kPa	27 – 125 kPa	32	38	46	28
Output Parameter	χ	Effective stress parameter	0.091 - 1	0.24 - 1	0.676	0.769	0.293	0.153

Table 2. Example models and the final selection based on inclusion of expected input parameters, performance, and complexity

Model (Polynomial equation)	Performance (Coefficient of Determination)		Includes all expected contributing parameters	Complexity (number of terms in the equation)	Selected as final model
	COD% [Training]	COD% [Testing]			
6	90.24	16.82	YES	13	NO (poor generalisation capability – low testing COD and higher complexity)
7	82.92	6.52	YES	14	NO (poor generalisation capability – low testing COD and higher complexity)
8	85.83	74.05	YES	10	YES (Best performance with regards to generalisation capability; optimum/balanced number of terms in relation with performance)

Accepted manuscript
doi: 10.1680/jsmic.21.00012

Figure 1. Flow diagram for EPR procedure

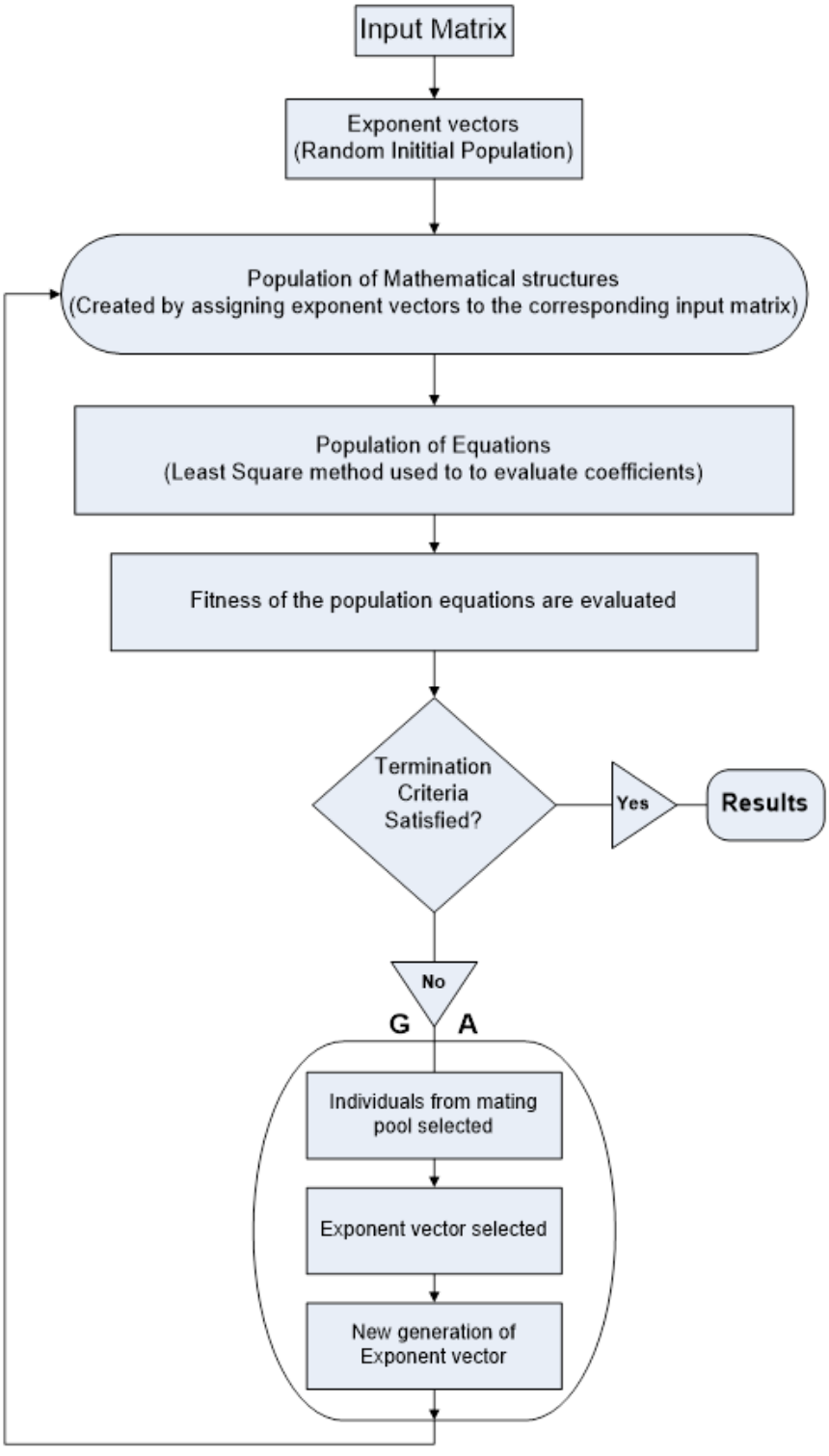
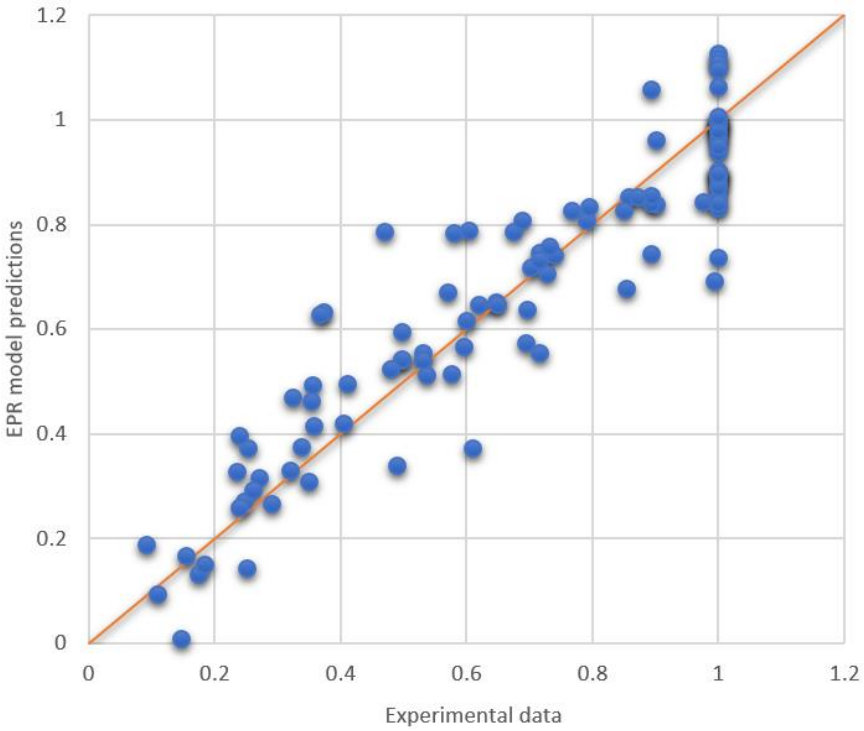
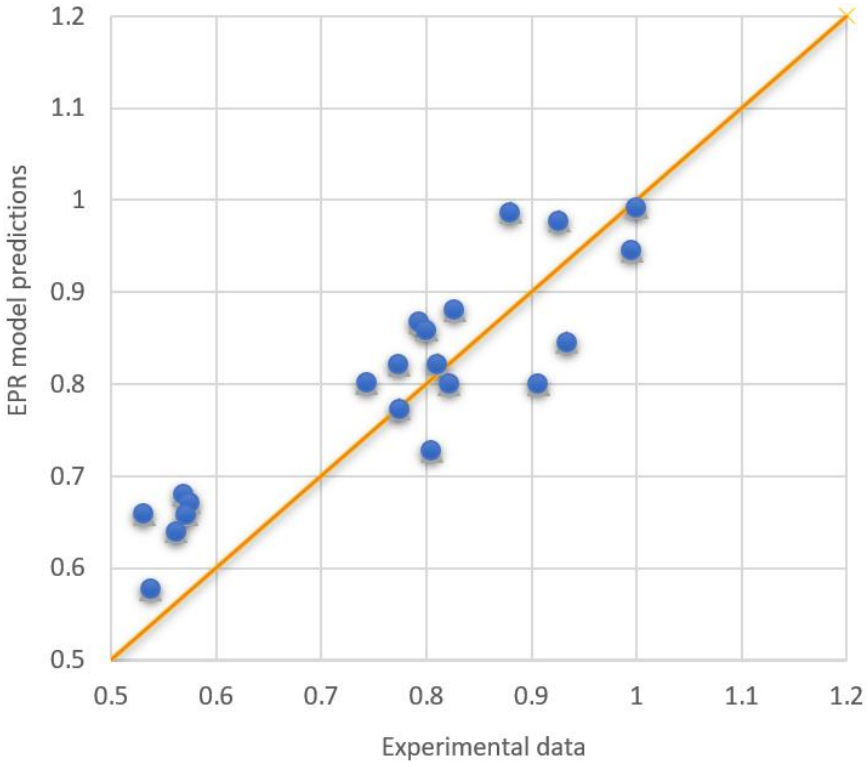


Figure 2. EPR model predictions against the experimental data for unsaturated soil effective stress parameter (Training data) - selected model [Equation 8]



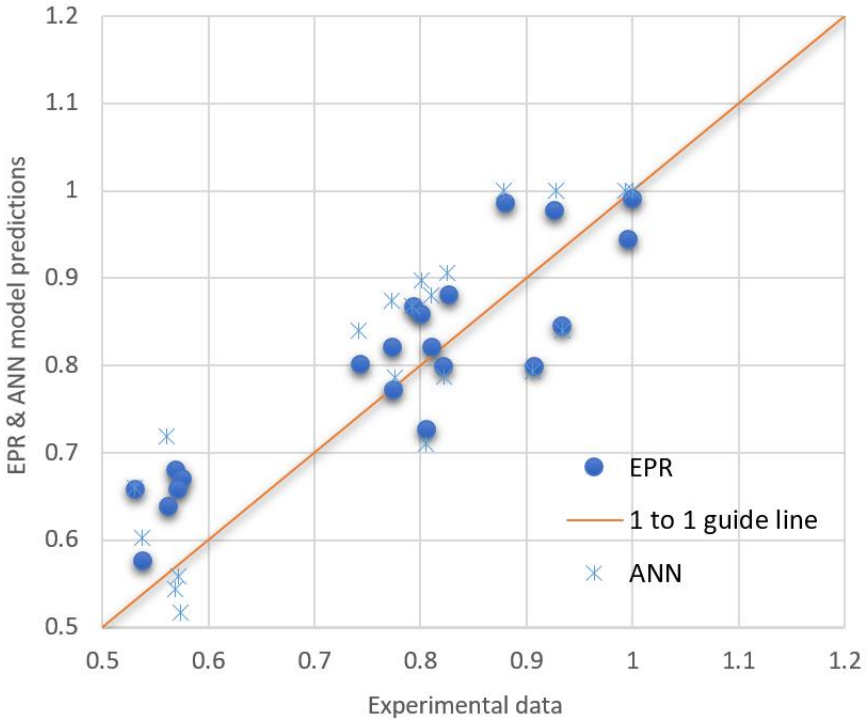
Accepted manuscript
doi: 10.1680/jsmic.21.00012

Figure 3. EPR model predictions against the experimental data for unsaturated soil effective stress parameter (Testing/Verification data) - selected model [Equation 8]



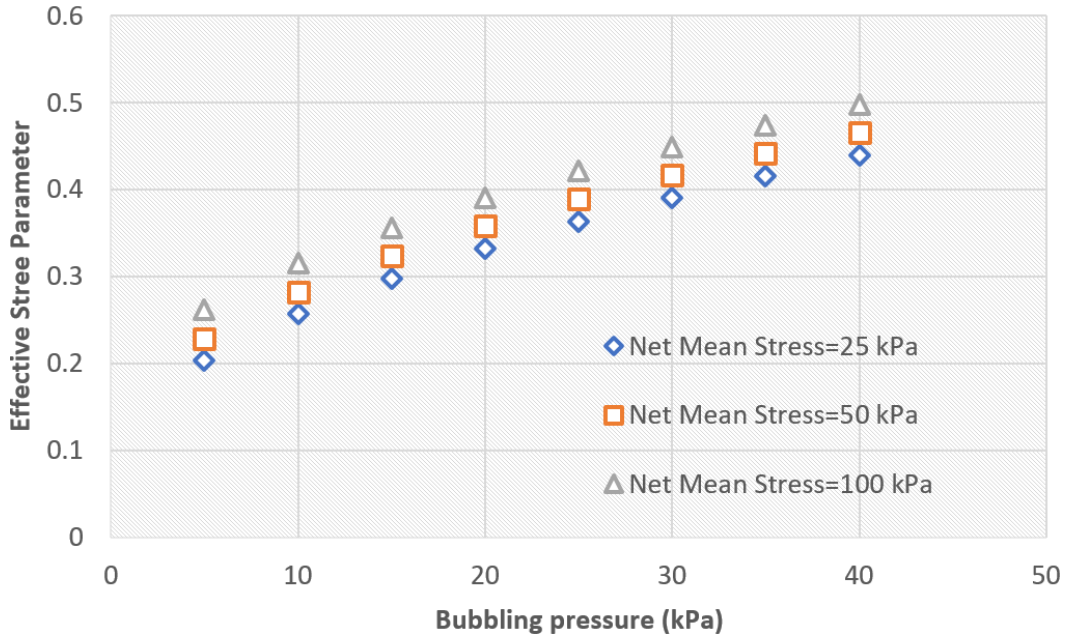
Accepted manuscript
doi: 10.1680/jsmic.21.00012

Figure 4. Comparison between EPR and ANN model predictions for unseen EPR validation / generalisation data



Accepted manuscript
doi: 10.1680/jsmic.21.00012

Figure 5. Changes in the effective stress parameter in relation to the bubbling pressure (suction=200 kPa, Volumetric water content at residual condition=20, Volumetric water content in saturated condition=35, Soil-water characteristic curve slope=1)



Accepted manuscript
doi: 10.1680/jsmic.21.00012

Figure 6. Changes in the effective stress parameter in relation to the net confining pressure (bubbling pressure=50 kPa, Volumetric water content at residual condition=0, Volumetric water content in saturated condition=45, Soil-water characteristic curve slope=0.1)