Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering– Rahman and Jaksa (Eds)

Prediction of the hydraulic conductivity based on the multivariate regression within machine learning

Prédiction de la conductivité hydraulique basée sur la régression multivariée en apprentissage automatique

Han-Saem Kim
Earthquake Research Center, Korea Institute of Geoscience and Mineral Resources, Korea, adoogen@kigam.re.kr

Hyun-Ki Kim
Department of Civil and Environmental Engineering, Kookmin University, Korea

ABSTRACT: This paper presents the non-stochastic regression modeling and its performance evaluation for predicting the hydraulic conductivity of sandy soils using a wide range of related index properties. The procedure was proposed with data preprocessing, modeling of regression algorithms, and optimization of models, and uncertainty estimation. The empirical data-dependent trends of the hydraulic conductivity with pore structure characteristics are derived by explicit model parameters. To recognize the prevailing relations between the hydraulic conductivity and multivariate influential index properties, the regression modeling in machine learning suggest the best combination of index properties and uncertainties depends on the prediction model performance. This study includes compilation of regression model and its tuning methods for the optimization of parametric representation. The prediction results highlight the best-fitting model and parameter combination having the lowest residuals.

RÉSUMÉ : Cet article présente la modélisation de la régression non stochastique et son évaluation de la performance pour prédire la conductivité hydraulique des sols sableux en utilisant un large éventail de propriétés d'indice connexes. La procédure a été proposée avec le prétraitement des données, la modélisation des algorithmes de régression, l'optimisation des modèles et l'estimation de l'incertitude. Les tendances dépendantes des données empiriques de la conductivité hydraulique avec les caractéristiques de la structure des pores sont dérivées par des paramètres de modèle explicites. Pour reconnaître les relations dominantes entre la conductivité hydraulique et les propriétés d'indice d'influence multivariée, la modélisation de régression en apprentissage automatique suggère que la meilleure combinaison de propriétés d'indice et d'incertitudes dépend des performances du modèle de prédiction. Cette étude comprend la compilation du modèle de régression et ses méthodes de réglage pour l'optimisation de la représentation paramétrique. Les résultats de la prédiction mettent en évidence le modèle le mieux adapté et la combinaison de paramètres ayant les résidus les plus bas.

KEYWORDS: hydraulic conductivity; regression model; machine learning; soil index properties; sand

1 INTRODUCTION.

Many complexities and uncertainties in geotechnical engineering related to the uncertainty of coherent soil composition, errors in in-situ and laboratory testing, and characterization of the index properties of geomaterials. The treatment of such problems have been simplified by the classical/conventional models of engineering modeling approaches. For the precise site characterization and management as well as geotechnical reliable design using large scale database, the accurate and multivariate regression models are essential. Especially, the generic modeling based on the artificial intelligence (AI) for predicting the engineering properties such as hydraulic conductivity (k) is appropriate when known the principal basic properties (e.g., void ratio, specific surface) and mathematical models. In this analysis, learning models were evaluated to predict the hydraulic conductivity and to provide a reliable regression models and principal features about index properties.

The hydraulic conductivity of a sediment, which is difficult to measure, is recognized to be associated with its pores structure (i.e., particle size distribution). The classical hydraulic conductivity (k) measurement equation of particle size data from Hazen (1911), in which the k was presented proportional to the squared grain size at 10% passing, it may be written as:

$$k = C_H(g/\mu)d_{10}^2 \quad (1)$$

where g is the gravitational acceleration; d_{10} denote the grain size at 10% passing; C_H is a coefficient about 6.54×10^{-4} (Harleman et al., 1963). The Kozeny-Carman's equation considers the sediment pore network as a bundle of tubes and assumes the laminar fluid flow of poiseuille in the tubes (Ren and Santamarina, 2018), and are as follows:

$$k = C_F\left(1/{S_S}^2\right)(\gamma_w/\mu{\rho_m}^2)(e^3/1+e) \quad (2)$$

where C_F is a dimensionless shape constant, with a value about $C_F \approx 0.2$ (Taylor, 1948); γ_w is unit weight of fluid (N/m^3); ρ_m (kg/m^3) is particle density of soil; μ (N·s/m^2) is fluid kinematic viscosity.

With the increasing availability of hydraulic conductivity database in respect to various soil index properties, which can be acquired effectively and proximally and open source algorithms freely available, machine learning techniques for soil analysis have been used quickly (Padarian et al., 2020). There are challenges applying the AI to the soil characterization: 1) sensitivity of the learning models; 2) input parameters for prediction; 3) lack of empirical case studies; 4) uncertainty of standard validation methods; 5) handling of missing data; 6) small size data for model training (Sharma et al., 2021). To solve this problem, several researches have tried to predict hydraulic conductivity using AI models such as artificial neural networks (Arshad et al., 2013; Tayfur et al., 2014; Sedaghat et al., 2016), fuzzy neural network (Arshad et al., 2013) and support vector machine (SVM) technique (Lamorski et al., 2008; Das et al.,

2012). Thus, measuring the hydraulic conductivity regarding with pore structures characteristics, the construction and refinement of the large scale database archived from verified data source and the robust optimization approach of machine learning algorithms should be considered.

In this study, the various regression models were trained to optimize the best-fitting model for predicting sandy soil's hydraulic conductivity using the relations with multivariate index/basic soil properties (Figure 1). Firstly, the hydraulic conductivity database was refined considering the major trends of relations of the hydraulic conductivity versus index properties. Second, the preprocessing procedures were performed because the original data from source have inconsistencies and errors, prior to regression modeling. In this phase, the index properties are composed of the combinations of principal components. Third, the best-fitting regression algorithms were modeled based on the K-fold cross-validation. For exploring the dataset, K-Fold cross validation was used as testing methodology. K-fold cross validation avoids overlapping by splitting data into K subsets and makes K iterations. For each iteration, a different subset was chosen for testing and the remainder for training. We picked K=10 because this value is considered appropriate to obtain an accurate estimation. There are uncertainty in prediction accuracy based on the trained model on certain data, due to the biased input relations of hydraulic conductivity with index properties. Finally, the uncertainties with regard to other relations from test datasets were evaluated. The classical model of k with index properties are also applied for comparing with the proposed fitting models.

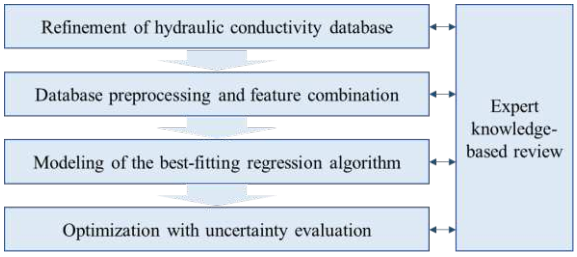

Figure 1. Conceptual procedure of machine learning for predicting k with index properties.

2 HYDRAULIC CONDUCTIVITY DATABASE OF SOIL INDEX PROPERTIES

The hydraulic conductivity (k) database containing index properties of sandy soil was used from Ren and Santamarina (2018)'s hydraulic conductivity database, which including natural and remolded sediments (gravels, sands, silts, and clays). There are 6,952 relations (data points) between k and index properties for 92 soils. In particular, the major components in the database were void ratio (2,879 points), specific surface (2,080), and percentage by weight passing the #200 sieve (1,360 points). In this study, for predicting k, there are six index properties: percentage by weight passing the #200 sieve (P#200); particle size distribution (D_{10}, D_{50}, D_{60}); specific surface (S_S); void ratio (e). Even though there are many relations, data are not evenly distributed with properties. Thus, target dataset having the similarity (cluster) of sand's relations were selected (Figure 2).

The hydraulic conductivity generally increases with increasing void ratio for silt, sand, and gravel. Although a plot of k versus $e^3/(1+e)$ must be a straight line that pass through the origin at the Kozeny-Carman relation (Equation 1), experimental studies in sandy soils do not always support such a linear association. There are two explicit groups corresponding to the range and increasing ratio of k with $e^3/(1+e)$. The R^2 of the linear relations using all data points is 0.6. The train-test split was conducted for evaluating the performance of a machine learning algorithm considering cluster in data points of k versus $e^3/(1+e)$. The major group (266 points) within ±1.5 standard deviation (σ) of the linear relation and larger k was defined as train and validation datasets. The secondary group having lower k relations was defined as test dataset. The R^2 of the linear relations increased to 0.9, only using train and validation datasets by excluding the test datasets (Figure 2). The counts of train and validation datasets (data sources: seven references), and test datasets are 266 and 75. The test datasets have the higher specific surface and void ratio than train data's index properties, despite having a higher k.

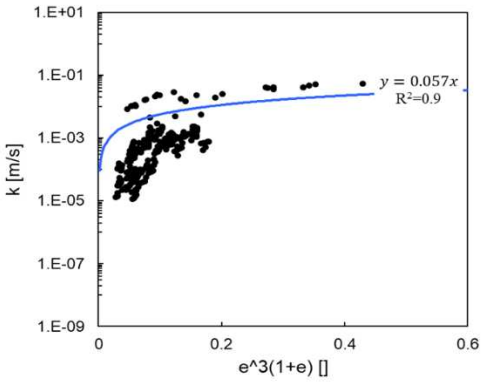

Figure 2. k and $e^3/(1+e)$ correlations (266 points) excluded the points far away from the central trend by 7 reference.

3 MODELING OF MACHNIE LEARNING REGRESSION

3.1 *Data preprocessing of the hydraulic conductivity database*

The selected index properties excluded the out-trending datasets have still atypical (i.e., null data, not labelled data, etc.) for directly applying the machine learning algorithms. Even though there are major relations between k and e, the multivariate correlations with other index properties influence the prediction accuracy of the trained models. The data preprocessing method is the first step to begin the process when it comes to building a machine learning model. The data preprocessing is also a method used for the conversions of raw data into a clean datasets. If the datasets is obtained from multiple sources, it is collected in a raw format that cannot be analyzed. In machine learning, this procedures are an integral step as the consistency of data and the geotechnical expert knowledge obtained directly affects the capacity of our algorithm to learn. Thus, before we input it in our model, it is highly necessary that we preprocess our data. In machine learning, there are five essential stages in data preprocessing: 1) handling null (missing) values; 2) scaling; 3) one-hot encoding; 4) feature selection; 5) splitting the dataset into training dataset and test dataset (Garcia et al., 2015).

There are always few null values in any real-world dataset. If it is a regression and classification, it doesn't matter whether any model can manage these values on its own so we have to interfere. There are many missing data treatments including the deletion methods (likewise deletion and pairwise deletion), and imputation methods (mean imputation, hot-deck imputation, cold-deck imputation, regression imputation, etc.) (Osman et al., 2018). In this study, deletion methods, particularly likewise deletion, which is commonly used to handle the missing values as default approaches then results in many datasets being discarded in cases and bias (Raymond, 1986), applied in this modeling.

The feature scaling is a technique of data preprocessing for normalizing datasets. This is useful for optimization algorithms used as gradient descent, algorithms that use distance measurements (i.e., K-nearest neighbors), and regression and

neural networks algorithms. There are two techniques of scaling: standardization, normalization. Standardization, or whitening of a sample requires rescaling the value distribution to 0 and standard deviation to 1. Normalization consists of a rescaling of the initial datasets such that all values are [0,1] or [-1,1], denoted as min-max scaling. As seen following equation (Angelov and Gu, 2019), the target of scaling are usually used for intervals of [0,1] and [-1,1].

Normalization:

$[0,1]\ interval: X' = (X - X_{min})/(\mathrm{X}_{max} - \mathrm{X}_{min})$ (3)

$[-1,1]\ interval: X' = [X - (X_{max} + X_{min})/2]/[(X_{max} - X_{min})/2]$ (4)

Standardization: $X' = (X - \mu)/\sigma$ (5)

Here, X_{max} and X_{min} are the maximum and the minimum values of the feature respectively. μ is the mean of the feature values and σ is the standard deviation of the feature values. Scaling transforms the characteristic value according to the Equations 3-5, which allows the same amount of control to be exercised by all scaled characteristics (Angelis and Stamelos, 2000) and thus immune to unit choice (Kosti et al., 2012). For any feature, the majority of algorithms use index encode, whereby the index code has a particular identifier for every feature string (Mishra et al., 2019).

The purpose of the collection of functions is to select the best subset of features for building models that have significant impact onto the predicting performance. High-dimensional databases nevertheless have irrelevant, noisy and redundant characteristics. The aim of reducing dimensionality is to reach optimized features faster, since the bigger the data size the slower the function optimization (Xue et al., 2016). The strategies for choosing principal features in the following categories can be categorized widely: filter methods, wrapper methods, embedded methods, hybrid methods (Ferreira and Figueiredo, 2012). Among these, the filter methods used to take the intrinsic characteristics of the characteristics calculated through univariate statistics rather than cross-validation. In the refined database of soil index properties, there are ln(k) value as label attribute and other original features: e, S_S, D_{10}, D_{50}, D_{60}, P#200. To evaluate the predicting performance depending on the feature combinations using the filter methods, the seven combinations having low to high dimensionality were defined (Table 1).

Table 1. Input combination and experimental results obtained while comparing the combinations of input attributes.

Combination	Input attributes	Number of data	Number of average epochs	Average runtime (min)
C#1	e	341	21	1
C#2	e, S_S	266	453	2
C#3	D_{10}, D_{50}, D_{60}	23	562	3
C#4	e, D_{10}, D_{50}, D_{60}	266	432	8
C#5	S_S, D_{10}, D_{50}, D_{60}	36	124	4
C#6	e, S_S, D_{10}, D_{50}, D_{60}	266	456	2
C#7	e, S_S, D_{10}, D_{50}, D_{60}, $P_{\#200}$, $e^3/(1+e)$	266	231	9

Each dataset must be divided into two distinct sets for the machine learning model composed of the training sets and testing sets. Since there are various number (23~341) of datasets with combinations (Table 1), the resample procedure such as K-fold cross-validation is necessary to build the best-fitting model on a limited data sample, excluded the performance effected by the constant dataset split. K-fold cross-validation eliminates the overlap of data division into K-folds and generates of K. Firstly, the datasets are shuffled randomly. If a certain value for K is determined, it can be used instead of K in the model relations so that K=10 is 10 times cross-validation. Each K-folds were permitted to be used as an end-of-the-line test datasets, and all other folds are used as a training datasets together.

3.2 *Modeling regression algorithms*

In this study, regression algorithms for predicting k value are compared with methods based on cost functions. In respect of the artificial intelligence, we designed the representative six regression algorithms: linear regression (LR), K-nearest neighbors (KNN), decision tree (DT), random forest (RF), support vector regression (SVR), multilayer perceptron (MLP).

The LR model is composed of the variable predictor and a variable dependent on each other linearly. We used the linear stepwise regression algorithms (Tibshirni et al., 2013), assuming that the linear relations between the ln(k) and combined feature groups (Table 1). Stepwise regression is a way to pick principal variables to obtain an interpretable model. In this algorithms, for each possible predictor, the one with the highest modified R2 is the starting point from the null model for a univariate linear regression model. Linear regression results in multiple linear regressions with less-squares, including selection of a combination, either by greedily reversing or by constructing a complete model from all the attributes and decreasing one-by-one terms from their uniform coefficient before an interruption criterion has been met. For LR with one combination (or feature), the object function is as follows:

$\mathrm{E}(y|x) = a + w_1x_1 + w_2x_2 + \cdots + w_nx_n + e$ (6)

$\mathrm{Min} \sum_{i=1}^{n}(y_i - w_ix_i)^2$ (7)

where y_i denote the target labeled data, w_i is the coefficient, x_i is the combination (feature), and e is the observed error. Least squares are based on the LR. Based on LR model, the equations can be derived to predict ln(k) for C#7 as follows:

$\ln(k) = -2.705P_{\#200} - 2.4355S_S + 7.70794\,e^3/(1+e) - 6.8585$ (8)

For the KNN method, the expected values are obtained as weighted averages from the values of adjacent measurements for interesting variables. One key drawback of this method is the need to choose a similarity metric sometimes ad hoc, in particular for heterogeneous datasets from which the extracted features are of various kinds and sizes and interrelated (Yao and Ruzzo, 2006). The vicinity of a given point in a high dimensional space becomes very sparse and induces a high variety. KNN algorithms are classified as two types: KNN classification and KNN regression. KNN regression is non-parametric approach which approximates in an intuitive way, integrating the observations in the same neighborhood, the relations between independent variables and continuous effects. KNN regression employs the distance functions (i.e., Euclidean distance, Manhattan distance, Minkowski distance, etc.) and an important method with less concern compared with KNN classification. In this study, Euclidean distance, only valid for continuous variables, was used as following:

$\mathrm{d}(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$ (9)

DT are a supervised non-parametric method for classification and regression learning. Classification and regression tree (CART) analysis, a type of decision tree method, is perfect for generating rules on engineering decision making, due to comparatively slowly accepted unlike other regression models (Lewis, 2000). CART are the if-then (split) specifications which allow for case prediction or classification (Razi and Athappilly, 2005). CART does not establish a probability equation, unlike logistic and linear regression. Datasets are instead partitioned in

sub-sets of homogeneous values of the dependent variable along the predictor axes and a method represented by a decision tree to forecast new observations (Krzywinski and Altman, 2017). CART is a binary recursive partitioning method that only two groups can be separated into each combination of index properties, represented in a decision tree by a node. CART can accommodate numeric or multi-modal information with an ordinary or non-ordinary structure, as well as categorical predictors.

RF is a bagging algorithm for supervised learning using a classification and regression ensemble learning system. In RF, each node is divided into a subset of randomly selected predictors using the best one. The method is a meta-estimator which combines various decision trees to require equal use of all possible predictive combination. The tree predictor assumes numerical values in contrast to RF classifier (Breiman, 1999). There are several methods to variable induction selection in the literature and the majority of approaches explicitly allocate a quality measure to the variable (Singh et al., 2017). Information gain ratio criterion (Quinlan, 1992) and Gini Index (Breiman et al., 1984) are the most widely used component selection tests. The Gini Index makes it possible to introduce larger distributions, while the information gain prefers smaller distributions with many unique values. In this study, the split criteria for regression tree is based on selecting the input variable with the smallest Gini Index:

$$I_G = 1 - \sum_{i=1}^{n} (P_i)^2 \quad (10)$$

where P_i is the likelihood of an element for a distinct class.

SVR, as a type of support vector machine (SVM), is linear or nonlinear regression method and denoted as support vector machine regression. SVM solves problems with binary classification by formulating them as problems with convex optimization (Vapnik, 1998; Dibike et al., 2001; Liong and Sivapragasam, 2002).). SVR approach uses linear quadratic programming techniques to deal with data in high dimension space (Lin et al., 2005). We strived to decrease the error rate in a LR. We try to fit the error within certain threshold during SVR. The former condition produces the objective function in equation, in which $\|w\|$ is approximated by the magnitude of the normal vector to the surface:

$$\text{Min } 1/2 \, \|w\|^2 \quad (11)$$

MLP is a static neural structure made up of layers that transmits and exchanges information by means of synaptic connections represented by weight adaptation. MLP is commonly used as an approximation method for regression functions. In MLP, a transfer function is used to transfer the weighted sum of inputs and bias terms to the activation level and the units are organized in a layered feed-forward topology (Venkatesan and Anitha 2006; Hornik, Stinchcombe, & White, 1989). A feed-forward neural network is an artificial neural network that doesn't form a cycle at a time. And each perceptron in a single layer is fully connected with all nodes. MLP modeling is typically composed of input, hidden, and output layers. In this study, the input layers for each combination (Table 1) are connected two hidden layers. The input values are weighted and generated in accordance with the activation function from the above layer (Jodouin, 1994).

As the activation function for the first and second hidden layers, rectified linear units (ReLu) and softmax are used, respectively. If the input is smaller than 0, and the raw output is different for ReLU, it has output 0 (Equation 12). ReLU has the advantage of being non-linear and has no backpropagation mistake as opposed to the sigmoid function (Li and Yuan, 2017). It does not just map output into a [0.1] range but also maps each output to the extent that the total value is 1. In the logistic regression model, softmax is used for multi-classification while sigmoid is used for binary classification in the logistic regression model, with one for softmax the number of probabilities. Softmax in which z is mathematically the input vector for the output layer and j indicates the output units (Equation 13).

$$f(x) = max(x, 0) \quad (12)$$

$$\sigma(\vec{z})_i = e^{z_i} / \sum_{j=1}^{K} e^{z_j} \text{ for } j = 1, \cdots, K \quad (13)$$

where all z_i values are input vector elements, and where any real value can be taken.

4 OPTIMIZATION OF REGRESSION ALGORITMS

4.1 *Verification of the best-fitting model and combination*

Using the six modelled regression algorithms, the seven input combinations were trained and verified with K-fold cross-validation, simultaneously. The averaged residuals between predicted and actual *k* value (Figure 3) were calculated for each cross-validation. The validation data for the model is then considered to be a single sub-sample and the remaining K-1 samplings are used as training data. Figure 3 presented the best-fitting model (blue dot) having the averaged lowest residuals for each combination after 10-fold cross-validation.

The prediction of *k* using C#1 shows generally the three or four linear clusters, which stationary estimations with measurements. Among the models, DT is the best-fitting model only considering e and *k* correlations. The relationship between *k* and e is reviewed and validated from classic geomechanics (i.e. Kozeny-Carman equation) for sandy soils. Using these regression algorithms, the high deviations of *e* influenced low prediction performance in comparison with other combinations, despite the 341 correlations.

On the contrary, the residuals in fitted *k* with C#2 are relatively low and show particularly denser cluster in application of MLP model. As Carman (1939) notes, "It is shown that the permeability of a water-saturated sand or fine powder can be calculated with considerable accuracy, if the porosity and the specific surface are known". Thus, the nonlinear correlation between *k* and combination of *e* and S_S provides better prediction performance using MLP model with much epoch (453).

When training with only grain size analysis results (D_{10}, D_{50}, D_{60}), C#3, LR is the best-fitting model, even so sparser cluster. Since the number of combination is smallest (23), the regression with K-fold cross-validation shows low performance. In addition to the grain size, the *k* value is more influenced by other parameters, such as degree of compaction, porosity and shape of the grains (Uma et al., 1989). The C#4 shows similar pattern with C#2 application. DT is best-fitting model. The *e* value, having more correlations than grain size, influenced more impact to regression model. The prediction of *k* using C#5, S_S only added to C#3, shows higher accuracy than the application C#3. Likewise, the S_S value influenced more impact to regression model. In this combination, MLP was determined as the best-fitting model.

The residuals in fitted *k* with *e*, S_S, D_{10}, D_{50}, D_{60} (C#6) are relatively low and similar with the application of C#2 and C#4. That is, *e* is the principal component for predicting *k*. Accordingly, the weighting of $e^3/(1+e)$, parameter in Kozeny-Carman relation proposed by Kozeny (Kozeny, 1927) and improved by Carman (Carman, 1937), was evaluated in C#6. As a result, the prediction of *k* was most plugged in the actual value based on the MLP model. To evaluate the performance of the model, the root mean squared error (RMSE) have been used (Figure 4). For determination of the best-fitting model, MLP and LR are show the best performance in terms of RMSE index. In addition, C#6 (*e*, S_S, D_{10}, D_{50}, D_{60}) and C#7 (*e*, S_S, D_{10}, D_{50}, D_{60}, P#200, $e^3/(1+e)$) show the best performance, which indicates

that global porosity and grain size distribution are required to be considered together for input information.

Moreover, the conventional correlations by references (Ren and Santamarina, 2018; Kozeny-Carman equation) are calculated. And the RMSE are also compared with the fitted regression models. As a results, the most of regression model's RMSEs are lower than Kozeny-Carman's RMSE, except the KNN method. The models having lower RMSE than in case of Ren and Santamarina (2018) were LR, SVR, and MLP, applying for C#1, C#2, C3, C#6, and C#7. Thus, these fitted models (LR, SVR, and MLP) were mostly verified as the better model than classical equations about i.

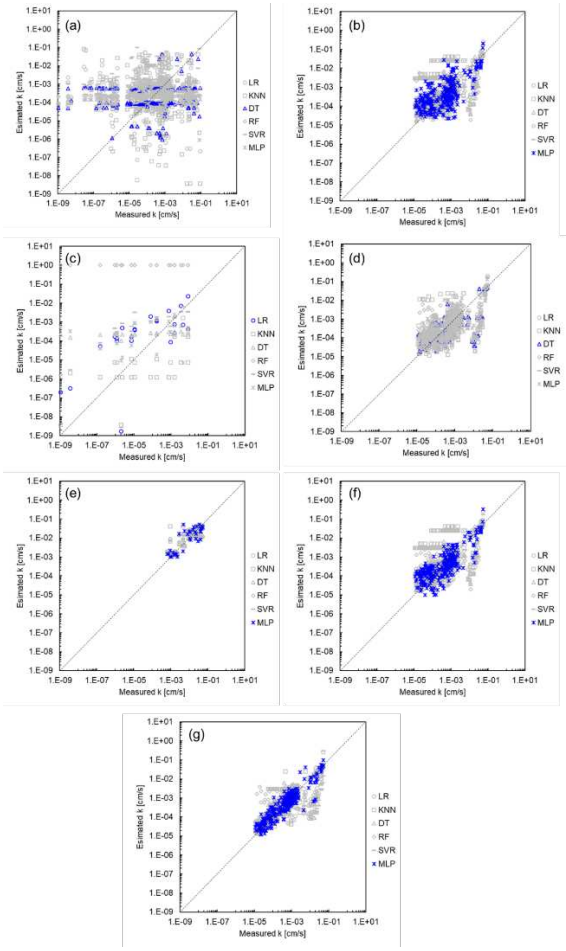

Figure 3. K-fold validation results based on the best-fitting model and input attribute combinations. Comparison between measured and predicted k with: (a) C#1; (b) C#2; (c) C#3; (d) C#4; (e) C#5; (f) C#6; (g) C#7. The blue symbol indicate the best-fitting model having the highest accuracy.

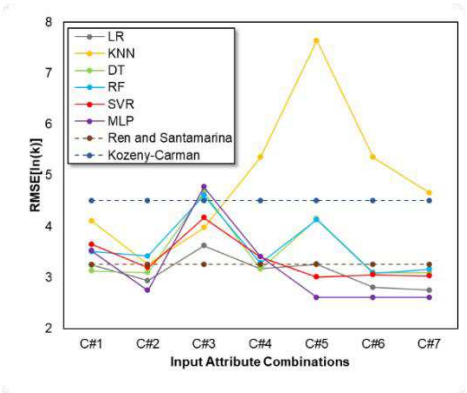

Figure 4. Prediction and validation results based on the best-fitting model and input attribute combination. Comparison of RMSE (ln(k)) according to seven combination.

4.2 *Performance evaluation with the randomized test data*

Using the six modelled regression algorithms, the seven input combinations were trained and verified with K-fold cross-validation, simultaneously. The averaged residuals between predicted and actual k value (Figure 3) were calculated for each cross-validation. The validation data for the model is then considered to be a single sub-sample and the remaining K-1 samplings are used as training data. Figure 3 presented the best-fitting model (blue dot) having the averaged lowest residuals d for each combination after 10-fold cross-validation.

The best regression model and combination of index properties for predicting k are determined in case of C#7 (e, S_S, D_{10}, D_{50}, D_{60}, P#200, $e^3/(1+e)$), after the likewise deletion method, using MLP based best-fitting model. In this study, to validate the best fitted model for the test datasets, which are out of the central trend of train and validation datasets (Figure 2), the k were predicted for test datasets. To de-trend the separated correlations between k and index properties in test datasets the large deviation of index properties according to the reference, ten datasets for each combination are randomly selected.

The best-fitting models for each combinations, having the lowest RMSE (Figure 4), are used for predicting k using ten test datasets. Since the test datasets have a relations of the relatively lower k when e is larger compared with train and validation datasets, the regression model for C#1 and C#2 predicted the lower k than measured value. The other regression results show the similar pattern with K-fold cross-validation for train datasets (Figure 3). The best model and combination of index properties in test datasets was MLP application for C#6 having the lowest average residuals.

Likewise the verification of the trained regression model, three error indices (RMSE, MSE, MAE) for each best model and combination were compared (Table 2). The k value predicted by MLP in case of C#6 has the lowest average residuals. The MLP model generally play a principal regression algorithm for predicting k. And the classical models of k of geomaterial were also computed. Then, the performance of these models are compared with the best regression model and combination. As a results, the most of average residuals applying regression models are lower than the application of classical equation.

Table 2. Metrics for ten test datasets applying the best-fitting model and the classical equation for determining k.

Combination	Best-fitting model	RMSE	MSE	MAE
C#1	Decision tree	0.34	0.12	0.23
C#2	MLP	0.74	0.55	0.36
C#3	Linear regression	0.28	0.08	0.26
C#4	Random forest	0.35	0.12	0.19
C#5	MLP	0.34	0.11	0.29
C#6	MLP	0.12	0.01	0.10
C#7	MLP	0.24	0.06	0.19
Hazen (1930)'s equation		0.89	0.66	0.29
Kozeny-Carman's equation		0.77	0.59	0.28

5 CONCLUSIONS

The hydraulic conductivity of sandy soils have been experimentally determined by stochastic functional computation of partial index properties. Multi-variables about pore structures are only used in archived resources and considered without the interactive weighting of the influential parameters. In this study, the data-driven methodology was proposed and applied for predicting the hydraulic conductivity of sandy soils using different references. The optimized models and related tuning procedures provides the relative reliable hydraulic conductivity in condition at various input combinations of index properties database. In specific, the investigation concludes with the following remarks:

- The data preprocessing of index properties database was conducted for refinement of database, handling missing values, feature scaling, feature selection, splitting the dataset into training and test datasets.
- The regression algorithms was modeled with linear regression, K-nearest neighbors, decision tree, random forest, support vector regression, multilayer perceptron.
- The best-fitting model and combination of index properties was multilayer perceptron application for e, S_S, D_{10}, D_{50}, D_{60} in training and test datasets, having the lowest average residuals.

6 ACKNOWLEDGEMENTS

This research was funded by the Basic Research Project of the Korea Institute of Geoscience and Mineral Resources (KIGAM).

7 REFERENCES

Angelis L. and Stamelos I. 2000. A simulation tool for efficient analogy based cost estimation. Empirical software engineering 5 (1), 35-68.

Angelov P.P. and Gu X. 2019. Empirical approach to machine learning. Cham: Springer.

Arshad R.R., Sayyad G., Mosaddeghi, M. and Gharabaghi, B. 2013. Predicting saturated hydraulic conductivity by artificial intelligence and regression models. International Scholarly Research Notices, 2013.

Arshad R.R., Sayyad G., Mosaddeghi M. and Gharabaghi B. 2013. Predicting saturated hydraulic conductivity by artificial intelligence and regression models. International Scholarly Research Notices, 2013.

Breiman L., Friedman J., Stone C.J. and Olshen R.A. 1984. Classification and regression trees. CRC press.

Carman P.C. 1939. Permeability of saturated sands, soils and clays. J. Agric. Sci. 29, 262-273.

Das S.K., Samui P. and Sabat A.K. 2012. Prediction of field hydraulic conductivity of clay liners using an artificial neural network and support vector machine. International Journal of Geomechanics 12 (5), 606-611.

Ferreira A.J. and Figueiredo M.A. 2012. Efficient feature selection filters for high-dimensional data. Pattern recognition letters 33 (13), 1794-1804.

Garcia S., Luengo J. and Herrera F. 2015. Data preprocessing in data mining (Vol. 72). Cham, Switzerland: Springer International Publishing.

Gribb M.M. and Gribb G.W. 1994. Use of neural networks for hydraulic conductivity determination in unsaturated soil. Proc., 2nd Int. Conf. Ground Water Ecology, J. A. Stanford and H. M. Valett, eds., Bethesda MD: Amer, Water Resources Assoc.

Harleman D.R.F. and Rumer R.R. 1963. Longitudinal and lateral dispersion in an isotropic porous medium. Journal of Fluid Mechanics 16 (3), 385-394.

Hazen A. 1911. Discussion of dams on sand foundations, Transactions, American Society of Civil Engineers 73, 199-203.

Kosti M.V., Mittas N. and Angelis L. 2012. Alternative methods using similarities in software effort estimation. In Proceedings of the 8th International Conference on Predictive Models in Software Engineering, 59-68.

Krzywinski M. and Altman N.S. 2017. Classification and regression trees. Nature Methods 14 (8), 757-758.

Lamorski K., Pachepsky Y., Sławiński C. and Walczak R.T. 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Science Society of America Journal 72 (5), 1243-1247.

Lewis R.J. 2000. An introduction to classification and regression tree (CART) analysis. In Annual meeting of the society for academic emergency medicine in San Francisco, California (Vol. 14).

Li Y. and Yuan Y. 2017. Convergence analysis of two-layer neural networks with relu activation. arXiv preprint arXiv:1705.09886.

Lin P.T., Su S.F. and Lee T.T. 2005. Support vector regression performance analysis and systematic parameter selection. In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. (Vol. 2, pp. 877-882). IEEE.

Mishra P., Varadharajan V., Tupakula U. and Pilli E.S. 2018. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys & Tutorials 21 (1), 686-728.

Osman M.S., Abu-Mahfouz A.M. and Page P.R. 2018. A survey on data imputation techniques: water distribution system as a use case. IEEE Access 6, 63279-63291.

Padarian J., Minasny B. and McBratney, A.B. 2020. Machine learning and soil sciences: A review aided by machine learning tools. Soil 6(1), 35-52.

Quinlan J.R. 1992. Learning with continuous classes. Proceedings of Australian Joint Conference on Artificial Intelligence. World Scientific Press, Singapore, 343-348.

Raymond M.R. 1986. Missing data in evaluation research. Evaluation & the health professions 9 (4), 395-420.

Razi, M. A. and Athappilly, K. 2005. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert systems with applications 29 (1), 65-74.

Ren X.W. and Santamarina J.C. 2018. The hydraulic conductivity of sediments: A pore size perspective. Engineering Geology 233, 48-54.

Ren X., Zhao Y., Deng Q., Kang J., Li D. and Wang, D. 2016. A relation of hydraulic conductivity—void ratio for soils based on Kozeny-Carman equation. Engineering Geology 213, 89-97.

Sedaghat A., Bayat H. and Sinegani A.S. 2016. Estimation of soil saturated hydraulic conductivity by artificial neural networks ensemble in smectitic soils. Eurasian Soil Science 49 (3), 347-357.

Sharma S., Ahmed S., Naseem M., Alnumay W.S., Singh S. and Cho, G.H. 2021. A Survey on Applications of Artificial Intelligence for Pre-Parametric Project Cost and Soil Shear-Strength Estimation in Construction and Geotechnical Engineering. Sensors 21 (2), 463.

Singh B., Sihag P. and Singh, K. 2017. Modelling of impact of water quality on infiltration rate of soil by random forest regression. Modeling Earth Systems and Environment 3 (3), 999-1004.

Tayfur G., Nadiri A.A. and Moghaddam A.A. 2014. Supervised intelligent committee machine method for hydraulic conductivity estimation. Water resources management 28 (4), 1173-1184.

Taylor H.M. and Gardner H.R. 1963. Penetration of cotton seedlingn taproots as influenced by bulk density, moisture content, and strength of soil. Soil Science 96 (3), 153-156.

Uma K.O., Egboka B.C.E. and Onuoha K.M. 1989. New statistical grain-size method for evaluating the hydraulic conductivity of sandy aquifers. Journal of Hydrology 108, 343-366.

Vapnik V.N. 1998. Statistical Learning Theory. New York: Wiley.

Venkatesan P. and Anitha S. 2006. Application of a radial basis function neural network for diagnosis of diabetes mellitus. Current Science 91, 1195-1199.

Xue B., Zhang M., Browne W.N. and Yao X. 2015. A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation 20 (4), 606-626.

Yao, Z. and Ruzzo W.L. 2006. A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. In BMC bioinformatics (Vol. 7, No. 1, pp. 1-11). BioMed Central.