KSCE Journal of Civil Engineering (2015) 19(7):2307-2316
Copyright ⓒ2015 Korean Society of Civil Engineers
DOI 10.1007/s12205-015-0210-x

Water Engineering

pISSN 1226-7988, eISSN 1976-3808
www.springer.com/12205

Support Vector Machine and Regression Analysis to Predict the Field Hydraulic Conductivity of Sandy Soil

Moussa S. Elbisy*

Received April 18, 2014/Revised September 26, 2014/Accepted October 27, 2014/Published Online February 6, 2015

Abstract

Saturated hydraulic conductivity is one of the key parameters in soil physics and hydrological modeling. This study explores the use of Support Vector Machine (SVM) and a nonlinear statistical regression approach for the purpose of predicting the field saturated soil hydraulic conductivity (K_{field}) of sandy soil based on basic soil properties of saline and alkaline soil data sets. Considering the significance of soil properties, both methods used the following levels of input soil data, which are easily measurable in the laboratory: hydraulic conductivity, clay/silt ratio, liquid limit, hydro carbonate anions, chloride ions, and calcium carbonate content. The influence of three kernel functions (linear, radial basis and sigmoid) on the performance of the SVM model was investigated. An adaptive genetic algorithm is used to determine the optimal free parameters of the SVM models. The results indicated that the SVM with the RBF model has better accuracy compared to the linear- and sigmoid-based models. The RBF model performed satisfactorily with a modeling efficiency of 0.972 and a correlation coefficient of 0.976. According to all of the performance measures, the different SVM models are a powerful tool and have better performance than statistical regression models. The excellent performance of the SVM with the RBF model demonstrated its potential to function as a useful tool for the indirect estimation of K_{field} to assess maximum obtainable prediction accuracy.

Keywords: *saturated soil hydraulic conductivity, statistical regressions, prediction, genetic algorithm, support vector machines*

1. Introduction

The saturated hydraulic conductivity is an important physical property of soil, particularly in determining the infiltration rate and irrigation practices, as well as designing functional subsurface drainage systems (Taskinen *et al.*, 2008; Elhakeem *et al.*, 2009). This parameter affects the economic and technical feasibility of large-scale subsurface drainage projects. Therefore, the design and evaluation of drainage projects require the most accurate determination of the saturated hydraulic conductivity. Unfortunately, the saturated hydraulic conductivity is one of the most difficult factors to measure for any drainage project (Schwab *et al.*, 1996). The value of the saturated hydraulic conductivity is spatially dependent, and it is difficult to determine a representative value to use in drain spacing calculations (Mohanty *et al.*, 1994; Gupta *et al.*, 1996).

Generally, the determination of a soil's saturated hydraulic conductivity is based on direct and indirect methods. A number of direct methods have been applied to the problem of determining the saturated hydraulic conductivity (Reynolds and Elrick, 1991; Reynolds and Zebchuk, 1996). However, most of these direct methods are difficult to use, labor-intensive, time-consuming, and expensive. Moreover, these methods require restrictive initial and boundary conditions (Libardi *et al.*, 1980). Despite a number of improvements, direct-measurement technology has only marginally advanced over the last decades. Conversely, indirect methods, which predict the hydraulic properties from more easily measured data, have received comparatively little attention. This situation is unfortunate because these indirect methods, which are known as 'predictive estimation methods', can provide reasonable estimates of hydraulic soil properties with considerably less effort and expense. Hydraulic conductivities determined with estimation methods may be accurate enough for a variety of applications (Wösten and Van Genuchten, 1988). Other important indirect methods are inverse methods to estimate parameters of analytical models that describe water retention and hydraulic conductivity. Aronovici (1947) presented a correlation between the content of silt and clay in subsoil materials in the Imperial Valley, California, U.S.A., and the results of hydraulic laboratory tests. Smedema and Rycroft (1983) created general tables with ranges of hydraulic conductivity values for certain types of soils. Ghulman *et al.* (2011) concluded that the saturated hydraulic conductivity is affected by soil properties and can be estimated from certain factors, such as chemical composition and physical properties. These researchers developed two regression equations to estimate the field saturated hydraulic conductivity

*Associate Professor, Civil Engineering Dept., Higher Technological Institute, 10th of Ramadan City 44629, Egypt (Corresponding Author, E-mail: msel-bisy@uqu.edu.sa)

(K_{field}) for a soil sample from its properties.

In the last decade, Neural Networks (NNs) and, more recently, Support Vector Machine (SVM) have emerged as two powerful tools for nonlinear modeling, particularly in situations where the development of phenomenological or conventional regression models becomes impractical or cumbersome. SVM has gained popularity in many traditionally NN dominated fields. Using an SVM eliminates the local minimum issue. In contrast to neural networks, SVMs automatically select their model size and are based on the principle of structural risk optimization, which prevents or reduces over-fitting (Vapnik, 1995). The main difference between SVMs and NNs is the principle of risk minimization. NNs implement empirical risk minimization to minimize the error on the training data. Alternatively, SVMs adhere to the principle of structural risk minimization, which seeks to establish an upper bound on the generalization error (Vapnik *et al.*, 1997).

NNs and SVMs have appeared in many applications in the field of water engineering. Minasny and Perfect (2004) used NNs coupled with bootstrap aggregation to predict soil-water retention and hydraulic conductivity characteristics from basic soil properties. Nakhaei (2005) used eight cumulative particle size fractions to predict log-transformed hydraulic conductivities for loamy sand soils with NNs, with individual modelling of different soil types being found superior to joint modelling. Elbisy (2006) applied NN models (feed-forward Back Propagation (BP) and Radial Basis Function (RBF)) to predict the K_{field} of sandy soil based on basic saline and alkaline soil data. This researcher found that the BP neural network is more accurate than the RBF neural network. Agyare *et al.* (2007) used, in addition to grain size and bulk density, organic carbon and nine different terrain parameters to predict K_{field}. Based on centrifuge data for silty sand and marine clay, Erzin *et al.* (2009) concluded that the NN approach is more efficient and reliable compared to the statistical method proposed by Benson *et al.* (1994). Lamorski *et al.* (2008) used an SVM to developed Transfer Functions (PTFs) and compared the results to NN results. These researchers found that the advantage of an SVM was more pronounced at soil matrix potentials, where larger relative errors have been encountered and the correlation between predicted and measured soil water content was lower. Rogiers *et al.* (2012) compared two data-driven modelling methods—Multiple Linear Regression (MLR) and NNS—that use the entire grain-size distribution data as input for the prediction of K_{field}. They combined NNs with a Generalized Likelihood Uncertainty Estimation (GLUE) approach to predict K_{field} from grain-size data. This GLUE-NN approach provided greater accuracy in the predicted K_{field}, and considerably smaller prediction intervals, with equal reliability. Das *et al.* (2012) used an NN and an SVM with a radial basis kernel function method to predict the value of K_{field} for clay liners based on compaction characteristics, lift thicknesses, number of lifts, Atterberg limits, and grain size. These researchers found the SVM model to be more efficient compared to the NN model. Arshad *et al.* (2013) used Radial Basis Function Neural Networks (RBFNN), Multi-Layer Perceptron Neural Networks (MLPNN), Adaptive Neuro-fuzzy Inference System (ANFIS) and MLR to predict the K_{field} in the Khuzestan province in southwest Iran. They used sand, silt, and clay percentages and bulk density as input variables. The results indicated that ANFIS and RBFNN are effective methods for K_{field} prediction and have better accuracy compared to the MLPNN and MLR models. Tayfur *et al.* (2014) applied Artificial Intelligence (AI) models of Sugeno Fuzzy Logic (SFL), Mamdani Fuzzy Logic (MFL), MLPNN associated with Levenberg–Marquardt (ANN), and Neuro-Fuzzy (NF) to estimate hydraulic conductivity using hydrogeological and geoelectrical survey data obtained from the Tasuj Plain Aquifer, northwest of Iran. The results revealed that SFL and NF produced acceptable performances, while ANN and MFL had poor predictions.

Several of the most important design choices are the SVM meta-parameters, which implicitly define the structure of the high-dimensional feature space where a maximal margin hyperplane will be found. A feature space that is too rich will cause the system to over-fit the data, but the system might not be capable of separating the data if the kernels are too poor (Cristianini *et al.*, 1998). Many techniques have been developed to select the most appropriate SVM parameter values. The most common optimization technique is the Genetic Algorithm (GA).

GA requires time to perform the simultaneous optimization of multiple SVM parameters. In this investigation, the SVM approach is proposed and adopted to forecast the K_{field} based on soil properties easily measured in the laboratory. GA is used to optimize the parameters of the SVM because the selection of parameters plays an important role in the performance of the SVM. This investigation aims to compare the accuracy of suitable nonlinear statistical regressions and the SVM approach with different kernel functions for such problems as considering the correlation of significant soil variables and the limitations on the selected structure factors.

2. Study Area and Data

Soil samples were collected from two areas; the first was the El-Nubaria area (sugar beet areas 1 and 2), which is located in the western delta of Egypt. The soil texture in this area varies between sandy loam and loamy sand for Sugar beet area 1, while the texture ranges from sandy loam to sandy clay loam up to 2.0 m below the soil surface at Sugar beet area 2. Soils in this area contain a certain amount of fine gravel and gypsum at depths varying between 0.40 and 0.60 m, and these soils are classified as calcareous soils. The second area is in Sinai, which has sandy unstable soil.

The selected areas are characterized by sandy soils with different physical and chemical properties. Data set (A) consists of 57 soil samples taken from the El-Nubaria area, and data set (B) consists of 28 soil samples taken from the Sinai area. Disturbed soil samples were collected from the selected areas and locations. The hydraulic conductivity was measured at each location using the auger-hole method (Abdel Hadi *et al.*, 2002).

The soil samples were analyzed in the laboratory to determine the physical and chemical properties (Abdel Hadi *et al.*, 2002). The physical properties of the soil samples included the contents of silt, clay, and sand, the d_{90} of the grains, the Liquid Limit (LL), the Plastic Limit (PL), and the Plasticity Index (PI). The chemical properties of the soil samples included soil pH, Electric Conductivity (EC), soil SP, SAR, and ESP, and the contents of calcium, magnesium, sodium, potassium, and chloride ions, along with hydrocarbonate anions (HCO_3), sulfate oxide cations (SO_4), and calcium carbonate ($CaCo_3$). A permeameter set-up was used to determine the laboratory hydraulic conductivity (K_{lab}).

The clay content of the soil varied between 0.00% and 37.4%, and the K_{field} ranged from 0.07 m/day to 6.06 m/day. Soil samples from the Sinai area contained more salts than those collected from the El-Nubaria area. The clay content of soil data set (A) varied between 0.10% and 27.00%, while soil data set (B) had clay content between 0.00% and 37.4%. The d_{90} of the samples ranged from 0.12 mm to 4.37 mm for soil data set (A) and from 0.12 mm to 0.34 mm for soil data set (B). The field hydraulic conductivity measured at sample locations in data set (A) varied between 0.07 m/day and 6.06 m/day, whereas the value varied between 0.35 m/day and 1.49 m/day at locations in data set (B). The hydraulic conductivities were determined in the laboratory using the permeameter set-up, and the disturbed soil samples were collected from the two areas. The electric conductivities of soil samples from data set (A) were less than 4.0 dS/m, which indicates that they are saline soils. The electric conductivities of the samples from data set (B) were more than 37.0 dS/m, and the sodium absorption ratios were more than 15.0. Therefore, samples in data set (B) are classified as alkaline soils.

3. Methodology

The methodology for this study consists of the following stages: data identification, data preprocessing, training and testing dataset preparation, SVM-model development, the development of nonlinear regression analysis based on soil properties, and model evaluation based on comparisons with regression analysis. In the data identification stage, an initial analysis of data is performed. The purpose of this stage is to identify the parameters that have a significant effect on a soil's saturated hydraulic conductivity. The number of these parameters is reduced by considering the internal coefficients that have high correlation with other parameters. To avoid calculation overflow and to accelerate the convergence rate of the learning and training process, the initial data have been normalized in the data preprocessing stage.

Based on the selected significant soil parameters, the models developed by the SVM technique are presented to produce better estimates of K_{field}. One of the most important design choices of the SVMs is the meta-parameters (C and ε) and the kernel function parameters (δ^2, k, and ν). This paper describes a technique that attempts to determine the optimal values of the SVM's free parameters by using a GA technique. Afterwards, the actual kernel must be chosen, and, as the results of this paper show, different kernels (RBF, linear function, and sigmoid function) may exhibit different performances. By dividing up the soil data sets (A and B), the development of regression equations is introduced.

The estimated K_{field} values obtained by nonlinear regression analysis based on soil properties are compared to the predictions of the SVM-K_{field} models. The agreement between the predictions and the observations can be checked statistically by calculating the following measures: the Root Mean Square Error (*RMSE*), the Mean Residual (*MR)*, the Mean Absolute Error (*MAE)*, the Mean Absolute Percentage Error (*MAPE)*, the coefficient of efficiency (E_f), and the correlation coefficient (R). These five criteria are defined as follows (Hack-ten Broke and Hegmans, 1996):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \tag{1}$$

$$MR = \frac{1}{N}\sum_{i=1}^{N}P_i - O_i \tag{2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i| \tag{3}$$

$$MAPE = \left[\frac{1}{N}\sum_{i=1}^{N}\left|\frac{(P_i - O_i)}{P_i}\right|\right] \times 100 \tag{4}$$

$$R = \frac{\sum_{i=1}^{N}(P_i - \overline{P})(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{N}(P_i - \overline{P})^2 \sum_{i=1}^{N}(O_i - \overline{O})^2}} \tag{5}$$

$$E_f = \left[\sum_{i=1}^{n}(O_i - \overline{O})^2 - \sum_{i=1}^{n}(P_i - O_i)^2\right] \Big/ \left[\sum_{i=1}^{n}(O_i - \overline{O})^2\right] \tag{6}$$

where, O_i is an observed K_{field} value, P_i is the predicted value, N is the total number of data points under validation, $\overline{O}$ is the mean value of the observations, and $\overline{P}$ is the mean value of the predictions. Each of the above performance statistics provides different information about the predictive ability of the models. The *RMSE* statistics indicates only the model's ability to predict away from the mean. The *RMSE* gives more weight to high K_{field} values because it involves the square of the difference between observed and predicted values. The *MR* is a measure of prediction bias, with a negative and positive value indicating under prediction and over prediction, respectively. The *MAE* is the most natural measure of the average error magnitude, and it is an unambiguous measure of the average error magnitude. It appears that all the dimensioned evaluations and inter-comparisons of average model performance error should be based on the *MAE*. The *MAPE* provides an unbiased error estimate because it gives appropriate weight to all magnitudes of the predicted variable. When the ratio of predicted to measured is closer to 100, the *MAPE* will be smaller. This aspect of relative error is found to give a more appropriate assessment and comparison of different models. R is estimated between the predicted and observed values of K_{field}. The model predictions are precise if the R-value equals one. The modeling efficiency, EF, is a measure that assesses the accuracy of the simulations. The maximum value for EF is one, which occurs when the simulated values perfectly match the measured values.

4. Field Hydraulic Conductivity Prediction Models Based on SVM

4.1 Support Vector Machine

In a regressive SVM, the basic idea is to map a low-dimensional input space x onto a higher dimensional feature space F via a nonlinear mapping φ. Then, the following estimation function is used to make linear regressions in that feature space:

$$f(x) = w.\varphi(x) + b \tag{7}$$

where, $\varphi(x)$ represents the high-dimensional feature space that has been nonlinear mapped from the input space, w is the weight vector, and b is the bias term. The coefficients w and b are estimated by minimizing the following regularized risk function:

$$R(C) = \frac{1}{2}\|w\|^2 + C\frac{1}{l}\sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i)) \tag{8}$$

$$L\varepsilon(y_i, f(x_i)) = \begin{bmatrix} |y_i - f(x_i)| - \varepsilon & |y_i - f(x_i)| \geq \varepsilon \\ 0 & |y_i - f(x_i)| < \varepsilon \end{bmatrix} \tag{9}$$

where, ε is a precision parameter representing the radius of the tube located around the regression function, $L\varepsilon(y_i, f(x_i))$ is the ε-insensitive loss function, and C is a regularization constant that determines the trade-off between the training error and the generalization performance. The term $\frac{1}{2}\|w\|^2$ measures the flatness of the function $L\varepsilon(y_i, f(x_i))$.

Introducing the slack variables ξ and ξ^* into Eq. (8), the overall optimization is formulated as follows:

Minimize

$$\varphi(w, \xi, \xi^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi + \xi^*) \tag{10}$$

subject to

$$\begin{cases} y_i - w.\varphi(x) - b \leq \varepsilon + \xi_i & \xi_i \geq 0 \\ w.\varphi(x) + b - y_i \leq \varepsilon + \xi_i^* & \xi_i^* \geq 0 \end{cases}$$

This constrained optimization problem is solved using the following Lagrangian form:

Maximize

$$H(\alpha, \alpha^*) = -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) + \sum_{i}^{l} y_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i}^{l}(\alpha_i + \alpha_i^*) \tag{11}$$

subject to

$$\sum_{i}^{l}(\alpha_i - \alpha_i^*) = 0 \quad 0 \leq \alpha_i,\ \alpha_i^* \leq C$$

where, α_i and α_i^* are Lagrangian multipliers.

Finally, the support vector machine regression function can be written as follows:

$$f(x) = \sum_{i}^{l} y_i(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{12}$$

where, $K(x_i, x) = \varphi(x_i)\varphi(x)$ is called the kernel function. Using the kernels, all the necessary computations can be undertaken directly in the input space without calculating the explicit map $\varphi(x)$. In this paper, three SVM kernel functions are employed and are defined as follows:

- the radial basis function (RBF): $K(x_i, x) = e^{(-\|x_i - x\|^2/2\delta^2)}$,
- the sigmoid function: $K(x_i, x) = \tanh(k(x_i x) + \nu)$ for $(k > 0, \nu < 0)$, and
- the linear function: $K(x_i, x) = x_i x$

where δ^2 is the kernel parameter of the radial basis function kernel, k is the scaling parameter of the input data, and ν is a shifting parameter that controls the mapping threshold.

4.2 Determination of SVM-Model Parameters

4.2.1 Analysis of SVM-Model Parameters

The SVM generalization performance depends on a good selection of the meta-parameters (C and ε), the kernel type, and the kernel function parameters (δ^2, k, and ν). The problem of optimal parameter selection is further complicated because the model complexity of an SVM depends on a combination of all the parameters. Thus, separately selecting each parameter is not adequate in obtaining an optimal model. Therefore, these parameters must be chosen carefully, and we have proposed using GA in this paper to select the SVM-model parameters. GAs are based on the principle of survival of the fittest member in a population, which retains genetic information by passing it from generation to generation.

4.2.2 Parameters of SVM Models Optimized by GA

The GA is a search algorithm for optimization based on the mechanics of natural selection and genetics (Goldberg, 1989). The GA is able to search very large solution spaces efficiently by providing lower computational costs through the use of probabilistic transition rules rather than deterministic ones. The GA has a number of components, or operators, which must be specified to define a particular GA. Fig. 1 depicts the operation of a GA. The specific steps in optimizing the parameters of an SVM model using a GA begin with an initial population of individuals (generation) that are randomly generated. Every individual (chromosome) encodes a single possible solution to the problem under consideration. The fittest individuals are selected by ranking them according to a pre-defined fitness function. In this study, the leave-one-out cross-validation method is used to evaluate

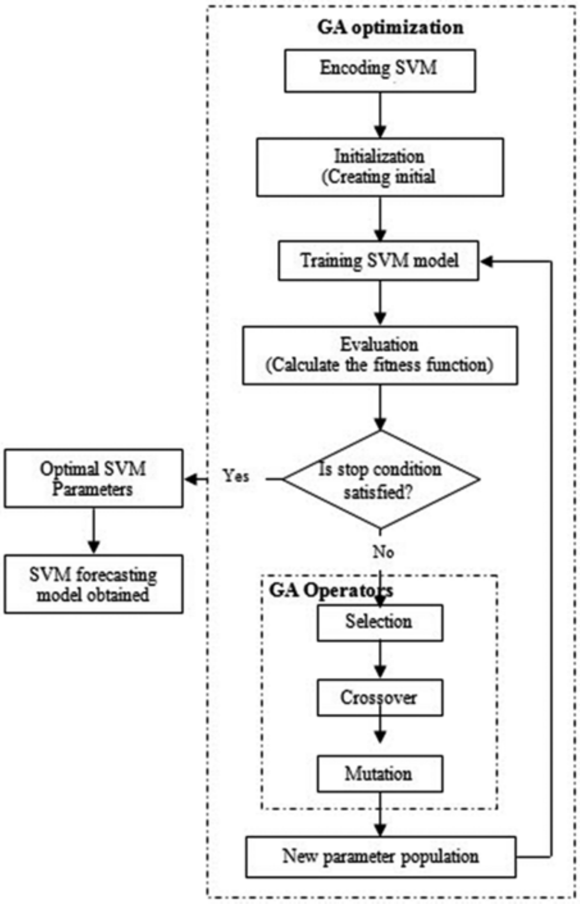


Fig. 1. Framework of Genetic Algorithms

fitness (Cawley and Talbot, 2004). A negative Mean Absolute Percentage Error (-MAPE) is used as the fitness function.

Selection, crossover, and mutation are the operators used to ensure reproduction in GAs. Selection is performed to choose excellent chromosomes to reduce. Based on the fitness function, chromosomes with higher fitness values are more likely to yield offspring in the next generation by means of the roulette wheel method. Following this initial process, the crossover and mutation operations are used to produce offspring from the individuals in the current population. The single-point crossover technique is used to randomly exchange genes between two chromosomes. Mutation is performed to alter the binary code from 0 to 1 or vice versa. The offspring replace the old population and form a new population in the next generation. The evolutionary process continues until stop conditions are satisfied.

5. Data Identification

Clearly, too many input parameters will drastically slow the learning process, and too few sets of training data can provide insufficient information regarding the localized features and cause the SVM to fail to generalize, which means the SVM response to unseen data will be poor. It is, therefore, essential to optimize the number of input parameters as much as possible. In this study, it is important to know the relative importance of all the input parameters with respect to their effects on the K_{field} value. Thus, a statistical regression of data was performed. This process provided information on the minimum descriptive input parameters required for the SVM models to predict the K_{field} of the sandy soil. It was found that the K_{field} of the El-Nubaria samples have high correlation coefficients with K_{lab}, PL, pH, Cl, HCO_3, Sp and $CaCO_3$. For the Sinai samples, it was found that the K_{field} had high correlation coefficients with K_{lab}, the clay/silt ratio, d_{90} of the grains, LL, Cl and $CaCO_3$. Tables 1 and 2 present the correlation coefficients for both areas.

Table 1. Correlation Coefficients between K_{field} and Soil Parameters of the El-Nubaria Area

	K_{lab} (m/day)	PL %	pH	HCO_3	Cl	Soil Sp	$CaCO_3$
K_{field} (m/day)	0.21	0.30	0.32	-0.69	-0.21	0.28	0.27

Table 2. Correlation Coefficients between K_{field} and Soil Parameters of the Sinai Area

	K_{lab} (m/day)	clay/silt	d_{90} (mm)	LL %	Cl	$CaCO_3$
K_{field} (m/day)	0.25	-0.21	-0.35	0.35	0.37	0.23

Table 3. Correlation Coefficients between Soil Parameters in the El-Nubaria

	PL%	pH	HCO_3	Cl	Soil Sp
PL%	1.00				
pH	0.19	1.00			
HCO_3	-0.08	-0.42	1.00		
Cl	0.24	0	0.33	1.00	
Soil Sp	0.57	0.54	-0.12	0.46	1.00
$CaCo_3$	0.76	0.32	0.03	0.38	0.77

The internal correlations between the soil parameters were determined to help eliminate dependent parameters. Table 3 presents the correlation coefficients between selected soil parameters for samples from El-Nubaria. From the table, soil Sp can be eliminated from the factors used to estimate K_{field} because it is highly correlated to $CaCO_3$; similarly, the Plastic Limit (PL) and pH can be eliminated because they are highly correlated to $CaCO_3$ and HCO_3, respectively. After eliminating some parameters, K_{field} is considered to be a function of K_{lab}, HCO_3, Cl and $CaCO_3$.

Table 4 presents correlation coefficients between the selected soil parameters for the samples from Sinai. From the results in the table, the d_{90} of the soil grains can be eliminated from the

Table 4. Correlation Coefficients between Soil Parameters in the Sinai Area

	clay/silt	d_{90} (mm)	LL %	Cl
clay/silt	1.00			
d_{90} (mm)	0.19	1.00		
LL %	0.07	-0.27	1.00	
Cl	-0.05	-0.53	0.32	1.00
$CaCO_3$	-0.05	-0.17	0.67	0.14

factors used to estimate K$_{\text{field}}$ because it is highly correlated to Cl, while CaCO$_3$ can be eliminated because it is highly correlated to the LL. Thus, K$_{\text{field}}$ is considered to be a function of the clay/silt ratio, LL, and Cl.

6. Analysis of Results and Discussion

6.1 SVM Methods

After identifying the parameters that have a significant effect on the field hydraulic conductivity, both groups of data (sets A and B) were integrated to develop an SVM-based model. The inputs to the SVM model included K$_{\text{lab}}$, HCO$_3$, Cl, CaCO$_3$, clay/silt ratio, and LL. The output of the SVM model was K$_{\text{field}}$. As the number of inputs differs greatly from the single output, the values are normalized using a normalization function to restrict the values within the range of 0 to 1:

$$S = (V - V_{\min})/(V_{\max} - V_{\min}) \tag{13}$$

where, S is the normalized value of variable V, while $V_{\min}$ and $V_{\max}$ are the variable's minimum and maximum values, respectively.

When applying an SVM, an appropriate kernel function is the first thing that needs be chosen because it is used internally in the SVM algorithm to map the input parameters to the highly dimensional feature space used in the internal computations of the algorithm. In the present study, we sought to investigate the influence of the kernel function on the SVM model performance. We tested three kernel functions: linear kernel, radial basis kernel and sigmoid kernel. The values of the SVM parameters in each kernel function represent the second thing that needs to be considered. The proper selection of the SVM parameters has an impact on model performance and the ability for generalisation. For the purpose of an automated search of the model parameters, genetic algorithms were used as an optimisation framework. The SVM parameters can be calculated according to the method discussed in Section 4.2. The selected parameter values for the different SVM models are shown in Table 5.

The performance statistics of different models in estimating K$_{\text{field}}$ during both training and testing, are represented in Table 6. During training, an SVM with a linear kernel model resulted in the *MAPE* of 15.976, an *MR* of -0.049, the *RMSE* of 0.302, the *R*-value of 0.947, with the E_f of 0.907 (Table 6); however, during testing, the corresponding values were 19.055, -0.013, 0.368, 0.938, and 0.891, respectively. Fig. 2 shows the correlations between the observed and predicted data for the linear kernel model. During training and testing, the sigmoid kernel (more complex) model performed better than the linear kernel model (less complex) model in all measures (*RMSE*, *MR*, *MAE*, *MAPE*, E_f, and *R*) (see Table 6). Both the linear and sigmoid kernel models slightly under predicted (negative MR) K$_{\text{field}}$. Figure 3 shows the correlations between the observed and predicted data for the sigmoid kernel model.

When the developed SVM model with an RBF was applied to the trained data, the maximum percentage error was 14.29%, the minimum percentage error was 1.03%, the *MAPE* was 5.293%, the *MR* was 0.023 m/day, and the *RMSE* was 0.107 m/day. For the tested data, the maximum percentage error was 15.79%, the minimum percentage error was 1.43%, the *MAPE* was 6.29%,

Table 5. Optimal Parameters for SVM Models

kernel function	free parameters		kernel parameters		
	C	ε	δ^2	k	ν
Linear	4.094	0.0022	-	-	-
Radial basis	89.968	0.0013	32	-	-
Sigmoid	35.571	0.0007	-	5.42	- 0.0309

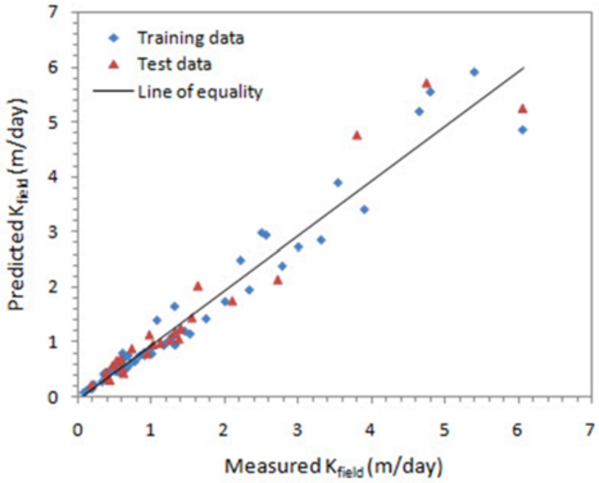


Fig. 2. Scatter of Predicted and Experimental Values of K$_{\text{field}}$ for the SVM Model with a Linear Function

Table 6. Evaluating the Performance of SVM Models

Kernel Function	Linear			Radial Basis			Sigmoid		
	Training	Testing	Total	Training	Testing	Total	Training	Testing	Total
RMSE (m/day)	0.302	0.368	0.322	0.107	0.145	0.118	0.255	0.309	0.271
MR (m/day)	-0.049	-0.013	-0.038	0.023	0.02	0.022	-0.019	-0.043	-0.036
MAE (m/day)	0.212	0.257	0.226	0.066	0.099	0.076	0.172	0.212	0.184
MAPE (%)	15.976	19.055	16.907	5.293	6.29	5.594	12.051	15.282	13.028
E_f	0.907	0.891	0.902	0.974	0.969	0.972	0.925	0.915	0.922
R	0.947	0.938	0.944	0.977	0.975	0.976	0.953	0.948	0.951
Min (%)	6.25	6.25	6.25	1.03	1.43	1.03	5.00	6.25	5.00
Max (%)	33.33	33.33	33.33	14.29	15.79	15.79	22.22	22.29	22.29

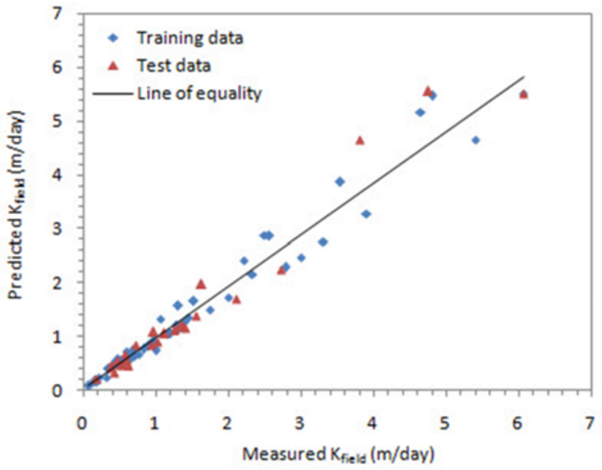


Fig. 3. Scatter of Predicted and Experimental Values of K_{field} for the SVM Model with a Sigmoid Function

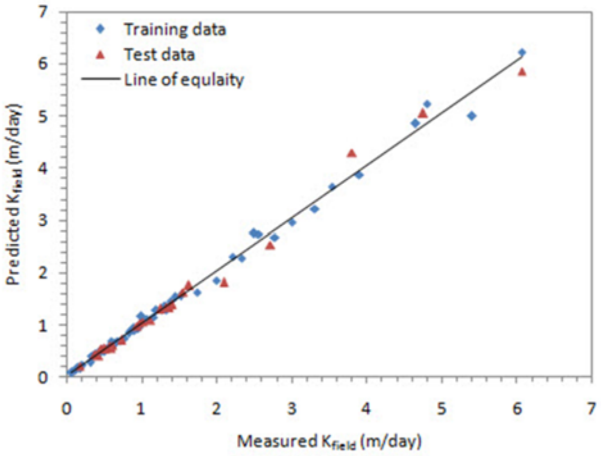


Fig. 4. Scatter of Predicted and Experimental Values of K_{field} for the SVM Model with RBF

the *MR* was 0.02 m/day, and the *RMSE* was 0.145 m/day. Efficiency coefficients E_f of 0.974 and 0.969 were obtained for the training and testing data, respectively. The variation of K_{field} shows the same trend for the experimental data and the present SVM model with an RBF. In general, it can be concluded that the SVM model with an RBF performed better than both the linear and sigmoid kernel models, in terms of *RMSE*, *MR*, *MAE, MAPE*, E_f, and *R.* Figure 4 shows the correlations between the observed and predicted data for the RBF kernel model.

A comparison between the observed K_{field} values and those predicted by the SVM models (linear, Radial Basis, and Sigmoid functions) is shown in Figs. 5 and 6. It is observed that the radial basis function model can significantly reduce the overall forecasting errors. The results showed that the model performance of the SVM with an RBF is superior to the linear- and sigmoid-based models at forecasting K_{field}. In other words, the predictions of K_{field} proposed by the SVM with an RBF are the most accurate. Moreover, it was noticed that while both the linear and sigmoid kernel models were slightly under-predicting (negative MR) K_{field}, the RBF kernel model was slightly over-predicting (positive MR) K_{field}, as shown in Table 6. From the results, the use of nonlinear kernel functions achieved better performance than the linear kernel. According to the six performance measures, the SVM (RBF) model produced the best performance among all the models in general.

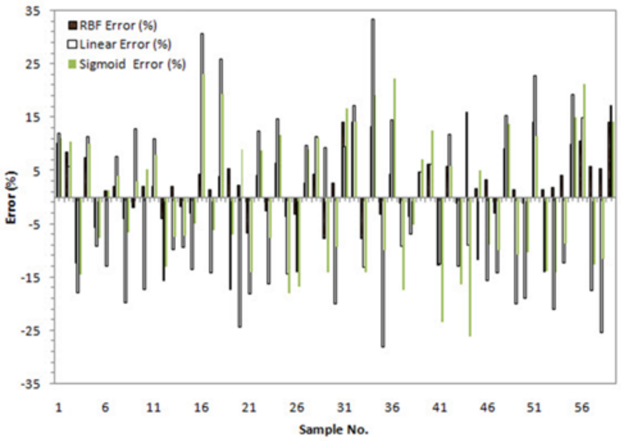


Fig. 5. Comparison of Percentage Errors Predicted by SVM Models according to the Experimental Data for Training Data

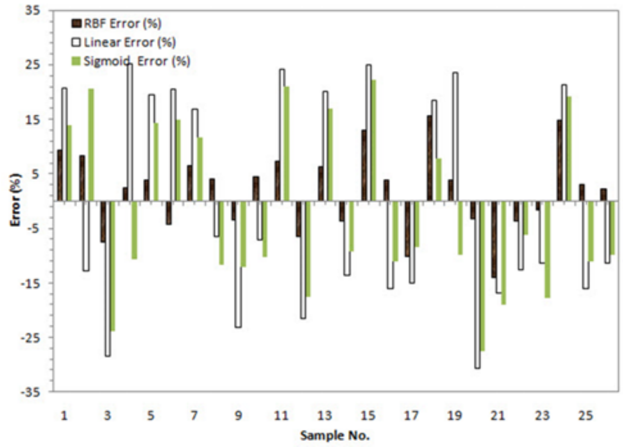


Fig. 6. Comparison of Percentage Errors Predicted by SVM Models according to the Experimental Data for Tested Data

6.2 Regression Method

To evaluate the performance of the proposed method, a regression method is conducted. According to the 5th methodology stage, data provided from the same datasets (A and B) were used to introduce two regression models. The formula for the nonlinear statistical regression approach was:

$$Y = a + \sum_{i=1}^{n}(b_iX_i + c_iX_i^2) \tag{14}$$

where, Y is the value of K_{field}, n is the number of soil parameters, X_i is the soil parameter and (a, b and c) are the sets of regression coefficients.

Two regression equations were developed according to the soil properties. These equations can be used to estimate the K_{field} for a soil sample from its properties after the soil type (alkaline or acidic) has been determined. For saline soils similar to dataset (A) (El-Nubaria Soil), the estimation of $K_{field-A}$ by regression analysis based on selected significant soil properties (1st stage of methodology) was developed as a function of independent variables K_{lab}, HCO_3, Cl and $CaCO_3$. The derived regression model ($K_{field-A}$) was as follows:

$$K_{field\text{-}A} = 30.14 + F(K_{lab}) + F(HCO_3) + F(Cl) + F(CaCO_3) \quad (15)$$

where,

$$F(K_{lab}) = 0.035K_{lab} + 0.0004\,K_{lab}^2$$
$$F(HCO_3) = -14.9HCO_3 + 1.88(HCO_3)^2$$
$$F(Cl) = -0.58Cl + 0.04Cl^2$$
$$F(CaCO_3) = 0.06CaCO_3 + 0.00096(CaCO_3)^2$$

For alkaline soils similar to dataset (B) (Sinai Soil), the estimation of $K_{field-B}$ by regression analysis based on soil properties was developed as a function of independent variables K_{lab}, the clay/silt ratio, LL, and Cl. The derived regression model ($K_{field-B}$) was as follows:

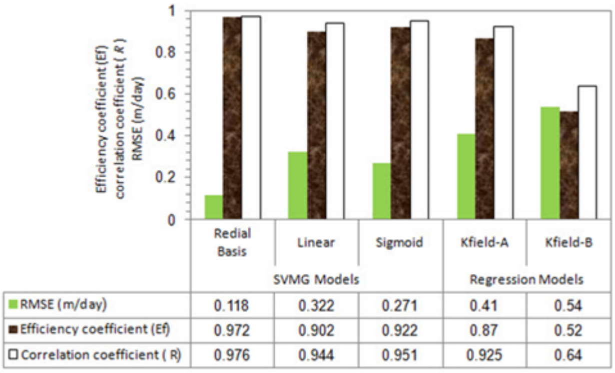
$$K_{field\text{-}B} = 0.781 + F(K_{lab}) + F(clay/silt) + F(Cl) + F(LL) \quad (16)$$

where,

$$F(K_{lab}) = 0.017K_{lab} - 0.00018\,K_{lab}^2$$
$$F(clay/silt) = 0.049(clay/silt) - 0.019(clay/silt)^2$$
$$F(Cl) = 0.00045Cl + 2.7 \times 10^{-7}\,Cl^2$$
$$F(LL) = -0.13LL + 0.0009LL^2$$

After determining the regression equations, the accuracy of the two regression models were evaluated by comparing its predicted field saturated soil hydraulic conductivity with the experimental data. The obtained values of the *RMSE*, *MAE*, *MAPE*, and *R* for $K_{field-A}$ (Eq. (15)) were 0.41 m/day, 0.289 m/day, 23.21%, and 0.925, respectively, compared to values of 0.54 m/day, 0.365 m/day, 30.32%, and 0.64, respectively, for $K_{field-B}$ (Eq. (16)). Merdun *et al.* (2006) obtained higher and RMSE values, which varied from 0.80 to 0.989 and from 0.013 to 0.938 for the regression method, respectively; therefore, the measured against the predicted field saturated soil hydraulic conductivity values obtained from the $K_{field-B}$ model for the test dataset had a poor correlation coefficient. Efficiency coefficients E_f of 0.87 and 0.52 were obtained for $K_{field-A}$ and $K_{field-B}$, respectively.

The goodness-of-fit statistical parameters indicated that differences existed in the accuracy of the K_{field} estimation by the applied approach with these soil datasets. The levels of *RMSE*, *MR*, *MAE*, *MAPE*, E_f, and *R* by different SVM models (linear kernel, radial basis kernel and sigmoid kernel) had higher accuracy than those derived by regressions models for predicting the field saturated soil hydraulic conductivity. This finding is observed because the SVM uses training data to build a forecast model, which works well in many learning situations because it generalizes to unseen data and is amenable to continuous and adaptive online learning, which is an extremely desirable property in network environments. The goodness-of-fit statistical parameters indicated that differences existed in the accuracy of the K_{field} estimation by the applied approach with these soil datasets. The statistical results showed that the SVM models performed well and were able to accurately estimate the K_{field} (Fig. 7). According to the measures (*RMSE*, *MR*, *MAE*, *MAPE*, E_f, and *R*), the SVM model with an RBF produced the best performance and was able to accurately estimate the K_{field}. However, the SVM models used datasets A and B together, and a comparison showed that it performed better than the regression equations. The performance of the $K_{field-B}$ equation was the poorest, as indicated by statistical analyses. The statistical parameters *RMSE*, *MR*, *MAE*, *MAPE*, and E_f were successfully used to directly compare the predictive ability and simulation accuracy among the developed models. The reason for the difference among these models arose from the various approaches to modeling and the types of data sets that were incorporated. This paper is intended to assist those authorities involved in designing guidelines for drainage systems and those participating in model induction from soil data.



	Redial Basis	Linear	Sigmoid	Kfield-A	Kfield-B
🟩 RMSE (m/day)	0.118	0.322	0.271	0.41	0.54
■ Efficiency coefficient (Ef)	0.972	0.902	0.922	0.87	0.52
□ Correlation coefficient (R)	0.976	0.944	0.951	0.925	0.64

Fig. 7. Comparison of *RMSE*, Efficiency Coefficient (E_f), and Correlation Coefficient (*R*) for the SVM and Regression Models

7. Conclusions

The saturated soil hydraulic conductivity is one of the most important parameters needed for designing drainage systems. This conductivity can be estimated from easily measurable soil parameters using the correlation methods. This paper investigated the utility of support vector regression with different kernel functions models in predicting the saturated soil hydraulic conductivity in the field based on basic soil properties that were easily measured in the laboratory. The optimal values of the SVM model parameters were selected using a genetic algorithm, and the leave-one-outcross-validation method was used for validation. The performance of the SVM models were compared with nonlinear statistical regression models. The accuracy of the predictions was evaluated using six measures (*RMSE*, *MR*, *MAE*, *MAPE*, E_f, and *R*).

The inputs to the SVM model included K_{lab}, HCO_3, Cl, $CaCO_3$, clay/silt ratio, and LL. The developed SVM model with an RBF ($E_F = 0.972$) was found to be more efficient compared to the linear ($E_F = 0.902$) and sigmoid ($E_F = 0.922$) kernel function models. The results show that the sigmoid kernel model performed better than the linear kernel model in all measures. It was noticed that while both the linear and sigmoid kernel models were slightly under-predicting K_{field}, the RBF kernel model was slightly over-predicting K_{field}.

Based on the nonlinear statistical regression approach, the $K_{field-A}$ and $K_{field-B}$ models were introduced using saline and alkaline soil samples, respectively. Using the saline dataset (K_{lab}, HCO_3, Cl and $CaCO_3$), the $K_{field-A}$ model had a modeling efficiency of 0.87 and performed nearly as well as the SVM with a linear kernel function model. The performance of the regression equation for $K_{field-B}$ on K_{lab}, clay/silt, LL and Cl was the poorest ($E_F = 0.53$) because it used the alkaline dataset.

In conclusion, the SVM models produced better performance with considerably less expense and effort needed to decide a priori on the class of input-output relationships. In summary, a comparison between the measurements and the results of the nonlinear statistical regression approach demonstrates that the SVM model with an RBF can function as a useful tool for analyzing the hydraulic soil properties from easily measurable soil data.

References

Abdel Hadi, A. M., Elbisy, M. S., and Ali, H. (2002). “Estimating the saturated soil hydraulic conductibility from easily measurable soil properties.” *III Regional Conference on Civil Engineering*, ASCE-EGS, Cairo, Egypt.

Agyare, W. A., Park, S. J., and Vlek, P. L. (2007). “Artificial neural network estimation of saturated hydraulic conductivity.” *Vadose Zone J.*, Vol. 6, No. 2, pp. 423-431.

Aronovici, V. S. (1947). “The mechanical analysis as an index of subsoil permeability.” *Proc., Soil Science Society of American*, Vol. 11, No. C, pp. 137-141, DOI: 10.2136/sssaj1947.036159950011000C0026x.

Arshad, R., Sayyad, G., Mosaddeghi, M., and Gharabaghi, B. (2013). “Predicting saturated hydraulic conductivity by artificial intelligence and regression models” *ISRN Soil Science*, Hindawi Publishing Corporation, Vol. 2013, pp. 1-8, DOI: org/10.1155/2013/308159.

Benson, C. H., Zhai, H., and Wang, X. (1994). “Estimating of hydraulic conductivity of compacted clay liners.” *J. Geotechnical Engineering*, ASCE, Vol. 120, No. 2, pp. 366-387, DOI: 10.1061/(ASCE)0733-9410(1994)120:2(366)

Cawley, G. C. and Talbot, N. L. C. (2004). “Fast exact leave-one-out cross-validation of sparse least-squares support vector machines.” Neural Networks, Vol. 17, No. 10, pp. 1467-1475, DOI:10.1016/j.neunet.2004.07.002

Cristianini, N., Campell, C., and Shawe-Taylor, J. (1998). *Dynamically adapting kernels in support vector machines*, NeuroCOLTH Technical Report NC-TR-98-017, Royal Holloway Collage, University of London, UK.

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*, Vol. 12, No. 5, pp. 606-6011, DOI: 10.1061/(ASCE)GM.1943-5622.0000129.

Elbisy, M. S. (2006). “Prediction of saturated hydraulic conductivity of sandy soil using neural network.” *Ain Shams Engineering Journal,* Ain Shams University, Vol. 41, No. 1, pp. 480-493.

Elhakeem, M., Chang, Y., Wilson, C. G., and Papanicolaou, A. N. (2009). *Field measurement of saturated hydraulic conductivity at the hillslope scale under different soil series and management practices*, AGU 2009 Fall Meeting, San Francisco, California.

Erzin, Y., Gumaste, S. D., Gupta, A. K., and Singh, D. N. (2009). “Artificial Neural Network (ANN) models for determining hydraulic conductivity of compacted fine-grained soils.” *Canadian Geotechnical Journal*, Vol. 46, No. 8, pp. 955-968, DOI: 10.1139/T09-035.

Ghulman, B., Elbisy, M. S., and Abdel Hadi, A. M. (2011). “Effect of soil properties on the saturated hydraulic conductivity of some Egyptian problematic soils.” *International Conference on Advances and Trends in Engineering Materials and their Applications*, Canada.

Goldberg, D. E. (1998). *Genetic algorithms in search optimization and machine learning*, Addison-Wesley Longman, New York.

Gupta, R. K., Rudra, R. P., Dickinson, W. T., Patni, N. K., and Wall, G. J. (1993). “Comparison of saturated hydraulic conductivity measured by various field methods.” *Transactions of the ASAE*, Vol. 36, No. 1, pp. 51-55, DOI: 10.13031/2013.28313.

Hack-ten Broke, M. J. D. and Hegmans, J. H. B. (1996). “Use of soil physical characteristics from laboratory measurements or standard series for modeling unsaturated water flow.” *Agricultural Water Management*, Vol. 29, No. 2, pp. 201-213, DOI: 10.1016/0378-3774(95)01190-0.

Lamorski, K., Pachepsky, Y., Slawihski, 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*, Vol. 72, No. 5, pp. 1243-1247, DOI:10.2136/sssaj2007.0280N.

Libardi, P. L., Reichardt, K., Nielsen, D. R., and Biggar, J. W. (1980). “Simple field method for estimating soil hydraulic conductivity.” *Soil Science Society of America Journal*, Vol. 44, pp. 3-7, DOI:10.2136/sssaj1980.03615995004400010001x.

Merdun, H., Cinar, O., Meral, R., and Apan, M. (2006). “Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity.” *Soil and Tillage Research*, Vol. 90 No. 1-2, pp. 108-116, DOI: 10.1016/j.still.2005.08.011.

Minasny, B. and Perfect, E. (2004). “Solute adsorption and transport parameters.” *Development of Pedotransfer Functions in Hydrology*, Elsevier, Amsterdam, pp. 195-224.

Mohanty, B. P., Kanvar, R. S., and Everts, C. J. (1994). “Comparison of saturated hydraulic conductivity measurement methods for a glacial-till soil.” *Soil Science Society of America Journal*, Vol. 58, No. 3, pp. 672-677, DOI: 10.2136/sssaj1994.03615995005800030006x.

Nakhaei, M. (2005). “Estimating the saturated hydraulic conductivity of granular material using artificial neural network based on grain size distribution curve.” *J. Sci. Islam Repub. Iran,* Vol. 16, No. 1, pp. 55-62.

Reynolds, W. D. and Elrick, D. E. (1991). “Determination of hydraulic conductivity using a tension infiltrometer.” *Soil Science Society of America Journal*, Vol. 55, No. 3, pp. 633639, DOI: 10.2136/sssaj1991.03615995005500030001x

Reynolds, W. D. and Zebchuk, W. D. (1996). “Hydraulic conductivity in a clay soil: Two measurement techniques and spatial characterization.” *Soil Science Society of America Journal,* Vol. 60, No. 6, pp. 1679-1685, DOI: 10.2136/sssaj1996.03615995006000060011x.

Rogiers, B., Mallants, D., Batelaan, O., Gedeon, M., Huysmans, M., and

Dassargues, A. (2012). "Estimation of hydraulic conductivity and its uncertainty from grain-size data using GLUE and artificial neural networks." *Math Geosci*, Vol. 44, pp. 739-763, DOI: 10.1007/s11004-012-9409-2.

Schwab, G. O., Frangmeier, D. D., and Elliot, W. J. (1996). *Soil and water management systems*, John Wiley & Sons, New York.

Smedema, L. K. and Rycrofrt, D. W. (1983). *Land drainage: Planning and design of agricultural drainage systems*, Batsford Academic Educational Ltd., London, 376 pp.

Taskinen, A., Sirviö, H., and Bruen, M. (2008). "Generation of two dimensionally variable saturated hydraulic conductivity fields: Model theory, verification and computer program." *Computers & Geosciences*, Vol. 34, No. 8, pp. 876-890, DOI: 10.1016/j.cageo.2007.04.010.

Tayfur, G., Nadiri, A., and Moghaddam, A. (2014). "Supervised intelligent committee machine method for hydraulic conductivity estimation" *Water Resour. Manage.*, Vol. 28, pp. 1173-1184, DOI: 10.1007/s11269-014-0553-y.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, John Wiley & Sons, New York.

Vapnik, V., Golowich, S., and Smola, A. (1997). "Support vector method for function approximation, regression estimation, and signal processing." *Advances in Neural Information Processing Systems 9*, MA, MIT Press, Cambridge, pp. 281-287.

Wösten, J. H. M. and Van Genuchten, M. Th. (1988). "Using texture and other soil properties to predict the unsaturated soil hydraulic functions" *Soil Science Society of America Journal*, Vol. 52, pp. 1762-1770, No. 6, DOI: 10.2136/sssaj1988.03615995005200060045x.