
Ali Mehmandoost Kotlar,* Bo V. Iversen, and Quirijn de Jong van Lier

Parametric and nonparametric supervised machine learning techniques were used to estimate saturated and near-saturated hydraulic conductivities ($K_s$ and $K_{10}$, respectively) from easily measurable soil properties including the name of the pedological horizon (HOR), soil texture (sand, silt, and clay), organic matter (OM), bulk density (BD), and water contents ($\theta_{pF1}$, $\theta_{pF2}$, $\theta_{pF3}$, and $\theta_{pF4.2}$) measured at four different matric heads ($-10$, $-100$, $-1000$, and $-15,848$ cm, respectively). Using a stepwise linear model (SWLM) and the Lasso regression as parametric methods with 316 data in training and 135 data in the testing phase, four pedotransfer functions (PTFs) were obtained in which water contents for both methods played an important role compared with other variables. The SWLM showed better performance than Lasso in the testing phase for $\log(K_s)$ and $\log(K_{10})$ prediction, with RMSE values of 0.666 and 0.551 cm d$^{-1}$ and $R^2$ of 0.26 and 0.65. Nonparametric supervised machine learning methods trained and tested with a similar data set significantly improved the accuracy of $K_s$ prediction, with $R^2$ of 0.52, 0.36, and 0.53 for Gaussian process regression (GPR), support vector machine (SVM), and ensemble (ENS) methods in the testing stage. These methods also described 74.9, 66.7, and 72.5% of the variation of $\log(K_{10})$. Bootstrapping validated the strong performance of nonparametric techniques. The feature selection capability of GPR determined that instead of using a model with all predictors, HOR, silt, $\theta_{pF1}$, and $\theta_{pF2}$ are sufficient for the prediction of $\log(K_s)$, while HOR, silt, and OM can predict $\log(K_{10})$ as accurate as the comprehensive model with all variables.

Abbreviations: ARD, automatic relevance determination; BD, bulk density; CV, coefficient of variance; ENS, ensemble; GPR, Gaussian process regression; HOR, pedological horizon; MLT, machine learning technique; OM, organic matter; PTF, pedotransfer function; SD, standard deviation; SVM, support vector machine; SWLM, stepwise linear model.

Appropriate estimation of soil hydraulic parameters, especially saturated ($K_s$) and near-saturated hydraulic conductivity, is an essential part of management practices including irrigation, drainage, flood protection, and erosion control in addition to water flow and transport modeling in soil. In both single-domain and dual-porosity models, near-saturated hydraulic conductivity at a tension of 10 cm ($K_{10}$) to a great extent excludes the effect of preferential or macropore flow (Jarvis et al., 2002). Therefore, $K_{10}$ can be used as a matching point instead of $K_s$, especially for precise estimation of unsaturated hydraulic conductivity in the dry range.

The most common in situ tools to measure $K_s$ directly, such as the double ring infiltrometer, rainfall simulator, or the Amoozemeter, are expensive and labor intensive (Duan et al., 2012). The same difficulties and high costs for measurement of unsaturated hydraulic conductivity (e.g., disk infiltrometers) make these direct methods less attractive as well. This is why mathematical relations are of interest for the estimation of $K_s$ and $K_{10}$ by using readily available or easily measurable physical and chemical properties of the soil such as the particle size distribution, bulk density, or organic matter content (Lilly et al., 2008; Zhao et al., 2016; Garcia-Gutiérrez et al., 2017). Soil hydraulic conductivity is also highly scale dependent, as shown for a very large database by Ghanbarian et al. (2017), and this scaling
The performance of PTTs can be improved by using powerful machine learning techniques (MLTs) which may lead to predictions that are more reliable. These techniques include parametric (regression models) and nonparametric methods such as artificial neural networks, support vector machine (SVM), Gaussian process regression (GPR), and Ensemble regressions, among others. In widely used parametric regressions, prior knowledge regarding the governing relationship between predictors and targets is required. On the other hand, nonparametric methods or memory-based learning algorithms do not require that knowledge, making them an interesting option due to the complexity of relationships among soil properties. Many data-mining-based PTTs developed to predict soil hydraulic conductivity use artificial neural networks combined with fuzzy logic (e.g., Iversen et al., 2011; Merdun et al., 2006; Arshad et al., 2013; Sihag et al., 2017a). Support vector machines with strong pattern recognition overcome local minima, which are conventional problems of neural networks in the training process (Lamorski et al., 2008; Haghverdi et al., 2014). In other words, the structural risk optimization assist in SVMs minimizes model size and prediction errors, while artificial neural networks have a predefined structure aiming solely to minimize the error in the training data (Vapnik et al., 1997; Elbisy, 2015).

Gaussian process regressions assume that adjacent observations should convey information about each other (Pal and Deswal, 2010), and they perform similarly to SVMs because they use kernel functions and also nonparametric basics but provide reliability of responses to the given input data, which makes GPR more reliable as a probabilistic model (Rasmussen and Nickisch, 2010). Use of this kernel-based machine-learning algorithm for evaluation of water flow in soils is still in the early stages. As an example, Li et al. (2017) predicted water inflow into tunnels using GPR, and Sihag et al. (2017b) assessed the application of GPR to determine the cumulative infiltration rate of a sandy soil.

The combination of several base models into one predictive model is called the Ensemble (ENS) method. The use of this method for regression-based studies has recently been adopted in soil science, e.g., in the simulation of water outflow from tile-drained agricultural fields by Kuzmanovski et al. (2015) and hydraulic conductivity prediction using boosted regression trees by Jorda et al. (2015).

To date, the simultaneous prediction of $K_s$ and $K_{10}$ using a variety of different MLTs has not been reported. As a novel contribution, this study explores the capability of MLTs for prediction of $K_s$ and $K_{10}$ in different soils from Denmark using common soil properties including soil pedological horizon, soil texture, organic matter, bulk density, and water contents at different pressure heads. In parametric regressions, using a Lasso regularization model compared with common stepwise regression is another new approach. Finally, three different nonparametric MLTs (GPR, SVM, and ENS) were trained to predict log($K_s$) and log($K_{10}$), and the effect of data variability on the models was analyzed by using bootstrapping data with and without replacement. If necessary, the internal parameters of nonparametric methods were obtained through Bayesian optimization using MATLAB.

**Materials and Methods**

**Study Area and Data**

The required soil properties including hydraulic conductivity and water contents were retrieved from the study of Iversen et al. (2011) determined on undisturbed soil columns with either 20- or 6.1-cm diameter and 20- or 3.5-cm height taken from different locations in Denmark covering a variety of Danish soil types. Large soil cores (6280 cm$^3$) were sampled from pedological horizons of hydrologic interest, and these large samples were used only for measurement of $K_s$ and $K_{10}$ as described below. A pedological description of the soil profile was performed and the different soil horizons (HOR) were identified and numbered from 1 to 6 starting from the top. Five large and five small soil cores were extracted from the same horizons. Samples were stored at 2 to 5°C to avoid biological activity that could change the structure of the soil. Measurement of hydraulic properties for all samples are performed once, and the temporal variation of these properties was not considered.

The particle size distribution was determined using the sieve and hydrometer methods. Originally the particle sizes in Iversen et al. (2011) were classified into seven fractions (<2, 2–20, 20–63, 63–125, 125–200, 200–500, and 500–2000 μm); however, we reclassified them as sand (50–2000 μm), silt (2–50 μm), and clay (<2 μm). The boundary between sand and silt fractions was chosen at 50 μm to correspond to results from USDA particle size classes. Accordingly, 80% of the coarse silt fraction (20–63 μm) was added to the silt component (20–50 μm) and the remaining 20% to the sand fraction (50–2000 μm) (Wagner et al., 2001).

In Iversen et al. (2011), unsaturated hydraulic conductivities at matric heads ranging from −2 to −100 cm were measured by drip infiltrometer with five tensiometers, and $K_{10}$ was found through fitted lines on a $K$–$b$ graph generated by the scaled conductivity model of Mualem (1976). Subsequently, soil cores were saturated overnight to be ready for measuring $K_s$ as described in detail by Iversen et al. (2004). In the current study, very comprehensive measured data (451 columns) including 10 predictors representing the physical and hydraulic properties of the soils were used for the prediction of log($K_s$) and log($K_{10}$). The small undisturbed soil cores (100 cm$^3$) were slowly saturated and then drained to predefined matric heads from −10 to −100 cm and a ceramic plate for potentials from −160 to −10,000 cm. Soil water characteristics at a matric head of −15,800 cm were measured after the soil had been ground and sieved through a 2-mm sieve (Klute, 1986). Water contents at matric heads of −10, −100, −1000, and −15,800 cm were determined and identified as $\theta_{pF1}$, $\theta_{pF2}$, $\theta_{pF3}$, and $\theta_{pF4}$, respectively. Bulk density (BD) was obtained by weighing the soil cores after drying for 24 h at 105°C.
Machine Learning Techniques

To estimate \( \log(K_p) \) and \( \log(K_{10}) \), various parametric MLTs including SWLM and Lasso (acronym for "least absolute shrinkage and selection operator") (Tibshirani 1996) besides nonparametric MLTs such as SVM, GPR, and ENS were compared. They were all trained and tested with the same input data composed of the 451 soil samples. The functions for GPR, SVM, and ENS, called `fitrgp`, `fitrsvm`, and `fitrensemble`, respectively, in MATLAB (R2018a), were used to form these comparative models. Based on literature recommendations and MATLAB documentation, different properties of the models were subject to change for each function to reach the best prediction. Finally, bootstrapping was used to assess the performance of the models.

Parametric Machine Learning Techniques

The Bayesian approach of the standard linear regression model in both function space and weight space views is

\[
y(X_i) = f(X_i) + \epsilon_i = \sum_{k=1}^{n} X_{ik} w_k + \epsilon_i \quad i = 1, 2, \ldots, n
\]

where \( w = (w_1, w_2, \ldots, w_n) \) is the coefficient vector of parameters (weights) to be estimated and \( X_{ik} \) is the transpose of the variables’ vector for \( k \) variables, while error \( \epsilon \) is zero mean normally and independently distributed with constant variance of \( \sigma^2 \). The learning function, is a priori specified model in parametric supervised methods. Parametric supervised machine learning optimizes the parameters of a priori known learning function \( f(X) \) or weights \( \omega \) to achieve the best fit to data by minimizing sum of the squared errors (SSE).

A SWLM determines which variable is retained or removed in the model based on \( F \) statistics (Liao et al., 2015). An \( F \) statistic is obtained for each predictor and is a ratio explaining the part of the variance of the dependent variable (response) the predictor. Predictors with a larger \( F \) value are more significant and strongly correlated to the response and subsequently more useful in representing the variation of the response.

Lasso is a variable selection method that minimizes the sum of loss function (squared residuals or SSE) under an upper constraint on the sum of the absolute values of the model parameters, \( w_k \) in Eq. [1] and [2], to reduce the complexity of the model. This shrinkage was originally introduced for use in least squares regression by penalizing coefficients of unimportant predictors, making them zero. The penalty factor controlling the amount of shrinkage is \( \lambda \). Cross-validation is used to adaptively find an optimal \( \lambda \) that minimizes the deviance. Two values of \( \lambda \) are usually taken into account: the value that minimizes the cross-validation mean squared error (denoted as \( \lambda_{\text{min}} \)) and the maximum value within one standard error from \( \lambda_{\text{min}} \) (denoted as \( \lambda_{\text{ISE}} \)). During this minimization of an optimization problem, some coefficients are shrunk [i.e., \( w_k(X) = 0 \)] for some values of \( k \) depending on \( \lambda \). Thus, the features with coefficients equal to zero are excluded from the model. For a given value of \( \lambda \), Lasso uses a quadratic programming algorithm to solve

\[
\min \left[ \sum_{i=1}^{n} y_i - \left( \sum_{k=1}^{n} X_{ik}^T w_k \right)^2 + \lambda \sum_{k=1}^{n} |w_k| \right]
\]

where \( n \) is the number of observations, with \( y_i \) responses. By increasing \( \lambda \), the number of non-zero components of \( w \) decreases, whereas stepwise regression is a systematic procedure for adding and removing predictors based on \( F \) statistics. Therefore, when \( \lambda \) is very small in Eq. [2], the Lasso solution should be very close to the ordinary least squared loss function. A Lasso variables trace plot shows the coefficients of variables \( (w_k) \) to predict the response based on \( \lambda \) variation.

Nonparametric Machine Learning Techniques

From a function space view to Eq. [1] and in a nonparametric MLT framework, any number of arbitrary \( f \) values in Eq. [1] for each pair of data \( D(X,y) \) including input \( (X = \text{soil parameters}) \) and response \( y = \log(K_p) \) or \( \log(K_{10}) \) can be generated and left unspecified. However, these functions, called latent functions, should be smooth and flexible. The Gaussian process (GP) defines distribution across a particular subset of random latent functions \( f = \{ f_1, \ldots, f_n \} \) corresponding to input \( f = \{ X_1, \ldots, X_n \} \), analogous to a Bayesian overview of linear regression for weights \( (w_k) \), but here instead of weights, functions are considered. In Gaussian process regressions, it is assumed that all \( f \) values, as a prior, have a normal distribution looking like \( f(X) \sim \text{GP}(m(x), k(x,x')) \), where \( m(x) \) is the mean and \( k(x,x') \) is the covariance or kernel function, computed as

\[
K(X, X) = \begin{bmatrix}
k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\
k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\
\vdots & \vdots & \ddots & \vdots \\
k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n)
\end{bmatrix}
\]

where the error has a normal distribution like \( N(\epsilon | 0, \sigma^2) \).

The GP begins with a prior distribution for \( f \) and updates this prior as new data are observed, producing the posterior distribution across functions. To avoid expensive posterior computations, a zero mean distribution for \( f(X) \) is assumed so that a GP prior on the \( f \) as \( p(f | X) = \text{GP}(f | 0, K) \) can be formed (Rasmussen and Nickisch, 2010; Hong et al., 2014), and the Gaussian likelihood of \( f \) is \( p(y | f, X) = \text{GP}(y | f, \sigma^2 I) \). The normality of likelihood and prior gives the posterior of \( f \) analytically as

\[
p(f | X, y) = \text{GP}(f | 0, K) \times \text{GP}(y | f, \sigma^2 I) \propto \text{GP}\left[ f | K + \sigma^2 I \right]^{-1} \text{GP}(y | K^{-1} + \sigma^{-2} I)^{-1}
\]

The posterior distribution of \( f \) is used to calculate the posterior predictive distribution \( \bar{f} \) at test point \( X^* \). Finally, the resulting posterior process \( \bar{f} \) for the prediction of test data is a Gaussian process with mean and covariance as (see Kang et al. [2015] and, for more details, Rasmussen and Nickisch [2010])

\[
\bar{f} = K(X^*, X) K(X, X) + \sigma^2 I^{-1} y
\]


\[
\text{cov}(f^*) = K(X^*, X^*) - K(X^*, X)[K(X, X) + \sigma^2 I]^{-1} K(X, X^*)
\]  

[6]

Like parametric methods, nonparametric techniques assume that similar predictor data sets correspond to similar target values, although this similarity is determined through covariance (or kernel) functions. Therefore, the proper selection of a kernel or covariance function is an important task because they determine the sample properties such as smoothness, length scale, and amplitude drawn from the GP to give a precise prediction for responses with inputs, which are close to trained data points. For example, the automatic relevance determination (ARD) squared exponential kernel function used in this study for GPR is

\[
k(x_i, x_j) = \sigma_f^2 \exp \left(-\frac{1}{2\sigma_k^2} \sum_{k=1}^{K} (x_{ik} - x_{jk})^T (x_{ik} - x_{jk}) \right)
\]  

[7]

where \(\sigma_f\) is amplitude and \(\sigma_k\) is the characteristic length scale, while both are kernel (hyper) parameters. The latter denotes the scale of the correlation in the input space or how far the input values could be from the hyperplane, so the targets become uncorrelated. The ARD length scale of each predictor identifies its role in the model and is sensitive to those with the shortest length scales (least invariant).

An SVM maps output from a labeled training input–output dataset. The input data through kernel functions are projected into a higher dimensional space called feature space to find the output \(y = f(x, w) + \text{noise}\) via \(f(x, w) = w \cdot \varphi(X) + b\), where \(\varphi(X)\) is the projected input data into feature space and \(w\) and \(b\) are the weight vector parameter and the bias of the searched regression function, respectively. Minimization of the following regularized risk function \(R(C)\) finds \(w\) and \(b\):

\[
R(C) = \frac{1}{2}\|w\|^2 + \frac{C}{n} \sum_{i=1}^{n} \epsilon_{i}\left(\|y_i - f(x_i, w)\| - \varepsilon \right)_{+}
\]

subject to

\[
\begin{align*}
&y_i - w \varphi(x_i) - b \leq \varepsilon + \xi_i \quad \xi_i \geq 0 \\
&w \varphi(x_i) + b - y_i \leq \varepsilon + \xi^*_i \quad \xi^*_i \geq 0
\end{align*}
\]

[8a,b]

Indeed, slack variables are those of under- and overprediction errors. Solving the convex quadratic problem Eq. [9] and introducing the Lagrangian multipliers \(\alpha\) and \(\alpha^\ast\) would yield in the ultimate regression function as

\[
f(x_i) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^\ast) \varphi(x_i) \varphi(x) + b
\]

[10]

where \(n\) is the number of training data, and the number of those training data points with nonzero coefficients in Eq. [9] is the number of support vectors (\(n_{sv}\)). Finally, the SVM function can be obtained as

\[
f(x_i) = \sum_{i=1}^{n_{sv}} (\alpha_i - \alpha_i^\ast) k(x_i, x) + b
\]

[11]

Sequential minimal optimization is used here to solve the SVM problem of Eq. [11] directly in the input space without calculation of the explicit transformed data \([\varphi(X)]\) by a Gaussian kernel function, which is \(k(x, x) = \exp(-\|x - x\|^2/2\sigma^2)\) (Rasmussen and Nickisch, 2010). A schematic diagram of the SVM is presented in Fig. 1.

![Schematic diagram of the support vector machine.](image-url)
Undoubtedly, input data variability and uneven soil properties cause uncertainty in PTF predictions. Availability of a large number of repeated datasets could be used to determine confidence intervals for the parameters describing performance of the model (e.g., RMSE) through the probability density function. Data resampling for PTF evaluation includes methods like bootstrap, jackknife, or cross-validation (Schaap and Leij, 1998; Lilly et al., 2008). One of the issues is the number of replacements or how the replaced samples mimic the real behavior of the dataset. To consider this issue, we first tested the models by bootstrapping, in which duplicates are sampled from the original dataset and these train and test the models. Furthermore, the original data were rearranged 250 times and the models were run with these new datasets. In all bootstraps, unobserved data were used for evaluation of the models' performance.

Results and Discussion

Data Analysis

The descriptive statistics of measured soil parameters show a high degree of variation, with coefficients of variation (CVs) varying between 21 and 107 (Table 1). Most of the soils, 76 ± 16%, are sandy. With an acceptable approximation, all parameters except OM, clay, and \( \theta_{pF4.2} \) are normally distributed (skewness and kurtosis are approximately between −1 to 1 and −3 to 3). These three parameters have large CVs of about 107, 80, and 79%. On the other hand, BD and \( \theta_{pF1} \) do not vary significantly, with CVs of 10 and 14. The CVs of water contents at higher matric potential vary in a wider range than water contents at lower potentials. As an example, the CV of \( \theta_{pF4.2} \) is about six times larger than the CV of \( \theta_{pF1} \). This is similar to the results of Baskan et al. (2013). The average \( \log(K_s) \) is 15.5 times higher than \( \log(K_{10}) \), the latter showing the greater variability.

<table>
<thead>
<tr>
<th>Table 1. Descriptive statistics of measured soil properties.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable†</td>
</tr>
<tr>
<td>Sand, %</td>
</tr>
<tr>
<td>Silt, %</td>
</tr>
<tr>
<td>Clay, %</td>
</tr>
<tr>
<td>OM, %</td>
</tr>
<tr>
<td>BD, g cm(^{-3})</td>
</tr>
<tr>
<td>( \theta_{pF1} ) m(^3) m(^{-3})</td>
</tr>
<tr>
<td>( \theta_{pF2} ) m(^3) m(^{-3})</td>
</tr>
<tr>
<td>( \theta_{pF3} ) m(^3) m(^{-3})</td>
</tr>
<tr>
<td>( \theta_{pF4.2} ) m(^3) m(^{-3})</td>
</tr>
<tr>
<td>( \log(K_{10}) ) log(cm d(^{-1}))</td>
</tr>
<tr>
<td>( \log(K_s) ) log(cm d(^{-1}))</td>
</tr>
</tbody>
</table>

† OM, organic matter; BD, bulk density; \( \theta_{pF1}, \theta_{pF2}, \theta_{pF3}, \theta_{pF4.2} \), water contents at −0.1, 1, 10, and 158 m tension, respectively; \( K_{10} \) and \( K_s \), unsaturated and saturated hydraulic conductivity.

<table>
<thead>
<tr>
<th>Table 2. Pairwise linear correlation coefficient for the data set with 451 soils.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter†</td>
</tr>
<tr>
<td>HOR</td>
</tr>
<tr>
<td>Sand</td>
</tr>
<tr>
<td>Silt</td>
</tr>
<tr>
<td>Clay</td>
</tr>
<tr>
<td>OM</td>
</tr>
<tr>
<td>BD</td>
</tr>
<tr>
<td>( \theta_{pF1} )</td>
</tr>
<tr>
<td>( \theta_{pF2} )</td>
</tr>
<tr>
<td>( \theta_{pF3} )</td>
</tr>
<tr>
<td>( \theta_{pF4.2} )</td>
</tr>
<tr>
<td>( \log(K_{10}) )</td>
</tr>
<tr>
<td>( \log(K_s) )</td>
</tr>
</tbody>
</table>

† HOR, pedological horizon of the column; OM, organic matter; BD, bulk density; \( \theta_{pF1}, \theta_{pF2}, \theta_{pF3}, \theta_{pF4.2} \), water contents at −0.1, 1, 10, and 158 m tension, respectively; \( K_{10} \) and \( K_s \), unsaturated and saturated hydraulic conductivity.
as described earlier by Nourbakhsh et al. (2004) and de Jong et al. (1983). When a dataset contains highly correlated variables, relatively small changes in the noise component of the regression model can greatly affect regression coefficients. Significant interrelations among variables facilitate using stepwise regression and Lasso regularization to reduce the number of predictors while maintaining high precision in the prediction. The positive interaction of OM with water contents is obvious (Table 2). The effect of OM on water retention is highly pronounced in sandy soils or soils with a low clay content, as also reported by Minasny and McBratney (2018), Ankenbauer and Lohseide (2017), and Hudson (2018). About 40% of log(K_{s}) was described by K_{s}, and there was a notable correlation between log(K_{s}) and the texture of the soil.

**Parametric Supervised Machine Learning**

For both parametric and nonparametric methods, a random 70% of the data were used for training the models, while the rest of the variables were left unvisited for the prediction of log(K_{s}) and log(K_{10}) in the testing stage. Note that similar data are attributed to all MLTs to allow comparison. The distribution of data selected for the training and testing setups is shown in Fig. 2.

The significant predictors are kept in SWLM based on their p value for F statistics of the hypothesis test. Because p values of all participating variables for PTFs of log(K_{s}) and log(K_{10}) were <0.05, they are not shown in Table 3. In the derived equations for the prediction of log(K_{s}) and log(K_{10}), HOR, \( \theta_{pF1} \), \( \theta_{pF2} \), \( \theta_{pF3} \), and \( \theta_{pF4.2} \) are common when predicting both parameters. Bulk density and silt are added as other parameters to predict \( K_{s} \), while only clay is used besides the common parameters for the prediction of log(K_{10}) by SWLM. Although PTFs are normally developed based on basic soil properties (such as BD, sand, silt, clay, and OM) (Xu et al., 2017), we added water retention data to trace their significance in the model. According to Table 3, stepwise regression showed the importance of water retention data for such predictions.

Lasso, as a regularization technique to perform regression, identified important predictors. In general, Lasso overestimated the value for \( K_{s} \) up to 10 cm d\(^{-1}\) and underestimated for higher values for both training and testing data (Fig. 3a and 3b). The performance of this method for test dataset prediction was relatively weak, with \( R^2 \) of 0.28 and 0.15 and RMSE of 0.80 and 0.71 cm d\(^{-1}\) (Table 4). This poor predictability by multiple linear models such as SWLM and Lasso

![Fig. 2. Distribution of data used for training models followed by data for the testing stage.](image)

**Table 3. Description of variables† in the supervised stepwise linear model (SWLM) and Lasso for prediction of saturated and near-saturated hydraulic conductivities (\( K_{s} \) and \( K_{10} \), respectively).**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient SWLM</th>
<th>Coefficient Lasso</th>
<th>Variable</th>
<th>Coefficient SWLM</th>
<th>Coefficient Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOR</td>
<td>4.254</td>
<td>5.453</td>
<td>HOR</td>
<td>2.795</td>
<td>2.219</td>
</tr>
<tr>
<td>silt</td>
<td>0.034</td>
<td>0.011</td>
<td>clay</td>
<td>0.101</td>
<td>-0.129</td>
</tr>
<tr>
<td>BD</td>
<td>-0.701</td>
<td>-0.870</td>
<td>( \theta_{pF1} )</td>
<td>-21.625</td>
<td>3.175</td>
</tr>
<tr>
<td>( \theta_{pF2} )</td>
<td>-1.798</td>
<td>-4.533</td>
<td>( \theta_{pF2} )</td>
<td>-29.303</td>
<td>-2.513</td>
</tr>
<tr>
<td>( \theta_{pF4.2} )</td>
<td>-1.248</td>
<td>-2.357</td>
<td>( \theta_{pF3} )</td>
<td>16.063</td>
<td>-3.382</td>
</tr>
<tr>
<td>( \theta_{pF3} )</td>
<td>5.686</td>
<td></td>
<td>( \theta_{pF4.2} )</td>
<td>-42.951</td>
<td>-1.960</td>
</tr>
<tr>
<td>( \theta_{pF3} )</td>
<td>-4.383</td>
<td></td>
<td>( \theta_{pF1} )</td>
<td>105.258</td>
<td></td>
</tr>
<tr>
<td>HOR, ( \theta_{pF1} )</td>
<td>1.253</td>
<td></td>
<td>( \theta_{pF2} )</td>
<td>-47.125</td>
<td></td>
</tr>
<tr>
<td>HOR, ( \theta_{pF2} )</td>
<td>-2.504</td>
<td></td>
<td>( \theta_{pF1} )</td>
<td>63.365</td>
<td></td>
</tr>
<tr>
<td>HOR, ( \theta_{pF3} )</td>
<td>3.232</td>
<td></td>
<td>( \theta_{pF1} )</td>
<td>-1.278</td>
<td></td>
</tr>
<tr>
<td>Clay</td>
<td>-0.031</td>
<td></td>
<td>( \theta_{pF4.2} )</td>
<td>-0.031</td>
<td></td>
</tr>
</tbody>
</table>

† HOR, pedological horizon of the column; OM, organic matter; BD, bulk density; \( \theta_{pF1} \), \( \theta_{pF2} \), \( \theta_{pF3} \), \( \theta_{pF4.2} \), water contents at -0.1, 1, 10, and 158 m tension, respectively.

† A constant term in the regression.
is consistent with Jorda et al. (2015), who reached a maximum $R^2$ of 0.13 for $K_s$ prediction. Here, SWLM performed better for prediction of $K_s$ rather than Lasso, as the RMSE dropped to 0.72 and 0.67 cm d$^{-1}$ from training to testing data. The other important distinction is the better performance of SWLM to predict smaller values of $\log(K_s)$, where the minimum value predicted by Lasso was 1.065 and 1.509 cm d$^{-1}$ in the training and testing datasets, respectively, while SWLM had values of −0.043 and 0.654 cm d$^{-1}$. Given $K_{10}$ in Fig. 3c and 3d, both models demonstrated better prediction because estimated $K_s$ may give values that are more independent of the macroporosity and structural properties of soils, which influence the $K_s$ but have a rather weak impact on the $K_{10}$. The $R^2$ for $K_{10}$ compared with $K_s$ improved by 44.9, 57.0, 60.8, and 74.8% for SWLM and Lasso training and testing datasets, respectively.

To feature selection or to underline which explanatory variables are mainly important for prediction of $\log(K_s)$ and $\log(K_{10})$, trace plots of coefficients are presented for the PTFs of $K_s$ and $K_{10}$ (Fig. 4). Lowering of $\lambda$ affects predictor coefficients. The penalty terms ($\lambda$) in Lasso constrain the size of the estimated coefficients $w$ in Eq. [1], as can be clearly seen in Fig. 4. For both targets (i.e., logarithms of $K_s$ and $K_{10}$), water contents play a critical role in prediction because not only do their coefficients not approach zero by reduction of $\lambda$ (Fig. 4), but also their values become larger. The $\lambda$ for minimum cross-validation error ($\lambda_{\text{min}}$) and minimum cross-validation error plus one standard deviation ($\lambda_{1SE}$) are 0.0280 and 0.0098, respectively, for $K_s$ and 0.1071 and 0.0087, respectively, for $\log(K_{10})$ (Fig. 4). For smaller values of $\lambda$ than $\lambda_{\text{min}}$, the coefficients of regression become closer to coefficients obtained through least

![Fig. 3. Measured vs. predicted logarithmic saturated and near-saturated hydraulic conductivities ($K_s$ and $K_{10}$) for parametric techniques including a stepwise linear model (SWLM) and Lasso with training and testing datasets.](image)

<table>
<thead>
<tr>
<th>Response</th>
<th>Model</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>RMSE</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$\log(K_s)$</td>
<td>SWLM</td>
<td>0.42</td>
<td>0.724</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>0.28</td>
<td>0.805</td>
</tr>
<tr>
<td>$\log(K_{10})$</td>
<td>SWLM</td>
<td>0.76</td>
<td>0.544</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>0.66</td>
<td>0.652</td>
</tr>
</tbody>
</table>
Increasing which was already criticized by Gamie and De Smedt (2018).

The remaining predictors have zero coefficients. Dashed vertical lines represent the value with minimal mean squared error \( (\lambda_{\text{min}}) \) and the \( \lambda \) value with minimal mean squared error plus one standard deviation \( (\lambda_{1\text{SE}}) \).

square estimation. To explore this, variation of \( \lambda \) shown on the horizontal axis of Fig. 4 (for \( K_s \)) shows that \( \theta_{\text{pF1}} \) and \( \theta_{\text{pF3}} \) have zero coefficients in \( \lambda_{1\text{SE}} \), while in \( \lambda_{\text{min}} \) they are in the model with values equal to 6.713 and 0.902, respectively. Other parameters such as BD, HOR, and silt, with coefficients close to zero as previously shown in Table 3, do not change significantly within the range of \( \lambda \) and remain close to zero. Therefore, the variation of BD is, however, plotted in Fig. 4.

For \( K_{10} \), the increasing impact of \( \theta_{\text{pF2}}, \theta_{\text{pF4.2}}, \) and \( \theta_{\text{pF1}} \) is associated with a reduced impact of \( \theta_{\text{pF3}} \) along with \( \lambda \) reduction. Increasing \( \lambda \) causes shrinkage of some variables; in other words, the number of nonzero components decreases. Thus, Lasso performs a kind of continuous subset selection that is preferable over stepwise methods. To summarize, both stepwise regression and Lasso previously shown in Table 3, do not change significantly within the range of \( \lambda \), with \( \theta_{\text{pF1}} \) and \( \theta_{\text{pF3}} \) associated with a reduced impact of \( \theta_{\text{pF2}} \) along with \( \lambda \) reduction.

Nonparametric Supervised Machine Learning

In supervised machine learning, similar predictors are expected to have close target values. This similarity, in kernel-based methods, is described by a covariance function between input variables. This covariance function is determined by different kernel functions, where appropriate selection of kernel parameters affects the accuracy of the ultimate prediction. Here, for each predictor, a separate length scale is specified due to the use of an ARD squared exponential kernel function in GPR, as indicated in Eq. [7]. The fully independent conditional approximation to approximate the true GPR kernel function is used, and subsequently three categories of estimated parameters including kernel parameters, noise variance, and the coefficient vector of basic functions are optimized by a quasi-Newton method.

The performance of GPR models for the prediction of \( \log(K_s) \) and \( \log(K_{10}) \) was evaluated with the same dataset used in the parametric MLTs. Figure 5 provides plots between observed and predicted values for \( \log(K_s) \) and \( \log(K_{10}) \), which proves the superiority of the GPR model over regression models.

The nonparametric GPR model explains 78.2 and 51.5% of \( \log(K_s) \) variations, with RMSE values of 0.412 and 0.664 cm d\(^{-1}\) (Table 5). Therefore, the weak performance of parametric methods (SWLM) to predict \( \log(K_s) \) has been improved by 46.2 and 50.3%, and overestimation of GPR for low values of \( \log(K_s) \) is not as severe as for parametric techniques. In the case of the use of GPR to predict \( \log(K_s) \), if the values of \( \log(K_s) \) are in the range of 2.52 ± 0.8 cm d\(^{-1}\), the prediction is the most precise. Furthermore, GPR failed to predict negative values with this dataset, which is expected because only a very few samples have negative values for \( \log(K_s) \). The trained GPR for \( K_s \) assigned the smallest length scale to the HOR, silt, \( \theta_{\text{pF1}} \), and \( \theta_{\text{pF3}} \) variables (summarized in Table 5), which makes them the most influential predictors. To examine this, the GPR model was trained with only these four parameters. The \( R^2 \) and RMSE for testing and training were obtained as 76.26 and 50.81%, respectively, and as 0.430 and 0.668 cm d\(^{-1}\), respectively, which are quite similar to values using all predictors in the model that are shown in Table 5.

Again, \( \log(K_{10}) \) prediction was more accurate than \( \log(K_s) \) because the \( R^2 \) and RMSE for training and testing were 93.3 and 72.5%, respectively, and 0.278, 0.554 cm d\(^{-1}\), respectively. For \( \log(K_{10}) \), overestimation occurred frequently for low values of \( \log(K_{10}) \) (less than zero) because the model was not trained enough for low-conductivity soils. Predictors with very low values of length scale, including HOR, silt, and OM, presented in Table 5, were used for new modeling by GPR; the \( R^2 \) for the training set did not change and the \( R^2 \) for the test set was reduced only by 4% compared with the GPR model with all predictors. There was also no significant change in RMSE for the new brief GPR model. Therefore, using GPR for prediction of \( \log(K_{10}) \) does not require data from the water retention curve for the dataset used in this study. Moreover, \( K_s \) prediction is also possible with information from horizon, silt percentage, and water contents at −10 and −1000 cm matric head. The positive effect of water content data or van Genuchten parameters in PTFs for the prediction of
hydraulic conductivity was also confirmed by Børgesen et al. (2008).

An SVM regression model composed of a Gaussian kernel function was trained by a similar training dataset for $K_s$ and $K_{10}$, and its performance is shown in Fig. 6. Appropriate scale factors were chosen using a heuristic procedure. Favorable performance of the radial basis function (Gaussian) was already demonstrated by Das et al. (2012). Internal parameters or hyperparameters of SVM are obtained through Bayesian optimization to try to minimize an objective function (Khlosi et al., 2016), which is a loss function in a cross-validated fit of an SVM model. While comparing RMSE and $R^2$ values with the results for GPR, a significant depletion can be observed especially in the prediction of log($K_s$) (Table 6); however, SVM still performed better than parametric PTFs. This weak performance, explaining only 36% of log($K_s$) in the testing stage with an RMSE of 0.761 cm d$^{-1}$, was caused by hydraulic conductivity was also confirmed by Børgesen et al. (2008).

An SVM regression model composed of a Gaussian kernel function was trained by a similar training dataset for $K_s$ and $K_{10}$, and its performance is shown in Fig. 6. Appropriate scale factors were chosen using a heuristic procedure. Favorable performance of the radial basis function (Gaussian) was already demonstrated by Das et al. (2012). Internal parameters or hyperparameters of SVM are obtained through Bayesian optimization to try to minimize an objective function (Khlosi et al., 2016), which is a loss function in a cross-validated fit of an SVM model. While comparing RMSE and $R^2$ values with the results for GPR, a significant depletion can be observed especially in the prediction of log($K_s$) (Table 6); however, SVM still performed better than parametric PTFs. This weak performance, explaining only 36% of log($K_s$) in the testing stage with an RMSE of 0.761 cm d$^{-1}$, was caused by

Table 5. Optimal Gaussian process regression (GPR) models and their variables and evaluation; $\sigma_1$ is the length scale of the predictor, and a very low value of $\sigma_1$ represents high influence of the predictor on the response.

<table>
<thead>
<tr>
<th>Target Variable†</th>
<th>$\log(\sigma_1)$</th>
<th>Training $R^2$</th>
<th>Testing $R^2$</th>
<th>Training RMSE</th>
<th>Testing RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>log($K_s$)</td>
<td>HOR</td>
<td>−0.312</td>
<td>0.78</td>
<td>0.412</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>silt</td>
<td>−0.041</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_{pF1}$</td>
<td>−2.088</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_{pF3}$</td>
<td>−2.397</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log($K_{10}$)</td>
<td>HOR</td>
<td>−35.99</td>
<td>0.93</td>
<td>0.278</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>silt</td>
<td>−74.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OM</td>
<td>−61.49</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† HOR, pedological horizon of the column; OM, organic matter; $\theta_{pF1}$ and $\theta_{pF3}$, water contents at −0.1 and 10 m tension, respectively.
the nature of the SVM. Support vector machines are a subset of sparse kernel machines, for which a separate length scale for each predictor cannot be determined (Das et al., 2012). Again, the zone in which log($K_s$) values are between 2 and 4 cm d$^{-1}$ were predicted better because the dataset contains more sandy soils.

Optimal machine learning hyperparameters using Bayesian optimization were determined by Ensemble regression to minimize cross-validation loss (error). However, the boosting method used in a neural network model resulted in better generalization by reducing both the bias and variance of the neural network models in Parasuraman et al. (2006). Here, optimization of bagging ensemble aggregation gave less error and higher $R^2$ for prediction of both log($K_s$) and log($K_{10}$) (Table 7; Fig. 7).

The performance statistics for the Ensemble method were more precise than for GPR methods, while $R^2$ for $K_s$ and $K_{10}$ in the testing set improved by >3% and RMSE decreased from 0.664 and 0.554 in GPR to 0.650 and 0.529 cm d$^{-1}$. To achieve this accuracy, a random forest with 495 and 498 regression trees for $K_s$ and $K_{10}$, respectively, were trained according to Table 7. This relatively high accuracy prediction by ENS and GPR methods, in terms of the RMSE, are presented in Fig. 9 for both log($K_s$) and log($K_{10}$). For $K_s$, it was confirmed that the GPR and ENS methods are more robust models than SVM because 95% of their RMSE values (confidence interval) vary between (0.452–0.700), (0.486–0.704), and (0.712–0.940) (cm d$^{-1}$), respectively (Fig. 9). This is also in agreement with RMSEs reported in Tables 5, 6, and 7, respectively, with values of 0.664, 0.650, and 0.761. This demonstrates that the random selection of data from the original data set as chosen earlier is reliable. As expected, RMSE values for predicted log($K_{10}$) reduced significantly when compared with predicted log($K_s$), confirming unsaturated hydraulic conductivity at −10 cm matric head to be less subject to variability. Here, the robustness of all models and lack of overfitting for log($K_{10}$) prediction was obviously evaluated by bootstrapping, where RMSE values

### Table 6. Optimal parameters of support vector machines and their performance evaluation for the prediction of logarithmic saturated and near-saturated hydraulic conductivities ($K_s$ and $K_{10}$).

<table>
<thead>
<tr>
<th>Target</th>
<th>Parameters</th>
<th>Bias $\epsilon$</th>
<th>No. of support vectors</th>
<th>Training $R^2$</th>
<th>RMSE</th>
<th>Testing $R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>log($K_s$)</td>
<td>2.149</td>
<td>0.0775</td>
<td>274</td>
<td>0.81</td>
<td>0.384</td>
<td>0.36</td>
<td>0.761</td>
</tr>
<tr>
<td>log($K_{10}$)</td>
<td>0.776</td>
<td>0.1069</td>
<td>255</td>
<td>0.93</td>
<td>0.287</td>
<td>0.67</td>
<td>0.609</td>
</tr>
</tbody>
</table>

### Table 7. Evaluation and parameterization of ensemble regressions for the prediction of saturated and near-saturated hydraulic conductivities ($K_s$ and $K_{10}$).

<table>
<thead>
<tr>
<th>Target</th>
<th>Algorithm</th>
<th>Learning cycles</th>
<th>Training $R^2$</th>
<th>RMSE</th>
<th>Testing $R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>log($K_s$)</td>
<td>Bag†</td>
<td>495</td>
<td>0.78</td>
<td>0.414</td>
<td>0.53</td>
<td>0.650</td>
</tr>
<tr>
<td>log($K_{10}$)</td>
<td>Bag</td>
<td>498</td>
<td>0.92</td>
<td>0.309</td>
<td>0.75</td>
<td>0.529</td>
</tr>
</tbody>
</table>

† Bagging ensemble aggregation algorithm.

As mentioned, the other advantage of the ENS method is to estimate predictor importance by permuting out-of-bag observations as shown in Fig. 8. Because larger values indicate predictors that have a greater influence on predictions, BD, $\theta_{pF3}$, and $\theta_{pF1}$ are the most influential predictors on $K_{10}$ and $\theta_{pF2}$, silt, and HOR on $K_s$. This also confirms that, based on the availability of parameters to estimate hydraulic conductivities, ENS or GPR can be practical.

### Evaluation of Models

Two hundred fifty bootstraps were performed on the original data with and without replacement. First, SVM, GPR, and ENS were trained with all bootstraps with replacements. Histograms of the performance of these models, in terms of the RMSE, are presented in Fig. 9 for both log($K_s$) and log($K_{10}$). For $K_s$, it was confirmed that the GPR and ENS methods are more robust models than SVM because 95% of their RMSE values (confidence interval) vary between (0.452–0.700), (0.486–0.704), and (0.712–0.940) (cm d$^{-1}$), respectively (Fig. 9). This is also in agreement with RMSEs reported in Tables 5, 6, and 7, respectively, with values of 0.664, 0.650, and 0.761. This demonstrates that the random selection of data from the original data set as chosen earlier is reliable. As expected, RMSE values for predicted log($K_{10}$) reduced significantly when compared with predicted log($K_s$), confirming unsaturated hydraulic conductivity at −10 cm matric head to be less subject to variability. Here, the robustness of all models and lack of overfitting for log($K_{10}$) prediction was obviously evaluated by bootstrapping, where RMSE values
for GRP, ENS, and SVM models are $0.457 \pm 0.050$, $0.471 \pm 0.055$, and $0.660 \pm 0.0560$ cm d$^{-1}$, all smaller than the reported values of RMSEs obtained through modeling of the original dataset.

In the second scenario of model evaluation, the resampling (bootstrapping) from the original dataset was performed without any replacement. In other words, the data were rearranged 250 times. As shown in Fig. 10, the average RMSE of predicted log($K_s$) values by ENS, GPR, and SVM are 0.674, 0.692, and 0.834 cm d$^{-1}$, respectively; for log($K_{10}$), the values are 0.518, 0.531, and 0.664 cm d$^{-1}$, respectively.

**Conclusions**

We achieved a substantial improvement in predictive accuracy using supervised machine-learning techniques. Gaussian process regression and ensemble modeling showed good performance...
particularly when the low values of prediction error were verified by bootstrapping. This indicates that machine-learning techniques can be effectively used in predicting soil hydraulic conductivity compared with common parametric pedotransfer functions. Prediction of $K_s$ was less accurate but still performed better when compared with commonly used PTFs. These methods are well calibrated for soils with relatively high conductivity when their values for log($K_s$) and log($K_{10}$) are between 1.8 to 3.8 and 0 to 2.9 cm d$^{-1}$, respectively. Among nonparametric methods, GPR and ENS (random forest) described more than two-thirds of the variation of $K_{10}$ in the testing stage, which later could be used for mapping targets. It was also found that parametric PTFs could be developed relying only on water content values at different matric potentials from small rings. Lasso, with a simple and practical equation including six parameters (clay, bulk density, and water content values), is a strong model to predict log($K_{10}$) performing even better than SVM.

**Data Availability**

The data set used in this study includes soil physical properties of columns (horizon, texture, bulk density), water contents at −0.1-, 1-, 10-, and 158-m tension, and finally saturated and unsaturated (at −0.1 m) hydraulic conductivities (Kotlar et al., 2019).

**Acknowledgments**

A.M. Kotlar is grateful to FAPESP Foundation, Brazil, for providing financial support (scholarship) for his internship study (Process no. 2016/18636-7) at Aarhus University, Department of Agroecology. This work has generously been supported by the project “Mapping of risk areas for phosphorus loss and phosphorus sensitive water areas” funded by the Danish Environment Agency.

**References**


Lilly, A., A. Nemes, W.J. Rawls, and Y.A. Pachepsky. 2008. Probabilistic...


