ESTIMATING SATURATED HYDRAULIC CONDUCTIVITY FOR SOUTHEASTERN SOILS USING DECISION TREE ANALYSIS

by

AUBREY C. SHIRLEY

(Under the Direction of David Radcliffe)

ABSTRACT

Pedotransfer functions (PTFs) are used to predict saturated hydraulic conductivity (K_s) from more easily measured soil properties. Our objective was to determine if soil morphology was an important factor in predicting K_s using PTFs. We used soil profile descriptions for nine soils from the S-124 regional project dataset describing soils of the southeastern United States. Our best decision-tree model predicted $\log_{10} K_s$ (cm day^{-1}) with an average $\log_{10}$ root mean square residual (RMSR) of 0.8017. The best models used bulk density and texture but not soil morphological descriptors. Sand textural class predicted the highest K_s. For the finer textured soils, the splits were based on bulk density. The NRCS method predicted K_s with a RMSR of 0.9562. Morphological descriptors of soil structure may not have been important because bulk density acted as a surrogate for structure.

INDEX WORDS: Regression tees, saturated hydraulic conductivity, soil structure

ESTIMATING SATURATED HYDRAULIC CONDUCTIVITY FOR SOUTHEASTERN SOILS USING DECISION TREE ANALYSIS

by

AUBREY C. SHIRLEY

BS, University of Georgia, 2008

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2011

ESTIMATING SATURATED HYDRAULIC CONDUCTIVITY FOR SOUTHEASTERN SOILS USING DECISION TREE ANALYSIS

by

AUBREY C. SHIRLEY

Major Professor: David Radcliffe

Committee: Miguel Cabrera
Todd Rasmussen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2011

ACKNOWLEDGEMENTS

I would like to thank my major professor Dr. David Radcliffe, and my committee members, Dr. Miguel Cabrera and Dr. Todd Rasmussen. I would also like to thank Dr. Attila Nemes for his guidance.

I want to thank my family for their continuous support and encouragement throughout my time at the University of Georgia. Finally, I want to thank God for allowing me to finish this chapter of my life.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER

1 INTRODUCTION 1

2 LITERATURE REVIEW 3

References 13

3 PEDOTRANSFER FUNCTION FOR ESTIMATING SATURATED HYDRAULIC CONDUCTIVITY IN STRUCTURED SOILS 18

Abstract 19

Introduction 20

Materials and Methods 22

Results and Discussion 29

Conclusions 40

References 57

LIST OF TABLES

Page

Table 1: Description and grouping of input variables and output variable.....43

Table 2: Correlations between variables in the S-124 data set 44

Table 3: Mean and coefficient of variance (CV) of untransformed and $\log_{10}$-transformed K_s by textural class for the S-124 data set 45

Table 4: Performance of decision tree models, in terms of root mean squared residual (RMSR), calculated by mean of 100 model runs using 247 different input combinations of input variables 46

Table 5: Input variables and their probabilities in each splitting level for the decision tree model considering all input variables. 47

Table 6: Input variables and their probabilities in each splitting level for the best decision tree model considering particle size distribution (PSD), grade (GRD), and bulk density (D_b)….....48

Table 7: Input variables and their probabilities in each splitting level for the decision tree model considering ped size (PED), crack orientation (CRK), grade (GRD), and consistence (CST).....49

Table 8: Input variables and their probabilities in each splitting level, for the decision tree model considering ped size (PED), consistence (CST), and bulk density (D_b).....50

LIST OF FIGURES

Page

Figure 1: Flow diagram of algorithm for developing regression trees 51

Figure 2: The Soil Survey method for estimating saturated hydraulic conductivity (K_s) using only texture and bulk density. 52

Figure 3: Most likely tree structure of best ranking regression tree model, the average saturated hydraulic conductivity (K_s)in bold and standard deviation (SD) in parenthesis, all values are $\log_{10}$ transformed……..53

Figure 4: Decision tree using only textural class (TXT) and horizon designation (HOR), the average hydraulic conductivity (K_s) in bold and standard deviation (SD) in parenthesis, all values are $\log_{10}$ transformed ..….....54

Figure 5: Most likely regression tree structure using only morphological descriptors of structure (PED, CRK, GRD, and CST), the average saturated hydraulic conductivity (K_s) in bold and standard deviation (SD) in parenthesis, all values are $\log_{10}$ transformed.....55

Figure 6: Most likely regression tree structure using only using ped size (PED), consistence (CST), and bulk density (D_b), the average K_s in bold and standard deviation (SD) in parenthesis, all values are $\log_{10}$transformed…….....56

CHAPTER 1

INTRODUCTION

Pedotransfer functions (PTFs) are useful tools to estimate saturated and unsaturated soil hydraulic properties that would otherwise be too expensive and too time consuming to measure (Lilly et al., 2008; Wösten et al., 2001). The large amount of qualitative data in soil surveys also makes PTFs desirable as a way to predict soil hydraulic properties. PTFs commonly use soil texture to estimate saturated hydraulic conductivity (K_s), even though some researchers have concluded that soil texture does not yield accurate predictions (Wagner et al., 1998; Tietje and Hennings,1996).This inability to predict K_s occurs because different climates can have similar soil texture, but the soil structure can vary greatly due to differing rainfall and temperature regimes (Wagner et al., 1998; Tietje and Hennings,1996). For example, an unstructured clay could have a low K_s, while a structured clay could have a higher K_s. Therefore, PTFs that combine both structure and texture could be more accurate than PTFs that rely solely on texture. Soil structure is difficult to quantify, but Lilly et al. (2008) developed a method to incorporate structure in a PTF. They used the HYPRES (Hydraulic Properties of European Soils database, which contains many different soils from Europe (Wösten et al., 1999). Lilly et al. (2008) identified the structural elements using a binary approach. If a pedon had a particular class of structure, then it was given a value of one (member). If it was not a member, then it was given a value of zero. This member/non-member method was also applied to both field-determined textural classes and laboratory-determined textural classes. Using decision trees, Lilly et al.

(2008) predicted K_s within one order of magnitude of the true value.

The S-124 dataset (published as a USDA Southeastern Cooperative Series Bulletin), similar to the HYPRES database, is a dataset that contains soil hydraulic properties. This is one of the few datasets that contain measured particle size, morphological descriptions of structure, and K_s of U.S. soils. These data include *in situ* field characterization and laboratory determined soil properties including K_s collected from 1977 to 1983. Twenty-one southeastern soils were included in the study: Troup and Lakeland (Dane et al., 1983); Norfolk, Dothan, Goldsboro, and Wagram (Quisenberry et al., 1987); Captina, Gigger, Grenada, Loring, Olivier, and Sharkey (Römkens et al., 1986); Bethany, Konawa, and Tipton (Nofziger et al., 1983); Vicksburg, Memphis, and Maury (Römkens et al., 1986); Cecil (Bruce et al., 1983), and Fullerton and Sequoia (Luxmore 1982). From this dataset, we developed PTFs using regression trees.

The objective of this study was to use the morphological descriptions of soil structure and the quantitative data (particle- size distribution and bulk density) that are available in these Southern Cooperative Series Bulletins to predict K_s and compare these predictions with the method used by the USDA-NRCS Soil Survey (USDA, 2010).

CHAPTER 2

LITERATURE REVIEW

Pedotransfer functions (PTFs) are important tools used for predicting soil hydraulic properties such as unsaturated hydraulic conductivity ($K(h)$) and saturated hydraulic conductivity (K_s). The term pedotransfer function was first introduced by Bouma (1989). PTFs are tools that allow for a translation of data from what is presently unavailable into some useful type of data. At present, most PTFs use readily available soil information such as texture, bulk density and organic matter for input variables (Tietje and Hennings, 1996).

Bouma (1989) stated that PTFs can be separated into two types: continuous and class. Continuous PTFs use numerical values, for instance, percentage of sand, silt, and clay. Class PTFs, however, use different categorical data to relate to other soil properties, e.g. soil textural class. From these two main PTF categories, many different statistical modeling methods for predicting K_s have been developed. According to McBratney et al. (2002), the most common methods include: multiple linear regression, artificial neural networks, generalized linear models, general additive models, group method of data handling, multiple adaptive regression splines, and regression trees. Wösten et al. (2001) also discuss the major techniques used for PTFs: regression analysis, artificial neural networks, group method of data handling and classification/regression trees. These will be described in more detail in the following pages.

Artificial Neural Networks

Artificial Neural Networks (ANNs) can mirror the behavior of complex systems because they are capable of varying both the strength and structure of component connections (Pachepsky

and Schaap, 2004). This method has been used in PTFs by Schaap and Leij (1998), Pachepsky et al. (1996), Tamari et al. (1996), and Minasny et al. (1999). ANNs have an advantage over traditional regression methods because no preconceived model concepts are needed. These functions find optimum relationships between the soil and hydraulic properties using an iterative calibration procedure (Pachepsky and Schaap, 2004).

ANNs are connected by processors called neurons; these neurons are further connected by positive or negative numbers (weights). These weights are used to describe the influence of a neuron on the computation of a particular scalar function (activation function or response; Siegelmann, 1999).

Bigus (1996) explains three unique types of neural networks: supervised, unsupervised, and reinforcement. Supervised networks are some of the most frequently used, followed by unsupervised, and reinforcement. Supervised networks can be thought of as "programming by example" (Bigus, 1996). Supervised networks work by being given a problem. From this problem they work to determine a correct response. If the response/answer is incorrect, the program will be shown the correct answer and continue to work until the correct solution is determined.

Unsupervised networks, are a clustering technique; however, there is no predetermined solution, making this a learning type of algorithm. This algorithm learns because it is capable of understanding inputs and outputs and how they are related to one another (Bigus, 1996).

Reinforcement learning is a type of neural network in which there are examples of the problems but the exact solution is not known. This type of neural network is capable of solving difficult time-related problems (Bigus 1996).

All neural networks can be arranged in three forms: single-layer feed-forward networks, multilayer feed-forward networks, and recurrent networks (Haykin, 1999). The single layer network consists of both input and output nodes. The input node layer is used to predict the output node layers. This process is not reversible, meaning the input layer is always used to predict an output node (Haykin, 1999). The multilayer feed-forward network is similar to the single layer feed-forward network; however, the difference arises due to the presence of one or more hidden nodes (Haykin, 1999). Increasing the number of hidden nodes allows for the determination of higher order statistics. (Haykin, 1999). The recurrent network differs from the feed-forward loop because it has at least one feedback loop. These feedback loops have a large impact on the learning capability of the network and performance

Multiple Linear Regression

Multiple linear regression, like simple linear regression, is a technique that develops a linear relationship between a prediction variable and a response variable. However, in the case of multiple linear regression there is more than one prediction variable, while there is still a single response variable. An example relating to soils would be using particle size distribution, bulk density, and organic matter to predict K_s. Multiple linear regression equations are in the form (Timm, 2002):

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 \ + \ldots + x_k\beta_k + e$$

Equation 1

where y is the response variable and $x_{1\ldots}\, x_k$ are the predictor variables. The coefficients for each of the variables are $\beta_1 \ldots \beta_k$ and β_0 is the intercept. The residual error is e.

Group Method of Data Handling

The GMDH is a process which involves neural network type equations to relate input and outputs. This method is made up of three different steps as explained by Pachepsky and Schaap (2004). They give the example of input data $x_1, x_2 \ldots, x_n$ and y being the output variable. Step one involves getting estimates of y by using quadratic regression equations in the form:

$$z_i = \beta_0 + \beta_1 x_i + \beta_2 x_{i+1} + \beta_3 x_i^2 + \beta_4 x_i x_{i+1} + \beta_5 x_{I+1}^2$$

Equation 2

where $\beta_0 \ldots \beta_5$ are regression coefficients and z_i is a preliminary estimate of y. All independent variables are taken two at a time forming x_i and x_{i+1}. The second step consists of eliminating the least effective preliminary estimates. Step three asks if these equations can be improved. If not, iteration ceases, and the network is built. If prediction improvement is made then steps one and two are continued until there is no improvement in the iteration from the previous set of equations.

Decision Trees (Classification and Regression)

Decision trees can be divided into two categories: classification and regression trees. The difference in these two statistical techniques arises from the fact that classification trees use only categorical data, while regression trees can have both numerical and categorical input. (Pachepsky and Schaap, 2004).

Regression and classification trees are non-parametric methods (they make no assumption about the probability distributions of the variables) that are used to uncover

relationships in data. Regression trees work by first forming splits, called nodes. These splits form from a parent node, and from this new node form two other nodes, then four, and so on. This is referred to as recursive partitioning. A pre-determined grouping size must be set before any splitting can begin. The grouping size is the minimum number of samples that must be present in a particular group (at a node) to continue splitting. The criteria for splitting is based on whether the predictor variable is numerical or categorical. For instance, if the predictor variable is numerical, then the split will be less than or equal to some specified value of the variable on the left node and the right node will be greater than the value. If the predictor variable is categorical, then the split will be based on a yes or no question. In the case of categorical variables, there can be 2^{k-1}-1 possible partitions, where k is the total number of levels (Pachepsky and Schaap, 2004).

Each split must be made so that it maximizes homogeneity within each split group while simultaneously maximizing the heterogeneity between groups. The homogeneity of each group is measured by the deviance (D) within the group:

$$\mathrm{D} = \sum_{\mathrm{i}} (\mathrm{y_i} - \bar{\mathrm{y}})^2$$

Equation 3

where y_i is a response variable and $\bar{y}$ is the mean of all response variables. In order to determine which partition to choose, the change in deviance (ΔD) is calculated by subtracting the deviance of the right group (D_R) and the left group (D_L) from the deviance before splitting. (Pachepsky and Schaap 2004). The partition with the greatest deviance is chosen as the split. This splitting

can continue until a maximum number of terminal nodes are formed (Prasad et al., 2006). The maximum number of nodes depends on the number of samples in the dataset. A tree could theoretically form until each sample has a corresponding node. This of course causes an over parameterization of the data so the decision tree must be pruned back to an optimal size. According to Sutton (2005), the most common way to prune a regression tree is referred to as cost-complexity pruning. This method cuts the nodes that arise from non-terminal nodes, the selected node is one in which the "pruned nodes provide the smallest per node decrease in the resubstitution misclassification rate". Sutton (2005) described two techniques that use this method. One method uses an independent dataset. The basis for this method involves using the least complex (fewer input variables), but accurate tree. If a less complex tree is within one standard error of a more complex tree which has the lowest estimated error, then the less complex tree can be chosen. Cross-validation is another pruning process (the process used in our study). This process grows a large tree and systematically prunes back the tree until the deviance is minimized with regard to the testing data set, and the tree with least error is selected (Sutton, 2005).

Soil Survey Method of Predicting K_s

The USDA Natural Resource Conservation Service (NRCS) Soil Survey Division has developed a PTF for estimating $K_{s,}$ using only bulk density and textural classes, with some over-riding conditions. The method first determines to which bulk density class a soil sample belongs. There are three classes: high (1.46 to 1.72 g cm^{-3}), medium (1.19 to 1.59 g cm^{-3}), and low (0.93 to 1.32 g cm^{-3}). There is overlap between the three groups, so samples falling in between two particular groups must be checked to ensure which group is suitable. Once the bulk density group

has been established, the textural class is used to predict K_s, by finding the K_s range for the soil sample.

Pedotransfer Functions using Structure and Other Hydraulic Properties

Many PTFs predict K_s without using morphological descriptors of soil structure (Jabro, 1992; Rawls et al., 1998; Loague, 1993). Including qualitative measures of soil structure as a predictor variable when estimating soil hydraulic properties could yield better results than using only quantitative predictor variables such as particle size and bulk density (Wösten et al., 2001; Rawls et al., 2004). The increase in prediction power of structure comes about because models based purely on texture do not take into account worm holes, cracks, or root channels (Rawls et al., 2004).

Other soil properties such as water retention and unsaturated hydraulic conductivity may also be predicted more accurately when a structural component is added. For example, Pachepsky and Rawls (2003) used regression trees to estimate water retention at -33 and -1500 kPa, which were taken from the US National Soil Characterization database, consisting of 2140 samples. They found that ped grade (which is a measure of the strength of the ped structure) was a strong predictor of water retention: the stronger the grade the greater the water retention**.** Pachepsky et al. (2006) developed a PTF using both regression and classification trees from the same database. They used these trees to group soil samples to predict the -33 kPa water content. Grade was a significant parameter for all intermediate textures, while ped shape was not. They found a stronger grade increased water retention. Structure can play an important role in predicting other soil properties (Levine et al., 1995; Crawford, 1994, and Anderson and Bouma,

1973). Moreover, state soil surveys contain suitable structural data for the use of PTFs in the prediction of hydraulic properties, including K_s.

McKeague et al. (1982) developed eight predictor variable classes to estimate K_s. These classes consisted of structure, porosity, texture, consistence, and the density of the horizons. They found that macroporosity and structure were major factors in predicting the K_s of many soils. Clayey soils, unlike coarser textured soils, had a higher prevalence of biopores. Low predicted K_s were associated with compacted, massive, clayey soils.

Another study conducted by McKenzie and Jacquier (1997) tested the predictive capability of morphological data. They found a relationship between grade and K_s, although the data were variable. They used regression trees to search for more accurate interactions between many of these variables. From their research they concluded that field texture, grade, areal porosity, dispersion index, and horizon type improved the prediction. They also concluded that using simple morphological data (i.e. grade) could be used for simple land evaluation.

Lin et al. (1999) proposed a method that would allow for the quantification of soil structure, thus enabling this structure to be used in PTFs. They used texture, moisture, pedality, macroporosity, and root density from 96 horizons. The term pedality was used to describe ped grade, size, and shape (Lin et al., 1999a). Their findings indicated that soils with prismatic shape had higher infiltration rates than those with blocky shape. The reason was that soils with prismatic structure had more macropores than the blocky soils (Lin et al., 1999). In a second paper published by Lin et al. (1999b), they used the data collected to develop both a class and continuous pedotransfer function. The predictor variables for the class PTF included morphometric indices, initial moisture state, pedality, macroporosity, and root density. The

continuous function variables included mass fractions of sand, silt, and clay, organic matter content, dry bulk density, initial soil gravimetric water content, field estimated macroporosity, and very fine porosity at the root-soil interface. Using multiple regression, Lin et al. (1999b) found that the class PTFs yielded results similar to that of the continuous PTF even though the class PTFs used qualitative morphological data. They also showed that structure impacted K_s and macropore flow , while, texture influenced micropore flow.

Lilly (2000) studied the effects of soil structure on predicting field K_s in Scottish soils. Six hundred twenty-seven samples were taken from topsoil and subsoil horizons. Saturated hydraulic conductivity ranged from 0.06 to 1036.8 cm day^{-1}. The author developed 49 unique structure classes, based on the FAO soil structure classes. Saturated hydraulic conductivity increased as the ped size decreased. Soils that had peds larger than 50 mm had K_s < 10 cm day^{-1}. Soils with peds < 20 mm had K_s ranging from 30 to 100 cm day^{-1}. With this data, Lilly (2000) concluded that the lowest conductivities occurred with no structure (massive) and vertical structure (coarse and very coarse prismatic).

Lilly et al. (2008) used regression trees to predict K_s, while incorporating structure. They were able to incorporate structure and other qualitative predictor variables into the PTF using a binary approach. Their data came from the HYPRES (Hydraulic Properties of European Soils) database. This is a database of soils from Europe containing both quantitative and qualitative data. There were a total of 5521 samples with replicates (Lilly et al., 2008).

A total of six input variables were used: three were qualitative and three were quantitative. The groups were ped size (PED), crack orientation (CRK), particle size distribution (PSD), field texture determined by hand (TXT), bulk density (BD), and horizon designation

(HOR). The dataset consisted of 502 samples, each with a measured K_s. The K_s for each of the samples was log-transformed (Lilly et al., 2008).

The regression technique used by Lilly et al. (2008) used the jackknife cross-validation technique, re-sampling without replacement, to prune the decision trees. This method systematically eliminated each of the input groups, until all combinations were evaluated. This resulted in 63 unique combinations of predictor variables. Each combination produced unique decision trees during 100 re-sampling steps. The optimization yielded a group size of 91, terminal node size of 5, and a training dataset size of 411 samples.

They found that the best model predictor of K_s consisted entirely of qualitative variables (HOR, PED, and TXT). The difference between the best and worst model was only 0.128 $\log_{10}$ K_s in cm d^{-1} (Lilly et al., 2008). This small difference showed that the models in the top quarter could perform as well as the models in the lowest quarter (Lilly et al., 2008). The full model that used all predictor variables was ranked 20^{th} and had a mean RMSR of 0.9696 $\log_{10}$ K_s, indicating that it could predict K_s within an order of magnitude. Lilly et al. (2008) found that field determined data along with structure were the most important predictor values.

References

Anderson, J.L. and J. Bouma. 1973. Division S-5 – Soil Genesis, Morphology, Classification: Relationships Between Saturated Hydraulic Conductivity and Morphometric Data of an Argillic Horizon. Soil Sci. Soc. Amer. Proc., 37:409-413.

Bigus, P. Joseph. 1996. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. McGraw (ed.).

Bruce, R., J. Dane, V. Quisenberry, N. Powell, and A. Thomas 1983. Physical Characteristics of Soils in the Southern Region: Cecil. S. Coop. Ser. Bull. 267.

Crawford, J.W. 1994. The relationship between structure and the hydraulic conductivity. European Journal of Soil Sci. 45: 493-502.

Dane, J., D.K. Cassel, J.M. Davidson, W.L. Pollans, and V.L. Quisenberry. 1983. Physical Characteristics of Soils of the Southern Region: Troup and Lakeland Series. S. Coop. Ser. Bull. 262. Alabama Agric. Exp. Stn., Auburn University.

Farlow, J. Stanley. 1984. Self-Organizing Methods in Modeling. GMDH Type Algorithms. Marcel Dekker (ed.) New York, New York.

Haykin, Simon. 1999. Neural Networks: A Comprehensive Foundation. McMaster University. Hamilton, Ontario, Canada. Prentice Hall (ed.). New Jersey.

Loague, Keith. 1993. Using Soil Texture to Estimate Saturated Hydraulic Conductivity and the Impact on the Impact on Rainfall-Runoff Simulations. Water Resource Bulletin. 28:687-693.

Lilly, A. 2000. The relationship between field-saturated hydraulic conductivity and soil structure: development of class pedotransfer functions. Soil Use and Management. 16: 56-60.

Lilly, A., and H.S. Lin. 2004. Using soil morphological attributes and soil structure in pedotransfer functions. p. 115-141. *In* Ya. Pachepsky and W.J. Rawls (ed.) Development of pedotransfer functions in soil hydrology. Dev. Soil Sci. 30. Elsevier Amsterdam.

Levin, E.R., D.S. Kimes, and V.G. Sigillito. 1996. Classifying soil structure using neural networks. Ecological Modelling. 92: 101-108.

Lin, H.S., K.J. McInnes, L.P. Wilding, and C.T. Hallmark. 1999. Effects of Soil Morphology on Hydraulic Properties: I. Quantifications of Soil Morphology. Soil Sci. Soc. Am. J. 63:948-954.

Lin, H.S., K.J. McInnes, L.P. Wilding, and C.T. Hallmark. 1999. Effects of Soil Morphology on Hydraulic Properties: II. Hydraulic Pedotransfer Functions. Soil Sci. Soc. Am. J. 63:955-9961.

Luxmoore, R. (1982). "Physical characteristics of soils of the southern region: Fullerton and Sequoia series." S. Coop. Ser. Bull. 268. ORNL-5868, Oak Ridge National Lab.TN, (USA).

McKeague, J.A.,C. Wang, and G.C. Topp. 1982. Estimating Saturated Hydraulic Conductivity from Soil Morphology. Soil Sci. Soc. Am. J. 46:1240-1244.

McKenzie, Neil and David Jacquier. 1997. Improving the field estimation of saturated hydraulic conductivity in soil survey. Aust. J. Soil Res. 35: 803-825.

Minasny, B., McBratney, A.B., Bristow, K., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93, 225-253.

Nofziger, D., J.R. Williams, A.G. Hornsby, and A.L. Wood. 1983. Physical characteristics of soils of the Southern Region-Bethany, Konawa, and Tipton Series. S. Coop. Ser. Bull. 265.

Pachepsky, Ya. And M.G. Schaap. 2004. Using soil morphological attributes and soil structure in pedotransfer functions. p. 115-141. *In* Ya. Pachepsky and W.J. Rawls (ed.) Development of pedotransfer functions in soil hydrology. Dev. Soil Sci. 30. Elsevier Amsterdam.

Pachepsky, Y.A., and W.J. Rawls. 2003. Soil structure and pedotransfer functions. European J. of Soil Sci. 54: 443-451.

Pachepsky, Ya. A., W.J. Rawls, and H.S. Lin. 2006. Hydropedology and pedotransfer functions. Geoderma. 131:308-316.

Pachepsky, Ya. A., Dennis Timlin, and G. Varallyay. 1996. Artificial Neural Networks to Estimate Soil Water Retention from Easily Measurable Data. Soil Sci. Soc. Am. 60: 727-733.

Prasad, M. Anantha. Louis R. Iverson. Andy Liaw. 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems. 9: 181-199.

Quisenberry, V., D. K. Cassel, J.H. Dane, and J.C. Parker. 1987. Physical characteristics of soils in the Southern Region: Norfolk, Dothan, Goldsboro, Wagram. S. Coop. Ser. Bull. 263.

Rawls, W.J., D. Gimenez, and R. Grossman. 1998. Use of Soil Texture, Bulk Density, and Slope of the Water Retention Curve to Predict Saturated Hydraulic Conductivity. Am. Soc. of Ag. Engineers. 41: 983-988.

Römkens, M. J. M., H.M. Selim, H.D. Scott, R.E. Phillips, and F.D. Whisler. 1986. Physical characteristics of soils in the southern region – Captina, Gigger, Grenada, Loring, Oliver, and Sharkey series. S. Coop. Ser. Bull. 264. Mississippi Agric. And Forestry Exp. Stn., Mississippi State.

Römkens, M.J.M., J.M. Selim, R.E. Phillips, and F.D. Whisler. 1985. Physical Characteristics of Soils in the Southern Region: Vicksburg, Memphis, Maury Series." S. Coop. Ser. Bull. 266. Alabama Agric. Exp. Stn., Auburn University.

Schaap, Marcel G. and Feike J. Leij. 1998. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil and Tillage. 47: 37-42.

Siegelmann, T. Hava. 1999. Progress in Theoretical Computer Science. Neural Networks and Analog Computation: Beyond the Turing Limit. *Birkhäuser Boston.*

Tamari, S., J.H.M. Wösten, and J.C. Ruiz-Suárez. 1996. Testing an Artificial Neural Network for Predicting Soil Hydraulic Conductivity. Soil Sci. Soc. Am. 60: 1732-1741.

Tietje, Olaf. And Volker Hennings. 1996. Accuracy of the saturated hydraulic conductivity prediction by pedo-transfer functions compared to the variability within FAO textural classes. Geoderma. 69:71-84.

Timm, Neil, H. 2002. Springer Texts in Statistics. Applied Multivariate Analysis. Springer-Verlag.George Casell, Stephen Fienberg, and Ingram Olkin (eds.) New York.

Tomasella, J. Ya. Pachepsky, S. Crestana, W.J. Rawls. 2003. Comparison of Two Techniques to

Develop Pedotransfer Functions for Water Retention. Soil Sci. Soc. Am. J. 67:1085-1092.

U.S. Department of Agriculture, Natural Resources Conservation Service. National Soil Survey Handbook, title 430-VI.

Wagner, B., V.R. Tarnawski, G. Wessolek, and R. Plagge. 1998. Suitability of models for the estimation of soil hydraulic parameters. Geoderma 86:229-239.

Wösten, J.H.M., A. Lilly, A. Nemes, and C. Le Bas. 1999. Development and use of a database of hydraulic properties of European soils. Geoderma 90:169-185.

Wösten, J.H.M., Ya.A. Pachepsky, and W.J. Rawls. 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic properties. J. of Hydrol. 251:123-150.

CHAPTER 3

PEDOTRANSFER FUNCTION FOR ESTIMATING SATURATED HYDRAULIC CONDUCTIVITY IN STRUCTURED SOILS[1]

---

[1] Shirley, A.C., A. Nemes, D.E. Radcliffe, and L.T. West, To be Submitted to the Journal of Soil and Water Conservation.

Abstract

Pedotransfer functions (PTFs) are useful tools that can predict saturated hydraulic conductivity (K_s) from more easily measured soil properties. However, most PTFs do not include soil structure. Soil structure can be categorized using morphological descriptors, found in many soil survey databases. The objective of our study was to determine if soil morphological descriptions of structure were important factors in predicting K_s using PTFs. We used soil profile descriptions for nine soil series from the S-124 regional project dataset describing soils of the Southeastern United States. The dataset contains qualitative morphological descriptions of soil structure as well as quantitative measures of particle size, bulk density, and K_s. We used six qualitative predictor variables: horizon position (HOR), textural class (TXT), ped size (PED), crack orientation (CRK), grade (GRD), and moist consistence (CST). We also used two quantitative predictor variables, bulk density (D_b) and particle size distribution (PSD). Our best decision tree model predicted $\log_{10} K_s$ (cm day^{-1}) with an average root mean square residual (RMSR) of 0.8017, indicating that the estimation was better than an order of magnitude. None of the best tree models used soil morphological descriptors. Instead they all used bulk density and texture. The top split in our tree models always separated out the high sand content soils and assigned a high predicted K_s. For the finer textured soils, the splits were based on bulk density. The NRCS Soil Survey method of predicting K_s did nearly as well as our best model with a RMSR of 0.9562 $\log_{10} K_s$ in cm day^{-1}. Morphological descriptors of soil structure may not have been important because bulk density acted as a surrogate for structure in the finer textural classes or because the descriptors did not capture the effect of structure on K_s.

Introduction

Pedotransfer functions (PTFs) are important tools for predicting soil properties such as saturated hydraulic conductivity (K_s). At present, most PTFs use a combination of readily available soil properties such as textural class, particle size distribution, bulk density, and organic matter for predictor variables (Tietje and Hennings, 1996). Pedotransfer functions can be broken into two categories: continuous and class PTFs (Bouma, 1989). Continuous PTFs use numerical values, for instance, percentage of sand, silt, and clay. Class PTFs use categorical data, e.g. soil textural class. From these two main PTF categories, different statistical modeling methods for predicting K_s have been developed. McBratney et al. (2002) and Wösten et al. (2001) discuss these including: multiple linear regression, artificial neural networks, generalized linear models, general additive models, group method of data handling, multiple adaptive regression splines, and decision trees. Decision trees are one of the methods suitable for PTFs based on qualitative predictor variables. Decision trees can be divided into two categories: classification and regression trees. Classification trees use only categorical data, while regression trees can use numerical as well as categorical input (Pachepsky and Schaap, 2004).

Regression and classification trees are both non-parametric methods that are used to uncover relationships in data. Decision trees work by recursively bifurcating. This splitting is referred to as recursive partitioning (Pachepsky and Schaap, 2004).

Soil structure is defined as the association of void space and peds, including void spaces within the peds themselves (Thomasson, 1978). Regression PTFs for predicting K_s, which include morphological descriptions of structural components have been shown to increase prediction capability (McKeague et al., 1982; Lin et al., 1999a; Lin et al., 1999b; Lilly, 2000;

O'Connell and Ryan, 2002; Lilly et al., 2008). Other soil properties such as water retention and unsaturated hydraulic conductivity can also be predicted more accurately when a structural component is considered (O'Neal, 1952; Pachepsky and Rawls, 2003; Pachepsky et al. 2006). Moreover, there is a large amount of information in state soil surveys on soil morphological descriptions that could be used in PTFs for the prediction of hydraulic properties such as K_s. In fact, the Soil Survey method uses this information to predict a range for K_s that is given in the NASIS database. However, the NRCS Soil Survey method uses only bulk density and textural class with some over-riding conditions, some of which are related to soil structure (these are described later). The method first determines the bulk density class: high (1.46 to 1.72 g cm^{-3}), medium (1.19 to 1.59 g cm^{-3}), or low (0.93 to 1.32 g cm^{-3}). There is overlap between the three groups, so samples falling into two groups must be checked to ensure which textural class and bulk density class they should be grouped in. Once the bulk density group has been established, the particle size distribution can be used to predict the K_s (USDA, 2010).

Many PTFs predict K_s using percentages of sand, silt, and clay and bulk density (continuous PTFs). Wösten et al. (2001) found that using texture or particle size distribution yielded accurate predictions of K_s. Wösten et al. (2001) and Rawls et al. (2004) reported that estimating soil hydraulic properties using structure could yield better results. The use of morphological descriptors could lead to better transferability of PTFs according to Lilly and Lin (2005). The increase in prediction power of using structural attributes may be expected because models based purely on texture and bulk density do not take into account the effect of mesopores and macropores (Rawls et al., 2004). A study conducted by Jabro (1992) using data from the S-124 project found a relationship between K_s, bulk density, and particle size distribution. Results

indicated that silt, clay, and bulk density performed well when comparing predictions to the measured K_s.

The objective of this study was to use qualitative morphological descriptors of soil structure as well as quantitative data on soil physical properties that are available in the Southern Cooperative Series Bulletins to predict K_s. These bulletins were part of the S-124 regional project. These predictions will be compared with the predictions using the Soil Survey method (USDA, 2010).

Materials and Methods

Input Data

The dataset used in this study was a subset of the S-124 dataset contained in the Southern Cooperative Series Bulletins 262-268. There were a total of 1664 samples, including replicates. The S-124 project included 21 different soil series from the Southeastern US (Bruce et al., 1983; Dane et al., 1983; Luxmore, 1982; Römkens et al., 1986a and 1986b; Nofziger et al., 1983; and Quisenberry et al., 1987). However, not all of these soils had sufficient data for inclusion in our data set. The bulletins contain data on K_s, particle size distribution (PSD), organic matter, morphological descriptions of soil structure, and horizon identification. We considered samples that contained close to all of these data, however, only a few samples had organic matter content determined. Soils that had no K_s data were not included.

With these data, we developed eight input groups: two quantitative and six qualitative. The qualitative groups included: horizon designation (HOR) to indicate if the sample was from a topsoil (A horizon) or subsoil horizon (soils below the A horizon), ped size (PED), crack orientation (CRK), grade (GRD), moist consistence (CST), and textural class (TXT). Unlike

Lilly et al. (2008), the textural class was solely based on the particle size distribution because there was not sufficient data on hand texture to allow for its inclusion. The two quantitative groups were particle size distribution (PSD) and bulk density (D_b). Table 1 shows a breakdown of all input variables. Group 1 (HOR) shows the horizon position of the sample: topsoil or subsoil. Group 2 (PED) indicates the size of a particular ped. In some cases the ped size was not given. In these instances the ped variable received a designation of PS8, indicating the ped size was unknown. Group 3 (CRK) described the cleavage planes of the structural cracks. Group 4 (GRD) indicated the particular grade to which a sample belonged. We included massive and single grain even though these classes can be considered structure-less. Group 5 (CST) indicated the moist consistence of a ped. This is simply a measure of how easy it was to crush a ped. Group 6 (TXT) indicated which of the 12 textural classes each sample belong to. In our dataset this was determined by using the PSD data taken from the bulletins. Group 7 (PSD) indicated the percent sand, silt, and clay. These were based on the USDA system (< 0.002 mm for clay, 0.002 to 0.05 mm for silt, 0.05 to 2 mm for sand). Group 8 (D_b) was bulk density. In all qualitative groups (i.e. one to six), membership of a sample was indicated by a binary dummy variable (1 for membership, 0 for non-membership).

The Decision Tree Technique

Our decision tree program was coded in MATLAB (The Math Works, 2009). In order to ensure that over fitting of the model did not occur, a stopping criterion had to be established. This was accomplished by using a pruning process. There are general guidelines for limiting combinations of prediction variables in the pruning processes. However, Lilly et al. (2008) found

that these settings are data-set specific Therefore, like Lilly et al. (2008), we found optimized settings by using all possible combinations of predictor variables.

The dataset had to be subdivided into a training data set for developing a full decision tree and a testing data set for pruning the full decision tree. We optimized the ratio between the sizes of the training and testing data sets, the maximum number of samples in each node before tree pruning, and the maximum number of terminal nodes after pruning. A trial and error approach was used in the optimization process. The size of the training data set was varied from 14 to 214 in steps of 10. Jackknife cross-validation (re-sampling without replacement) was used 100 times to generate 100 training and testing data set pairs for the optimization procedure. The maximum number of samples within each node before pruning was varied from 10 to 150 by steps of 10. The development of tree models was stopped when the number of samples in each node became smaller than the allotted value. This optimization process yielded 315 different tree models, each of which had 100 replicates which were used to estimate the mean $\log_{10}$-transformed K_s. In each of the 100 replicates, estimations were made for samples of the test data set, first using the full tree models. Subsequently, the root mean square residuals (RMSRs) were calculated separately for each of the 100 replicates using equation 4:

$$RMSR = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}[\log(K_{s\,i,meas}) - \log(K_{s\,i,est})]^2}$$

Equation 4

where n is the total number of samples in the test dataset, log ($K_{s\,i,\,meas}$), and log($K_{s\,i,\,pred}$) refer to the measured and predicted log transformed K_s. We then pruned each of the trees by removing non-terminal nodes, one-by-one in the order of adding the least deviance to the tree. This pruning process continued until there were only two terminal nodes remaining (one partition). Statistical measures of the tree models' performance were evaluated at each step of pruning. Once the tree models were completed, the RMSRs for each of the 100 replicates were averaged for every combination of the training and testing data set sizes, the maximum number of samples in each terminal node, and the number of terminal nodes. These values were then ordered using the mean RMSR + 1 standard deviation and the running average of the above three factors were calculated over the best 100 models. These running averages yielded the ratio of developmental and testing data set, number of samples in each terminal node, and the total number of terminal nodes that we used in the further steps of our study. The optimum ratio between the developmental and testing data set was 160/55 (74%/26%), which is near to what was found optimal by Lilly et al. (2008) (82%/18%). The optimal maximum number of samples in each node was 25 and the optimal number of terminal nodes was found to be five.

Determining the Best Combination of Input Variables using the Optimized Settings

These optimized settings were then used to run our data set, using all combinations of input variables, repeating the jackknife cross-validation approach, which generated 100 replicates of every input combination that we considered (Figure 1). All the replicates and the corresponding predictor variables were used to predict the $\log_{10} K_s$ for the training dataset. The RMSR was calculated for the test dataset of each of the 100 replicates using the input parameters from the pruned model. The RMSR values were then averaged over the 100 replicates for each

tree model. Using 100 replicates of the tree models allowed for probability estimates of the best input combinations to be made.

Once the 100 runs were made for each input combination, the pruned decision tree models were used to make predictions of K_s. There were a total of eight unique input groups: ped size (PED), crack orientation (CRK), grade (GRD), consistence (CST), particle size distribution (PSD), textural class (TXT), bulk density (D_b), and horizon designation (HOR). When looking at these input groups, there is an obvious overlap between PSD and TXT. However, the systematic elimination of inputs showed which of these were the most important in our model in a similar manner as used by Lilly et al. (2008).

Using the optimized settings from the full model, we evaluated the performance of the model after systematic elimination of all possible combinations of input variables. When all the possible combinations of input variables had been evaluated, we analyzed the model structure of the best ranking model, and also described some models with practical applications. It should be noted that we did not use only one input from some groups because that would not have allowed for our optimum numbers of terminal nodes. Take for instance, HOR. There were only two options, either a topsoil or subsoil, yielding only two terminal nodes. However, textural class (TXT) and ped size (PED) were run as the only inputs because these two groups allowed for the optimum number of terminal nodes.

Soil Survey Method

The Soil Survey method uses texture and bulk density, as shown in Figure 2. Bulk density is divided into three main groups: low (0.93 to 1.32 g cm^{-3}), medium (1.19 to 1.59 g cm^{-3}), and high (1.46 to 1.72 g cm^{-3}). There is overlap between the three groups, so samples that fall

into two of those groups must be checked to find out which group is most suitable. The low density group has a predicted K_s range from 0.1 μm s^{-1} to 100 μm s^{-1} (0.864 to 864 cm day^{-1}) depending on the textural class. The medium bulk density group has a predicted K_s range of 0.01 μm s^{-1} to 100 μm s^{-1} (0.0864 to 864.0 cm day^{-1}). One of the main differences between the medium and low density ranges is that the medium bulk density group predicts lesser K_s values than the low bulk density group for the same textural classes. This shift is even more pronounced in the high bulk density group where the predicted K_s ranges from 0.01 μm s^{-1} to 10 μ ms^{-1} (0.0864 to 86.4 cm day^{-1}).

Although the general approach of the Soil Survey method uses texture and bulk density, there are over-riding conditions, which deal mainly with the soil structure. These over-riding conditions take precedent over the method involving bulk density and texture, and assign predicted values as follows:

-Soils with all fragmental, cindery, or pumiceous particle size classes (≥ 864 cm day^{-1})

-Soils with many medium or coarser vertical pores that extend through the layer. Medial-pumiceous, medial-skeletal, ashy-pumiceous, ashy-skeletal, or hydrous-pumiceous material that is very friable, friable, soft, or loose (≥ 864 cm day^{-1})

-When material is moderately moist or wetter, structure is moderate or strong granular, strong blocky, or prismatic smaller than very coarse; no stress surfaces or slickensides (86.4 to 864 cm day^{-1})

- Soils with common medium or coarser vertical pores extend through the layer (86.4 to 864 cm day^{-1})

- Soils with strong very coarse blocky or prismatic structure and no stress surfaces or slickensides (86.4 to 864 cm day^{-1})

- Soils with ≥ 35 percent clay that is soft, slightly hard, very friable or friable; no stress surfaces or slickensides and the clay is subactive after subtracting the quantity [2 x (OC x 1.7)] (8.64 to 86.4 cm day^{-1})

- Soils with few stress surfaces and/or slickensides (8.64 to 86.4 cm day^{-1})

- Soils with massive structure and very firm or extremely firm consistence or weakly cemented (0.864 to 8.64 cm day^{-1})

- Soils with continuously moderately cemented (0.864 to 8.64 cm day^{-1})

- Soils with common or many stress surfaces and/or slickensides (0.0864 to 0.864 cm day^{-1})

- Soils with continuously indurated or very strongly cemented (<0.0864 cm day^{-1})

Comparing Methods

Once the decision trees had been developed, we wanted to compare the decision trees and the Soil Survey method. The Soil Survey method assigns a range, whether using the bulk density and textural triangles (Figure 2) or the overriding conditions. When calculating the RMSR, we used the midpoint of the range as the predicted K_s for the Soil Survey method. The RMSR was calculated by taking 55 test samples (subset of the 215 total samples) and re-sampling 100 times (the same re-sampling procedure as the decision tree technique). The RMSR for each run was determined and averaged for the 100 replicates.

Results and Discussion

The S-124 Data Set

Calculating correlations between variables can aid in understanding how variables interact and show if any relationships between structure and K_s exist. We found statistically significant correlations between many combinations of variables in our dataset, which led to a large matrix (45 by 45), making it impractical to show all correlations. The 45 by 45 matrix was created out of 8 different input groups. Each input group had several members, which led to this large matrix. For example, the grade (GRD) group is made up of weak, moderate, strong, single grain, massive, and none. Table 2 shows some of the significant correlations, ranking from highest to lowest (absolute value of R). The first major correlation was between topsoil (TOP) and 2-5 mm ped size (PS2) (R = 0.64). This correlation showed that samples that were topsoils tended to have smaller ped sizes. Topsoil and PS1 had a similar relationship, however, 1-2 mm peds were not as strongly correlated with topsoil as the 2-5 mm peds. Another positive correlation was between K_s and sand content, as one would expect. Saturated hydraulic conductivity and clay content were negatively correlated. An interesting positive correlation between bulk density (D_b) and very firm consistence (VERY FIRM) was found. Bulk density and K_s had a weak positive correlation. This relationship is counter-intuitive because higher bulk densities have slower K_s, however, in our dataset a weak positive relationship between the two was found probably due to the number of coarse textured soils (sands, loamy sands, sandy loams). Bulk density (D_b) and K_s had a weak positive correlation (R = 0.15).

However, if the data set was limited to clayey and intermediate textured soils, a different relationship was found. For clay soils, bulk density and K_s were inversely correlated (R = -0.508)

with a significance level of 0.001. The loam, silt loam, silt, sandy clay loam, clay loam, silty clay loam textural classes had a similar correlation with K_s (R = -0.307) with a significance level of 0.001.

Trying to predict K_s based solely on texture can be difficult because finer textures do not always yield smaller K_s values. This can be seen in table 3 which shows the average K_s, with and without the $\log_{10}$ transform, of all textural classes in our data set. The textural classes are listed in the order of the mean K_s. It is interesting to see that the clay textural class had the fourth highest K_s of the 12 classes. In the data base of 1323 U.S. samples compiled by Rawls et al. (1982), the clay textural class had the lowest mean K_s. However, the order of textural classes in our data set was nearly identical to that in the Rosetta data base of over 2000 soils compiled by Schaap et al. (2001) (see Table 1 of Radcliffe and West, 2009). The values for mean K_s were also remarkably similar to the Rosetta database. The only real exception was the silt textural class which was higher in the order and in mean K_s in the Rosetta data base. It is also interesting to note the pattern in coefficient of variation (CV) of the textural classes. The CV was quite small for the sandy textures but increased sharply for the clay, sandy clay loam, silt loam, silty clay loam, and clay loam classes. In finer textured soils, we expect that soil structure plays an important role in determining the K_s. One of the distinguishing characters of the S-124 data base was the high incidence of samples from clayey textural classes. Proportionally fewer samples from these classes occurred in the Rosetta data base (Schaap et al., 2001).

It should also be noted that soils in the clay textural class had a relatively large ped size, most falling within the ped size range of the 10-20 mm (PS4). The remaining soils fell in the ped

size range of 5-10 mm (PS3). Each of these soils also had grade of either weak, moderate, or strong, indicating structure development.

Decision Tree Models

Table 4 shows a list of some of the 247 decision tree models (each using a different combination of input groups), ranked in increasing order of the mean RMSR. Included in the table is the minimum RMSR, maximum RMSR, and standard deviation of the RMSR (taken from 100 replicates), all $\log_{10}$-transformed K_s (cm day^{-1}). The table shows what the RMSR was in 100 replicates when the input variables listed were made available. It is important to note that not all of the available input variables were used in the model as the decision trees were developed, however. This will be seen later when we discuss common decision trees that were developed from a given set of input variables. The full model that used all 8 input variables was ranked 62nd. The best and worst model had a mean RMSR difference of 0.2008. By comparison, Lilly et al. (2008) had a mean difference between best and worst models of 0.128. Our best model performed better than that of Lilly et al. (2008) with an average of 0.8017 $\log_{10}$-transformed K_s (cm day^{-1}) while Lilly et al. (2008) had an average value of 0.9510.

When looking at the standard deviation, the model ranked 196th was within one standard deviation of the best model. It could be argued that all models ranked higher than this model were equal in their prediction power. All the possible number of input groups, from eight input groups (full model) to two input groups, were represented in the top quarter of the models. The best model with seven input groups available included all but GRD as a predictor. The best model with six input groups available included all inputs except TXT and HOR. The best model with five input groups available was ranked eighth, and did not use PED, CST, or HOR. The best

model with four input groups available (CRK, GRD, PSD, and D_b) was ranked fourth. The best model with two input groups available was ranked first and included PSD and D_b. From Table 4, the best models with fewer inputs performed better than most models with more inputs. Unlike Lilly et al. (2008), there were no models that simply used qualitative (structure) variables in the best 20 models.

The mean RMSR of many of the models ranked close to each other or often had the same value, to four decimal places (Table 3). These models were probably using the same splits even though the available input groups differed slightly. For example, the models ranked second (using GRD, PSD, and D_b) and first (using PSD and D_b) had the same mean RMSR, SD, Min, and Max values indicating that they used the same splits to form trees, even though the model ranking second had more inputs. The addition of the GRD input group in this case had no effect on the estimations. Not only did GRD not seem to be an important variable in this case, but the same pattern with GRD appeared in the best 87 models. For these particular cases it showed that GRD was not important for improving K_s predictions. Another interesting point about all of these trees was that each used a measure of texture (PSD or TXT) and D_b. This pattern appeared in all models except for those where the PSD and D_b were excluded. These results are contrary to those of Lilly et al. (2008). The best models in their study included structure, whereas in our study we found that structural components only appeared in models where PSD, D_b, and TXT were all excluded, indicating a weak relationship between soil morphological measures of structure and K_s. However, when PSD, D_b, and TXT were excluded these measures became important. Some of these trees will be discussed further.

Decision Tree Full Model

The first model we discuss is the full decision tree model which resulted from using all of the available input variables. This model was used to identify which were the most important input parameters and to set the number of partitioning levels. Lilly et al. (2008) only went to the fourth partitioning level because five terminal nodes could be obtained at the third or fourth splitting level. Since our study also had an optimized value of five terminal nodes, analyzing four splitting levels of the tree structures was also sufficient in our study.

Table 5 shows a list of partitioning variables used in the full model decision trees (model number 62 in Table 4). It also shows the average splitting value, frequency, and probability of occurrence at each level. The frequency is the probability of a variable occurring at least once at that particular splitting level. The sum of probabilities within a splitting, unlike frequency, can be greater than one because variables can occur more than once at a particular level due to the fact that the variable can be present in different nodes at a particular level.

For the full model, the first split was sand content and, on average, the split occurred at a value of 65.8% sand content. The right branch of this split (sand > 65.8%) included the sandy loam, loamy sand, and sand textural classes. Sand content was the primary splitting variable in all cases at the first level, because these types of soils had the highest K_s (see Table 3). At the second partitioning level, D_b was the primary splitting variable with a frequency of 0.72. This was followed by percent clay (CLAY) with a frequency of 0.16, and VERY FRIABLE consistence with a frequency of 0.06. For the first two splitting levels the main predictors were quantitative input variables, with the exception of consistence. At the third level, the percentage

of silt (SILT) was the top splitting variable with a frequency of 0.38, followed by percentage of sand (SAND) with a frequency of 0.30, and D_b with a frequency of 0.17.

Best Decision Tree Models

The best ranking model used only PSD, and D_b. The only important splitting variables were percentages of sand, silt, clay, and D_b. Table 6 provides a breakdown of the most important variables at each partitioning level. At the first partitioning level, soils were split based on sand content (on average at 65.9%). This was the same split that appeared in the full model. At the second splitting level, D_b was the most frequent splitting variable. At level three, percentage of silt was the most important variable, having a frequency of 0.42. SAND was an alternative important splitting variable at this level, with a frequency of 0.31. Figure 3 shows the most likely tree structure of the best model (with input variables PSD, and D_b). The test RMSR that resulted from a run with 100 randomly selected subsets from the data set using this model, is also shown. For each terminal node, the average predicted K_s and SD is given. The test RMSR (0.7663 in Figure 3) for this particular tree was slightly less than the RMSR for all tree models using this combination of inputs (0.8017 in Table 4). There were four terminal nodes in this tree with predicted log-transformed K_s of 2.45, 1.66, 0.84, and 0.73 in cm day^{-1}. At the first level, soils with sand content greater than 65.9 % (part of the sandy clay loam and sandy loam textural classes and all of the loamy sand and sand textural classes) formed the right branch which leads to a terminal node. The reason for this split was these textural classes tended to have high K_s (see Table 3). On the left branch, the next partitioning split soils with D_b greater than 1.51 g cm^{-3} to the right branch. This branch leads to a terminal node with the lowest predicted K_s (0.73 cm day^{-1}). Soils in this group (sand < 65.9% and D_b > 1.51 g cm^{-3}) ranged from sandy loams to clays.

However, 63% of these soils were silt loam, sandy clay loam, or clay textural classes. It should be noted that the clays in this section were well structured, (no massive, single grain, or structureless designations) and still are in the group with the lowest K_s. This implies that D_b was a better predictor of K_s than morphological descriptors in finer textural classes. As was shown earlier, within the finer textured soils, a high D_b is associated with a low K_s.

Going to the third splitting level (26.4% < sand ≤65.9% and $D_b \leq 1.51$ g cm^{-3}), both branches from this level lead to terminal nodes with intermediate predicted K_s. Half of all clays in this data set were routed to the right branch terminal node with the second highest K_s. There were a few clay samples that were massive or structureless for ped shape. The remaining samples had morphological descriptors that indicated well structured soils. This node had the second highest K_s average value. The remaining soils in this node came from the silt loam, silt, sandy clay loam, and clay loam textural classes. The final left branch node consisted of 70% silt loams, 25% silty clay loams and the remaining samples were silt and clay loam. Only 11% of the samples had a massive or structureless morphology.

The morphological descriptors of structure did not seem to perform as well in predicting K_s as the combination of D_b and texture. When looking at our data set, the ped size 2 to 5 mm (PS2) had the lowest average K_s, while ped size 50 to 100 mm (PS6) had the highest K_s. The remaining ped size classes had similar values (around 0.7 $\log_{10}$ transformed cm day^{-1}). The wide range in morphological descriptors with similar values for K_s made structure seem less important in our data set, which can be seen throughout most of the decision tree models.

Many of the best models in Table 4 had RMSRs that were within one SD of each other. Therefore, using a model that was not ranked the highest can still yield predictions that are not

significantly different from the highest ranking model. In this study, we wanted to see if easily obtainable morphological data (structural data) could be used to predict K_s. Therefore we investigated models that used only easily categorical data.

As noted earlier, the model ranking 69th was the highest ranking model when only qualitative input variables were made available as inputs, in this case TXT and HOR. However, the actual decision tree models developed used only TXT (Figure 4). The splits simply segregated the first four textural classes in the order shown earlier in Table 3 where the mean K_s for each textural class in our data set was shown.

An example of a model where qualitative variables appeared in the decision tree is the one that had PED, CRK, GRD, and CST available as inputs. This model ranked 202nd overall (Table 4). Table 7 provides a breakdown of the partitioning variables. At the first partitioning, loose consistence was the main splitting variable with a frequency of 0.83, followed by single-grained crack orientation and very friable consistence. Single grain crack orientation was included in the crack orientation, as well as in the grade class, because single grain is given as a structural component. These variables replaced the high sand content that was present in the first splitting level of the best model (see Table 7). Loose consistence, single-grain crack orientation, and very friable consistence all correlated strongly with a high sand content (see Table 2).

The variable of highest frequency of occurrence at the second splitting level was very friable consistence (probability of occurrence 0.87), followed by loose consistence, 2-5 mm ped size (PS2), and 10-20 mm ped size (PS4), with frequencies of 0.07, 0.06, and 0.04, respectively. These splits resulted in sorting of the coarser material at partitioning levels one and two. At the lower levels, further splitting mostly occurred within groups of the finer material. In the third

partition level, the primary partitioning variable was 10-20 mm ped size (PS4) with a frequency of 0.44. This was followed by very firm consistence, with a frequency 0.27, showing that at this level finer-textured soils (silt loam and finer) were being grouped together. There was a correlation between PS4 and textural class: 30% of soils that were PS4 size were also in the clay textural class. The remaining soils were in the silt loam, silt, sandy clay loam, clay loam, silty clay loam, and clay textural classes. The forth partitioning primary split was the 2-5 mm ped size (PS2) with a frequency of occurrence of 0.38.

Figure 5 shows the most likely tree structure of the model using only morphological variables (PED, CRK, GRD, and CST), along with the RMSR, average predicted K_s, and SD for each terminal node. The test RMSR (0.8534 in Figure 5) for these hard coded splits were slightly less than the RMSR for all tree models using this combination of inputs (0.8235 in Table 4). As expected, the first split separated soils with high sand content, (high sand content and loose consistence was positively correlated, R = 0.47, in Table 2) grouping all sands from the dataset in the right branch and predicting the highest $\log_{10}$-transformed K_s (2.75). The terminal node with the second highest K_s (1.85) separated out textures with high silt content (loamy sand, sandy loam, and silt loams). Very friable consistence was positively correlated with loamy sands and sandy loams. Many of these soils had morphological descriptors indicating a lack of structure (massive, single grained, or structure-less), but this might be expected in high-silt soils.

The terminal node with the third highest K_s (0.95) selected samples that were silt loams, silt, sandy clay loam, clay loam, silty clay loam, and clay. The silty loam and sandy clay loams were weakly correlated with friable consistence. The clay textural class was correlated with the firm consistence (each of these two soil consistence groups were in this node). It is interesting to

note that this group included all clays in the data set (18% of the S-124 data set were clays). As noted earlier, all of the soils in the clay class had evidence of structure development (not massive or single grained). The mean K_s of all the clays was 1.25 $\log_{10}$ cm day^{-1}, showing that the clay textured soil actually increased the mean of this node (refer to Table 3). Even though the clay textured soils had greater K_s, the reason for this node's lower K_s (compared to the highest and second highest terminal nodes) was because of the consistence. The soils at this node had a consistence of firm or friable.

The very friable consistence soils were segregated at level two and routed to the terminal node with the lowest K_s (-0.51). This group only contained two samples, both of which were C or B/C horizons. Although both soils were well structured, the low K_s could be due to the fact that these horizons can have little pore inter-connectivity.

The second lowest K_s was probably due to these soils having a designation of a fragipan (defined by a restriction of water flow). Just as with the lowest K_s node these were well structured but the fragipan designation (most of the soils in this node had this designation) is probably the cause for such a low K_s.

Another model of interest was one that used both qualitative and quantitative variables. An example of such a model is the one that used only PED, CST, and D_b as inputs. This model ranked 209th overall (Table 4). This model was significantly different from the top ranking model, however, the prediction capabilities between the two is only 1.26 cm day^{-1}. Table 8 provides a breakdown of the partitioning variables. At the first partitioning, loose consistence was the predominant splitting variable with a frequency of 0.91 This variable replaced the sand content (sand content and loose consistence was positively correlated, R = 0.47, Table 2) that

was present in the first splitting level of the best model and can be considered a qualitative descriptor for the sands (see Table 8).

The second splitting level with the highest frequency of occurrence was very friable consistence (frequency of occurrence 0.82), followed by loose consistence, 2-5 mm ped size (PS2), and 10-20 mm ped size (PS4), with frequencies of 0.07, 0.06, and 0.04, respectively. These splits resulted in sorting of the coarser material at partitioning levels one and two. The lower levels yielded less uniform groups, which can be seen by the wider distribution of splitting variables at those levels. At the third partition level, the primary partitioning variable was 10-20 mm ped size (PS4) with a frequency of 0.43. This was followed by D_b with a frequency 0.35.

Figure 6 shows the most likely tree structure for this model, along with the RMSR, average predicted K_s, and SD for each terminal node. The test RMSR (0.8482) for these hard coded splits were slightly less than the RMSR for all the tree models with this combination of inputs available (0.9009 in Table 4). As expected, the first split separated soils with high sand content, grouping all sands from the dataset in the right branch and predicting the highest $\log_{10}$-transformed K_s (2.75 cm day^{-1}). The terminal node with the second highest K_s (1.85 cm day^{-1}) were soils that were very friable and had a ped size other than 10-20 mm (this combination probably selected the smaller ped sizes). This again selected coarser textured soils (i.e. loamy sands and sandy loams).

The terminal node with the third highest K_s (1.62 cm day^{-1}) selected samples that were not of loose consistence or very friable and had a lower bulk density. The other node at the lowest splitting level with the next to lowest K_s had a higher bulk density, which probably were the compacted soils. It is difficult to understand (because the samples in this node were well

structured) why the node with the lowest K_s was lower than the other nodes (over an entire order of magnitude different from the next lowest node). This node only contained two values, which could be an explanation for the low value. The soils in this node were deep in the soil profile (B/C and C horizons); this could also be an explanation for such a low K_s (low interconnectivity between the pores).

Comparing Soil Survey Predictions to Best Decision Tree Model Predictions

The Soil Survey PTF model is based on bulk density and textural classes for the prediction of K_s. Using equation 1, we calculated the RMSR for our dataset using the Soil Survey method. The RMSR in $\log_{10}$-transformed K_s (cm day^{-1}) was 0.9562 compared to 0.8017 for our best model and 0.8911 for our best model using only qualitative data. As mentioned earlier, neither the NRCS method nor our best model to estimate K_s take into account structure directly. There are, however, overriding factors in the NRCS method which are used in some cases to estimate K_s instead of the bulk density and texture method. Our dataset used such conditions for only 15 samples, all of which were related to structure. These samples all had strong sub-angular blocky structure. If the overriding conditions were not taken into account then the RMSR was 0.9152. In the case of our dataset, using the overriding conditions increased the error.

Conclusions

Even though much of the literature indicates that using only texture and bulk density does not yield accurate results (Loague, 1992; Wagner et al., 1998) there are some studies showing that one can use texture and bulk density to adequately predict K_s. Jabro (1992) found that texture and bulk density were in fact good predictors of K_s using the S-124 data set. Since we used soils from the same data set, it should not be surprising that we found that the strongest

predictors of K_s were texture and bulk density. McKenzie and Jacquier (1997) found that morphological descriptors of structure, bulk density, and texture could be used to predict K_s, however, all of these relationships showed variability

Using decision tree analysis, we found that texture and bulk density were the most important variables in the prediction of K_s in a subset of the S-124 data set (selected based on data availability) consisting of soils from the Southeastern U.S. Contrary to the study by Lilly et al. (2001) on European soils, morphological descriptors of structure did not appear in the decision trees of our best models. However, it is important to note that models that relied solely on morphological descriptors of structure could be used to predict with similar results with our top model (PSD and D_b).

We found that high sand contents were the most important splitting variable in our best models. This was due to the consistently high K_s present in coarse-textured soils. When looking at the models that used only morphological descriptors of structure, the best model was ranked 202nd. The worst of these models was ranked 238th. The only time that a structural component appeared was when texture and bulk density were excluded as potential input variables. The Soil Survey method PTF performed nearly as well as our region-specific best model with a RMSR of 0.9562 for the Soil Survey method and a RMSR of 0.8017 for our best model. However, our model best model using only morphological descriptors of structure outperformed the Soil Survey method with an RMSR of 0.8911. It is interesting that both our best model and the Soil Survey method used only bulk density and texture. This is probably because bulk density in combination with texture acted as a surrogate for soil structure or because the descriptors did not capture the effect of structure on K_s. For instance, finer-textured soils with high bulk densities

could indicate compacted soils with no structure, whereas finer textured soils with low bulk densities could indicate well-structured soils. Future research should examine larger datasets (spatially) which may increase a PTFs prediction capability by taking into account soils from different climatic regions. The study relied on soils from similar climatic regions making it more difficult for estimations outside these regions.

Table 1. Description and grouping of input variables and output variable.

Group	Variable	Number of Members	Description
1-HOR	HOR	169(0) ; 46(1)	0 for subsoil; 1 for topsoil
2-PED	PS1	12	1-2 mm ped size class
	PS2	24	2-5 mm ped size class
	PS3	10	5-10 mm ped size class
	PS4	89	10-20 mm ped size class
	PS5	14	20-50 mm ped size class
	PS6	9	50-100 mm ped size class
	PS7	0	>100 mm ped size class
	PS8	57	ped size not determined
3-CRK	BOTH	154	horizontal and vertical cracks
	TRANS	1	Horizontal cracks
	VERT	21	Vertical cracks
	MASSIVE	16	Massive
	SINGLE	9	Single
	NONE	14	Structureless
4-GRD	MASSIVE	16	Massive grade
	SINGLE	9	Single grade
	WEAK	91	Weak grade
	MODERATE	43	Moderate grade
	STRONG	15	Strong grade
	NONE	41	No grade given
5-CST	LOOSE	11	Loose consistence
	VERY FRIABLE	47	Very Friable consistence
	FRIABLE	64	Friable consistence
	FIRM	19	Firm consistence
	VERY FIRM	4	Very Firm consistence
	NONE	70	No consistence given
6-TXT	S	13	sand textural class (USDA)
	LS	17	Loamy sand textural class (USDA)
	SL	18	Sandy loam textural class (USDA)
	L	5	Loam textural class (USDA)
	ZL	76	Silt loam textural class (USDA)
	Z	2	Silt textural class (USDA)
	SCL	19	Sandy clay loam textural class (USDA)
	CL	9	Clay loam textural class (USDA)
	ZCL	28	Silty clay loam textural class (USDA)
	ZC	0	Silty clay textural class (USDA)
	SC	0	Sandy clay textural class (USDA)
	C	28	Clay textural class (USDA)
7-PSD	CLAY	215	Clay content (<2µm; USDA)
	SILT	215	Silt content (2-50µm; USDA)
	SAND	215	Sand content (50-2000µm; USDA)
8-BD	D(b)	215	Bulk density
Output	LOGKS	215	Transformed K_s, $\log_{10}$ (cm d^{-1})

Qualitative Variable (1 if member, 0 if not a member)

Table 2. Correlations between variables in the S-124 data set.

*Variables	R	Significance Level	Absolute value of R
TOP vs. PS2	0.64	0.001	0.64
SILT vs. D_b	-0.53	0.001	0.53
K_s vs. Sand	0.53	0.001	0.53
PS1 vs. TOP	0.47	0.001	0.47
SAND vs. LOOSE CONSISTENCE	0.47	0.001	0.47
K_s vs. Clay	-0.44	0.001	0.44
SAND vs. SINGLE GRAIN CRACK	0.43	0.001	0.43
SAND vs. VERY FRIABLE	0.43	0.001	0.43
MODERATE vs. PS4	0.38	0.001	0.38
STRONG vs. C	0.38	0.001	0.38
PS4 vs. TOP	-0.37	0.001	0.37
MODERATE vs. C	0.36	0.001	0.36
K_s vs. SILT	-0.35	0.001	0.35
STRONG vs.PS4	0.33	0.001	0.33
WEAK vs. BOTH	0.31	0.001	0.31
WEAK vs.PS2	0.29	0.001	0.29
D_b vs. VERY FIRM	0.17	0.05	0.17
K_s vs. D_b	0.15	0.05	0.15
WEAK vs. D_b	0.14	0.05	0.14

* D_b (bulk density), TOP (topsoil), (PS2) ped size 2-5 mm, PS4 (ped size 10-20 mm), VERY FRIABLE (very friable consistence, VERY FIRM (very firm consistence), SAND (percentage of sand), LOOSE CONSISTENCE (loose consistence, SINGLE GRAIN CRACK (single grain crack orientation, WEAK (weak grade), MODERATE (moderate grade), STRONG (strong grade), K_s ($\log_{10}$-transformed K_s).

Table 3. Mean and coefficient of variance (CV) of untransformed and $\log_{10}$-transformed K_s by textural class for the S-124 data set.

Textural Class	Number of members	Mean K_s cm day^{-1}	CV K_s %	Mean $\log_{10}$ transformed K_s cm day^{-1}
Sand	13	508.74	0.5	2.71
Loamy Sand	17	258.52	0.7	2.41
Sandy Loam	18	66.74	16.8	1.82
Clay	28	17.58	71.4	1.25
Silt	2	12.26	9.8	1.09
Loam	5	10.44	38.7	1.02
Sandy Clay Loam	19	8.41	101.4	0.93
Silt loam	76	7.12	86.2	0.85
Silty Clay Loam	19	4.41	109.5	0.64
Clay Loam	9	3.24	435.0	0.51
Silty Clay*	0	N/A	N/A	N/A
Sandy Clay*	0	N/A	N/A	N/A

Table 4. Performance of decision tree models, in terms of root mean squared residual (RMSR), calculated by the mean of 100 model runs, using 247 different combinations of input variables.

Rank	Input Groups								Mean RMSR	SD RMSR	Minimum RMSR	Maximum RMSR
									-------Log_{10} transformed K_s--------			
1**					PSD		D_b		0.8017	0.0855	0.5680	0.9710
2**			GRD		PSD		D_b		0.8017	0.0855	0.5680	0.9710
3		CRK			PSD		D_b		0.8034	0.0854	0.5680	0.9831
4**		CRK	GRD		PSD		D_b		0.8034	0.0854	0.5680	0.9831
5			GRD		PSD	TXT	D_b		0.8039	0.0881	0.5680	1.0149
6					PSD	TXT	D_b		0.8039	0.0881	0.5680	1.0149
7		CRK			PSD	TXT	D_b		0.8056	0.0880	0.5680	1.0149
8**		CRK	GRD		PSD	TXT	D_b		0.8056	0.0880	0.5680	1.0149
17			GRD		PSD		D_b	HOR	0.8120	0.0806	0.6587	0.9710
18					PSD		D_b	HOR	0.8120	0.0806	0.6587	0.9710
35	PED	CRK		CST	PSD		D_b		0.8152	0.0896	0.5680	1.0166
36**	PED	CRK	GRD	CST	PSD		D_b		0.8152	0.0896	0.5680	1.0166
37	PED			CST	PSD		D_b		0.8160	0.0892	0.5680	1.0166
38	PED		GRD	CST	PSD		D_b		0.8160	0.0892	0.5680	1.0166
60	PED		GRD	CST	PSD		D_b	HOR	0.8221	0.0841	0.6587	1.0166
61**	PED	CRK		CST	PSD	TXT	D_b	HOR	0.8235	0.0867	0.6587	1.0166
62**	PED	CRK	GRD	CST	PSD	TXT	D_b	HOR	0.8235	0.0867	0.6587	1.0166
75		CRK				TXT	D_b	HOR	0.8387	0.0746	0.6862	1.0790
76		CRK	GRD			TXT	D_b	HOR	0.8387	0.0746	0.6862	1.0790
77			GRD			TXT	D_b	HOR	0.8387	0.0746	0.6862	1.0790
196	PED	CRK		CST			D_b		0.8868	0.0656	0.7539	1.1334
197		CRK	GRD				D_b	HOR	0.8886	0.0629	0.7442	1.0439
198			GRD				D_b	HOR	0.8886	0.0629	0.7442	1.0439
199		CRK		CST			D_b	HOR	0.8899	0.0660	0.7565	1.0731
200			GRD	CST			D_b		0.8899	0.0670	0.7555	1.0929
201	PED			CST			D_b	HOR	0.8900	0.0657	0.7641	1.0945
202	PED	CRK	GRD	CST					0.8911	0.0654	0.7450	1.0746
203		CRK					D_b	HOR	0.8915	0.0633	0.7513	1.0524
204		CRK		CST			D_b		0.8932	0.0636	0.7648	1.0692
205	PED	CRK		CST			D_b	HOR	0.8952	0.0662	0.7641	1.0945
206	PED		GRD				D_b	HOR	0.8965	0.0678	0.7726	1.0964
207	PED	CRK	GRD				D_b	HOR	0.8990	0.0677	0.7726	1.0964
208	PED		GRD	CST			D_b	HOR	0.9006	0.0714	0.7641	1.0945
209	PED			CST			D_b		0.9009	0.0718	0.7738	1.1916
219	PED	CRK					D_b	HOR	0.9054	0.0596	0.7925	1.0574
245		CRK	GRD				D_b		0.9673	0.0651	0.8498	1.1787
246		CRK					D_b		0.9766	0.0627	0.7759	1.1787
247			GRD				D_b		1.0025	0.0604	0.8254	1.1787

*PED (ped size), CRK (crack orientation), GRD (grade), CST (consistence), PSD (particle size distribution), TXT (textural class), D_b (bulk density), HOR (topsoil or subsoil).

** Indicates the best performing model using the same number of input groups.

Table 5. Input variables and their probabilities in each splitting level for the decision tree model considering all input variables.

Splitting level	Variable*	†Average splitting value	†SD of splitting value	Total Count within level	Frequency as partitioning in level	Probability of occurrence in level
1	SAND	65.8	3.7	100	1	1
2	D_b	1.5	0.1	105	0.72	0.76
	CLAY	26.8	14.0		0.16	0.17
	VERY FRIABLE	N/A	N/A		0.06	0.06
	SAND	1	N/A		0.04	0.04
	SILT	26.5	N/A		0.01	0.01
	CLAY LOAM	N/A	N/A		0.01	0.01
	VERY FRIABLE	N/A	N/A		0.01	0.01
3	SILT	35.4	15.7	112	0.38	0.43
	SAND	26.3	14.1		0.30	0.34
	D_b	1.5	0.1		0.17	0.19
	CLAY	40.0	7.1		0.11	0.12
	PS2	N/A	N/A		0.03	0.03
4	D_b	1.5	0.1	83	0.57	0.47
	SILT	43.5	19.6		0.11	0.09
	SAND	20.1	21.7		0.11	0.09
	CST(NONE)	N/A	N/A		0.07	0.06
	GRD (NONE)	N/A	N/A		0.05	0.04
	CLAY	38.7	2.5		0.04	0.03
	PS3	N/A	N/A		0.01	0.01
	PS8	N/A	N/A		0.01	0.01
	VERY FRIABLE	N/A	N/A		0.01	0.01
	FIRM	N/A	N/A		0.01	0.01
	VERY FIRM	N/A	N/A		0.01	0.01

* SAND, sand content, (>50μm); SILT, silt content (2-50 μm); and CLAY, clay content (<2 μm) values based on the USDA classification; PS2, ped size 2-5mm; PS3, ped size 5-10mm; PS8, no structure; CLAY LOAM, refers to the textural class, clay loam; VERY FIRM, FRIM, VERY FRIABLE, and NONE refers to the consistence; D_b refers to the bulk density; CRK (SING and NONE) refers to the single grained crack orientation and no structure reported, respectively; GRD (SING and NONE) refers to the single grained grade and no grade reported. †Average of value, N/A refers to the qualitative variables; the numerical values indicate splitting values for each group

Table 6. Input variables and their probabilities in each splitting level, for the best decision tree model considering particle size distribution (PSD), grade (GRD), and bulk density (D_b).

Partitioning level	Variable*	†Average splitting value	†SD of splitting value	Total Count within level	Frequency as partitioning within level	Probability of occurrence within level
1	SAND	65.9	3.7	100	1	1
2	D_b	1.5	0.1	106	0.74	0.79
	CLAY	28.2	14.7		0.19	0.20
	SAND	1.0	N/A		0.07	0.07
3	SILT	39.3	15.9	108	0.42	0.45
	SAND	26.4	14.1		0.31	0.33
	D_b	1.5	0.1		0.19	0.20
	CLAY	37.3	8.0		0.09	0.10
4	D_b	1.5	0.1	86	0.52	0.56
	SILT	37.3	20.7		0.16	0.14
	SAND	15.4	19.5		0.11	0.12
	CLAY	25.4	13.2		0.05	0.04

* SAND, sand content, (>50μm); SILT, silt content (2-50 μm); and CLAY, clay content (<2 μm) values based on the USDA classification; PS2, ped size 2-5mm; PS3, ped size 5-10mm; PS8, no structure; CLAY LOAM, refers to the textural class, clay loam; VERY FIRM, FRIM, . VERY FRIABLE, and NONE refers to the consistence; D_b refers to the bulk density; CRK (SING and NONE) refers to the single grained crack orientation and no structure reported, respectively; GRD (SING and NONE) refers to the single grained grade and no grade reported. †Average of value, N/A refers to the qualitative variables; the numerical values indicate splitting values for each group

Table 7. Input variables and their probabilities in each splitting level for the decision tree model considering ped size (PED), crack orientation (CRK), grade (GRD), and consistence (CST).

Splitting level	Variable*	Total Count within level	Frequency as partitioning within level	Probability of occurrence within level
1	LOOSE	100	0.83	0.83
	SINGLE (CRK)		0.08	0.08
	VERY FRIABLE		0.05	0.05
	PS8		0.04	0.04
2	VERY FRIABLE	108	0.81	0.87
	LOOSE		0.07	0.08
	PS2		0.06	0.07
	PS4		0.04	0.04
	MASSIVE (CRK)		0.01	0.01
	FRIABLE		0.01	0.01
3	PS4	176	0.44	0.77
	VERY FIRM		0.27	0.48
	PS2		0.11	0.19
	PS6		0.05	0.09
	VERTICAL		0.03	0.06
	WEAK		0.03	0.06
	VERY FRIABLE		0.03	0.06
	BOTH		0.01	0.02
	PS8		0.01	0.01
	SINGLE (CRK)		0.01	0.01
	MODERATE		0.01	0.01
4	PS2	16	0.38	0.06
	PS4		0.19	0.03
	BOTH		0.19	0.03
	VERY FIRM		0.13	0.02
	PS6		0.06	0.01
	WEAK		0.06	0.01

* PS2, ped size 2-5mm; PS4, ped size 10-20mm; PS6, ped size 50-100mm , PS8, no ped size given; VERY FIRM,VERY FRIABLE, LOOSE, and refers to the consistence; WEAK and MODERATE refers to the grade; CRK SINGLE, MASSIVE, BOTH, and VERTICAL) refers to the crack orientation, single, massive, both, and vertical, respectively.

Table 8. Input variables and their probabilities in each splitting level, for the decision tree model considering ped size (PED), consistence (CST), and bulk density (D_b).

Splitting level	Variable*	Average splitting value	SD of splitting value	Total Count within level	Frequency as partitioning in level	Probability of occurrence in level
1	LOOSE	N/A	N/A	100	0.91	0.91
	VERY FRIABLE	N/A	N/A		0.06	0.06
	PS8	N/A	N/A		0.03	0.03
2	VERY FRIABLE	N/A	N/A	107	0.82	0.88
	LOOSE	N/A	N/A		0.07	0.08
	PS2	N/A	N/A		0.06	0.06
	PS4	N/A	N/A		0.04	0.04
	FRIABLE	N/A	N/A		0.01	0.01
3	PS4	N/A	N/A	176	0.43	0.75
	D_b	1.35	0.07		0.35	0.61
	VERY FIRM	N/A	N/A		0.13	0.22
	PS6	N/A	N/A		0.05	0.08
	PS2	N/A	N/A		0.04	0.07
	VERY FRIABLE	N/A	N/A		0.02	0.03
4	VERY FIRM	N/A	N/A	17	0.29	0.05
	PS2	N/A	N/A		0.24	0.04
	D_b	1.38	0.16		0.24	0.04
	PS4	N/A	N/A		0.12	0.02
	PS6	N/A	N/A		0.06	0.01
	VERY FRIABLE	N/A	N/A		0.06	0.01

*VERY FIRM,VERY FRIABLE, LOOSE, and refers to the consistence; * PS2, ped size 2-5mm; PS4, ped size 10-20mm; PS6, ped size 50-100mm , PS8, no ped size; D_b refers to bulk density.

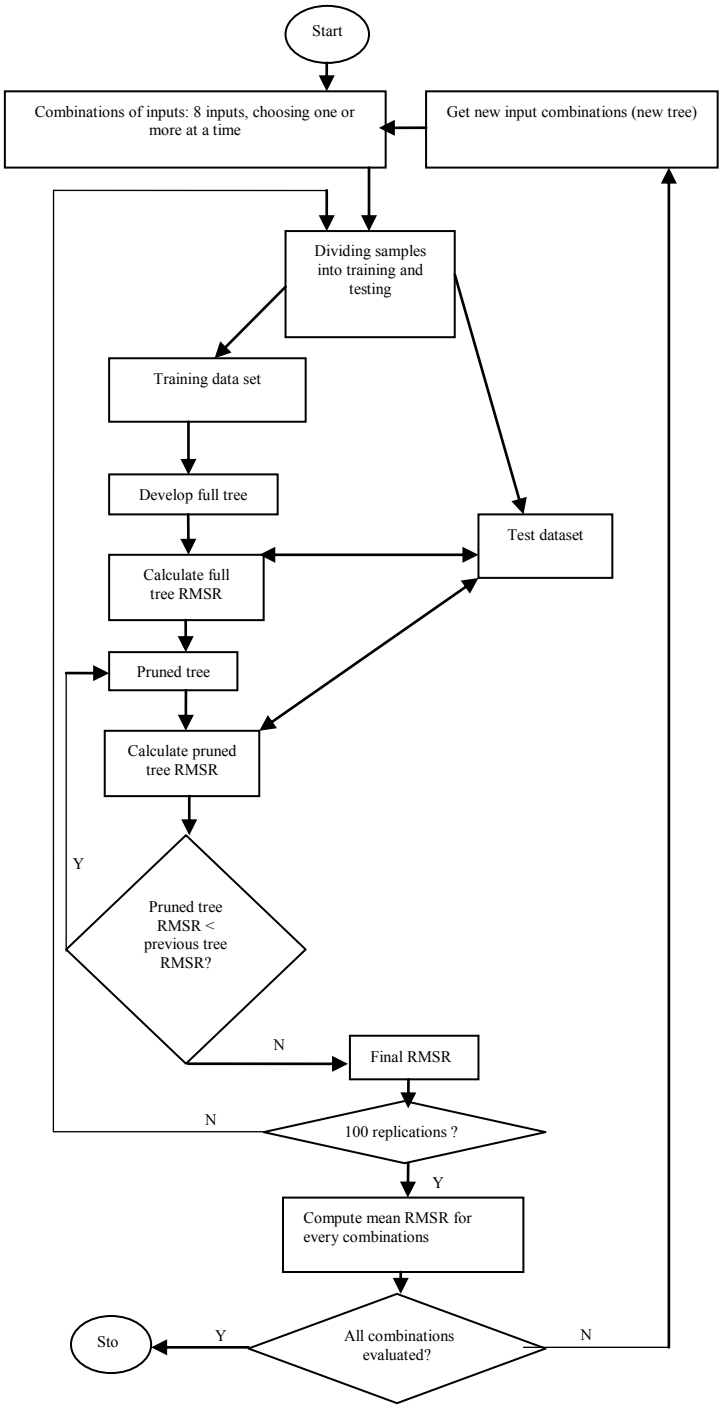


Figure 1. Flow diagram of algorithm for developing decision trees.

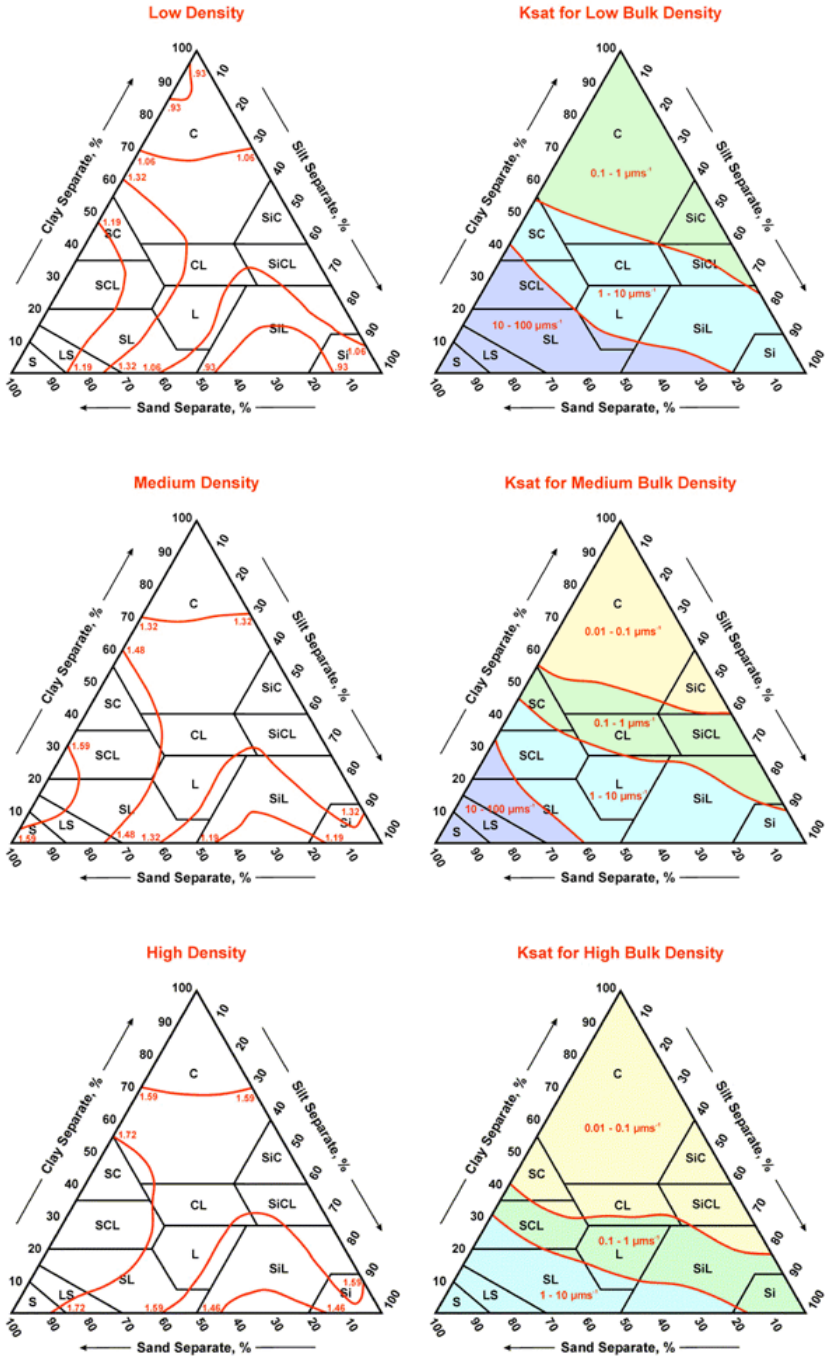
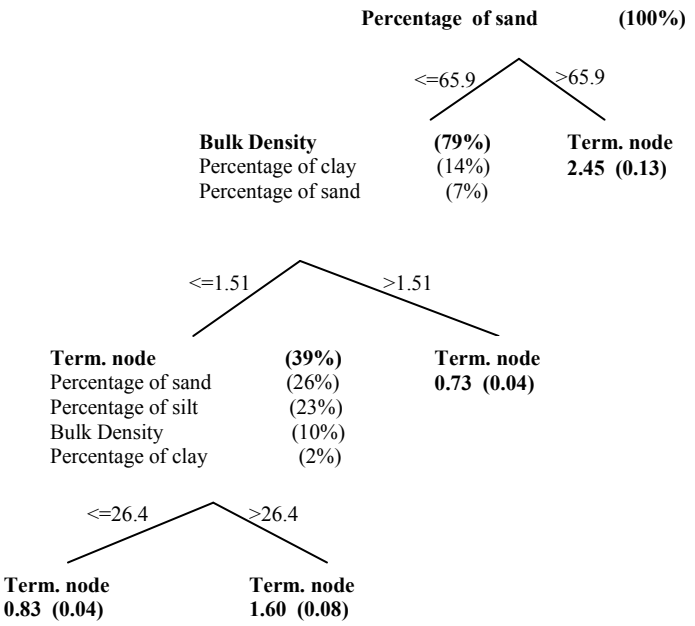
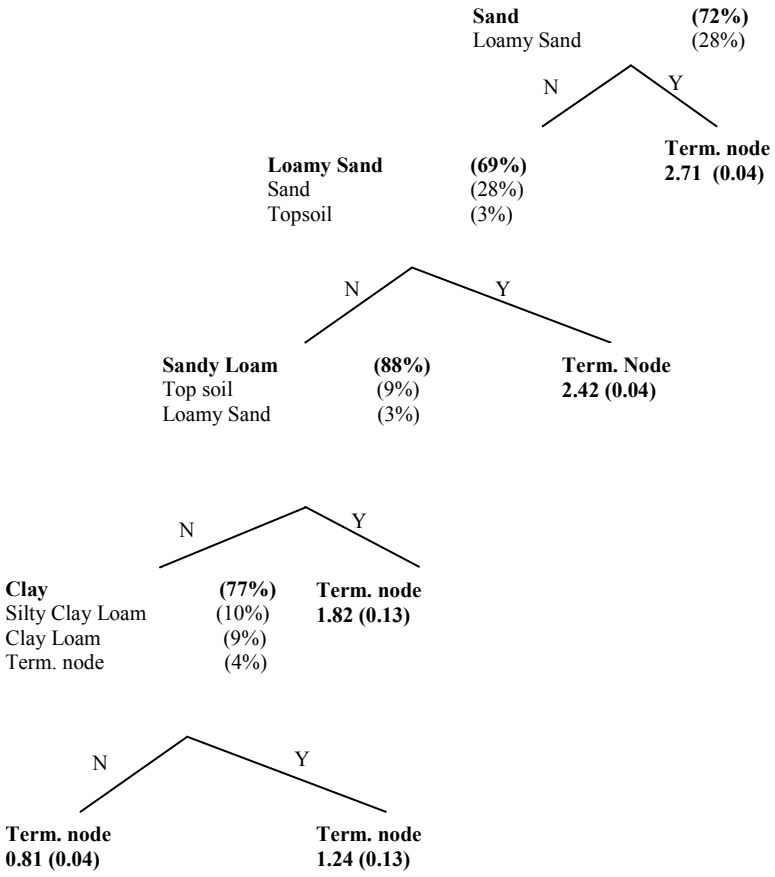


Figure 2. The Soil Survey method for estimating K_s using only texture and bulk density (USDA, 2010). The low, medium, and high density triangles in the left column indicate D_b for particular soil textures. The triangles in the right column predict the range for K_s.



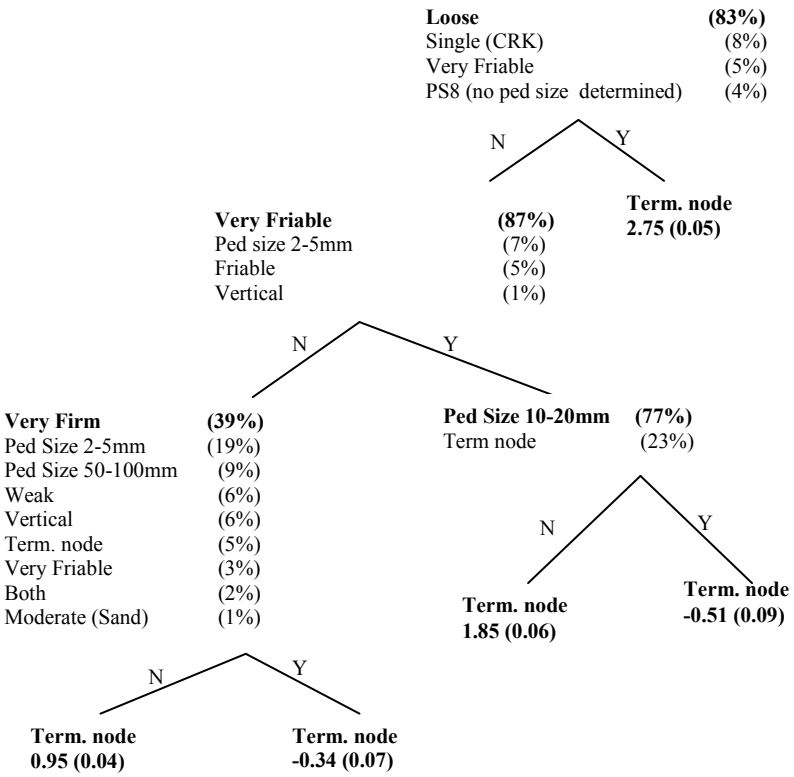
Test RMSR: 0.7663 (0.0617)

Figure 3. Most likely tree structure of best ranking decision tree model, the average saturated hydraulic conductivity (K_s) in bold and standard deviation (SD) in parenthesis, all values are $\log_{10}$ transformed.



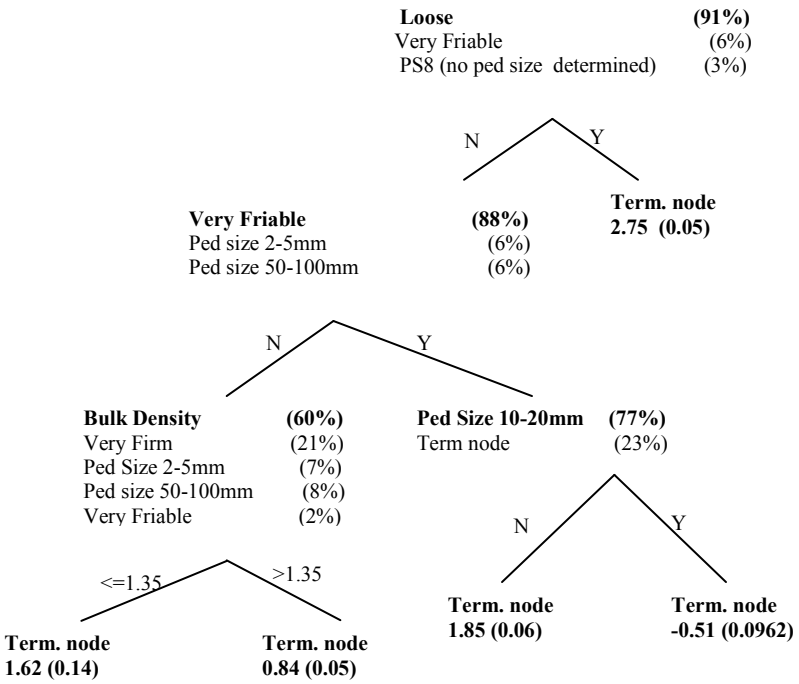
Test RMSR: 0.8104 (0.0627)

Figure 4. Decision tree considering only textural class (TXT) and horizon designation (HOR), the average saturated hydraulic conductivity (K_s) in bold and standard deviation (SD) in parenthesis, all values are $\log_{10}$ transformed.



Test RMSR: 0.8534 (0.0612)

Figure 5. Most likely decision tree structure considering only morphological descriptors of structure (PED, CRK, GRD, and CST), the average saturated hydraulic conductivity (K_s) in bold and standard deviation (SD) in parenthesis, all values are $\log_{10}$ transformed.



Test RMSR : 0.8482 (0.0593)

Figure 6. Most likely decision tree structure considering only using ped size (PED), consistence (CST), and bulk density (D_b), the average saturated hydraulic conductivity (K_s) in bold and standard deviation (SD) in parenthesis, all values are $\log_{10}$ transformed.

References

Anderson, J.L. and J. Bouma. 1973. Division S-5 – Soil genesis, morphology, classification: Relationships between saturated hydraulic conductivity and morphometric data of an argillic horizon. Soil Sci. Soc. Amer. Proc., 37:409-413.

Bigus, P. Joseph. 1996. Data mining with neural networks: Solving business problems from application development to decision support. McGraw (Ed.).

Breiman, Leo. Bagging predictors. 1994. Technical report no. 421. Department of statistics. University of California.

Bruce, R., J. Dane, V. Quisenberry, N. Powell, and A. Thomas 1983. Physical characteristics of soils in the southern region: Cecil. S. Coop. Ser. Bull. 267.

Crawford, J.W. 1994. The relationship between structure and the hydraulic conductivity. European Journal of Soil Sci. 45: 493-502.

Dane, J., D.K. Cassel, J.M. Davidson, W.L. Pollans, and V.L. Quisenberry. 1983. Physical characteristics of soils of the southern region: Troup and Lakeland series. S. Coop. Ser. Bull. 262. Alabama Agric. Exp. Stn., Auburn University.

D.A. O'Connell and P.J. Ryan. 2002. Prediction of three key hydraulic properties in a soil survey of a small forested catchment. Aust. J. Soil Res. 40: 191-206.

Farlow, J. Stanley. 1984. Self-organizing methods in modeling. GMDH type algorithms. Marcel Dekker (ed.) New York, New York.

Haykin, Simon. 1999. Neural networks: A comprehensive foundation. McMaster University. Hamilton, Ontario, Canada. Prentice Hall (ed.) New Jersey.

J.D. Jabro. 1992. Estimation of saturated hydraulic conductivity of soils from particle size distribution and bulk density data. Am. Soc. of Ag. Eng. 35:557-560.

Lilly, A., A. Nemes, W.J. Rawls, and Ya. A. Pachepsky. 2008. Probabilistic approach to the identification of input variable to estimate hydraulic conductivity. Soil Sci. Soc. Am. J. 72:16-24.

Lilly, A. 2000. The relationship between field-saturated hydraulic conductivity and soil structure: development of class pedotransfer functions. Soil Use and Management. 16: 56-60.

Lilly, A., and H.S. Lin. 2004. Using soil morphological attributes and soil structure in pedotransfer functions. p. 115-141. *In* Ya. Pachepsky and W.J. Rawls (ed.) Development of pedotransfer functions in soil hydrology. Dev. Soil Sci. 30. Elsevier Amsterdam.

Levin, E.R., D.S. Kimes, and V.G. Sigillito. 1996. Classifying soil structure using neural networks. Ecological Modelling. 92: 101-108.

Lin, H.S., K.J. McInnes, L.P. Wilding, and C.T. Hallmark. 1999a. Effects of soil morphology on hydraulic properties: I. quantifications of soil morphology. Soil Sci. Soc. Am. J. 63:948-954.

Lin, H.S., K.J. McInnes, L.P. Wilding, and C.T. Hallmark. 1999b. Effects of soil morphology on hydraulic properties: II. Hydraulic pedotransfer functions. Soil Sci. Soc. Am. J. 63:955-961.

Luxmoore, R. 1982. Physical characteristics of soils of the southern region: Fullerton and sequoia series. S. Coop. Ser. Bull. 268. ORNL-5868, Oak Ridge National Lab., TN (USA).

McKeague, J.A.,C. Wang, and G.C. Topp. 1982. Estimating saturated hydraulic conductivity from soil morphology. Soil Sci. Soc. Am. J. 46:1240-1244.

Nofziger, D., J.R. Williams, A.G. Hornsby, and A.L. Wood. 1983. Physical characteristics of soils of the southern region-Bethany, Konawa, and Tipton series. S. Coop. Ser. Bull. 265.

O'Neal, A.M. 1952. A key for evaluating soil permeability by means of certain field clues. Proc. Sci. Soc. Am. 16:312-315.

Pachepsky, Y.A., and W.J. Rawls. 2003. Soil structure and pedotransfer functions. European J. of Soil Sci. 54: 443-451.

Pachepsky, Ya. A., W.J. Rawls, and H.S. Lin. 2006. Hydropedology and pedotransfer functions. Geoderma. 131:308-316.

Pachepsky, Ya. and M.G. Schaap. 2004. Using soil morphological attributes and soil structure in pedotransfer functions. p. 115-141. *In* Ya. Pachepsky and W.J. Rawls (ed.) Development of pedotransfer functions in soil hydrology. Dev. Soil Sci. 30. Elsevier Amsterdam.

Quisenberry, V., D. K. Cassel, J.H. Dane, and J.C. Parker. 1987. Physical characteristics of soils in the southern region: Norfolk, Dothan, Goldsboro, Wagram. S. Coop. Ser. Bull. 263.

Radcliffe, D.E., and L.T. West. 2009. Design hydraulic loading rates for on-site wastewater systems. Vadose Zone Journal. 8: 64-74.

Rawls, W.J., D.L. Brakensiek, and K.E. Saxton. 1982. Estimation of soil water properties. Trans. ASAE. 1316-1320, 1328.

Römkens, M. J. M., H.M. Selim, H.D. Scott, R.E. Phillips, and F.D. Whisler. 1986a. Physical characteristics of soils in the southern region - Captina, Gigger, Grenada, Loring, Oliver,

and Sharkey series. S. Coop. Ser. Bull. 264. Mississippi Agric. and Forestry Exp. Stn., Mississippi State.

Römkens, M.J.M., J.M. Selim, R.E. Phillips, and F.D. Whisler. 1985b. Physical characteristics of soils in the southern region: Vicksburg, Memphis, Maury Series. S. Coop. Ser. Bull. 266. Alabama Agric. Exp. Stn., Auburn University.

Schaap, Marcel G. and Feike J. Leij. 1998. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil and Tillage. 47: 37-42.

Schaap, M.G. and F.J. Leij. 1998. Database related accuracy and uncertainty of pedotransfer functions. Soil Sci. 163: 765-779.

Schaap M.G.., F.J. Leij, and M.Th. van Genuchten (2001) ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions, J. Hydrology, 251, 163-176.

Schoeneberger, P., and A. Amoozegar. 1990. Directional saturated hydraulic conductivity and macropore morphology of a soil-saprolite sequence. Geoderma 46:31-49.

Soil Survey Division Staff. 1993. Soil survey Manual. USDA Agric. Handbk. 18. USDA-NRCS, vol.18. U.S. Government Printing Office, Washington D.C.

Siegelmann, T. Hava. 1999. Progress in Theoretical Computer Science. Neural Networks and Analog Computation: Beyond the Turing Limit. Birkhäuser Boston.

Sutton, Clifton D. 2005. Classification and regression trees, bagging, and boosting. Handbk. of Statistics. 24. Elsevier B.V.

Tietje, Olaf. and Volker Hennings. 1996. Accuracy of the saturated hydraulic conductivity

prediction by pedo-transfer functions compared to the variability within FAO textural classes. Geoderma. 69:71-84.

Timm, Neil, H. 2002. Springer Texts in Statistics. Applied Multivariate Analysis. Springer-Verlag.George Casell, Stephen Fienberg, and Ingram Olkin (eds.) New York.

Thomasson, A.J. 1978. Towards an objective classification of soil structure. J. of Sci. 29:38-46.

U.S. Department of Agriculture, Natural Resources Conservation Service. National Soil Survey Handbk., title 430-VI. Available online at: http://soils.usda.gov/technical/handbook/ accessed [July/8/2010].

Wagner, B., V.R. Tarnawski, G. Wessolek, and R. Plagge. 1998. Suitability of models for the estimation of soil hydraulic parameters. Geoderma 86:229-239.

Wösten, J.H.M., Ya.A. Pachepsky, and W.J. Rawls. 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic properties. J. of Hydrol. 251:123-150.