Conversion Between Soil Texture Classification Systems Using the Random Forest Algorithm

Authors: Cisty, Milan, Celar, Lubomir, and Minaric, Peter

Source: Air, Soil and Water Research, 8(1)

Published By: SAGE Publishing

URL: https://doi.org/10.1177/ASWR.S31924

Conversion Between Soil Texture Classification Systems Using the Random Forest Algorithm

Milan Cisty, Lubomir Celar and Peter Minaric

Faculty of Civil Engineering, Slovak University of Technology in Bratislava, Bratislava, Slovakia.

ABSTRACT: This study focuses on the reclassification of a soil texture system following a hybrid approach in which the conventional particle-size distribution (PSD) models are coupled with a random forest (RF) algorithm for achieving more generally applicable and precise outputs. The existing parametric PSD models that could be used for this purpose have various limitations; different models frequently show unequal degrees of precision in different soils or under different environments. The authors present in this article a novel ensemble modeling approach in which the existing PSD models are used as ensemble members. An improvement in precision was proved by better statistical indicators for the results obtained, and the article documents that the ensemble model worked better than any of its constituents (different existing parametric PSD models). This study is verified by using a soil dataset from Slovakia, which was originally labeled by a national texture classification system, which was then transformed to the USDA soil classification system. However, the methodology proposed could be used more generally, and the information provided is also applicable when dealing with the soil texture classification systems used in other countries.

KEYWORDS: soil texture, particle-size distribution, data-driven modeling, ensemble model, random forests

CITATION: Cisty et al. Conversion Between Soil Texture Classification Systems Using the Random Forest Algorithm. *Air, Soil and Water Research* 2015:8 67–75 doi:10.4137/ASWR.S31924.

TYPE: Original Research

RECEIVED: July 17, 2015. **RESUBMITTED:** October 13, 2015. **ACCEPTED FOR PUBLICATION:** October 16, 2015.

ACADEMIC EDITOR: Carlos Alberto Martinez-Huitle, Editor is Chief

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 993 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by the by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/0665/15, and also by the European Commission's Seventh Framework project RECARE, Contract No. 603498. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: milan.cisty@stuba.sk

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

The term soil texture indicates the distribution of soil particles (mineral grains) in soil according to their size (diameter). The range of individual soil particle diameters is classified into discrete intervals, which are also known as grain size fractions or categories (for instance, labeled as sand, silt, or clay). There are many classification systems in the world, which differ according to the limits of the diameter sizes for each grain fraction or by the number of fractions.

The most preferred representation of such a classification is a grading curve. A grading curve is a cumulative function describing the relationship between the soil fraction percentages and particle diameters, where the vertical axis (y-axis) defines the percentage of each fraction and the horizontal axis (x-axis) defines the soil particle sizes on a logarithmic scale. A point on the curve gives the percentage according to the weight of material smaller in size than the diameter at the given point on the graph's x-axis.

Many environmental problems in which soil data serve as an input to simulation models are not restricted to national boundaries and therefore require international cooperation if solutions are to be found. The classification of soils according to their texture is one of the basic methods used for a soil description. However, only a few countries use the same particle-size fractions in their classification systems for soil textures. Therefore, the transformation of particle-size texture descriptions between various systems is needed.[1]

Through a soil texture description, also known as a particle-size distribution (PSD), it is possible to predict various important soil properties (eg, saturated hydraulic conductivity, the soil water retention curve, available water capacity, thermal conductivity, etc.). The so-called pedotransfer functions (PTFs) are often based on the sand, silt and clay fractions[2–4] of a particular classification system,[5] eg, the USDA classification system. Not all countries use this classification system; as a consequence, databases from these countries cannot provide us with the inputs for such computations or models. An example of existing tools that have been developed for the above-mentioned tasks is the Rosetta model, which was designed for PTF evaluations and is based on neural networks.[6] This model works exclusively using the USDA classification system, so if the available data are not classified in this system, it is often desirable to accomplish a reclassification. Also, in other tasks, it is often necessary to carry out the transformation of textural classifications when data from different sources should be merged and used together.

The present study deals with a description of a texture system reclassification by the proposed model on a dataset from Slovakia, which was originally labeled by the National Classification System. However, the authors of the article

assume that the proposed methodology could be used more generally and that the information provided is also applicable when dealing with other soil texture classification systems and in other countries. Besides the classification systems that we studied in this article (the Kopecky classification system used in the Czech Republic and Slovakia and the USDA system), various other classification systems are commonly known in the soil scientific community, eg, the FAO soil texture classification (also known as the *European Soil map* or *HYPRES*), the French "Aisne" soil texture classification, the French "GEPPA" soil texture classification, the German "Bodenartendiagramm" soil texture classification, the German "Standortserkundungs-Anweisung" soil texture classification for forest soils, the German "Landwirtschaftliche Boden" soil texture classification for arable soils, the UK Soil Survey of England and Wales texture classification, the Australian soil texture classification, the Belgian soil texture classification, etc.[7]

Some researchers have already proposed fitting the measured PSDs in various continuous parametric curves. When achieving such a relationship, it is possible to obtain a granular fraction's percentage ratio in the sample under consideration for any size of the particle's diameter, which means that it is possible to get the values necessary for accomplishing a translation from one texture classification system to another. Several authors have conducted comparative studies on various PSD models in order to determine the best model for the soil groups selected for their studies.[7–10]

The reported findings of the abovementioned works somewhat differ from each other, and there is no generally suitable PSD model available. In some of these models, there are also various optional parameters, the selection of which is based on a researcher's know how. If this is not accomplished correctly, the results of the computations may be biased. As the transformation of a soil texture system is usually only a prerequisite for solving some subsequent tasks, this bias is propagated in the subsequent modeling or other work. Therefore, for the sake of achieving more general and precise outputs while solving tasks dealing with transformations between various soil texture systems, the authors of the present article propose a hybrid approach, which has the potential for obtaining improved results. Although the authors continue recommending the use of the mentioned parametric PSD models in the proposed methodology, the final prediction is made by an ensemble machine learning algorithm based on regression trees, ie, the so-called random forest (RF) algorithm,[11] which is built on top of the outputs of models that serve as ensemble members.

Materials and Methods

Description of the study area and available datasets. The area of interest—the Zahorska Lowland—is located in central Europe, more specifically in the western Slovak Republic. It is bounded by the river Morava in the west, and the Little Carpathians mountain range forms a natural boundary in the east. Most of the Zahorska Lowland is composed of Neogene sediments of a marine origin and younger Quaternary sediments covering the surface of the plains, which are mainly represented by clayey sands, drift sands, and sandy clays.[12] The main pedogenetic factors of the lowlands are azonal, such as the accumulation activity of streams and soil-disrupting floods along with soil erosion. The most widespread soils in this area are chernozem, arenosol, and fluvial soils on the fluvial plains of river Morava. An intense accumulation process of organic soil matter can be observed in the chernozems; therefore, they are appropriate for a large number of different plant species.

On the contrary, arenosols are soils at an early stage of development, and they contain almost no continuous vegetation on their surface, due to which the organic matter content is very low. They are suitable for growing crops with lower demands such as rye. Fluvisols are periodically disrupted by floods. Their profile is often constantly loaded with new layers of flood sediment (sludge).[12]

Classification systems used in this study. Among the various soil classification systems based on the soil texture, the ones most used in the Slovak Republic are as follows:

1. The Kopecky texture classification system, which distinguishes four categories of particle classes. The first category (clay) contains particles with diameter less than 0.01 mm, the second category (silt) particles whose diameter is limited to an interval of 0.01–0.05 mm, the third category (powder sand) particles with diameter 0.05–0.1 mm, and the fourth category (sand) particles with sizes in the interval 0.1–2.0 mm.
2. A well-known classification very frequently used (also in Slovakia) is the USDA classification system, which is based on the classification of soils according to the percentage of clay (up to 0.002 mm), silt (0.002–0.05 mm), and sand (0.05–2 mm) particles. The visual representation of the USDA classification is a triangular classification diagram. Within this diagram, 12 basic grain classes are marked (clay, silty clay, sandy clay, sandy clay loam, clay loam, silty clay loam, sand, loamy sand, sandy loam, loam, silt loam, and silt) in which it is possible to classify the samples.

A comparison of these two classification systems clearly shows various amounts of fractions in each of them and the discordant limits for each fraction. Namely, the fraction of clay particles in the USDA classification system ranges up to a value of 0.002 mm, while the Kopecky classification sets this limit at 0.01 mm. When solving various tasks for which it is necessary to have soil texture data in the USDA classification (eg, the mentioned example with the application of the Rosetta model), datasets using the Kopecky classification system are not compatible, which can be a problem. As we have already mentioned, there are many different texture classification

systems in the world. Hence, similar situations could arise more often, and the methodology which the authors would like to propose in this article may be generally useful.

Description of the datasets used. Two datasets were used in this article:

1. Data A—data for which the whole grading curve is available on the basis of which it is possible to calculate the PSD values for both the Kopecky and USDA classifications.
2. Data B—data with only texture grain intervals under the Kopecky classification are available.

Dataset A. The samples were taken from the Zahorska Lowland.[13] The number of samples was 43. They were air-dried and sieved; textural and other analyses were performed. After these analyses were accomplished, the dataset contained the following parameters: grain curve, reduced bulk density ρ_d, and the points of the drying branches of the water retention curve. Additionally, the data for each sample also contain the volume of humus in the soil, the value of the saturated hydraulic conductivity *K*, and the geographic coordinates of the individual samples. Dataset A was recently obtained, and for each sample, a complete grading curve from which the readings of the percentages of the Kopecky and USDA classifications were made is also available. Because textural information for both classifications (Kopecky and USDA) is available for this dataset, these data were used to create and verify a model that serves for converting the soil textural description from the Kopecky classification to the USDA classification system.

Dataset B. This dataset contains data obtained from a previous work that was conducted in the area of the Zahorska Lowland in Slovakia. A total of 140 soil samples were taken from the various localities in this area, but the exact geographic location was not recorded when the samples were taken.[14]

The soil samples were evaluated by similar laboratory methods as in the previous dataset. The soil samples were air-dried and sieved for a physical analysis. A particle-size analysis according to four grain categories was performed using Cassagrande's methods. Category I means the percentages of the clay (diameter $<$ 0.01 mm), category II those of silt (0.01–0.05 mm), category III those of fine sand (0.05–0.1 mm), and category IV those of sand (0.1–2.0 mm). The dry bulk density, particle density, porosity, and saturated hydraulic conductivity were also measured for the soil samples. The points of the drying branches of the PTFs for the pressure head values of –2.5, –56, –209, –558, –976 and –3060 cm were estimated using an overpressure equipment (set for pF-determination with ceramic plates). As they were in a sufficiently large quantity, these data served later (not in this work) for the derivation of the PTFs by using a data-driven model.

PSD models and their fitting. The nine parametric models of PSD involved in this study were evaluated and compared in order to derive the PSD functions. The models were developed by using an optimization procedure in order to choose the most suitable set of parameters of each model. The models chosen are the Fredlund models with three and four parameters (FR3, FR4), the Weibull model with three parameters (WB3), the Andersson model with four parameters (AND4), the van Genuchten model with three parameters (VG3), two Gompertz models (GP2, GP4), and a logarithmical (LG) and an exponential (EXP) model. These models were previously applied in various works, eg, in Zhao et al,[15] for determining the hydrological properties of soils adjacent to dams constructed in China. The parts of the models were chosen according to the results obtained by Botula et al.[7]

In the FR3 Fredlund model, three parameters needed to be optimized[16] ie, *a*, *b*, and *c*, and it uses the predefined parameters d_f = 0.001 mm and d_m (d_f = 0.0001 mm is a parameter related to the amount of fine particles in a soil and d_m is the diameter of the minimum allowable particle size). The FR4 model contains four parameters, ie, *a*, *b*, *c*, and d_f, which should be found by the optimization procedure and the predefined parameter d_m. The WB3 model with three parameters, ie, *a*, *b*, and *c* (and two predefined parameters: d_{min} = 0.002 mm and d_{max} = 2.0 mm), was previously used in Refs. 17 and 18 for the creation of PSD curves of several diverse soils. AND4 is a four-parameter model (parameters *a*, *b*, *c*, *e*) developed by Jauhiainen,[19] based on an original theory of textural and water retention soil properties presented by Andersson.[20] The VG2 and VG3 models were proposed by Haverkamp and Parlange[2] on the basis of van Genuchten's original soil water retention curve model (developed in 1980). The VG3 model was subsequently developed from the first derived VG2 version with two parameters, when the mutual relation $m = 1 - 1/n$ was used. VG3 considers both fitting parameters *m* and *n* as different and independent of each other and uses three additional parameters (*a*, *b*, and *c*). The GP2 and GP4 models are two forms of the Gompertz model with two and four parameters (*a*, *b*) and (*a*, *b*, *c* and *e*). Their curves present specific cases of a logistic curve, which is more general than the normal one. The equation of this curve constitutes an asymmetric closed form. Both models were previously used for extracting PSD curves by Silva et al[21] for soils in Brazil and by Nemes[8] for soils in Germany and the Netherlands. The soil particles expressed in these models follow the Gompertz distribution. Finally, the logarithmic model and EXP model contain two parameters—LG parameters *a* and *b* and EXP parameters *c* and *b*.

All the models listed were used for determining the PSD functions by the optimization method. The *L-BFGS-B* method used box constraints, which means that for each variable (the model's parameter), a lower bound and an upper bound are given. This method is a limited-memory modification of the quasi-Newton method.[22] It was implemented in R language.[23] The purpose of this optimization was to predict the points of the grain curve by each model as closely

as possible to the observed data by searching for the proper parameters of the model. The problem to be solved should be defined by the objective function, which in this article is proposed to have the following form:

$$\text{Minimize}\left(\sum_{d=1}^{n} W_d \cdot \left(\left(F_d^{\text{act}} - F_d^{\text{comp}}\right)/F_d^{\text{act}}\right)^2\right) \quad (1)$$

where W_d are the weights assigned to the grain diameters, by which it is possible to stress the precision of fitting of the particular points of the grain curve; n is the number of grain diameters; F_d^{act} is the actual (measured) percentage of the material with a diameter d or the smallest one in the sample; and F_d^{comp} is the percentage computed by the corresponding equation of the particular model from Table 1. This objective function is proposed to be minimized. In the case of an ideal model, the value of the objective function is zero.

Description and tuning of the RF model. The usual process for finding the best model for obtaining a proper theoretical parametric PSD function means applying more methods, eg, the models from Table 1, comparing their predictive ability with the help of the observed data and some statistical goodness-of-fit indicator, and then finally choosing the best performing model. However, there is usually no best parametric model that is superior under all circumstances. Various parametric models frequently show different degrees of precision in different soils and different environments, so the application of a single parametric model often leads to a functional relationship that could be more precise in one part of a textural domain but less suitable in other parts. One of the possible solutions of this problem is the application of the ensemble methodology, which uses the best features of various parametric models for achieving more general results from fitting the data to the actual values measured. Moreover, as was proved in this study, such a meta-model usually has the capability to fix systematic errors, if they are produced by the individual models (underestimation, overestimation, multiplicative error, etc.).

The goal of the ensemble methodology is to combine the outputs of several models in order to improve the generalizability/robustness that could be obtained from any of the constituent models. The nine parametric models described hereinbefore were used for obtaining the parametric PSD

Table 1. PSD models used in this study.

NAME	MODEL	PARAMETERS
Fredlund (FR3)	$F(d) = \frac{1}{\ln\left[\exp(1) + \left(\frac{a}{d}\right)^b\right]^c}\left\{1 - \left[\frac{\ln\left(1+\frac{d_f}{d}\right)}{\ln\left(1+\frac{d_f}{d_m}\right)}\right]^7\right\}$	a, b, c, $d_f = 0.001$ mm, $d_m = 0.0001$ mm
Fredlund (FR4)	$F(d) = \frac{1}{\ln\left[\exp(1) + \left(\frac{a}{d}\right)^b\right]^c}\left\{1 - \left[\frac{\ln\left(1+\frac{d_f}{d}\right)}{\ln\left(1+\frac{d_f}{d_m}\right)}\right]^7\right\}$	a, b, c, d_f, $d_m = 0.0001$ mm
Weibull (WB3)	$F(d) = c + (1-c)\{1 - \exp(-aD^b)\}$ where $D = (d - d_{\min})/(d_{\max} - d_{\min})$	a, b, c, $d_{\min} = 0.002$ mm, $d_{\max} = 2$ mm
Andersson (AND4)	$F(d) = a + barctg\left[c \log\left(\frac{d}{e}\right)\right]$	a, b, c, e
Van Genuchten (VG3)	$F(d) = \left[1 + (a/d)^b\right]^{-c}$, where b and c are independent to each other	a, b, c
Gompertz (GP2)	$F(d) = \exp\left(-\exp\left[-a(d-b)\right]\right)$	a, b
Gompertz (GP4)	$F(d) = c + e\exp\left(-\exp\left[-a(d-b)\right]\right)$	a, b, c, e
Logarithmic (LG)	$F(d) = a\ln d + b$	a, b
Exponential (EXP)	$F(d) = cd^{-b}$	c, b

curves selected for the present study. The outputs of these parametric models are inputs for the ensemble model, namely, a data-driven model based on the RF algorithm, which means that a stacking type of ensemble was used. Stacking introduces the concept of (1) base models (PSD models) and (2) a meta-model, which computes the final results and replaces the averaging procedure used, eg, in bagging. In such a way, stacking tries to learn which base models are more reliable than others, using the mentioned meta-model (RFs in the solved task) to discover how best to combine the output of the base models to achieve the final results. The results of the base models are de facto new data for the learning problem, and an RF algorithm is used to solve this problem.

The RF algorithm consists of a set of regression trees (in this study, the authors are addressing a regression problem; it could also be populated with classification trees in the case of studying a classification problem). The resulting RF prediction is an average of the values of these many tree outputs, each one of which is grown on a bootstrap sample of the training data. The user chooses the number of trees that will be in the RF ensemble. A bootstrap sample means that each tree is trained using a sample obtained by randomly drawing N cases with replacements from the original dataset, where N is the number of variables in that dataset. With each of these bootstrapped training sets, a different tree is obtained. For the regression, the values predicted by each tree are averaged to obtain the RF prediction. More details and a more mathematically based explanation of the RF algorithm could be found in Breiman.[11]

Optimization of the RF model. Data-driven models must be optimized to obtain reliable results that are as precise as possible. The optimization of the model mainly means finding the optimal model parameters. An RF has three tunable parameters: *ntree* (the number of trees to grow), *mtry* (the number of variables randomly sampled as candidates at each tree split), and *nodesize* (the minimum size of the terminal nodes), which has the main effect on the final precision of the model. Two concepts are applied in this work for tuning an RF: grid search and repeated cross-validation.

The grid search is designed in the optimization process to choose the values for each parameter of the model by consecutively picking them from a grid of predefined values and then calculating with these parameters. The best combination of the parameters is chosen from that iteration in which the highest degree of precision of the model was achieved.

This precision is evaluated as an average value of more runs of the so-called cross-validation process. The so-called repeated cross-validation is used in the present article.[18] In this process, the initial step consists of randomly dividing the training data into several approximately equal-sized datasets called *folds*. The training process uses all the folds except one as the inputs to the model, and the one unused fold is used as the validating data. This process runs as many times as the number of folds created with each combination of parameters. *Repeated cross-validation* means that the initial random splitting of the training data into folds is repeated more than once. This repetition is applied to obtain a more generalized evaluation of the model. The precision of the model in each iteration of the grid search is in fact the average value of the assessed statistic (eg, root mean square error (RMSE)) from all runs of the model with one set of parameters, eg, if there are two repetitions and five folds, the resulting statistic is the average value from 10 values.

This tuning concept involves two purposes in this study: (1) finding the best parameters of the RF algorithm and (2) estimating the precision of the proposed model, which could be expected for future data.

Results and Discussion

Fitting of the parametric PSD functions. When dealing with the transformation of a soil texture classification from the Kopecky system to the USDA classification, it is only necessary to extrapolate one point of the PSD, ie, to find the grain curve value for a soil particle diameter of 0.002 mm. Other USDA fractions could be derived from the Kopecky classification system by basic arithmetic operations. Nevertheless, from a more general point of view, when it is necessary to deal with other classification systems, and more points or different points of a grain curve are needed to be modeled, the methodology is the same as proposed in this article for this one point. Moreover, in previous works, when finding the percentages of unknown fractions in a soil sample, interpolation problems were mainly solved.[7,15] In this work, it is necessary to solve the extrapolation problem on the left tail of a PSD distribution, which is more complicated, so the methodology will surely be suitable while solving relatively easier interpolation problems.

Figure 1 shows the results of fitting various PSD functions to the soil texture data measured by the approach described in the "Materials and methods" section. As can be seen from the graphic representation on the left side of the correlation matrix and the correlation coefficients on its right side, the majority of models work quite well. This is especially apparent in the last column of the correlation matrix, where the correlation of the model with the measured data is evaluated. The results are similar to those that have been obtained in other works.[7,15] Additional statistical coefficients, which serve for an evaluation of the model's fitting, are available in Table 2. These coefficients assess the correspondence between measured and computed values. The statistics used are mean error (ME), mean absolute error (MAE), mean squared error (MSE), RMSE, the percent bias between simulated and observed values (PBIAS%), and the Pearson correlation coefficient (r).

Figure 2 shows an evaluation of the best models for various soils by their absolute errors. In every sample of the data (dataset A), the model with the lowest error, while predicting the grain curve point of 0.002 mm needed for translation to the USDA classification, is indicated by the color key. As can be seen, there is no single best model that could be preferred, either for a whole set of data samples or for samples of various

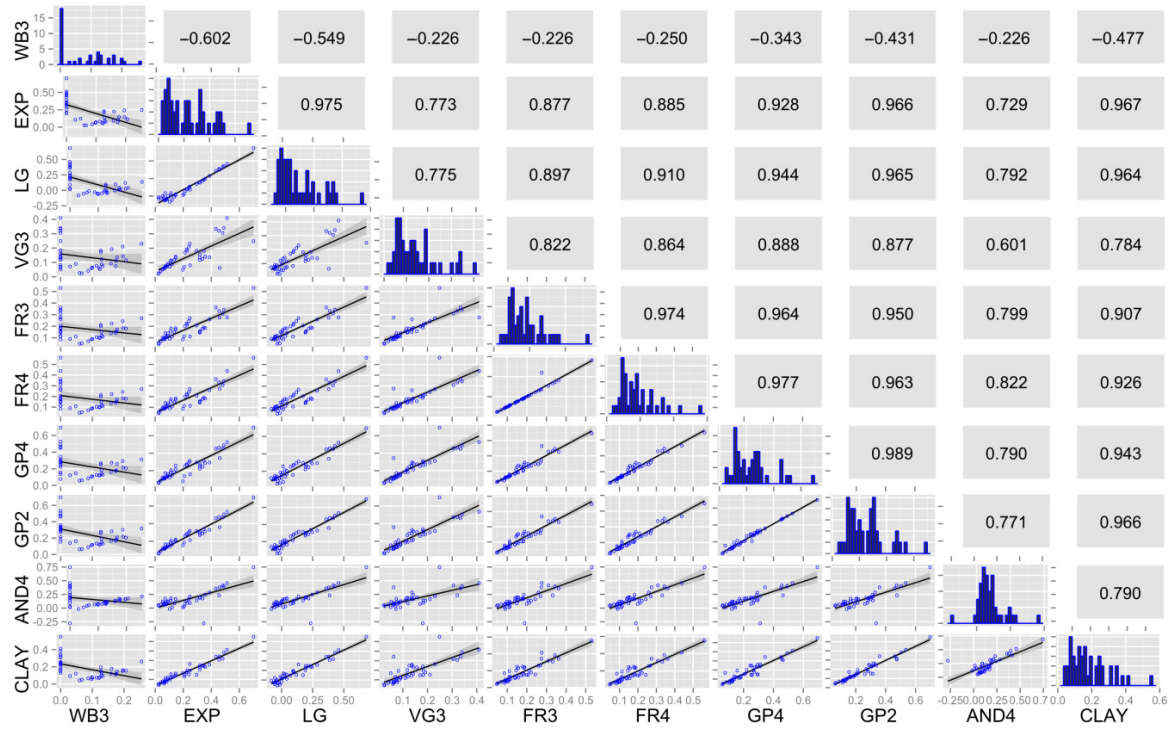

Figure 1. Correlation matrix of PSD models and measured data (in the clay column).

soil types. This means that the proposed ensemble methodology, which is a combination of all the successful models, could be useful in this task.

As can be seen in Table 2, some models are not suitable for extrapolation problems regarding PSD fitting, so they were excluded from the final ensemble modeling. This inappropriateness for extrapolation is especially clear for the Weibull model. This is the case not only in the study presented but also could be considered as a general result. It can be explained by the following: Besides the empirical results, which were obtained by the computational experiment accomplished in this work and evaluated in Table 2, an important issue for this model is that parameter d_{min} should be set to represent the minimal diameters of the soil particles, which are assumed to be present in the sample. However, in the case of the left tail extrapolation, this is the unknown information for which one is searching, so it cannot be correctly set in advance. The second model that was excluded from the final modeling was the LG model, because negative values of the clay content were computed by this model for some samples. This means that the results of the seven models finally served as inputs to the ensemble modeling.

According to the statistical values in Table 2, evaluations of the results by various statistical coefficients differ, eg, the EXP model is evaluated as the best model by the correlation coefficient (r), but the FR4 model is evaluated as the best model by the RMSE. Various goodness-of-fit statistics evaluate different aspects of fitting, eg, the emphasis of one is more from a perspective of the variances and other statistics capture the bias better. For example, although it is possible to see in Figure 1 that the best correlation coefficient is for the EXP model, its prediction has multiplicative errors, which are not evident if one is only using an evaluation by the correlation coefficient (see Fig. 1, left side).

In this article, the authors propose a methodology inspired by the idea of ensemble learning, in which the RF algorithm is built on the top of the predictions computed by various

Table 2. Goodness-of-fit measures for the PSD models.

	AND4	EXP	FR3	FR4	GP2	GP4	LE	LG	VG2	VG3	WB3
ME	−0.023	0.044	−0.004	−0.0002	0.069	0.055	0.044	−0.064	−0.138	−0.043	−0.105
MAE	0.055	0.058	0.036	0.033	0.070	0.061	0.058	0.088	0.139	0.050	0.137
MSE	0.010	0.006	0.002	0.002	0.007	0.006	0.006	0.010	0.034	0.007	0.037
RMSE	0.100	0.076	0.047	0.042	0.081	0.074	0.076	0.099	0.185	0.081	0.192
PBIAS%	−12.8	24.2	−2.4	−0.1	38.0	30.2	24.2	−35.1	−75.7	−23.7	−57.9
R	0.790	0.967	0.907	0.926	0.966	0.944	0.967	0.964	0.261	0.784	−0.477

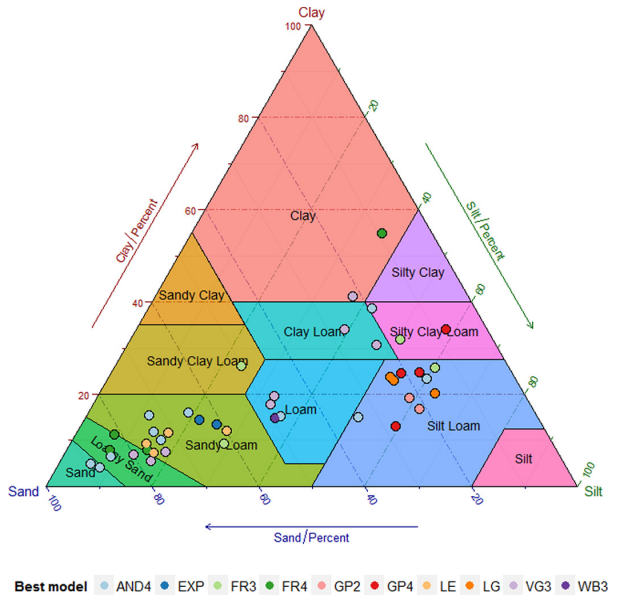

Figure 2. Evaluation of the PSD fittings for various USDA soil types.

parametric PSD models (they are inputs to the RF) and the optimal final result is obtained with this ensemble.

Fitting of the RF algorithm. While producing the final model for fitting the PSD, there are two basic tasks that are necessary to deal with: (1) to find the optimal model (eg, the optimal parameters of the RF algorithm that are suitable for the task to be solved) and (2) to evaluate the model's expected performance. The predicted values of the USDA clay fractions in the seven PSD models derived from the dataset A were used for the model's calibration or so-called training. This dataset is used because in the training phase of data-driven modeling, it is necessary to know not only the input data (Kopecky grain fractions) but also the measured outputs (USDA clay fractions), which are known in this case as mentioned in the descriptions of dataset.

The basic problem with the training dataset in this task is that it is relatively small (43 samples). The usual, so-called validation set approach, which involves randomly dividing the available set of samples into two parts, ie, a training set and a validation or hold-out set,[18] is not appropriate to apply here. Instead of this method, the authors used the repeated cross-validation approach described in the "Materials and methods" section of this article. Through the know-how of the data mining community as expressed in various books and papers,[18] fivefold cross-validation was principally used with two repetitions. The resulting RF model is based on the best parameters obtained from the cross-validation evaluations. The expected precision of the model is computed by using the computed and observed data from the folds held out in each iteration of the cross-validation.

The fitted RF model has the following parameters: 500 trees, four variables randomly sampled as candidates at each tree split, and terminal nodes with a minimum size of 5. As has already been stated in the "Materials and methods" section, the purpose of the cross-validation was not only to find these optimal parameters but also to evaluate the precision of the proposed model, which could be expected for future data. The precision using regression coefficient r was evaluated for the ensemble model of the soil texture transformation from the Kopecky to the USDA classification as 0.971 and using RMSE as 0.0343. When these values are compared with the results of the individual models of the ensemble model (see Table 2), it can be seen that no model achieved such a degree of precision; hence, the usefulness of the ensemble approach for this study was verified.

Some data-driven models, such as artificial neural networks, have often been criticized because of their black box character. It is true that these models are generally not meant to be descriptive and are usually not well-suited for inferences. Because of this, the authors used RF model in this work, as it not only generates very accurate estimates and is considered to be one of the most effective data-driven algorithms but also offers some information that helps understanding the modeled task. The RF model also included the ability to measure the importance or influence of each of its variable.[11] For each tree, the MSE on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The differences between the two are then averaged over all the trees, and in such a way, the importance of the variable is obtained (eg, a decrease in accuracy after the permuting of the variables over all the trees is measured). The importance of each variable of the proposed ensemble model (eg, the predicted values by the 10 parametric PSD models) is scaled from 0 to 100 and displayed in Figure 3. With this evaluation, it is possible to see which model is more important and useful for the final prediction. The most important ones are the EXP model and the Gompertz model with two parameters. Although the authors would like to underscore the general usefulness of the proposed methodology, it has to be said that in the case of the other tasks, especially in the case of interpolation problems, other models may have a greater impact on the ensemble model. This result allows us to assume that in the context of ensemble modeling, the correlation coefficient is more important for the model selection than the other statistical measures that evaluate

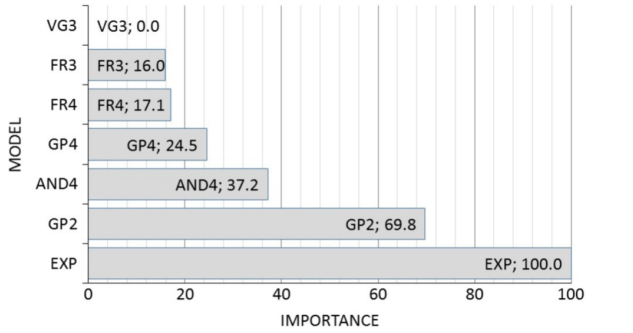

Figure 3. Importance of individual PSD models in the final RF ensemble.

errors (such as MSE, RMSE, etc.), because these error coefficients are better, for instance, for both Fredlund models (see Table 2). This is true only in the context of ensemble modeling; otherwise, the Fredlund models should be chosen (when one is deciding only between the individual models for the final modeling). This could be explained by the better ability of the ensemble to repair systematic errors than the individual inaccuracies. Surprisingly, the Andersson model, which is not the most accurate one, plays quite an important role in the final ensemble. This is due to the fact that the Andersson model has a small degree of correlation with both of the best models (Fig. 2), so this means it is different. An efficient ensemble should be composed of predictors that are not only sufficiently accurate but also dissimilar, in the sense that the predicted errors occur in different regions of the input space.[24] Obviously, combining several identical models results in no gain in precision. From the evaluation in Figure 3, it can be seen that the ensemble mechanism is also capable of excluding a model if it is redundant (the Van Genuchten model in our case).

The authors would like to emphasize the following practical aspect about ensemble modeling. According to the so-called *no free lunch* theorem, it is never clear in advance which PSD model best suits a particular task. For this reason, it is usually necessary to try more models. On the basis of the results of this article, it could be said that when more PSD models are already fitted, instead of selecting and using only the best one, it is better to compose an ensemble prediction based on all of these already fitted PSD models (or on the basis of a subset of these models). Forming an ensemble usually brings an improvement in precision as has also been confirmed in the case study in this article, and ensemble prediction is relatively easy to accomplish when the fitted models for a particular task are already available.

Conclusion

In this work, the authors investigated whether an ensemble paradigm could bring some improvement in the soil texture transformation task, when the existing PSD models are used as ensemble members. This paradigm was verified by using a soil dataset from Slovakia; however, the methodology proposed is also appropriate when dealing with soil texture classification systems used in other countries. An improvement in precision was demonstrated in the mentioned case study, and it is documented in the article that the ensemble model worked better than any of its constituents. The precision was evaluated by a regression coefficient for the ensemble model of the soil texture transformation from the Kopecky to the USDA classification as 0.971 (very close to 1). When these values are compared with the results of the individual parametric PSD models of the ensemble model (see Table 2), which could eventually be used separately for such a transformation, it can be seen that no model achieved such a degree of precision as the proposed RF ensemble. The results should also be verified in the future on other datasets and for the transformation of other classification systems. However, in this work, the extrapolation problem was solved (the computed margin of the target distribution is on the left tail of the original PSD distribution), which is quite complicated, so the authors assume that the proposed methodology will generally have even better results when solving easier and more frequent interpolation problems, eg, when the computed margin of a target distribution is between two known soil texture margins of the original distribution.

Author Contributions

Conceived and designed the experiments: MC. Analyzed the data: MC. Wrote the first draft of the manuscript: MC. Contributed to the writing of the manuscript: PM and LC. Agree with manuscript results and conclusions: MC, LC and PM. Jointly developed the structure and arguments for the paper: MC, LC and PM. Made critical revisions and approved final version: MC. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Minasny B, McBratney AB. Digital soil mapping: a brief history and some lessons. Available online, accessed 13.11.2015: https://www.researchgate.net/publication/280733900_Digital_soil_mapping_A_brief_history_and_some_lessons.
2. Haverkamp R, Parlange JY. Predicting the water-retention curve from particle-size distribution: 1. Sandy soils without organic matter. *Soil Sci*. 1986;142:325–339.
3. Smettem KRJ, Gregory PJ. The relation between soil water retention and particle size distribution parameters for some predominantly sandy Western Australian soils. *Aust J Soil Res*. 1996;34:695–708.
4. Wu L, Vomocil JA, Childs SW. Pore size, particle size, aggregate size, and water retention. *Soil Sci Soc Am J*. 1990;54:952–956.
5. Bouma J. Using soil survey data for quantitative land evaluation. *Adv Soil Sci*. 1989;9:177–219.
6. Shaap MG, Feike JL, Martinus Thv G. Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *J Hydrol*. 2001;251:163–176.
7. Botula YD, Cornelis WM, Baert G, Mafuka P, Van Ranst E. Particle size distribution models for soils of the humid tropics. *J Soils Sediments*. 2013;13:686–698.
8. Nemes A, Wosten JHM, Lilly A, Voshaar JHO. Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. *Geoderma*. 1999;90:187–202.
9. Hwang S. Effect of texture on the performance of soil particle-size distribution models. *Geoderma*. 2004;123:363–371.
10. Shangguan W, Dai Y, García-Gutiérrez C, Yuan H. Particle-size distribution models for the conversion of Chinese data to FAO/USDA System. *Sci World J*. 2014;2014(2014):11.
11. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
12. Buol SW, Southard RJ, Graham RC, McDaniel PA. Soil Genesis and Classification. ISBN-13: 978-0813807690, Chichester UK, Wiley-Blackwell, 2011.
13. Skalova J, Cisty M, Bezak J. Comparison of three regression models for determining water retention curves. *J Hydrol Hydromech*. 2011;59(4):275–284.
14. Stekauerova V, Skalova J, Sutor J. Using of pedotransfer functions for assessment of hydrolimits. *Plant Soil Environ*. 2002;48:407–412.
15. Zhao P, Shao M, Horton R. Performance of soil particle-size distribution models for describing deposited soils adjecent to constructed dams in the China Loess Plateau. *Acta Geophysica*. 2011;59:124–138.
16. Fredlund MD, Fredlund DG, Wilson GW. An equation to represent grain-size distribution. *Can Geotech J*. 2000;37:817–827.
17. Assouline S, Tessier D, Bruand A. A conceptual model of the soil water retention curve. *Water Resour Res*. 1998;34:223–231.
18. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York: Springer; 2013.
19. Jauhiainen M. Relationships of Particle Size Distribution Curve, Soil Water Retention Curve and Unsaturated Hydraulic Conductivity and Their Implications on Water Balance of Forested and Agricultural Hillslopes [Ph.D dissertation]. Finland: Helsinki University of Technology; 2004.

20. Andersson S. Markfysikaliska undersökningar i odlad jord, XXVI. Om mineraljordens och mullens rumsutfyllande egenskaper. En teoretisk studie. Uppsala: Swedish University of Agricultural Sciences; 1990.
21. Silva EM, Lima JEFW, Rodriguez LN, Azevedo JA. Comparação de modelos matemáticos para o traçado de curvas granulométricas. *Pesqui Agropecu Bras*. 2004; 39:363–370.
22. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput*. 1995;16:1190–1208.
23. R Development Core Team, 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R foundation for Statistical Computing; 2013.
24. Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn*. 2003;51:181–207.