Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

Hydrology and Earth System Sciences
Discussions

Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data

Mo Zhang[1,2], Wenjiao Shi[1,3]

[1]Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[2]School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China
[3]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

Correspondence to: Wenjiao Shi (shiwj@lreis.ac.cn)

Abstract. Soil texture and soil particle size fractions (psf) play an increasing role in physical, chemical and hydrological processes. Digital soil mapping using machine-learning methods was widely applied to generate more detailed prediction of qualitative or quantitative outputs than traditional soil-mapping methods in soil science. As compositional data, interpolation of soil psf combined with log ratio approaches was developed to improve the prediction accuracy, which also can be used to indirectly derive soil texture. However, few reports systematically analyzed and compared the classification and regression, the accuracies of original (untransformed) and log ratio approaches, and the performance of direct and indirect soil texture classification using machine-learning methods. In this total, a total of 45 evaluation models generated from five different machine-learning models combined with original and three log ratio approaches—additive log ratio, centered log ratio and isometric log ratio (ALR, CLR and ILR, respectively), to evaluate and compare the performance of soil texture classification and soil psf interpolation. The results demonstrated that log ratio approaches modified the soil sampling data more symmetrically, and with respect to soil texture classification, random forest (RF) and extreme gradient boosting (XGB) showed notable consequences. For soil psf interpolation, RF delivered the best performance among five machine-learning models with lowest root mean squared error (RMSE, sand: 15.09 %, silt: 13.86 %, clay: 6.31 %), mean absolute error (MAE, sand: 10.65 %, silt: 9.99 %, clay: 5.00 %), Aitchison distance (AD, 0.84) and standardized residual sum of squares (STRESS, 0.61), and highest coefficient of determination (R^2, sand: 53.28 %, silt: 45.77 %, clay: 53.75 %). STRESS was improved using log ratio approaches, especially CLR and ILR. There is a pronounced improvement (21.3 %) in the kappa coefficient using indirect soil texture classification compared to the direct approach. Our systematic comparison helps to elucidate the processing and selection of compositional data in spatial simulation.

1 Abbreviations: psf, soil particle-size fractions; HRB, Heihe River Basin; DSM, digital soil mapping; KNN, k-nearest neighbor; MLP, multilayer perceptron neural network; RF, random forest; SVM, support vector machines; XGB, extreme gradient boosting; ALR, additive log-ratio; CLR, centered log-ratio; ILR, isometric log-ratio; ORI, original; ROC, receiver

1 Introduction

Soil texture, classified by ranges of soil particle-size fractions (psf), is one of the most important attributes affecting the soil properties and the physical, chemical and hydrological processes covering soil porosity, soil fertility, water retention, infiltration, drainage and aeration. Measuring soil texture can be used for soil fertility management (Pahlavan-Rad and Akbarimoghaddam, 2018), water management (Thompson et al., 2012), maintenance of organic carbon (Bationo et al., 2007) and provision of ecosystem services (Adhikari and Hartemink, 2016). The soil psf, i.e., sand, silt and clay, are vital in most hydrological, ecological, and environmental risk assessment models (Liess et al., 2012). The spatial distributions of soil texture and soil psf affect and control runoff generation, slope stability, depth of accumulation, and soluble salt content (McNamara et al., 2005; Follain et al., 2006; Yoo et al., 2006; Gochis et al., 2010; Crouvi et al., 2013).

Previous reports revealed that there are close correlations between the spatial variations of soil texture and landscape and topography (Gobin et al., 2001; Brown et al., 2004; Zhao et al., 2009; Liess et al., 2012). Compared with traditional soil mapping methods, digital soil mapping (DSM) has an obvious advantage in that it is considerably more economical and efficient; additionally, soil maps using DSM yielded more details because of the development of data-mining algorithms and GIS tools and more extensive application of spatial remote sensing data, particularly in the regional and continental scale. DSM methods were applied by an increasing number of soil scientists to map soil properties using ancillary data (McBratney et al., 2003; Zeraatpisheh et al., 2017), the so-called environmental covariates, which can be obtained from digital elevation models (DEM), remote sensing data, and categorical or geomorphology maps (Krasilnikov et al., 2011). Furthermore, some soil physicochemical attributes, such as soil organic carbon (SOC) and pH, were also permissible to obtain as environmental covariates (Camera et al., 2017). Wang and Shi (2017) also recommended that the soil psf prediction should consider the ancillary data, which can enhance the performance of interpolation.

Different machine-learning methods, such as boosting regression trees (Jafari et al., 2014; Yang et al., 2016), random forests (Hengl et al., 2015; Zeraatpisheh et al., 2017) and artificial neural networks (Bagheri Bodaghabadi et al., 2015; Taalab et al., 2015), have been most commonly employed in DSM models for both regression and classification combined with environmental covariates in soil science. Hengl et al. (2015) contrasted the performance of spatial predictions of soil properties, such as soil psf, using random forests and linear regression, and the results demonstrated that the random forests were superior

operating characteristics; PRC, precision-recall curve; AUC, area under the ROC curve; AUPRC, area under the PRC; RMSE, root mean squared error; MAE, mean absolute error; R^2, coefficient of determination; MAD, median absolute deviation; AD, Aitchison distance; STRESS, standardized residual sum of squares; KNN_ALR, KNN_CLR, KNN_ILR, KNN_ORI, MLP_ALR, MLP_CLR, MLP_ILR, MLP_ORI, RF_ALR, RF_CLR, RF_ILR, RF_ORI, SVM_ALR, SVM_CLR, SVM_ILR, SVM_ORI, XGB_ALR, XGB_CLR, XGB_ILR, XGB_ORI, KNN, MLP, RF, SVM, XGB combined with ALR, CLR, ILR, ORI respectively; ClLo, clay loam; Lo, loam; LoSa, loamy sand; Sa, sand; SaClLo, sandy clay loam; SaLo, sandy loam; Si, silt; SiClLo, silty clay loam; SiLo, silt loam.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

to the linear regression with remarkable advantages of not only robust to noise but also low bias and variance. Hengl et al. (2017) improved the prediction of organic carbon, bulk density, pH and soil texture fractions on a global scale using machine-learning models – random forest, gradient boosting and multinomial logistic regression – indicating that random forest and gradient boosting outperformed linear models in large data sets. Taghizadeh-Mehrjardi et al. (2015) investigated the predictive power of soil classes using six machine learning-based classifiers and found that artificial neural network and decision trees performed better than any other models they mentioned with relatively high overall accuracies and kappa coefficients. Heung et al. (2016) evaluated a suite of 10 machine-learning models for predicting soil taxonomic units, and the consequences suggested that although the k-nearest neighbor and support vector machine had the highest accuracy, "tree learners" were preferred because of the interpretability of the results and the speed of parameterization. Most previous studies selected one or more machine-learning algorithms to simulate soil category or continuous variables for classification or regression problems. From this perspective, however, few studies systematically analyzed both soil texture classification and soil psf interpolation using multiple machine-learning methods.

The soil psf, which can be classified as soil texture, are not only continuous variables but also compositional data. We need to pay more attention to the latter case. Numerous different interpretations of the interpolation of compositional data in soil science have been suggested (Gobin et al., 2001; Lark and Bishop, 2007; Salazar et al., 2015), and the most extensively used were a combination of log ratio approaches involving the additive log ratio (ALR) and the centered log ratio (CLR) put forward by Aitchison (1982), as well as the isometric log ratio (ILR) from Egozcue et al. (2003). However, most studies using log ratio approaches to simulate the spatial variation of soil psf were kriging methods (so-called geostatistics), rather than machine-learning methods. Huang et al. (2014) combined multiple linear regression with ALR to improve the prediction precision of soil psf using electromagnetic data on a 1-m transect. Odeh et al. (2003) proposed that modified ALR ordinary kriging transcended compositional kriging and cokriging. Sun et al. (2014) contradistinguished compositional kriging, log ratio cokriging, cokriging, and ALR-cokriging, and produced proximate results. In contrast, Walvoort and de Gruijter (2001) thought compositional kriging had better performance than ALR ordinary kriging. Zhang et al. (2013) suggested compositional kriging was more appropriate for soil texture prediction than symmetry log ratio ordinary (or regression) kriging. Wang and Shi (2018) developed log ratio kriging combined with robust variogram estimation, which was preferable to compositional kriging methods. However, few studies combined log ratio with machine-learning models for soil psf interpolation in soil science. Aside from those mentioned above, the lack of systematic comparison of accuracy, strengths and weaknesses between original (untransformed) and log ratio approaches should be considered, especially in terms of combining with machine-learning methods.

Soil texture classification using machine-learning methods can be classified as a dependent variable; furthermore, it also can be derived indirectly from soil psf. Camera et al. (2017) reported that random forests were more remarkable than multinomial logistic regression in the direct soil texture classification. Wu et al. (2018) compared the support vector machines (SVM), artificial neural network (ANN), and classification tree (CT) models, demonstrating better prediction performance generated from SVM than from CT and ANN. For the indirect classification of soil texture, Poggio and Gimona (2017) combined hybrid geostatistical generalized additive models with ALR and modeled soil particle classes at medium resolution (250 m) in Scotland, expecting that vegetation index, morphological features and information about the phenological season were of vital significance as environmental covariates. Considering the particularity of compositional data, the consequences of soil psf classification and regression (indirect soil texture classification and soil psf interpolation, respectively) could be compared from the direct and indirect soil texture classification as a result of the relationship between soil texture and soil psf. Nevertheless, few studies systematically compared these using different machine-learning methods combined with original (untransformed) and log ratio transformed data for both direct and indirect soil texture classification.

In our study, five machine-learning models – k-nearest neighbor (KNN), multilayer perceptron neural network (MLP), random forest (RF), support vector machines (SVM), and extreme gradient boosting (XGB) – were included and applied for DSM of soil texture classification and soil psf interpolation. Furthermore, the original (untransformed) and log ratio transformed data were also combined with the machine-learning algorithms mentioned above for soil psf interpolation. Hence, the objectives of this study are (i) to compare different performance of five machine-learning models in the direct soil texture classification, (ii) to evaluate the accuracies of different log ratio approaches and original (untransformed) method applied for soil psf from the perspective of compositional data using machine-learning models, and (iii) to estimate whether the accuracies of indirect soil texture classification using original (untransformed) data and log ratio transformed data were improved compared with the direct soil texture classification.

2 Data and methods

2.1 Study area

The Heihe River Basin (HRB, 97 °6 '-102 °3 ' E, 37 °43 '- 42 °40 ' N) is situated in the Hexi Corridor, northwest of China, covering the Inner Mongolia Autonomous Region, Gansu and Qinghai provinces (Fig. 1a), which is the second largest inland river basin in China with an area of 146,700 km^2. The elevation and three reaches (i.e., upper, middle and lower) of the study area are shown in Fig. 1b. For the upper reaches of HRB, the climate changes significantly with altitude; the mean annual precipitation is 350 mm, the mean annual temperature is from -5-4 °C and the annual average evaporation is 1000 mm. For the middle reaches of HRB, the mean annual precipitation declines between 250 and 50 mm, the annual average evaporation increases from 2000 (east) to 4000 mm (west), and the mean annual temperature is from 2.8 to 7.6 °C. The lower reaches of HRB are situated in Ejina Banner on the Alxa Plateau, which is an arid desert climate with annual precipitation under 50 mm and annual average evaporation above 3500 mm; the mean annual temperature is from 8 to 10 °C.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

Hydrology and Earth System Sciences
Open Access
Discussions

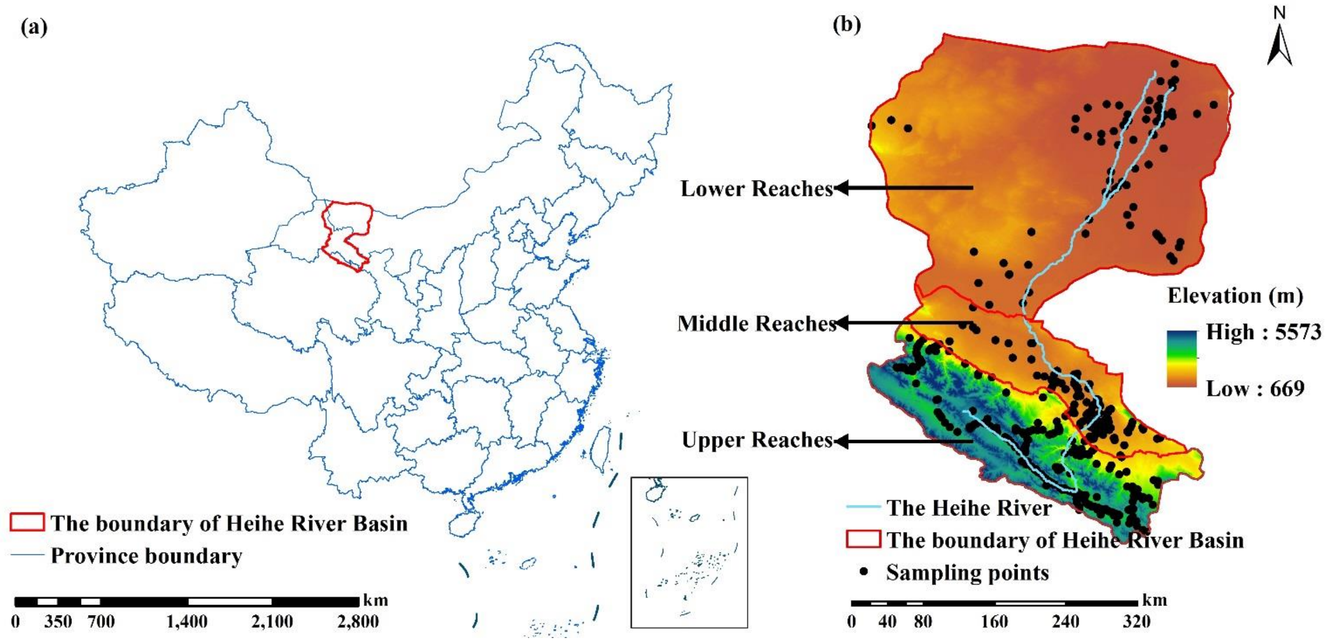


Figure 1. The (a) geographical location, (b) Heihe River, elevation and soil sampling points of Heihe River Basin, China.

The vegetation of the upper reaches of HRB is influenced from the southeast to northwest by hydrothermal conditions. The main vegetation types are alpine vegetation (4000-5000 m), alpine meadow vegetation belt (3000-4000 m), alpine shrub meadow (3200-3800 m), mountain forest meadow belt (2400-3200 m), mountain grassland belt (1800-2400 m), and desert base belt (less than 1800 m). The main vegetation types of the middle and lower reaches of the HRB are relatively fewer, including cultivated vegetation and desert, and the areas near the Heihe River on the lower reaches are shrub and steppe.

The main soil types are frigid desert soils (less than 4000 m), alpine meadow soil and alpine steppe soil (3600-4000 m), gray cinnamon soil and chernozem (3200-3600 m), sierozem and chestnut soil (2600-3200 m), chestnut soil (2300-2600 m) and sierozem (1900-2300 m) on the upper reaches of the HRB. The main soil types on the middle reaches of HRB are aeolian sandy soil, frigid frozen soil and gray brown desert soil. The main soil types in the lower reaches of HRB are aeolian sandy soil, gray brown desert soil (northwest) and lithosol (northeast).

The main types of geomorphology on the upper reaches of HRB are modern glaciers, alpine and hilly, and plimatic basins. Narrow plains are distributed on the middle reaches of HRB. For the lower reaches, the main types of geomorphology are hilly (northwest), plain, sandy land and platform (east), and the area near Heihe River is a flood plain.

2.2 Soil sampling

A total of 640 soil sampling points was collected in the HRB from the Science Data Center of Cold and Arid Regions (WestDC) in China (http://westdc.westgis.ac.cn/), involving 392 soil sampling points on the upper reaches and 248 soil sampling points on the middle and lower reaches of the HRB. The soil types, vegetation types, distribution of DEM and geomorphology types of the HRB were considered in soil sample collection according to the location and proportion of these types for the purpose of more representative spatial characteristics of soil psf using limited soil samples. There were more soil sampling points on the middle and upper reaches of HRB due to the more complicated soil types and vegetation types in these areas. In contrast, the types on the lower reaches are relatively similar with more desert in the northwest. Hence, the east of the lower reaches of the HRB contained more soil sampling points. All soil samples had information about soil psf (i.e., sand, silt and clay) and related environmental covariates using a laser diffraction approach and the extraction tool in ArcGIS, respectively, and the global position system (GPS) recorded the position information.

2.3 Environmental covariates and pre-processing

The environmental covariates, such as topographic attributes, remote sensing attributes, climate and position attributes, soil physicochemical attributes and categorical maps, are logically related to the distributions of soil psf. System for Automated Geoscientific Analysis (SAGA) GIS (Conrad et al., 2015) was used to compute their topographic attributes from DEM, including slope, aspect, convergence index, curvature, plane curvature, profile curvature and valley depth. Remote sensing attributes, including the normalized difference vegetation index (NDVI, Huete et al., 2002), the Brightness index (BI, Metternicht and Zinck, 2003), and the soil adjusted vegetation index (SAVI, Huete, 1988) were derived from the Landsat 7 based on band operation. We also collected climate attributes from the National Meteorological Information Center (NMIC, http://data.cma.cn/), such as the mean annual precipitation and the mean annual temperature. Latitude and longitude were also considered because of the large scale of the HRB. Mean annual surface evapotranspiration data (Wu et al., 2012) were gathered from WestDC (http://westdc.westgis.ac.cn/), as well as soil physicochemical attributes, such as soil organic carbon, saturated water content, field water holding capacity, wilt water content, saturated hydraulic conductivity, and soil thickness (Yi et al., 2015; Song et al., 2016; Yang et al., 2016), which can also address the distributions of soil psf. Additionally, the categorical maps were of significance, such as geomorphology types, soil types, land cover and vegetation types. For slope, the method of dividing the hierarchy rotates clockwise from the north (0 °), and each 45 ° was an interval, including north (337.5-22.5 °), northeast (22.5-67.5 °), east (67.5-112.5 °), southeast (112.5-167.5 °), south (167.5-202.5 °), southwest (202.5-247.5 °), west (247.5-292.5 °), and northwest (292.5-337.5 °).

2.4 Machine learning methods and parameters optimization

2.4.1 K-nearest neighbor (KNN)

K-nearest neighbor (KNN) is a simple non-parametric classifier based on known instance to label unknown instance (Cover and Hart, 1967). For the test set, k-nearest training set vectors were found, and maximum summed kernel densities were computed for classification. Moreover, continuous variables can also be predicted for regression with the average values of k-

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

nearest neighbors. Weighted KNN is an extended version of KNN that considers the distances of the nearest neighbors; therefore, the parameters of KNN contain the maximum value of k (kmax), the distances of the nearest neighbors (distance) and the types of kernel function (kernel). The KNN model is available in the R package "kknn" (Schliep and Hechenbichler, 2016).

2.4.2 Multilayer perceptron neural network (MLP)

Multilayer perceptron neural network (MLP), which is currently one of the most popular multilayer feed forward backpropagation networks, was selected to train artificial neural network (ANN) models in our study due to its rapid operation, small set of training requirements and ease of implementation (Subasi, 2007). MLP neurons can perform classification or regression depending on whether the response variable is categorical or continuous. The MLP has three sequential layers: input layer, hidden layer and output layer. The resilient backpropagation algorithm was chosen because the learning rate of this algorithm is adaptive, avoiding oscillations and accelerating the learning process (Behrens and Scholten, 2006). The range of the data set should be standardized because MLPs operate in terms of the scale 0 to 1. MLP can be run using the R package "RSNNS" (Bergmeir and Benitez, 2012).

2.4.3 Random forest (RF)

Random forest (RF) was developed by Breiman (2001), combining the bagging method (Breiman, 1996) with the random variable selection, and the principle was to merge a group of "weak learners" together to form a "strong learner". Bootstrap sampling is used for each tree of RF, and the rules to binary split data are different for regression and classification problems. For classification, the Gini index is used to split the data; for regression, minimizing the sum of the squares of the mean deviations can be selected to train each tree model. Benefits of using RFs are that the ensembles of trees are used without pruning. In addition, RF is relatively robust to overfitting, and standardization or normalization are not necessary because it is insensitive to the range of value. Two parameters should be adjusted for RF model: the number of trees (ntree) and the number of features randomly sampled at each split (mtry). The RF model is available in the R package "randomForest" (Liaw and Wiener, 2002).

2.4.4 Support vector machines (SVM)

The support vector machine (SVM), proposed by Cortes and Vapnik (1995), is a type of generalized linear classifier that is widely applied for classification and regression problems in soil science (Burges, 1998). The main principle of SVM is to classify different classes by constructing an optimal separating hyperplane in the feature space (so called "structural risk minimization"). Regression problems also can be solved by minimization of the structural risk using loss functions (Vapnik, 1998) in SVM, named support vector regression. The advantages of SVMs are that they are effective in high dimensional

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

spaces. Radial basis function was selected for SVM as the kernel function in our study, and two other parameters need to be tuned, i.e., cost and gamma, controlling the tradeoff between the classification accuracy and complexity, and the ranges of radial effect, respectively. The SVM model is available in the R package "e1071" (Meyer et al., 2017).

2.4.5 Extreme gradient boosting (XGB)

Extreme Gradient Boosting, put forward by Chen and Guestrin (2016), is an efficient method of implementation for gradient boosting frames, tree learning algorithms and efficient linear model solvers to solve both classification and regression problems (Chen et al., 2018). Like the boosted regression trees (Elith et al., 2008), it follows the principle of gradient enhancement; however, more regularized model formalization is applied to XGB to control over-fitting, making it more remarkable. In addition, parallel calculations can be automatically executed during the training phase of the XGB model, presenting a great advantage in large data sets, as the XGB can be more than ten times faster than the existing gradient boosting model (Chen and Guestrin, 2016). There are seven parameters should be tuned in XGB, containing the learning rate (eta), the maximum depth of a tree (max_depth), the max number of boosting iterations (nrounds), the subsample ratio of columns (colsample_bytree), the subsample ratio of the training instance (subsample), the minimum loss reduction (gamma) and the minimum sum of instance weight (min_child_weight). The XGB model is available in the R package "xgboost" (Chen et al., 2018).

2.4.6 Parameters optimization

The parameters of machine-learning models we mentioned above need to be adjusted, and the numbers of these parameters of models are different. For instance, XGB has seven parameters and is one of the most complicated models; on the other hand, for the MLP, in the case where we have chosen the algorithm, the only parameter that should be tuned is the size of the MLP model.

R package "caret" (Kuhn, 2018) provides an effective grid-search method that can automatically adjust the parameters by setting the adjustment grid, avoiding the uncertainty of artificial adjustment for some models (e.g., XGB) with more parameters. A set of parameters with the lowest RMSE or the highest R^2 for regression and the highest overall accuracy or kappa coefficient for classification by cross-validation can be selected to be the best parameters. However, in the presence of many adjustment parameters, it may be inefficient due to the long training time. Thus, we used the other package of "randomForest" for RF and "kknn" for KNN, which can also restructure the parameters for these two models.

In our study, eleven dependent variables (i.e., ten for regression and one for classification) were trained with environmental covariates (independent variables) for the sake of parameter adjustment for each model, including "sand, silt, clay, ilr1, ilr2, alr1, alr2, clr1, clr2, clr3" and "class". Subsequently, the parameters were definitely computed; here, we just give the relative ranges of the parameters after adjustment for most dependent variables; for example, in KNN the kmax was 15, the distance was 1, and the kernel was rectangular; in MLP, the size fluctuated between 5 and 10; in RF, the ntree was 1000 and mtry fluctuated from 9 to 11; in SVM, gamma was 0.01 and cost was 1; and in XGB, the range of parameters of max_depth (3-4), eta (0.05-0.1), colsample_bytree (0.6-0.8), nrounds (30), subsample (0.8-1), gamma (0-0.4), and min_child_weight (0.6-0.8) were obtained after conditioning.

2.5 Log-ratio transformation methods

For compositional data, the sum of the components is 1 (or 100 %), which should be guaranteed. Soil particle size data (i.e., sand, silt and clay), including three dimensions, are typical compositional data. The closed number system can be explained as follows: the individual variables in the data set are not independent of each other; moreover, they are related by being expressed as a percentage (Filzmoser et al., 2009). In the Euclidean space, one dimension (variable) would be omitted for the original method to guarantee no information loss because of the constant-sum constraint. Therefore, the Euclidean space is not appropriate for the analysis of soil psf data. The most widely used approaches are log ratio approaches (Aitchison, 1982), consisting of the additive log ratio, centered log ratio and isometric log ratio (ALR, CLR and ILR for short, respectively) from Aitchison (1982) and Egozcue et al. (2003).

For the composition of D elements $x = [x_1, \ldots, x_D]$, $x_j > 0$, $\forall j = 1, \ldots j-1, j+1, \ldots D$, and $\sum_{j=1}^{D} x_j = 1$, the transformation equation for ALR is defined as follows:

$$alr(x) = (ln\frac{x_1}{x_j}, \ldots, ln\frac{x_{j-1}}{x_j}, ln\frac{x_{j+1}}{x_j}, \ldots, ln\frac{x_D}{x_j}), \quad (1)$$

For soil psf ($D = 3$) in our study, the transformation equations for ALR are:

$$alr(1) = ln\frac{sand}{clay}, \quad (2)$$

$$alr(2) = ln\frac{silt}{clay}, \quad (3)$$

All of the information regarding the soil psf was contained in *alr(1)* and *alr(2)*; however, the ALR has been criticized because the choice of denominator is subjective, which can influence the results (Bacon-Shone, 2011). The CLR transformation method can remove this arbitrariness, and the equation is defined as follows:

$$clr(x) = (y_1, \ldots, y_j, \ldots, y_D) = (ln\frac{x_1}{\sqrt[D]{\prod_{j=1}^{D} x_j}}, \ldots, ln\frac{x_D}{\sqrt[D]{\prod_{j=1}^{D} x_j}}), \quad (4)$$

where y_j is the *j*th component. Similarly, for the soil psf, the transformation equations for CLR are:

$$clr(1) = ln\frac{sand}{\sqrt[3]{sand \times silt \times clay}}, \quad (5)$$

$$clr(2) = ln\frac{silt}{\sqrt[3]{sand \times silt \times clay}}, \quad (6)$$

$$clr(3) = ln\frac{clay}{\sqrt[3]{sand \times silt \times clay}}, \quad (7)$$

In the CLR approach, the geometric mean composed of all compositions of soil psf is the denominator, and one-to-one mapping of equations and soil psf could be implemented. Nevertheless, the CLR is inapplicable for multivariate analysis

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

because the sum of the dimensions of CLR is 0, and thus the results are collinear. These problems can be overcome by using ILR approach, which transforms all the information into D-1 orthogonal log contrasts (Abdi et al., 2015). The transformation equations for ILR are defined as follows:

$$z = (z_1, \ldots z_{D-1}) = ilr(x), \tag{8}$$

$$z_i = \sqrt{\frac{D-i}{D-i+1}} ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \tag{9}$$

where z_i is the *i*th component. The ILR transformation equations for soil psf in our study can also be defined as follows:

$$ilr(1) = \sqrt{\frac{2}{3}} ln \frac{sand}{\sqrt{silt \times clay}}, \tag{10}$$

$$ilr(2) = \sqrt{\frac{1}{2}} ln \frac{silt}{clay}, \tag{11}$$

For a more uniform comparison of the descriptive statistics, the ordering of three components of soil psf followed sand-silt-clay, and we added the third equation for the ALR and ILR. Although all the information could be included in the first two equations, note that in the process of interpolation, only the first two equations were used for ALR and ILR:

$$alr(3) = ln \frac{clay}{sand}, \tag{12}$$

$$ilr(3) = \sqrt{\frac{2}{3}} ln \frac{clay}{\sqrt{sand \times silt}}, \tag{13}$$

The equations for $alr(1), alr(2), alr(3)$ were equivalent to $alr(sand), alr(silt), alr(clay)$ in ALR, the same as in ILR. The back-transformed equations for ALR, CLR and ILR were recommended in our previous research (Wang and Shi, 2017), and were computed in the "compositions" R package (van den Boogaart and Tolosana-Delgado, 2008).

For the original (untransformed) method, the standardization function was used to ensure predictions of soil psf were between 0 and 100 and that their sum was 100%:

$$sand_s = \frac{sand}{(sand+silt+clay)} \times 100, \tag{14}$$

where, $sand_s$ is the content of sand after standardization, the same as silt and clay fractions.

2.6 Validation

2.6.1 Validation method

A total of 45 methods that we simulated are presented in Table 1; five machine-learning models were combined with one original (ORI) and three log ratio approaches (ALR, CLR, ILR). Five machine-learning methods were applied for the direct soil texture classification; additionally, these methods were combined with original (untransformed) and log ratio transformed data for a total of 40 methods for the indirect soil texture classification (20) and soil psf interpolation (20). The data were randomly divided into two sets to guarantee prediction accuracy and persuasion; for instance, one (70 % = 448 soil samples) was employed for training models and the other (30 % = 192 soil samples) was set aside for validation. This process was repeated 30 times for soil texture classification and soil psf interpolation, and different indicators were chosen to evaluate different performances of models (or methods).

Table 1. The method system of soil texture classification and soil psf interpolation.

Methods	Soil texture classification		Soil psf interpolation
	Direct classification	Indirect classification	—
Original data (ORI)	KNN, MLP, RF, SVM, XGB	KNN_ORI, MLP_ORI, RF_ORI, SVM_ORI, XGB_ORI	
Log-ratio transformed data (ALR, CLR, ILR)	—	KNN_ALR, KNN_CLR, KNN_ILR, MLP_ALR, MLP_CLR, MLP_ILR, RF_ALR, RF_CLR, RF_ILR, SVM_ALR, SVM_CLR, SVM_ILR, XGB_ALR, XGB_CLR, XGB_ILR,	

2.6.2 Validation indicators for soil texture classification

The overall accuracy (Brus et al., 2011) and kappa coefficient were selected to evaluate the overall effects of soil texture types predicted by different models. Moreover, the receiver operating characteristic (ROC) curve, precision-recall curve (PRC), area under the ROC curve (AUC), area under the precision-recall curve (AUPRC) and abundance index were applied to evaluate the performance of different soil texture types.

The overall accuracy represents all samples of soil texture types correctly classified by machine-learning models, divided by the total number of samples of soil texture types used in the validation. The higher overall accuracy, the more accurate soil map (Brus et al., 2011):

$$Overall\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{15}$$

where T, F, P and N denote True, False, Positive, and Negative and TP, TN, FP, FN were true positive, true negative, false positive, and false negative. When the numbers of samples in different classes are imbalanced in the data set, the kappa coefficient can explain the agreement of classes (Marchetti et al., 2011), which is calculated based on the confusion matrix, the equation is defined as:

$$kappa = \frac{p_o - p_e}{1 - p_e}, \tag{16}$$

where, p_o is the probability of observed agreement (overall accuracy) and p_e is the probability of agreement when two classes are unconditionally independent. The strength of the kappa coefficients is interpreted in the following manner: 0.01-0.20: slight, 0.21-0.40: fair, 0.41-0.60: moderate, 0.61-0.80: substantial, 0.81-1.00: almost perfect (Landis and Koch, 1977).

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

Hydrology and Earth System Sciences
Open Access
Discussions

The probabilities of different soil texture types (sum to 1) obtained during the training and predicting processes of machine-learning models were selected to calculate the sensitivity, specificity, precision and recall:

$$Sensitivity = recall = \frac{TP}{TP+FN}, \tag{17}$$

$$Specificity = \frac{TN}{TN+FP}, \tag{18}$$

$$Precison = \frac{TP}{TP+FP}, \tag{19}$$

In general, sensitivity, precision and recall indicate the extent of identifying positive cases, and specificity demonstrates the extent of identifying the negative cases of models. ROC analysis is commonly used in two-class problems. However, soil texture types are more than two classes. In our point of view, a one-vs-rest strategy was employed to produce different ROC graphs for each soil texture type.

$$P_i = c_i, \tag{20}$$

$$N_i = \cup j \neq ic_j \in C, \tag{21}$$

where C is the set including all classes, P_i is the positive class, N_i is the negative class, including all classes except c_i in ROC graph i (Fawcett, 2006).

In practice, the weakness of the ROC curve is that it cannot indicate the differences among the models in the cases of imbalanced samples between positive and negative. Soil texture data are a class-imbalanced data set of positive and negative, and the negative classifier would be overvalued under these circumstances because of the overabundance of majority (negative) examples, additionally revealing overly optimistic findings (Davis and Goadrich, 2006). However, precision and recall curves (PRC) are more informative than ROC curves in dealing with class-imbalanced data (Fu et al., 2017). The R package "precrec" (Saito and Rehmsmeier, 2017) generated ROC and PRC curves and computed AUC and AUPRC for each soil texture type. This process was repeated 30 times and eventually, the average ROC and PRC curves with their average areas under these curves were obtained.

Abundance index was applied to describe the proportion of all soil texture types and correctly predicted soil texture types in the prediction map, which was defined as follows:

$$Abundance\ index = p/t, \tag{22}$$

where p is correctly predicted soil texture type(s) and t is all soil texture types in the prediction map. For the sake of ensuring the balance of the soil texture types, all nine soil texture types were involved in test sets, covering clay loam (ClLo: 12), loam (Lo: 57), loamy sand (LoSa: 18), sand (Sa: 23), sandy clay loam (SaClLo: 4), sandy loam (SaLo: 58), silt (Si: 31), silty clay loam (SiClLo: 37), and silt loam (SiLo: 400); most were SiLo (62.5%) and the fewest were SaClLo (0.63%).

2.6.3 Validation indicators for soil psf interpolation

The accuracy and performance of machine-learning models mentioned above for the original (untransformed) and different log ratio transformation approaches were evaluated using five statistical indicators, containing coefficient of determination

(R^2), root mean square error (RMSE), mean absolute error (MAE), Aitchison distance (AD, Aitchison, 1992), and standardized residual sum of squares (STRESS, Martin-Fernandez et al., 2001). The equations for the validation indicators R^2, RMSE, MAE, AD and STRESS are as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(Y_{i,m}-Y_{i,e})}{\sum_{i=1}^{n}(Y_{i,m}-\bar{Y}_{i,m})}, \tag{23}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{i,m}-Y_{i,e})^2}, \tag{24}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_{i,m}-Y_{i,e}\right|, \tag{25}$$

where $Y_{i,m}, Y_{i,e}, \bar{Y}_{i,m}$ and n are the measured, predicted and the mean of measured soil psf and the number of observations (soil sampling points for validation). Closer to 1 and higher values of R^2 and the lower values of RMSE and MAE show better performance of models and methods.

$$AD = \left[\sum_{i=1}^{D}\left[log\frac{x_i}{g(x)} - log\frac{X_i}{g(X)}\right]^2\right]^{1/2}, \tag{26}$$

$$STRESS = \left[\frac{\sum_{i<j}(AD_{x,ij}-AD_{X,ij})^2}{\sum_{i<j}(AD_{x,ij})^2}\right]^{1/2}, \tag{27}$$

where x is the observed value; X is the predicted value; D is the number of dimensions (for soil psf is 3); $g(x)$ denotes the geometric mean $(x_1 \ldots x_D)^{1/D}$; $AD_{x,ij}$ and $AD_{X,ij}$ are the ADs between the observed soil psf and the predicted soil psf at sites i and j. Both present that model performances are better when the values are lower.

2.6.4 Indirect soil texture classification by soil psf interpolation

Seventy percent of the 640 soil sampling points were used for training each machine-learning model, and the remaining 30 % were used for the soil psf interpolation; thereafter, we transformed the content of three components (sand, silt and clay) into the soil texture types in the USDA soil texture classification triangles using the R package "soiltexture" (Moeys, 2018). Eventually, the overall accuracy and kappa coefficient were computed and evaluated. This process was repeated 30 times, and the averages of these consequences were employed to compare the classification performance of each model. The direct and indirect soil texture classifications were also compared with the overall accuracy and kappa coefficient. The training and testing sets for each time were the same by setting seeds, and all calculations and analysis were performed with the freely available software R (R Core Team, 2018).

2.7 Statistical analysis for the original and log ratio transformed data

The mean, median, minimum (Min), maximum (Max), median absolute deviation (MAD), skewness (Skew), kurtosis and Kolmogorov-Smirnov test (p>0.05) were employed for descriptive statistical analysis of the original (untransformed) and log ratio transformed soil psf data. The arithmetic mean of log-ratio transformation data should be back-transformed to the real space. For $X = [X_1, \ldots, X_n]$, the MAD can be calculated according to the Eq. (28) as below:

$$MAD(X) = median(|X_i - median(X)|). \tag{28}$$

3 Results

3.1 The descriptive statistics for the original and log ratio transformed soil psf data

With respect to the original (untransformed) data of sand, the mean fraction (30.64 %) was much higher than that of median fraction (25.10 %); conversely, both silt and clay were the opposite, with lower mean fractions (silt: 55.79 %, clay: 13.57 %) than median fractions (silt: 59.47 %, clay: 13.78 %). For the log ratio transformed data, the means of sand (28.69 %) and silt (60.54 %) were closer to the median values of the original data, aside from clay, with mean of 10.78 %.

All MADs of log ratio transformed data were much smaller than those of the original data in all cases; for instance, ILR contained the best value of MAD for sand (0.66) and clay (0.44), and CLR generated the lowest MAD for silt (0.43) among different log ratio approaches (Fig. 2). All log ratio approaches had lower skews (ALR: 0.77, CLR: 0.88, ILR: -1.20) than those of the original data (1.24) for sand. Moreover, CLR (-0.4) declined the original skew (-0.93) for silt. However, it was negligible for log ratio transformation data compared with the original skew of clay (0.4). The kurtosis of all log ratio approaches was much higher compared with the consequences generated from original (untransformed) data. In terms of the k-s test ($p < 0.05$), although the p values of the original (untransformed) and different log ratio transformed data were not significant and all histograms were not subject to normal distribution, log ratios made the images of the data more symmetric (Fig. 2).

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

Hydrology and Earth System Sciences
Open Access
Discussions

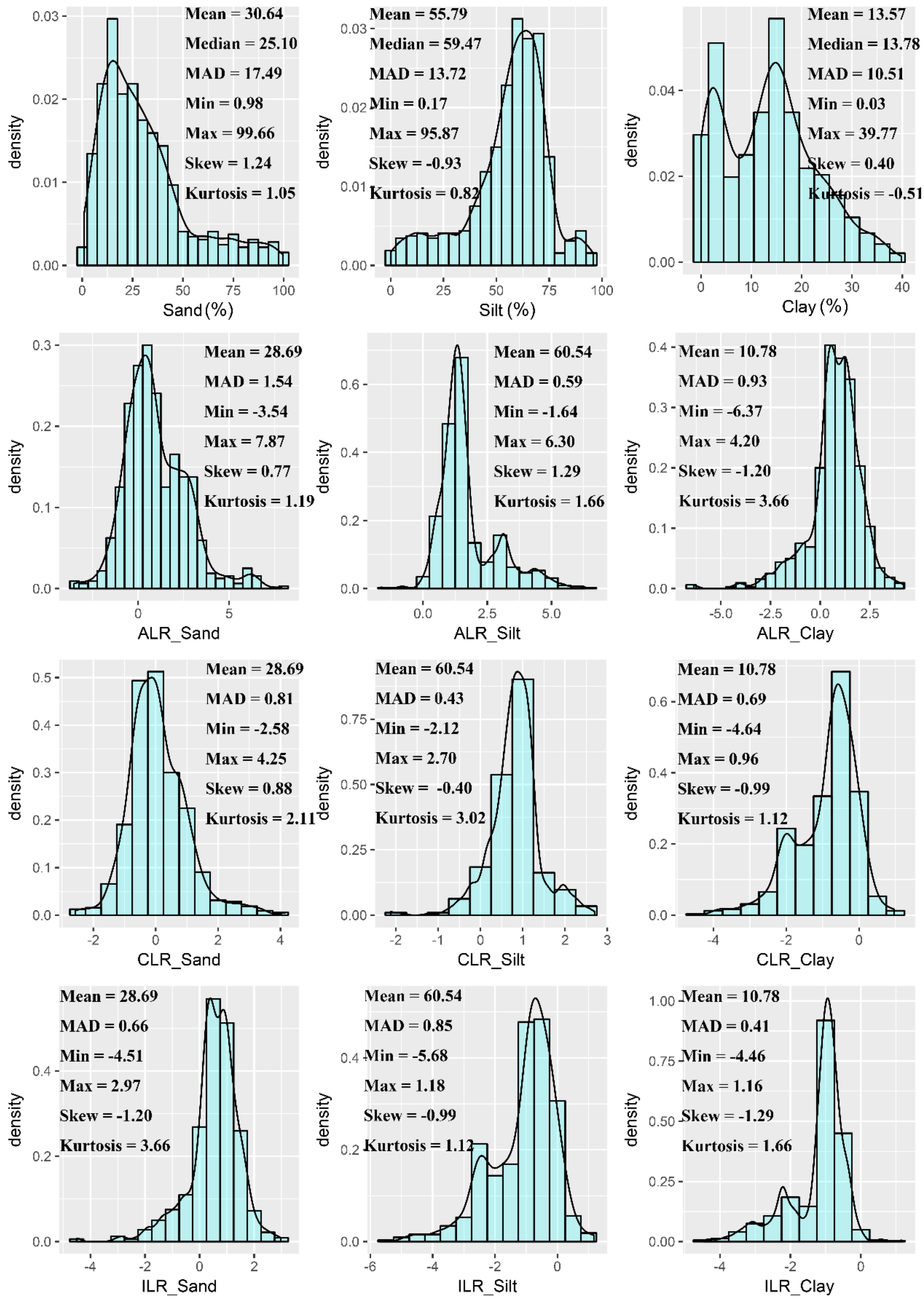


Figure 2. Descriptive statistical analysis for the original (untransformed) and logratio transformed soil sampling data. Note that the means generated from log ratio transformed data were back-transformed to the real space.

3.2 Comparison of the machine learning models in the classification of soil texture types

3.2.1 Comparison of the validation indicators for soil texture classification

The overall accuracy of each model ranged from 0.610 to 0.647 (Fig. 3a). SVM had the highest overall accuracy (0.647) among the five models, followed closely by the accuracies of KNN (0.631) and RF (0.629). XGB (0.611) and MLP (0.610) were relatively lower among these models. The highest kappa coefficient was generated from XGB (0.240), followed by RF (0.238), KNN (0.234) and MLP (0.230), and the worst performer was SVM, with kappa coefficient dropping to 0.186 (Fig. 3b).

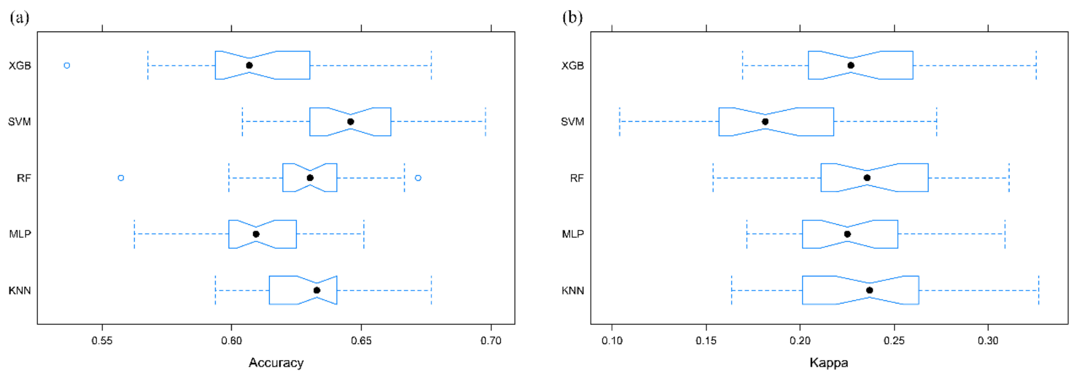


Figure 3. (a) The overall accuracies and (b) kappa coefficients for different machine learning models of KNN, MLP, RF, SVM and XGB.

The AUC with regard to each soil texture type of 640 soil sampling points predicted from five different models demonstrated that the ranking of the AUC was RF>XGB>SVM>KNN>MLP in the case of fewer soil sampling points (ClLo, LoSa, Sa, SaClLo and Si). However, in the case of the types with more soil sampling points (Lo, SaLo, SiLo, SiClLo), the ROC curves exhibited roughly the same shape for each model (Fig. 4); therefore, the order of performance was as follows: RF>SVM>XGB>MLP>KNN.

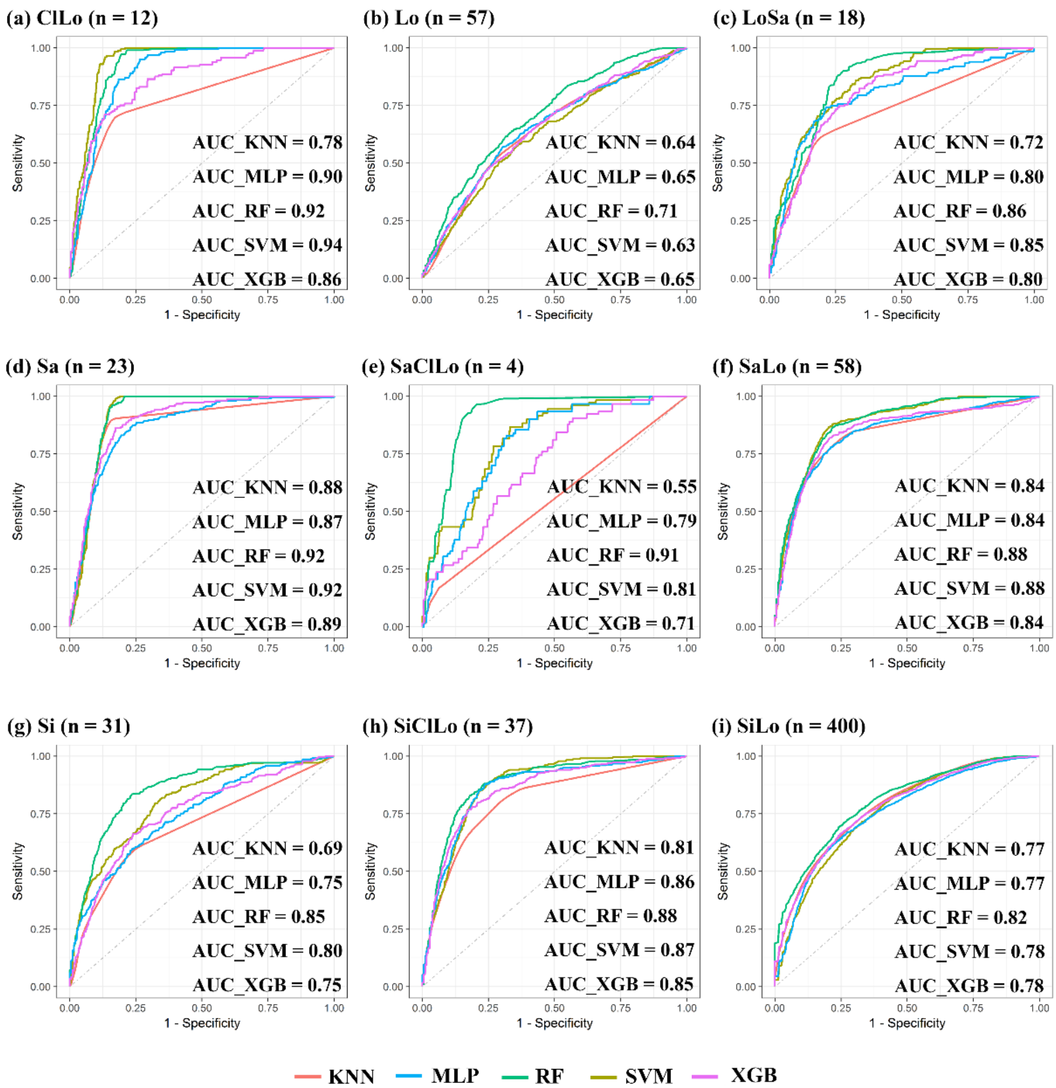


Figure 4. The AUC for different machine-learning methods of each soil texture type (a) ClLo (b) Lo (c) LoSa (d) Sa (e) SaClLo (f) SaLo (g) Si (h) SiClLo (i) SiLo; n was the sampling points for different soil texture types.

We combined the PRCs with five machine-learning methods to evaluate the performance of these models with respect to predicting each soil texture type using soil psf imbalanced data with different samples of soil texture types (Fig. 5). We found that the AUPRC of types with fewer positive examples were typically small, especially in the case of SaClLo (only four samples), which resulted in unsatisfying consequences because the lack of soil sampling points made models learn poorly during the training process. Hence, the soil texture types (Lo, SaLo, SiLo, SiClLo) with more positive examples delivered superior results to those with fewer positive examples. Moreover, these soil texture types had significant differences in AUPRCs. For example, SiLo, which had the largest number of samples, was the most effective among these nine types. The total AUPRC calculated by the weights of samples for AUPRC of each type was applied to evaluate the effect of each model, and the order was as follows: RF (0.646)>XGB (0.616)>KNN (0.601)>MLP (0.600)>SVM (0.599).

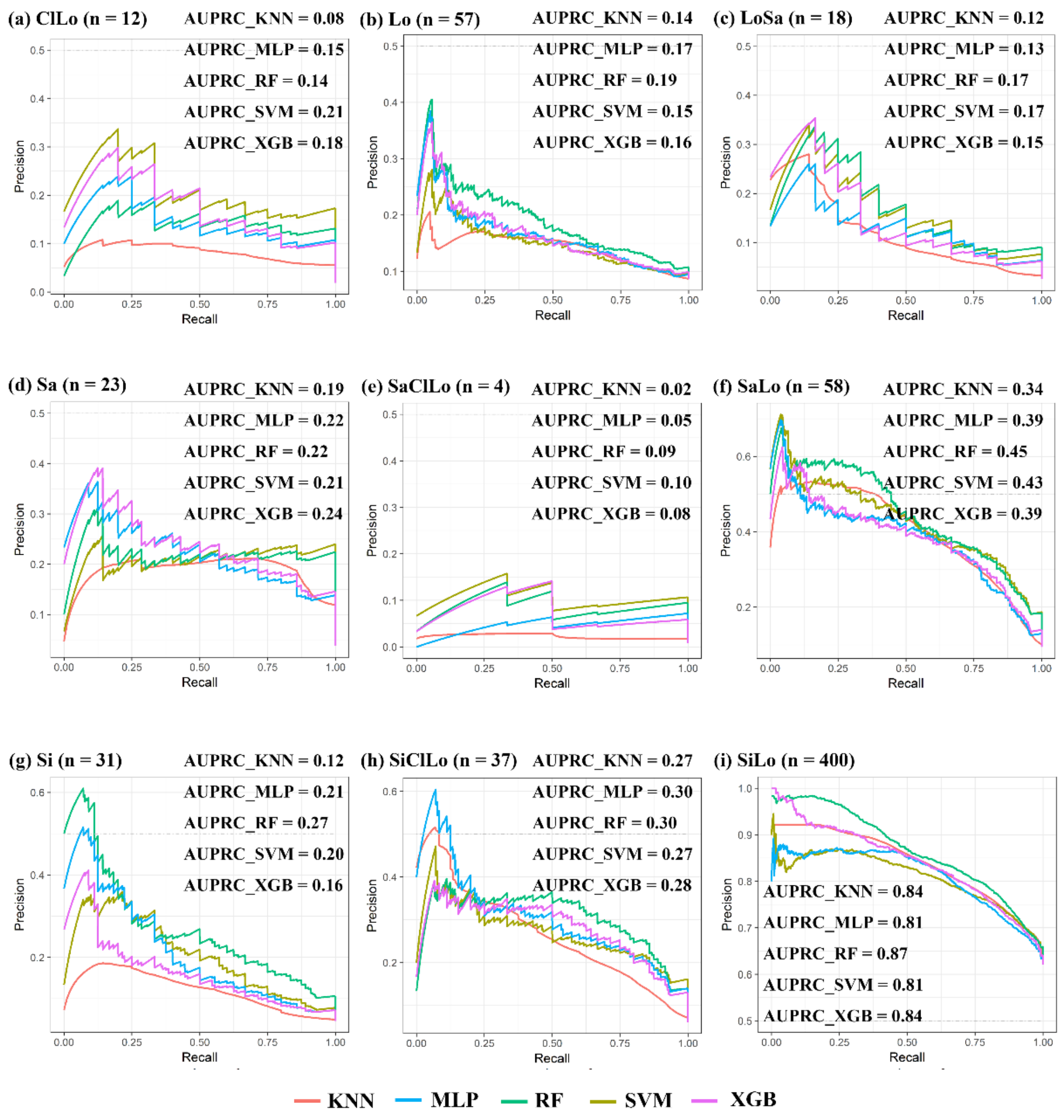


Figure 5. The AUPRC for different machine-learning methods of each soil texture type (a) ClLo (b) Lo (c) LoSa (d) Sa (e) SaClLo (f) SaLo (g) Si (h) SiClLo (i) SiLo; n was the sampling points for different soil texture types.

3.2.2 Comparison of the prediction maps for soil texture classification

Prediction maps of soil texture types in the HRB using machine-learning models delivered quite different spatial distributions in the overall performance of different models (Fig. 6). The abundance indices pointed out that all models could not predict the type of SaClLo; in other words, KNN and XGB predicted 8 of 9 types, followed closely by RF (7 of 9 types) and MLP (6 of 9 types). However, SVM predicted only two types, which was an unsatisfactory result associated with the lowest kappa coefficient (Fig. 3). Additionally, the prediction effects of different models were different in the distributions of soil texture types in the HRB. The consequences of RF and XGB illustrated that the main soil texture types in the northwest of the lower reaches of HRB were mostly LoSa, while other prediction models produced SaLo. On the upper reaches of the HRB, soil texture types generated from RF were more abundant and more in accordance with the real environment.

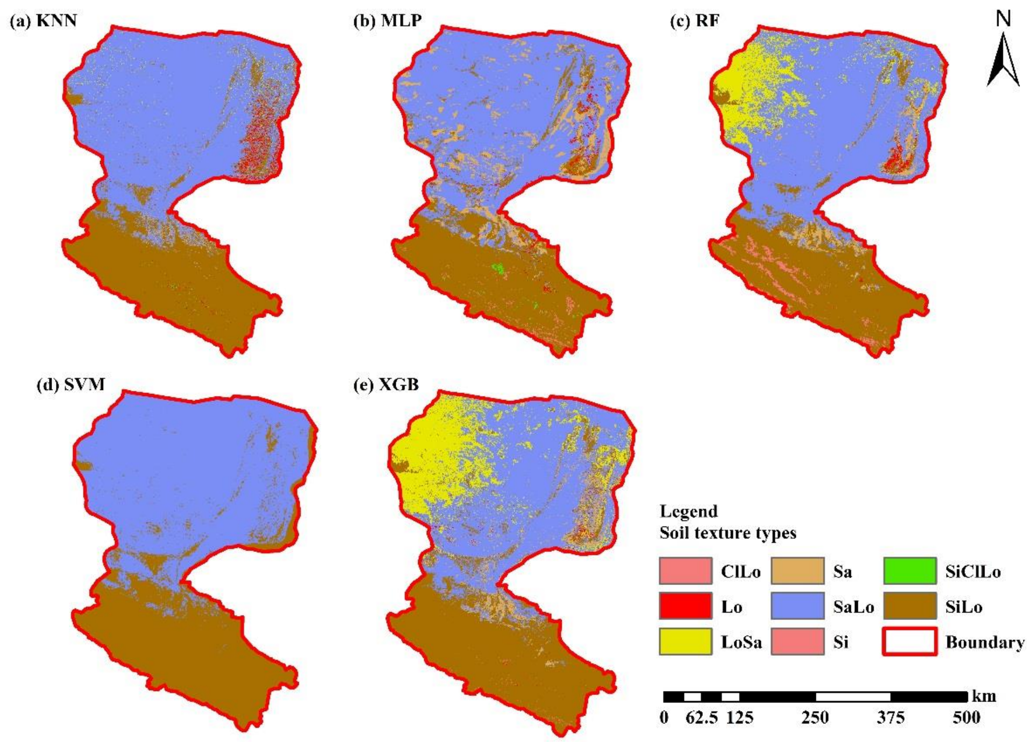


Figure 6. Soil texture classification prediction map of different soil texture types of (a) KNN, (b) MLP, (c) RF, (d) SVM and (e) XGB.

3.3 Comparison of the machine-learning models combined with log ratio approaches in the interpolation of soil psf

3.3.1 Comparison of the validation indicators for interpolation of soil psf

We compared the performance of each machine-learning model combined with the original (untransformed) and the log ratio transformed data of soil psf. The results indicated that the accuracies of STRESS of the methods combined with log ratio transformed data were superior to other approaches using original (untransformed) data (Table 2). With respect to KNN, MLP, RF and XGB, values of RMSE, MAE, R^2 and AD generated from original (untransformed) data outperformed log ratio transformed data; for SVM, log ratio transformed data delivered superior improvement. For instance, SVM_CLR and SVM_ILR had higher R^2 and lower RMSE and MAE than SVM_ORI of sand, silt and clay.

By comparison among different log ratio approaches of the same machine-learning model, ILR and CLR outperformed ALR in these models, other than MLP, showing a slight difference. As shown in Table 2, KNN_CLR demonstrated the most remarkable performance among the three KNN models using different log ratio transformed data with highest R^2 (sand: 48.48 %; silt: 38.37 %; clay: 41.43 %) and lowest RMSE (sand: 15.82 %; silt: 14.77 %; clay: 7.09 %) and MAE (sand: 11.21 %; silt: 10.74 %; clay: 5.58 %). Furthermore, CLR and ILR generated relatively similar consequences for each model of RF and SVM; with respect to XGB, XGB_ILR showed the best performance with all indicators we measured, aside from RMSE (6.75 %) and MAE (5.36 %) of clay, and STRESS (0.63).

We also compared five different machine-learning models using the same log ratio approaches. In the case of ALR, ALR_RF had talent, with the lowest RMSE (sand: 15.50 %; silt: 14.43 %; clay: 6.62 %) and MAE (sand: 10.90 %; silt: 10.52 %; clay: 5.24 %), the highest R^2 (sand: 50.57 %; silt: 41.23 %; clay: 48.90 %), and the lowest AD (0.86) and STRESS (0.61), followed by SVM_ALR, XGB_ALR, KNN_ALR and MLP_ALR. Regarding CLR and ILR, RF also produced the most preferable performance followed by SVM, XGB, KNN and MLP. For ORI approach, RF outperformed other models in accordance with log ratio approaches, and the next were XGB, SVM, KNN and MLP. Therefore, it is clear that RFs demonstrated the most extraordinary indicators of RMSE, MAE, R^2 and AD from the ORI model and the best STRESS from the log ratio models (RF_ALR, RF_CLR and RF_ILR).

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

Table 2. The comparisons of accuracies of different machine-learning models combined with original (untransformed) and transformed data.

	RMSE (%)			MAE (%)			R^2 (%)			AD	STRESS
	Sand	Silt	Clay	Sand	Silt	Clay	Sand	Silt	Clay		
KNN_ALR	16.05	15.04	7.12	11.35	10.93	5.59	47.02	36.11	41.07	0.90	0.62
KNN_CLR	15.82	14.77	7.09	11.21	10.74	5.58	48.48	38.37	41.43	0.88	0.62
KNN_ILR	15.82	14.82	7.14	11.22	10.84	5.60	48.46	37.88	40.74	0.88	0.64
KNN_ORI	15.51	14.47	7.05	11.12	10.51	5.49	50.59	40.92	42.24	0.84	0.66
MLP_ALR	15.83	15.07	7.43	11.42	11.06	5.97	48.50	35.82	35.79	0.92	0.66
MLP_CLR	15.84	15.07	7.41	11.45	11.05	5.96	48.42	35.86	36.19	0.92	0.66
MLP_ILR	15.84	15.07	7.40	11.46	11.04	5.95	48.40	35.85	36.32	0.92	0.66
MLP_ORI	15.80	14.72	6.96	11.50	10.85	5.52	48.75	38.84	43.72	0.90	0.68
RF_ALR	15.50	14.43	6.62	10.90	10.52	5.24	50.57	41.23	48.90	0.86	0.61
RF_CLR	15.28	14.22	6.61	10.70	10.25	5.21	51.95	42.89	49.16	0.86	0.61
RF_ILR	15.27	14.25	6.66	10.66	10.26	5.26	51.99	42.60	48.28	0.86	0.61
RF_ORI	15.09	13.86	6.31	10.65	9.99	5.00	53.28	45.77	53.75	0.84	0.66
SVM_ALR	15.66	14.59	6.76	11.66	10.88	5.34	49.61	39.87	46.89	0.88	0.66
SVM_CLR	15.27	14.36	6.87	11.01	10.41	5.41	52.12	41.85	45.14	0.87	0.65
SVM_ILR	15.29	14.37	6.84	10.92	10.43	5.42	51.99	41.69	45.58	0.87	0.65
SVM_ORI	15.30	14.38	6.92	10.94	10.32	5.43	51.98	41.71	44.45	0.87	0.67
XGB_ALR	15.82	14.92	6.72	11.32	11.01	5.35	48.57	37.23	47.44	0.88	0.64
XGB_CLR	15.70	14.80	6.75	10.96	10.67	5.39	49.23	38.10	46.90	0.88	0.62
XGB_ILR	15.45	14.57	6.75	10.91	10.52	5.36	50.88	40.01	47.01	0.88	0.63
XGB_ORI	15.15	14.05	6.47	10.88	10.15	5.15	52.85	44.27	51.36	0.86	0.68

3.3.2 Comparison of the interpolation maps of soil psf

Interpolation maps of soil psf (sand, silt and clay) using log ratio transformed data (ILR) and original (untransformed) data were represented in Figs. 7, S1 and S2. At first glance, there was a negligible difference between ILR and ORI based on the same machine-learning model. However, the maps generated from models combined with ILR transformed data showed closer ranges to the original soil sampling data in the case of sand (0.98-99.66 %), silt (0.17-95.87 %) and clay (0.03-39.77 %), and the texture features were more suitable for the distributions of the real environment (Figs. 7, S1 and S2). With respect to different machine-learning models, RF and XGB delivered more detailed information about texture features in prediction maps than did KNN, SVM and MLP.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

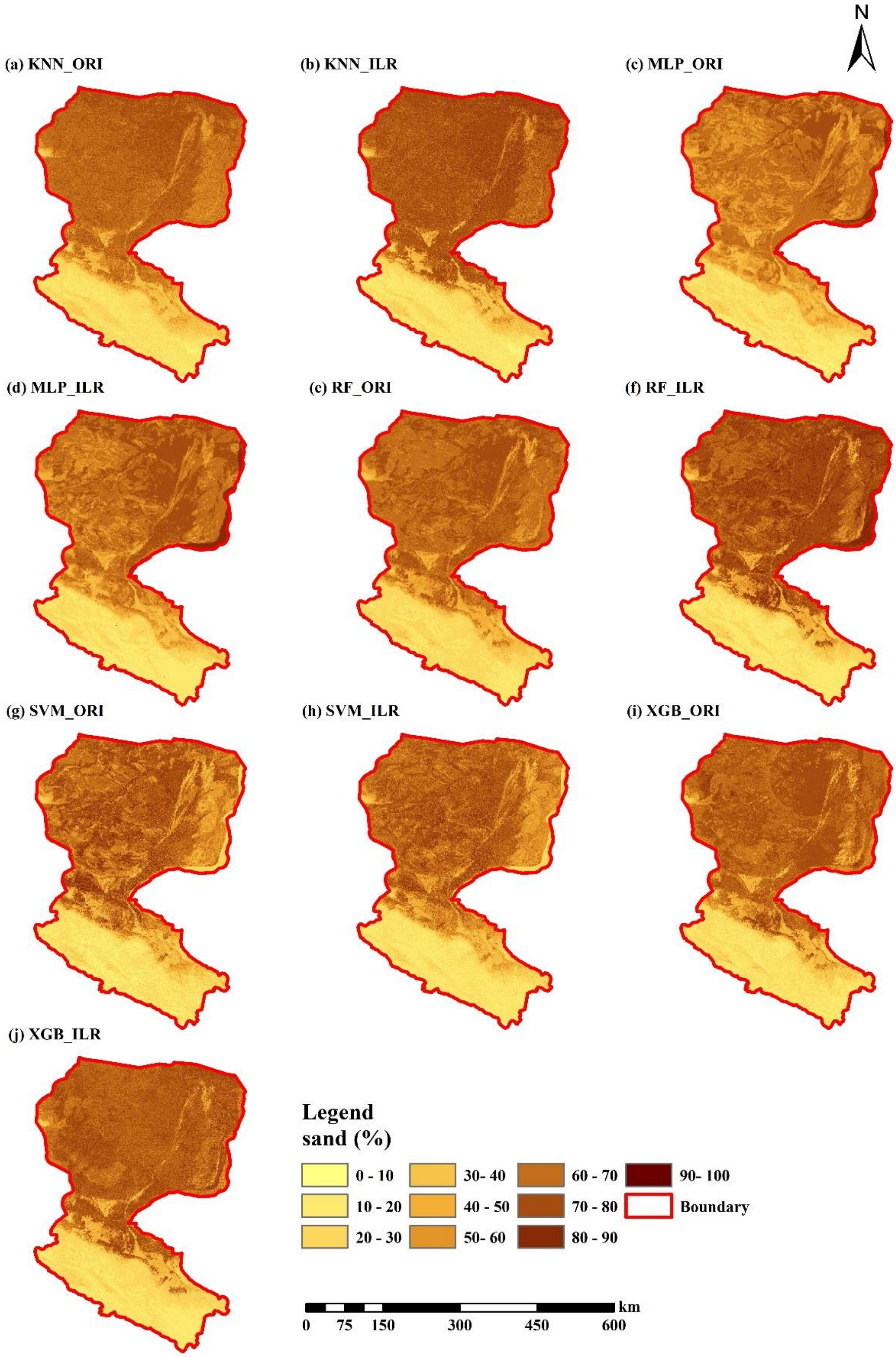


Figure 7. The interpolation maps of sand fraction.

3.4 Comparison of direct and indirect soil texture classification

3.4.1 Comparison of the validation indicators for direct and indirect soil texture classification

Compared with the classification performance of the five machine-learning models using original (untransformed) data, the overall accuracies and kappa coefficients of models using log ratio transformed data were improved, especially RF and XGB, which combined with all three log ratio approaches were superior to the ORI approach. Table 3 shows that the overall accuracy (0.631) and kappa coefficient (0.245) of the original method in KNN models were better than any other log ratio approach. In summary, the ILR transformation method of five machine-learning models showed the highest overall accuracy among three log ratio transformation approaches (KNN: 0.628; MLP: 0.614; RF: 0.631; SVM: 0.631; XGB: 0.632), which also demonstrated the best performance with regard to kappa coefficients (KNN: 0.244; RF: 0.291; SVM: 0.239; XGB: 0.252), except for MLP (ALR: 0.216; CLR: 0.216; ILR: 0.214). We also compared direct classification (Fig. 3) with indirect classification and found that the highest values of overall accuracy of indirect classification (KNN: 0.631; MLP: 0.614; RF: 0.628; SVM: 0.638; XGB: 0.632) were slightly decreased in comparison of the direct classification (KNN: 0.631; MLP: 0.610; RF: 0.629; SVM: 0.647; XGB: 0.611) for RF and SVM, and improved or kept stable for MLP and XGB, and KNN, respectively. In turn, the kappa coefficients were greatly modified using indirect classification (KNN: 0.245; MLP: 0.216; RF: 0.291; SVM: 0.239; XGB: 0.252) compared with direct classification (KNN: 0.234; MLP: 0.230; RF: 0.238; SVM: 0.186; XGB: 0.240), other than MLP; peculiarly, RF_ILR increased the kappa coefficient to 0.291 (21.3 % improvement) while keeping accuracy stable, which showed the highest kappa coefficient among these methods.

Table 3. Overall accuracies and kappa coefficients calculated from soil texture classification by the interpolated maps from five models using original (untransformed) data and log ratio transformed data.

Methods	Overall accuracy	Kappa coefficient
KNN_ALR	0.623	0.236
KNN_CLR	0.627	0.241
KNN_ILR	0.628	0.244
KNN_ORI	0.631	0.245
MLP_ALR	0.614	0.216
MLP_CLR	0.614	0.216
MLP_ILR	0.614	0.214
MLP_ORI	0.611	0.216
RF_ALR	0.619	0.284
RF_CLR	0.625	0.276

RF_ILR	0.628	0.291
RF_ORI	0.619	0.279
SVM_ALR	0.591	0.205
SVM_CLR	0.630	0.227
SVM_ILR	0.631	0.239
SVM_ORI	0.638	0.232
XGB_ALR	0.610	0.226
XGB_CLR	0.612	0.240
XGB_ILR	0.632	0.252
XGB_ORI	0.619	0.239

3.4.2 The prediction performance of soil texture types from different methods

The distributions of soil texture classes using original (untransformed) data and ILR transformed data are illustrated in the USDA soil texture triangle (Fig. 8). The triangle of the original soil psf data (Fig. 8a) shows wider ranges of spatial dispersion than the interpolation data using machine-learning models, revealing the properties of aggregate from the sides to the center of triangles. With respect to these machine-learning models, RF showed the most dispersed feature in accordance with the original soil psf data. The distributions predicted from models combined with ILR transformed data were more discrete and more associated with the original soil psf data than those resulting from ORI approaches. The results of prediction represented striking differences in that the error ratio (red color) of soil sampling points on types of LoSa, SaLo and Lo (left side of triangles) were significantly more than those on types of SiLo and Si (the right side of triangles) for most models, especially KNN and MLP. The log ratio approaches overestimated the content of silt in the process of transformation (Fig. 2); in this way, these points were biased to the right of the USDA soil texture triangle based on overall contraction (regression smoothing effects), crossing the classification boundary and becoming other soil texture types. RF_ILR (Fig. 8f) delivered the highest right ratio (RR) among these models, and the classification accuracy was enhanced using the ILR approach (83.9%) compared with the ORI approach (81.7%). In the case of other models, the differences between the ORI and log ratio approaches were negligible. We also compared the RRs of indirect classification models with those of direct classification, demonstrating all RRs of direct classification were higher (KNN: 67.97 %; MLP: 75.16 %; RF: 100 %; SVM: 66.09 %; XGB: 81.09 %), especially RF and XGB; however, we removed this evaluation indicator because the same data sets were employed in the processes of training and predicting.

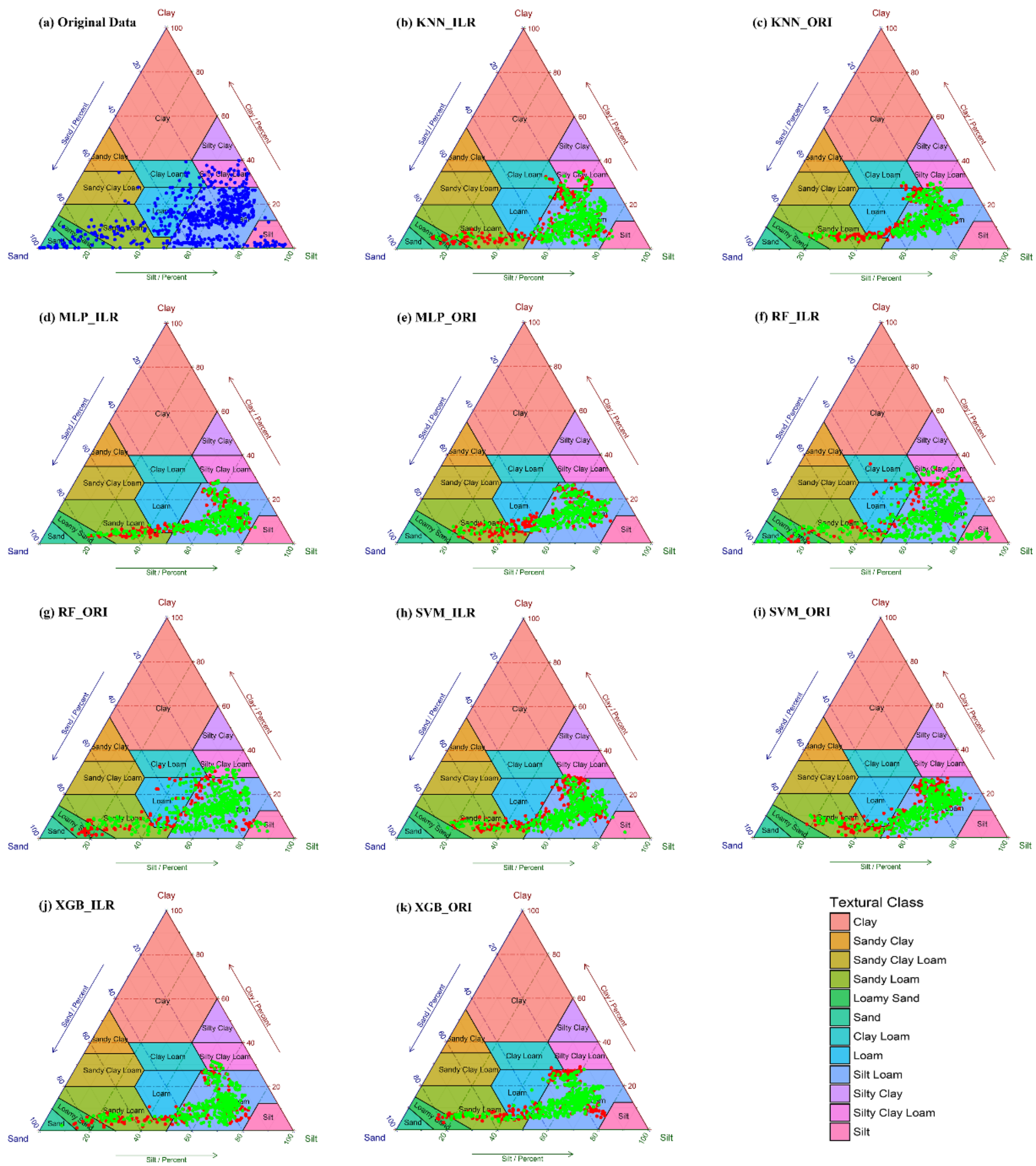


Figure 8. Soil texture types of 640 soil samples shown in USDA texture triangle. The results of soil psf were generated from (a) original (untransformed) data, (b) KNN_ILR (65.0 %), (c) KNN_ORI (65.9 %), (d) MLP_ILR (63.3 %), (e) MLP_ORI (63.6 %), (f) RF_ILR (83.9 %), (g) RF_ORI (81.7 %), (h) SVM_ILR (66.1 %), (i) SVM_ORI (66.4 %), (j) XGB_ILR (67.8 %), and (k) XGB_ORI (68.0 %). Note that the predicted right-ratios (RRs) of the soil texture types were in the bracket after interpolators

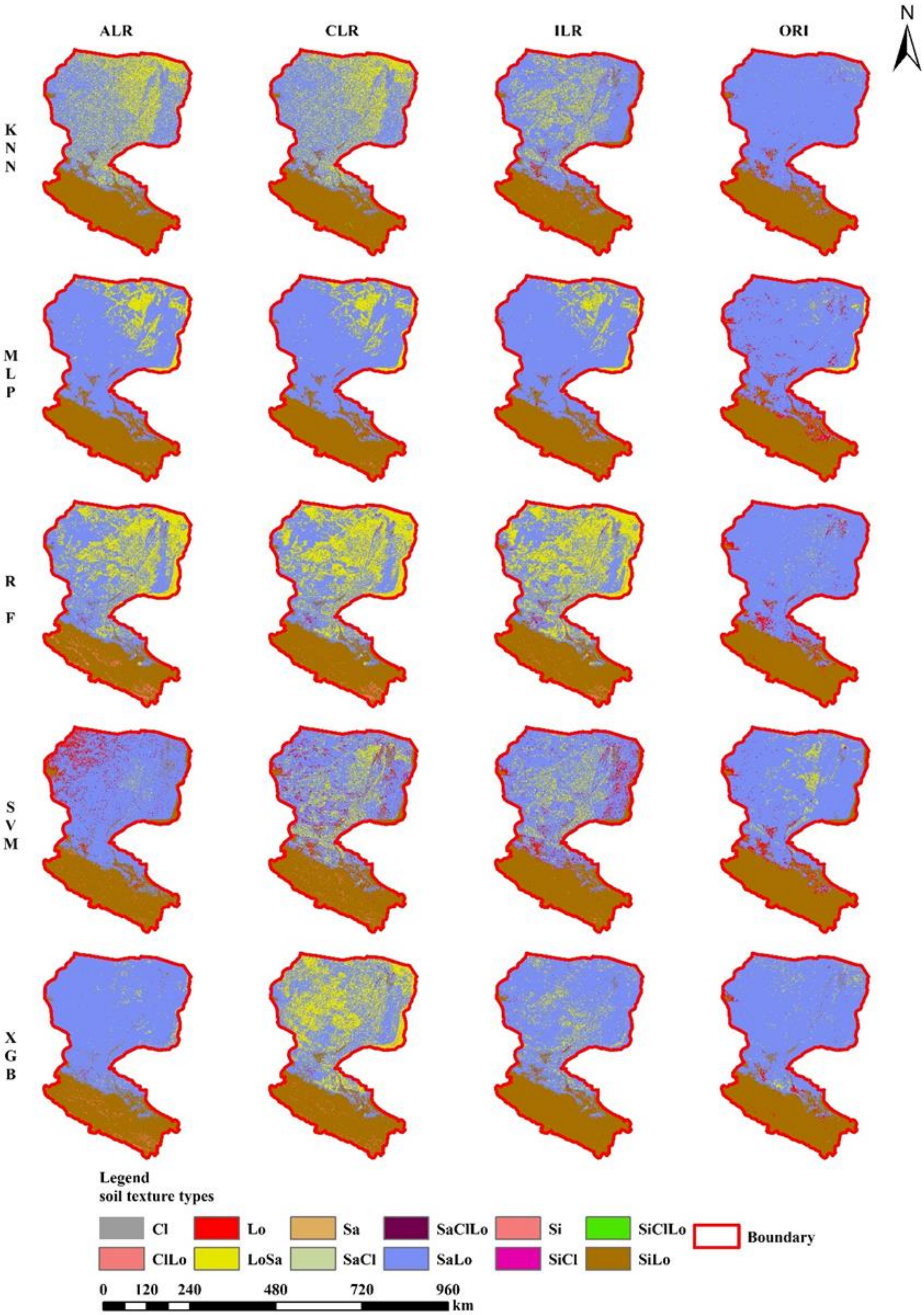
Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

3.4.3 Comparison of prediction maps of direct and indirect soil texture classification

Fig. 9 shows the similarity of the three log-ratio approaches. The soil texture maps predicted using original (untransformed) data is different from those generated by log ratio transformed data, and the classification maps from the machine learning models combined the log ratio transformed data had more detailed information. Three log-ratio transformation methods of the same machine learning model are similar in the number of each type predicted; however, there are some differences between methods using original (untransformed) data and those using log-ratio transformed data. All machine learning models combined with original data predicted more types of Lo and SaLo, and less types of LoSa and Si, which could also be presented in Fig. 9. The performance of different machine-learning models, especially in the lower reaches of the HRB was also compared, for log ratio approaches, for KNN, KNN_ALR and KNN_CLR predicted more type of LoSa than KNN_ILR in the north of lower reaches; for each model of MLP and RF, the differences were slight; more types of Lo in the northwest of lower reaches and less LoSa near the Heihe River were generated by SVM_ALR, compared with SVM_CLR and SVM_ILR; for XGB, the performance of three maps were different due to the prediction of LoSa. We also compared the prediction of the soil texture types by direct classification (Fig. 6) with those generated by indirect classification using the same machine-learning model, and found completely difference between them on the lower reaches of HRB, such as the distribution of LoSa; on the middle and upper reaches of HRB, all the prediction maps were similar, mainly distributed with SiLo.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

Hydrology and Earth System Sciences
Open Access
Discussions



Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-584
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 11 February 2019

Figure 9. Soil texture classification prediction maps by soil psf interpolation (ALR, CLR, ILR log-ratio and ORI approaches) of KNN, MLP, RF, SVM and XGB.

3.4.4 Comparison of time-spending for each model in soil texture classification and soil psf interpolation

Time spending for models was computed to compare the efficiency of different machine-learning models in soil texture classification and soil psf interpolation (Fig. 10). Because the differences in time spending among ORI and log ratio approaches were similar, ILR was selected for soil psf interpolation. For the different models, RFs required the longest time for both classification (453.73 s) and regression (188.87 s), which may cause it to lose advantages when dealing with big data sets. KNN (classification: 4.2 s, regression: 23.6 s) and SVM (classification: 4.15 s, regression: 12.4 s) both showed shorter time in not only classification but also regression. Likewise, XGB (classification: 21.6 s, regression: 17.13 s) was much more stable and used less time, and the data processes were simpler compared with MLP (classification: 47.28 s, regression: 152.31 s). Moreover, it delivered better performance than KNN and SVM in prediction maps of HRB, demonstrating an effective way of dealing with larger data.

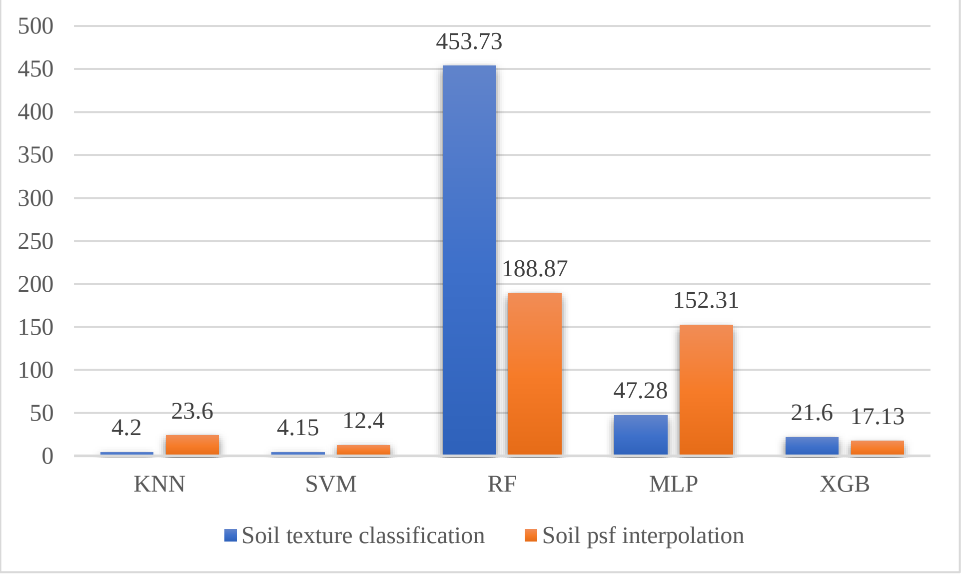


Figure 10. Average time spent running 30 times for KNN, MLP, RF, SVM and XGB of soil texture classification and soil psf interpolation.

4 Discussion

4.1 The systematic comparison of the five machine learning models

As mentioned previously, we compared the performance of different machine-learning methods containing KNN, MLP, RF, SVM and XGB. The results demonstrate that SVM had the highest overall accuracy and XGB generated the highest kappa coefficient with respect to direct soil texture classification; considering the comprehensive evaluation of AUC and AUPRC, RF showed the best performance among these machine-learning models. In the case of soil psf interpolation, the indicators of RMSE, MAE, R^2, AD and STRESS showed that RFs outperformed other machine-learning models, which also indicated more detailed information in prediction maps of sand, silt and clay as well as XGB. For the indirection classification of soil texture, the USDA soil texture triangles generated from RF were the closest to the distribution of the original soil psf data (Fig. 8a), with the highest classification right ratio. Prediction maps of indirect soil texture classification were also considered, demonstrating RF and MLP models were more suitable for the real environment, especially the models combined with log ratio transformation approaches. Time spending of different machine-learning models showed that KNN, SVM and XGB required less time than RF and MLP to fit large data sets.

The comparisons of machine-learning models were also mentioned in previous reports. Heung et al. (2016) demonstrated that tree learners, such as RFs, delivered better performance than KNN and SVM due to the advantages of the interpretability of the results for classification problems in soil science; tree learners (decision trees) were also shown by Taghizadeh-Mehrjardi et al. (2015), indicating that the decision trees and ANN outperformed KNN, RF and SVM. ANNs, however, were typically complicated, which was true for our study due to the standardization and back transformation of MLP. In contrast, Wu et al. (2018) proposed that SVM revealed reliable consequences in the direct soil texture classification, which was quite different from our results. In general, as binary classifiers, multi-class tasks can be handled as well using SVM; however, this is no longer the case in our study, as only 2 types of soil texture were generated from SVM, showing unsatisfactory results in both kappa coefficients (Fig. 3b) and prediction maps (Fig. 6d). The consequences may be explained by the imbalanced data of soil texture types. For more information about tree learners in soil science for regression, Hengl et al. (2017) found lower R^2 using XGB than RF on a global-scale prediction. Zeraatpisheh et al. (2018) put forward the lowest RMSE and the highest R^2 using RF compared with multiple linear regression and regression trees for the prediction of clay, and this conclusion was similar to our study.

4.2 The systematic comparison of the models combined with three log-ratio transformed data and original data

We compared the performance of models combined with three types of log ratio transformed data and original (untransformed) data for soil psf interpolation and indirect soil texture classification, and the results showed that the models using original data performed better in the case of indicators, such as RMSE, MAE, R^2 and AD, while the models using log ratio transformed data improved the STRESS. The interpolation maps of soil psf using ILR approach illustrated closer ranges of soil sampling data than those based on the ORI approach. With respect to the indirect soil texture classification, models using log ratio transformed data improved the overall accuracies and kappa coefficients, such as RF and XGB. The USDA soil texture triangles showed more discrete distribution and more accordance with soil sampling data using the ILR approach. Better performance was shown in soil texture classification prediction maps generated from log ratio transformed data. Among the three log ratio approaches, ILR and CLR were superior to ALR for the reason of more accurate indicators of soil psf interpolation and indirect soil texture classification, as well the performance of prediction maps. Additionally, log ratio approaches modified soil sampling data to become more symmetric (Filzmoser et al., 2009); however, this improvement was not greatly effective. Fig.2 illustrated that soil sampling data for sand and clay were right-skewed, and silt was left-skewed because the silt component was predominant. The ALR approach enhanced soil sampling data of sand; nevertheless, the ALR_sand was still right-skewed, similar to the CLR_sand, presenting the lack of adjustment. In contrast, the ILR_sand changed from right-skewed to left-skewed; from this point of view, the over-adjustment was revealed. Similarly, the lack of adjustments was also shown in CLR_silt and ILR_silt; over-adjustments included ALR_silt, ALR_clay, CLR_clay and ILR_clay, making images that were different from normal distribution, and the *p* values of k-s tests were not significant. In our previous research (Wang and Shi, 2017), the ILR approach had better performance than ALR and CLR, with the highest R^2 and lowest AD. The CLR approach also performed well due to the lowest RMSE and mean error (ME) among the three log ratio approaches. When comparing the original (untransformed) and log ratio approaches, kriging approaches based on the log ratio delivered slightly decreased accuracies, which was similar to the conclusion in our study.

4.3 The systematic comparison of the direct and indirect classification for soil psf

Indirect classification showed not only better performance with respect to accuracy evaluation but also more accordance with the real environment than direct classification. The highest kappa coefficient generated from indirect classification (RF_ILR: 0.291) demonstrated obvious improvement (approximately 21.3 %) compared with that of direct classification (XGB: 0.240), keeping the highest overall accuracy stable (-1.4 %) at the same time (direct: 0.647; indirect: 0.638, respectively).

Compared with the real soil texture distribution and environment of the HRB, SiLo overlaid the upper reaches of HRB, and SaLo and Lo were in the south of the upper reaches of HRB showed strip distribution. Moreover, an uncovered area was detected in the northwest of the lower reaches of HRB, where it cannot be predicted due to a lack of information (soil samples) input in the process of model training. The main soil texture types of the lower reaches of the HRB were SiLo, LoSa and small amounts of SaLo and Lo distributed in uncovered area. The main soil texture types predicted by direct classification using machine-learning models were SaLo and SiLo; RF and XGB delivered much more LoSa than other direct classification models. However, all these models predicted that the main soil type of the lower reaches of HRB was SaLo, which was not fitted for the real environment (LoSa). In addition, because of the limitation of the train sets, direct classification can only predict types in the training data; in contrast, indirect classification broke such limitations, and new prediction types arose due to the transformation from soil psf to soil texture types. Moreover, more suitable matching performance with the real environment should be considered, such as the log ratio approaches of MLP, RF, KNN_ ALR, KNN_ ILR and XGB_CLR. The direct soil texture classification generated relative unsatisfactory consequences. Although the indirect soil texture classification outperformed the direct one, kappa coefficients for indirect classification at fair-level (0.21-0.40) also need to be enhanced. Hence, soil sampling data appear to be comprehensively meaningful, considering accuracy improvement. In the case of soil sampling data, the laser diffraction approach we mentioned above was applied to obtain the discrete representation of particle size curves based on the given quantiles of these curves, i.e., soil particle size fractions (psf, sand, silt and clay). Subsequently, soil psf data were separately modeled for prediction and validation. Another perspective of soil psf should be considered, i.e., the probability density functions of particle size curves (so-called functional compositions), which are non-negative values that integrate to 1 (or 100 %) and can be considered as compositional data with infinitesimal parts (Menafoglio et al., 2014). Unlike conventional approaches, the viewpoints of functional compositions are beneficial to acquiring complete and continuous information rather than discrete information (sand, silt and clay), and soil texture and soil particle size fractions can be extracted using the stochastic simulation of soil particle-size curves (Menafoglio et al., 2016b). Previous studies applied such functional-compositional data for the simulation of particle size curves combined with geostatistical or machine-learning methods such as kriging and bayes approaches (Menafoglio et al., 2016a) in hydrogeology, demonstrating more remarkable results compared with traditional methods. Therefore, which data should be used is the key points of accuracy improvement in future research.

5 Conclusion

We systematically compared a total of 45 models for direct and indirect soil texture classification, and soil psf interpolation using five machine-learning approaches combined with original (untransformed) and three different log ratio transformed data in the HRB. The results indicate that as flexible and stable models, tree learners such as RF delivered powerful performance in both classification and regression and were superior to other machine-learning models mentioned above. As a new and sub-optimal machine-learning method in our study, XGB appeared to be more meaningful and more computationally efficient when dealing with large data sets. In addition, the log ratio approaches had advantages of modifying STRESS in soil psf interpolation. Moreover, the indirect soil texture classification outperformed the direct one, especially when combined with the log ratio approaches. The indirect soil texture classification generated preferable consequences in both cases of accuracy indicators and prediction maps. More appropriate environmental covariates, more symmetric distribution of soil sampling data (or multiple perspectives of compositional data selection), and systematic parameter adjustment algorithms of compositional data are key to improving accuracy in the future.

Data availability. The soil sampling data (DOIs are: 10.3972/heihe.009.2013.db; 10.3972/heihe.009.2013.db; DOI:10.3972/heihe.00135.2016.db; 10.3972/hiwater.147.2013.db; 10.3972/heihe.037.2014.db; 10.3972/heihe.0034.2013.db;

10.3972/heihe.093.2013.db), remote sensing data and part of environmental covariates data can be accessed through http://westdc.westgis.ac.cn/ (last access: 29 October 2018). The meteorological data can be accessed through http://data.cma.cn/ (last access: 29 October 2018).

Author contributions. WS contributed to soil data sampling, oversaw the design of the entire project. MZ performed the analysis and wrote the manuscript. Both authors contributed to writing this paper and interpreting data.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This study was supported by the National Natural Science Foundation of China (Grant No. 41771111 and 41771364), Fund for Excellent Young Talents in Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (2016RC201), and the Youth Innovation Promotion Association, CAS (No. 2018071).

References

Abdi, D., Cade-Menun, B. J., Ziadi, N., and Parent, L.-E.: Compositional statistical analysis of soil p-31-nmr forms, Geoderma, 257, 40-47, https://doi.org/10.1016/j.geoderma.2015.03.019, 2015.

Adhikari, K., and Hartemink, A. E.: Linking soils to ecosystem services - a global review, Geoderma, 262, 101-111, https://doi.org/10.1016/j.geoderma.2015.08.009, 2016.

Aitchison, J.: The statistical-analysis of compositional data, Journal of the Royal Statistical Society Series B-Methodological, 44, 139-177, 1982.

Aitchison, J.: On criteria for measures of compositional difference, Mathematical Geology, 24, 365-379, https://doi.org/10.1007/bf00891269, 1992.

Bacon-Shone, J. H.: A short history of compositional data analysis, in: Compositional data analysis: Theory and applications, Wiley, Chicester, West Sussex, 3, 2011.

Bagheri Bodaghabadi, M., Antonio Martinez-Casasnovas, J., Salehi, M. H., Mohammadi, J., Esfandiarpoor Borujeni, I., Toomanian, N., and Gandomkar, A.: Digital soil mapping using artificial neural networks and terrain-related attributes, Pedosphere, 25, 580-591, https://doi.org/10.1016/s1002-0160(15)30038-2, 2015.

Bationo, A., Kihara, J., Vanlauwe, B., Waswa, B., and Kimetu, J.: Soil organic carbon dynamics, functions and management in west african agro-ecosystems, Agricultural Systems, 94, 13-25, https://doi.org/10.1016/j.agsy.2005.08.011, 2007.

Behrens, T., and Scholten, T.: Chapter 25 a comparison of data-mining techniques in predictive soil mapping, in: Developments in soil science, edited by: Lagacherie, P., McBratney, A. B., and Voltz, M., Elsevier, 353-617, 2006.

Bergmeir, C., and Benitez, J. M.: Neural networks in R using the stuttgart neural network simulator: RSNNS, Journal of Statistical Software, 46, 1-26, 2012.

Breiman, L.: Bagging predictors, Machine Learning, 24, 123-140, https://doi.org/10.1023/a:1018054314350, 1996.

Breiman, L.: Random forests, Machine Learning, 45, 5-32, https://doi.org/10.1023/a:1010933404324, 2001.

Brown, D. J., Clayton, M. K., and McSweeney, K.: Potential terrain controls on soil color, texture contrast and grain-size deposition for the original catena landscape in uganda, Geoderma, 122, 51-72, https://doi.org/10.1016/j.geoderma.2003.12.004, 2004.

Brus, D. J., Kempen, B., and Heuvelink, G. B. M.: Sampling for validation of digital soil maps, European Journal of Soil Science, 62, 394-407, https://doi.org/10.1111/j.1365-2389.2011.01364.x, 2011.

Burges, C. J. C.: A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, 2, 121-167, https://doi.org/10.1023/a:1009715923555, 1998.

Camera, C., Zomeni, Z., Noller, J. S., Zissimos, A. M., Christoforou, I. C., and Bruggeman, A.: A high resolution map of soil types and physical properties for cyprus: A digital soil mapping optimization, Geoderma, 285, 35-49, https://doi.org/10.1016/j.geoderma.2016.09.019, 2017.

Chen, T., and Guestrin, C.: Xgboost: A scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y.: Xgboost: Extreme gradient boosting, R package version 0.71.2, available at: https://CRAN.R-project.org/package=xgboost (last access: 18 November 2018), 2018.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J.: System for automated geoscientific analyses (saga) v. 2.1.4, Geosci. Model Dev., 8, 1991-2007, https://doi.org/10.5194/gmd-8-1991-2015, 2015.

Cortes, C., and Vapnik, V.: Support-vector networks, Machine Learning, 20, 273-297, https://doi.org/10.1023/a:1022627411411, 1995.

Cover, T. M., and Hart, P. E.: Nearest neighbor pattern classification, Ieee Transactions on Information Theory, 13, 21, https://doi.org/10.1109/tit.1967.1053964, 1967.

Crouvi, O., Pelletier, J. D., and Rasmussen, C.: Predicting the thickness and aeolian fraction of soils in upland watersheds of the mojave desert, Geoderma, 195, 94-110, https://doi.org/10.1016/j.geoderma.2012.11.015, 2013.

Davis, J., and Goadrich, M.: The relationship between precision-recall and roc curves, Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA, 2006.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C.: Isometric logratio transformations for compositional data analysis, Mathematical Geology, 35, 279-300, https://doi.org/10.1023/a:1023818214614, 2003.

Elith, J., Leathwick, J. R., and Hastie, T.: A working guide to boosted regression trees, Journal of Animal Ecology, 77, 802-813, https://doi.org/10.1111/j.1365-2656.2008.01390.x, 2008.

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., and Xiang, Y.: Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in china, Energy Conversion and Management, 164, 102-111, https://doi.org/10.1016/j.enconman.2018.02.087, 2018.

Fawcett, T.: An introduction to roc analysis, Pattern Recognition Letters, 27, 861-874, https://doi.org/10.1016/j.patrec.2005.10.010, 2006.

Filzmoser, P., Hron, K., and Reimann, C.: Univariate statistical analysis of environmental (compositional) data: Problems and possibilities, Science of the Total Environment, 407, 6100-6108, https://doi.org/10.1016/j.scitotenv.2009.08.008, 2009.

Follain, S., Minasny, B., McBratney, A. B., and Walter, C.: Simulation of soil thickness evolution in a complex agricultural landscape at fine spatial and temporal scales, Geoderma, 133, 71-86, https://doi.org/10.1016/j.geoderma.2006.03.038, 2006.

Fu, G.-H., Xu, F., Zhang, B.-Y., and Yi, L.-Z.: Stable variable selection of class-imbalanced data with precision-recall criterion, Chemometrics and Intelligent Laboratory Systems, 171, 241-250, https://doi.org/10.1016/j.chemolab.2017.10.015, 2017.

Gobin, A., Campling, P., and Feyen, J.: Soil-landscape modelling to quantify spatial variability of soil texture, Physics and Chemistry of the Earth Part B-Hydrology Oceans and Atmosphere, 26, 41-45, https://doi.org/10.1016/s1464-1909(01)85012-7, 2001.

Gochis, D. J., Vivoni, E. R., and Watts, C. J.: The impact of soil depth on land surface energy and water fluxes in the north american monsoon region, Journal of Arid Environments, 74, 564-571, https://doi.org/10.1016/j.jaridenv.2009.11.002, 2010.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., de Jesus, J. M., Tamene, L., and Tondoh, J. E.: Mapping soil properties of africa at 250 m resolution: Random forests significantly improve current predictions, Plos One, 10, https://doi.org/10.1371/journal.pone.0125814, 2015.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: Soilgrids250m: Global gridded soil information based on machine learning, Plos One, 12, https://doi.org/10.1371/journal.pone.0169748, 2017.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., and Schmidt, M. G.: An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping, Geoderma, 265, 62-77, https://doi.org/10.1016/j.geoderma.2015.11.014, 2016.

Huang, J., Subasinghe, R., and Triantafilis, J.: Mapping particle-size fractions as a composition using additive log-ratio transformation and ancillary data, Soil Science Society of America Journal, 78, 1967-1976, https://doi.org/10.2136/sssaj2014.05.0215, 2014.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G.: Overview of the radiometric and biophysical performance of the modis vegetation indices, Remote Sensing of Environment, 83, 195-213, https://doi.org/10.1016/s0034-4257(02)00096-2, 2002.

Huete, A. R.: A soil-adjusted vegetation index (savi), Remote Sensing of Environment, 25, 295-309, https://doi.org/10.1016/0034-4257(88)90106-x, 1988.

Jafari, A., Khademi, H., Finke, P. A., Van de Wauw, J., and Ayoubi, S.: Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern iran, Geoderma, 232, 148-163, https://doi.org/10.1016/j.geoderma.2014.04.029, 2014.

Krasilnikov, P. V., Garcia-Calderon, N. E., Ibanez-Huerta, A., Bazan-Mateos, M., and Hernandez-Santana, J. R.: Soilscapes in the dynamic tropical environments: The case of sierra madre del sur, Geomorphology, 135, 262-270, https://doi.org/10.1016/j.geomorph.2011.02.013, 2011.

Kuhn, M.: Caret: Classification and regression training, R package version 6.0-80, available at: https://CRAN.R-project.org/package=caret (last access: 18 November 2018), 2018.

Landis, J. R., and Koch, G. G.: Measurement of observer agreement for categorical data, Biometrics, 33, 159-174, https://doi.org/10.2307/2529310, 1977.

Lark, R. M., and Bishop, T. F. A.: Cokriging particle size fractions of the soil, European Journal of Soil Science, 58, 763-774, https://doi.org/10.1111/j.1365-2389.2006.00866.x, 2007.

Liaw, A., and Wiener, M.: Classification and regression by randomforest, R News, 2, 18-22, 2002.

Liess, M., Glaser, B., and Huwe, B.: Uncertainty in the spatial prediction of soil texture comparison of regression tree and random forest models, Geoderma, 170, 70-79, https://doi.org/10.1016/j.geoderma.2011.10.010, 2012.

Marchetti, A., Piccini, C., Santucci, S., Chiuchiarelli, I., and Francaviglia, R.: Simulation of soil types in teramo province (central italy) with terrain parameters and remote sensing data, Catena, 85, 267-273, https://doi.org/10.1016/j.catena.2011.01.012, 2011.

Martin-Fernandez, J. A., Olea-Meneses, R. A., and Pawlowsky-Glahn, V.: Criteria to compare estimation methods of regionalized compositions, Mathematical Geology, 33, 889-909, https://doi.org/10.1023/a:1012293922142, 2001.

McBratney, A. B., Santos, M. L. M., and Minasny, B.: On digital soil mapping, Geoderma, 117, 3-52, https://doi.org/10.1016/s0016-7061(03)00223-4, 2003.

McNamara, J. P., Chandler, D., Seyfried, M., and Achet, S.: Soil moisture states, lateral flow, and streamflow generation in a semi-arid, snowmelt-driven catchment, Hydrological Processes, 19, 4023-4038, https://doi.org/10.1002/hyp.5869, 2005.

Menafoglio, A., Guadagnini, A., and Secchi, P.: A kriging approach based on aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers, Stoch. Environ. Res. Risk Assess., 28, 1835-1851, https://doi.org/10.1007/s00477-014-0849-8, 2014.

Menafoglio, A., Guadagnini, A., and Secchi, P.: Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a bayes space approach, Water Resources Research, 52, 5708-5726, 10.1002/2015wr018369, 2016a.

Menafoglio, A., Secchi, P., and Guadagnini, A.: A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers, Math Geosci., 48, 463-485, https://doi.org/10.1007/s11004-015-9625-7, 2016b.

Metternicht, G. I., and Zinck, J. A.: Remote sensing of soil salinity: Potentials and constraints, Remote Sensing of Environment, 85, 1-20, https://doi.org/10.1016/s0034-4257(02)00188-8, 2003.

Meyer, D., Dimitriadou, E., Hornik, K., Andreas, W., and Friedrich, L.: E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien, R package version 1.6-8, available at: https://CRAN.R-project.org/package=e1071 (last access: 18 November 2018), 2017.

Moeys, J.: Soiltexture: Functions for soil texture plot, classification and transformation, R package version 1.4.6, available at: https://CRAN.R-project.org/package=soiltexture (last access: 18 November 2018), 2018.

Odeh, I. O. A., Todd, A. J., and Triantafilis, J.: Spatial prediction of soil particle-size fractions as compositional data, Soil Science, 168, 501-515, https://doi.org/10.1097/00010694-200307000-00005, 2003.

Pahlavan-Rad, M. R., and Akbarimoghaddam, A.: Spatial variability of soil texture fractions and ph in a flood plain (case study from eastern iran), Catena, 160, 275-281, https://doi.org/10.1016/j.catena.2017.10.002, 2018.

Poggio, L., and Gimona, A.: 3d mapping of soil texture in scotland, Geoderma Regional, 9, 5-16, https://doi.org/10.1016/j.geodrs.2016.11.003, 2017.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. available at: https://www.R-project.org/ (last access: 18 November 2018), 2018.
2018.

Saito, T., and Rehmsmeier, M.: Precrec: Fast and accurate precision-recall and roc curve calculations in r, Bioinformatics, 33, 145-147, https://doi.org/10.1093/bioinformatics/btw570, 2017.

Salazar, E., Giraldo, R., and Porcu, E.: Spatial prediction for infinite-dimensional compositional data, Stochastic Environmental Research and Risk Assessment, 29, 1737-1749, https://doi.org/10.1007/s00477-014-1010-4, 2015.

Schliep, K., and Hechenbichler, K.: Kknn: Weighted k-nearest neighbors, R package version 1.3.1, available at: https://CRAN.R-project.org/package=kknn (last access: 18 November 2018), 2016.

Song, X.-D., Brus, D. J., Liu, F., Li, D.-C., Zhao, Y.-G., Yang, J.-L., and Zhang, G.-L.: Mapping soil organic carbon content by geographically weighted regression: A case study in the heihe river basin, china, Geoderma, 261, 11-22, https://doi.org/10.1016/j.geoderma.2015.06.024, 2016.

Subasi, A.: Eeg signal classification using wavelet feature extraction and a mixture of expert model, Expert Systems with Applications, 32, 1084-1093, https://doi.org/10.1016/j.eswa.2006.02.005, 2007.

Sun, X.-L., Wu, Y.-J., Wang, H.-L., Zhao, Y.-G., and Zhang, G.-L.: Mapping soil particle size fractions using compositional kriging, cokriging and additive log-ratio cokriging in two case studies, Mathematical Geosciences, 46, 429-443, https://doi.org/10.1007/s11004-013-9512-z, 2014.

Taalab, K., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M. J., Hannam, J. A., and Creamer, R.: On the application of bayesian networks in digital soil mapping, Geoderma, 259, 134-148, https://doi.org/10.1016/j.geoderma.2015.05.014, 2015.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., and Triantafilis, J.: Comparing data mining classifiers to predict spatial distribution of usda-family soil groups in baneh region, iran, Geoderma, 253, 67-77, https://doi.org/10.1016/j.geoderma.2015.04.008, 2015.

Thompson, J. A., Roecker, S., Grunwald, S., and Owens, P. R.: Chapter 21 - digital soil mapping: Interactions with and applications for hydropedology, in: Hydropedology, edited by: Lin, H., Academic Press, Boston, 665-709, 2012.

van den Boogaart, K. G., and Tolosana-Delgado, R.: "Compositions": A unified r package to analyze compositional data, Computers & Geosciences, 34, 320-338, https://doi.org/10.1016/j.cageo.2006.11.017, 2008.

Vapnik, V.: The support vector method of function estimation, Nonlinear modeling: Advanced black-box techniques, edited by: Suykens, J. A. K., and Vandewalle, J., 55-85 pp., 1998.

Walvoort, D. J. J., and de Gruijter, J. J.: Compositional kriging: A spatial interpolation method for compositional data, Mathematical Geology, 33, 951-966, https://doi.org/10.1023/a:1012250107121, 2001.

Wang, Z., and Shi, W.: Mapping soil particle-size fractions: A comparison of compositional kriging and log-ratio kriging, Journal of Hydrology, 546, 526-541, https://doi.org/10.1016/j.jhydrol.2017.01.029, 2017.

Wang, Z., and Shi, W.: Robust variogram estimation combined with isometric log-ratio transformation for improved accuracy of soil particle-size fraction mapping, Geoderma, 324, 56-66, https://doi.org/10.1016/j.geoderma.2018.03.007, 2018.

Wu, B., Yan, N., Xiong, J., Bastiaanssen, W. G. M., Zhu, W., and Stein, A.: Validation of etwatch using field measurements at diverse landscapes: A case study in hai basin of china, Journal of Hydrology, 436, 67-80, https://doi.org/10.1016/j.jhydrol.2012.02.043, 2012.

Wu, W., Li, A.-D., He, X.-H., Ma, R., Liu, H.-B., and Lv, J.-K.: A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest china, Computers and Electronics in Agriculture, 144, 86-93, https://doi.org/10.1016/j.compag.2017.11.037, 2018.

Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., and Li, D.-C.: Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem, Ecological Indicators, 60, 870-878, https://doi.org/10.1016/j.ecolind.2015.08.036, 2016.

Yi, C., Li, D., Zhang, G., Zhao, Y., Yang, J., Liu, F., and Song, X.: Criteria for partition of soil thickness and case studies, Acta Pedologica Sinica, 52, 220-227, 2015.

Yoo, K., Amundson, R., Heimsath, A. M., and Dietrich, W. E.: Spatial patterns of soil organic carbon on hillslopes: Integrating geomorphic processes and the biological c cycle, Geoderma, 130, 47-65, https://doi.org/10.1016/j.geoderma.2005.01.008, 2006.

Zeraatpisheh, M., Ayoubi, S., Jafari, A., and Finke, P.: Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in iran, Geomorphology, 285, 186-204, https://doi.org/10.1016/j.geomorph.2017.02.015, 2017.

Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., and Finke, P.: Digital mapping of soil properties using multiple machine learning in a semi-arid region, central iran, Geoderma, https://doi.org/10.1016/j.geoderma.2018.09.006, 2018.

Zhang, S.-w., Shen, C.-y., Chen, X.-y., Ye, H.-c., Huang, Y.-f., and Lai, S.: Spatial interpolation of soil texture using compositional kriging and regression kriging with consideration of the characteristics of compositional data and environment variables, Journal of Integrative Agriculture, 12, 1673-1683, https://doi.org/10.1016/s2095-3119(13)60395-0, 2013.

Zhao, Z., Chow, T. L., Rees, H. W., Yang, Q., Xing, Z., and Meng, F.-R.: Predict soil texture distributions using an artificial neural network model, Computers and Electronics in Agriculture, 65, 36-48, https://doi.org/10.1016/j.compag.2008.07.008, 2009.