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

Mo Zhang[1,2], Wenjiao Shi[1,3], Ziwei Xu[4]

[1]Key Laboratory of Land Surface Pattern and Simulation, State Key Laboratory of Resources and Environmental Information System, 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
[4]State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

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

Abstract. Soil texture and soil particle size fractions (PSFs) play an increasing role in physical, chemical and hydrological processes. Many previous studies have used machine-learning and log ratio transformation methods for soil texture classification and soil PSF interpolation to improve the prediction accuracy. However, few reports systematically compared their performance in both classification and interpolation. Here, five machine-learning models – K-nearest neighbor (KNN), multilayer perceptron neural network (MLP), random forest (RF), support vector machines (SVM), extreme gradient boosting (XGB), combined with original and three log ratio methods – additive log ratio (ALR), centered log ratio (CLR) and isometric log ratio (ILR), were applied to evaluate using both raw and log ratio transformed data from 640 soil samples in the the Heihe River Basin (HRB) in China. The results demonstrated that the log ratio transformations have decreased the skewness of soil PSF data. For soil texture classification, RF and XGB showed better performance with higher overall accuracy and kappa coefficient. They were also recommended to evaluate classification capacity of imbalanced data according to the area under

Abbreviations: PSFs, particle size fractions; HRB, Heihe River Basin; 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; PRC, precision-recall curve; AUPRC, area under the PRC; RMSE, root mean square error; MAE, mean absolute error; RCC, Spearman rank correlation coefficient; MAD, median absolute deviation; AD, Aitchison distance; STRESS, standardized residual sum of squares; SD, standard deviation; 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.

the precision-recall curve (AUPRC) analysis. For soil PSF interpolation, RF delivered the best performance among five machine-learning models with the lowest root mean square 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 the highest Spearman rank correlation coefficient (RCC, sand: 0.69, silt: 0.67, clay: 0.69). STRESS was improved using log ratio methods, especially for CLR and ILR. Prediction maps from both direct and indirect classification were similar on middle and upper reaches of the HRB. However, indirect classification maps using log ratio transformed data had more detailed information in the lower reaches of the HRB. There is a pronounced improvement with 21.3 % of kappa coefficient using indirect methods for soil texture classification compared with the direct ones. RF was recommended as the best strategy among the five machine-learning models, according to the accuracy evaluation of soil PSF interpolation and soil texture classification, and ILR was recommended for component-wise machine-learning models without multivariate treatment, considering the constrained nature of compositional data. In addition, XGB was preferred than other models when trade-off of accuracy and time was considered. Our findings can provide a reference for future works about spatial prediction of soil PSFs and texture using machine-learning models with skewed distributions of soil PSF data in a large area.

1 Introduction

Soil texture, classified by ranges of soil particle size fractions (PSFs), 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, aeration, etc. Soil texture distribution can be used for soil fertility management (Pahlavan-Rad and Akbarimoghaddam, 2018; Bationo et al., 2007), water management (Thompson et al., 2012) and ecosystem service provision (Adhikari and Hartemink, 2016). The soil PSFs – 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 PSFs affect runoff generation, slope stability, soluble salt content and estimation of evaporative fraction (McNamara et al., 2005; Follain et al., 2006; Yoo et al., 2006; Gochis et al., 2010; Crouvi et al., 2013; Xu et al., 2019).

The ancillary data should be considered in the prediction especially in a large study area to enhance the interpolation performance (Wang and Shi, 2017). Machine-learning models such as boosting regression trees (Jafari et al., 2014; Yang et al., 2016), random forests (RF) (Hengl et al., 2015; Zeraatpisheh et al., 2017) and artificial neural networks (Bagheri Bodaghabadi et al., 2015; Taalab et al., 2015) have been commonly employed in both interpolation and classification combined with environmental covariates for soil properties. Machine-learning models such as RF and gradient boosting had better performance than statistical linear models (e.g., multiple linear regression) in the prediction of soil properties, because they are robust to noise and have low bias when dealing with large data sets (Hengl et al., 2015; Hengl et al., 2017). Among machine-learning models, artificial neural network and “tree learners” (e.g., decision trees) were preferred because of relatively high overall accuracy and kappa coefficients, the interpretability of the results and the speed of parameterization in the prediction

of soil classes (Taghizadeh-Mehrjardi et al., 2015; Heung et al., 2016). Most previous studies used machine-learning algorithms to simulate soil category or continuous properties for classification or regression problems. However, few studies systematically analyzed both soil texture classification and soil PSF interpolation using different machine-learning models.

The soil PSFs, which can be classified as soil texture, are not only continuous variables but also compositional data – the constant sum (1 or 100 %) should be guaranteed. Soil PSF data are typical compositional data with three components, which are not independent of each other but expressed in percentage (Filzmoser et al., 2009). Because of the spurious correlations between components, different results would occur on different measurement scales (Abdi et al., 2015; Reimann and Filzmoser, 2000). Indicators and statistical methods based on Euclidean distances can reveal misleading or biased results (Butler, 1979). Numerous different interpretations of compositional data were suggested in soil science (Gobin et al., 2001; Salazar et al., 2015; Tolosana-Delgado et al., 2019; Hengl et al., 2018), and the most extensively used method was a combination of log ratio transformation methods, including the additive log ratio (ALR) (Aitchison, 1982), the centered log ratio (CLR) (Aitchison, 1982), as well as the isometric log ratio (ILR) (Egozcue et al., 2003). Soil PSFs were predicted using multiple linear regression (Huang et al., 2014) and kriging (Wang and Shi, 2018; Zhang et al., 2013) combined with log ratio transformation methods. Moreover, multivariate treatment of soil PSFs can be realized using the probability density functions of soil particle size curves (PSCs), since non-negative values integrating to 1 (or 100 %) can be considered as compositional data with infinitesimal parts (so-called functional compositions) (Menafoglio et al., 2014). Functional compositions are beneficial to acquiring complete and continuous information rather than discrete information, and soil texture and soil PSFs can be extracted from the stochastic simulation of soil PSCs (Menafoglio et al., 2016a), applying jointly to the fractions and exploiting fully the richness of information. Menafoglio et al., (2016b) applied such functional-compositional data for the stochastic simulation of PSCs based on geostatistical Monte Carlo and Bayes space approach combined with CLR transformation method in heterogeneous aquifer systems in hydrogeology, demonstrating more remarkable improvement of characterizations of the spatial variability and uncertainty compared with traditional methods. However, most soil PSFs data of studies are discrete (i.e., sand, silt and clay), and few studies conducted systematic comparison of accuracy, strengths and weaknesses for different machine-learning models using original data and different log ratio transformed data.

Soil texture classification can be predicted by machine-learning models directly, and it also can be derived indirectly from soil PSFs. For the direct soil texture classification, tree-based models such as RF and classification tree (CT) performed better than multinomial logistic regression, support vector machines (SVM) and artificial neural network (ANN) (Camera et al., 2017; Wu et al., 2018). For the indirect classification of soil texture, Poggio and Gimona (2017) combined hybrid geostatistical generalized additive models with ALR and modeled PSFs at 250 m resolution in Scotland. Considering the particularity of compositional data, the results of soil PSF classification and interpolation could be compared using the direct and indirect methods. Nevertheless, few studies systematically compared the different machine-learning models for both direct and indirect soil texture classification.

In our study, five machine-learning models – K-nearest neighbor (KNN), multilayer perceptron neural network (MLP), RF, SVM, and extreme gradient boosting (XGB) – were applied for soil texture classification and soil PSF interpolation.

Furthermore, the log ratio transformed data were also combined with these five machine-learning models for soil PSF interpolation. The objectives of this study are (i) to compare the performance of five machine-learning models for soil texture classification and soil PSF interpolation, (ii) to evaluate the performance of machine-learning models using original and different log ratio transformed data for soil PSF interpolation, and (iii) to estimate the performance of direct and indirect soil texture classification using these methods.

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 northwest of China, covering the Inner Mongolia Autonomous Region, Gansu and Qinghai provinces, which is the second largest inland river basin in China with an area of 146,700 km^2 (Fig. 1a). The elevation ranges from 669 m to 5573 m (Fig. 1b). For the upper reaches of the HRB, the mean annual precipitation is 350 mm; the annual mean temperature is from -5 to 4 °C; the annual average evaporation is 1000 mm. For the middle reaches of the HRB, the mean annual precipitation is between 50 and 250 mm; the annual average evaporation increases from 2000 (east) to 4000 mm (west); the mean annual temperature is from 2.8 to 7.6 °C. The lower reaches of the HRB are situated in Ejina Banner on the Alxa Plateau, which has 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.

The vegetation in the upper reaches of the HRB (Fig. 1c) is influenced by hydrothermal conditions from the southeast to northwest. 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, and the shrub and steppe are mainly located in the area near the lower reaches of Heihe River.

The main soil types (Fig. 1d) are frigid desert soils (higher than 4000 m), alpine meadow soil and alpine steppe soil (3600 – 4000 m), gray cinnamon soil and chernozem (3200 – 3600 m), sierozem and gray cinnamon soil (2600 – 3200 m), gray cinnamon 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 the HRB are aeolian sandy soil, frigid frozen soil and gray brown desert soil. The main soil types in the lower reaches of the HRB are aeolian sandy soil, gray brown desert soil (northwest) and lithosol (northeast).

The main geomorphology types on the upper reaches of the HRB are modern glaciers, alpine, hilly, and intermountain basin (Fig. 1e). Narrow plains are distributed on the middle reaches of the HRB. For the lower reaches, the main types of geomorphology are hilly (northwest), plain, sandy land and platform (east), and a flood plain located in the area near the Heihe River.

The main land use types of upper reaches, middle reaches and lower reaches were forest land and grassland, cultivated land, unused land, respectively (Fig. 1f). Water area and construction area were mainly distributed near the river on the middle reaches of the HRB.

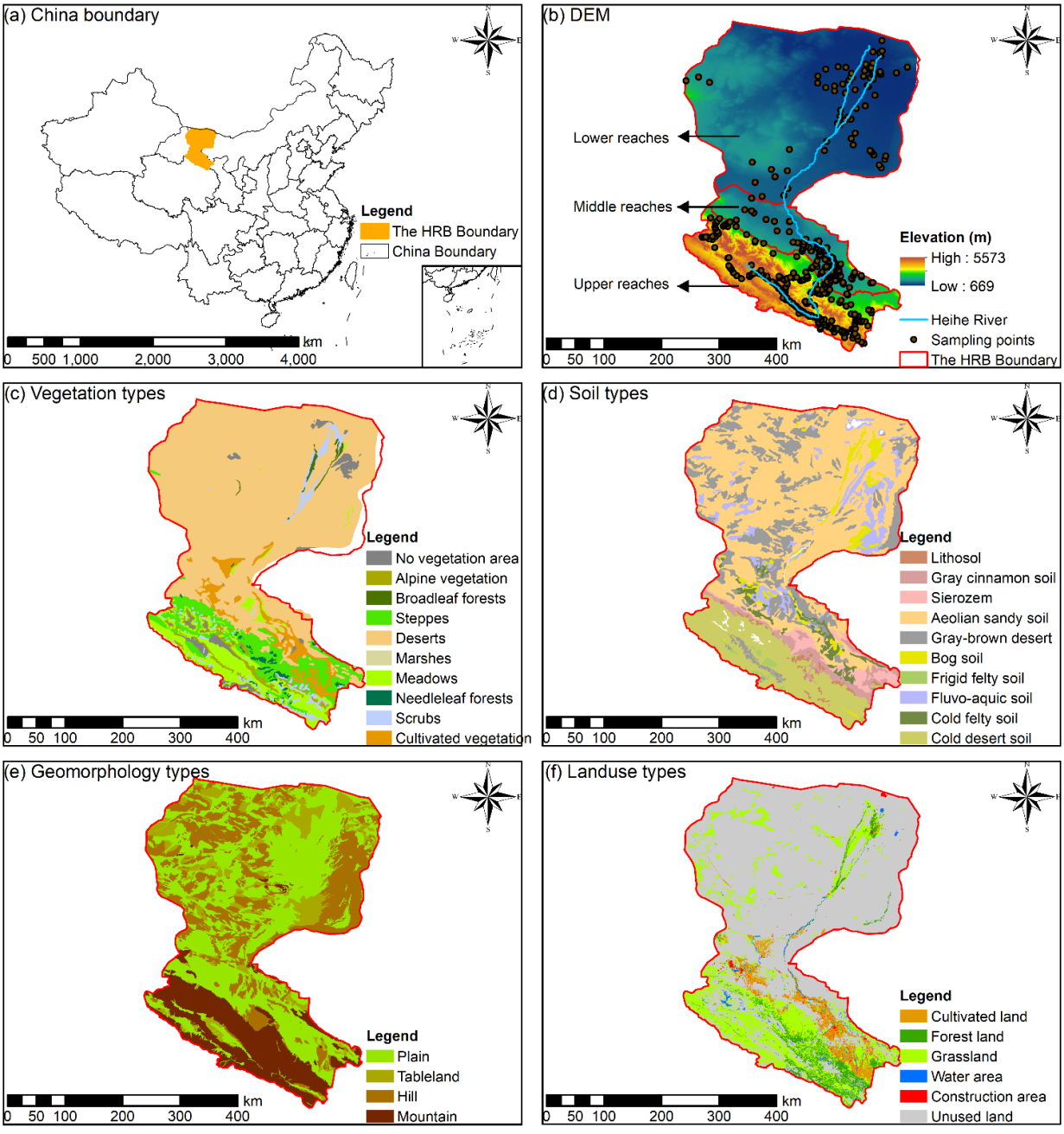


Figure 1. (a) The location of the Heihe River Basin (HRB) in China, (b) the distributions of Heihe River, elevation and soil sampling points, (c) vegetation types, (d) soil types, (e) geomorphology types, and (f) land use types of the HRB.

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/), including 392 samples on the upper reaches and 248 samples on the middle and lower reaches of the HRB (Fig. 1b). The soil types, vegetation types, distribution of DEM and geomorphology types of the HRB were considered in soil sample collection in terms of their location and proportion for more representative of soil PSFs using limited soil samples. Purposive sampling was used as the sampling strategy to collect soil samples and to characterize the spatial variability of soil PSFs. In this strategy, sample sites were chosen based on the variability of soil formation factors, which represented the heterogeneity of the soil PSFs in the HRB such as the distribution of climate, categorical maps, etc. Due to more complicated soil types and vegetation types on the middle and upper reaches of the HRB, there were more soil sampling points in these areas. In contrast, fewer samples were collected because of relatively similar vegetation types on the lower reaches. To reduce the noise effect of soil samples, the average of mixed 3 – 5 topsoil (0 – 20 cm) samples for each soil sample and its parallel sample was used as the final measurement. The global position system (GPS) information and related environmental covariates were recorded. Subsequently, the samples were dried, analyzed and measured for soil PSFs (approximately 30 g of each sample). Soil PSFs were analyzed using Malvern Mastersizer 2000 laser diffraction particle size analyzer (average measurement error is less than 3 %).

2.3 Environmental covariates

The environmental covariates, such as topographic variables, remote sensing variables, climate and position variables, soil physicochemical variables and categorical maps, are related to the distributions of soil PSFs. System for Automated Geoscientific Analysis (SAGA) GIS (Conrad et al., 2015) was used to compute the topographic variables from DEM, including slope, aspect, convergence index, general curvature, plane curvature, profile curvature and valley depth. Remote sensing variables, 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 variables such as the mean annual precipitation and the mean annual temperature from the National Meteorological Information Center (http://data.cma.cn/). Latitude and longitude were also considered because of the large area of the HRB. Mean annual surface evapotranspiration (Wu et al., 2012) were gathered from WestDC (http://westdc.westgis.ac.cn/), as well as soil physicochemical variables including 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). Additionally, the categorical maps were also used, such as geomorphology types, soil types, land use types and vegetation types (Fig. 1).

2.4 Machine-learning models and parameters optimization

2.4.1 K-nearest neighbor

K-nearest neighbor (KNN) is a simple and non-parametric classifier based on the known instance to label unknown instance (Cover and Hart, 1967). For the test set, K-nearest training set vectors (k) were found based on distance, and maximum summed kernel densities were computed for classification. Moreover, continuous variables can also be predicted for regression with the average values of K-nearest neighbors. The parameters of KNN contain the maximum value of k (kmax), the distances of the nearest neighbors (distance) and the types of a kernel function (kernel). The KNN model is available in the R package "kknn" (Schliep and Hechenbichler, 2016).

2.4.2 Multilayer perceptron neural network

Multilayer perceptron neural network (MLP), which is one of the most commonly multilayer feedforward backpropagation networks (Zhang et al., 2018), was selected to train artificial neural network (ANN) models due to the rapid operation, the 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 was 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 scale 0 to 1. MLP can be run using the R package "RSNNS" (Bergmeir and Benitez, 2012).

2.4.3 Random forest

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. The standardization or normalization is not necessary because it is insensitive to the range of input values. Two parameters should be adjusted for the 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 machine

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 can also be solved by minimization of the structural risk using loss functions (Vapnik, 1998) in SVM, named support vector regression. The SVM is more effective in high dimensional spaces. Linear function was selected for SVM as the kernel function in our study. Additionally, cost and gamma are two other parameters needed to be tuned, 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

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 perform better in terms of accuracy assessment. The residuals of the first tree can be fitted by the second tree to enhance the model accuracy and the sum of the prediction of each tree generates the ultimate prediction. There are seven parameters in XGB – 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 equation description of five machine-learning models can be found in the Supplementary Section S1. R package "caret" (Kuhn, 2018) for MLP, SVM, XGB, "randomForest" for RF and "kknn" for KNN were used to adjust the above parameters. A set of parameters with the lowest RMSE for regression and the highest kappa coefficient for classification by cross-validation will be selected as the best parameters. There are 11 dependent variables (i.e., "sand, silt, clay, ilr1, ilr2, alr1, alr2, clr1, clr2, clr3" for regression and "class" for classification) trained with environmental covariates (independent variables). All methods were applied independently on these 11 components, and all adjusted parameters for different models were listed in Table S2.1. More details about parameters optimization and independent modeling were demonstrated in the Supplementary Section S2.

2.5 Log ratio transformation methods

For the composition of D elements $\mathbf{x} = [x_1, \ldots, x_D]$, $x_j > 0$, $\forall j = 1, 2, \ldots, D$, and $\sum_{j=1}^{D} x_j = 1$, the transformation equation for ALR, CLR and ILR are defined as follows:

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

$$clr(\mathbf{x}) = (ln\frac{x_1}{\sqrt[D]{\prod_{j=1}^{D} x_j}}, \ldots, ln\frac{x_D}{\sqrt[D]{\prod_{j=1}^{D} x_j}}), \quad (2)$$

$$\mathbf{z} = (z_1, \ldots, z_{D-1}) = ilr(\mathbf{x}), \quad (3)$$

$$z_i = \sqrt{\frac{D-i}{D-i+1}}\, ln\frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \text{ for } i = 1, \ldots, D-1, \quad (4)$$

where z_i is the *i*th component. The inverse transformation equations for ALR, CLR and ILR were computed in the "compositions" R package (van den Boogaart and Tolosana-Delgado, 2008), which were defined as follows:

$$\overline{alr}(x_j) = \frac{exp(alr(x_j))}{\sum_{j=1}^{D} exp(alr(x_j))}, \quad (5)$$

$$\overline{clr}(x_j) = \frac{exp(clr(x_j))}{\sum_{j=1}^{D} exp(clr(x_j))}, \quad (6)$$

$$Y(x_j) = \sum_{j=1}^{D} \frac{ilr(x_j)}{\sqrt{j \times (j+1)}} - \sqrt{\frac{j-1}{j}} \times ilr(x_j), \quad (7)$$

$$ilr(x_0) = ilr(x_D) = 0, \quad (8)$$

$$\overline{ilr}(x_j) = \frac{exp(Y(x_j))}{\sum_{j=1}^{D} exp(Y(x_j))}, \quad (9)$$

For original data, the standardization function was used to ensure predictions of soil PSFs were between 0 and 100 and that their sum was 100 %:

$$sand_s = \frac{sand}{(sand+silt+clay)} \times 100, \quad (10)$$

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

2.6 Validation

2.6.1 Validation method

We used five machine-learning models combined with original data (ORI) and three log ratio methods (ALR, CLR, ILR) in this study, including five machine-learning models for direct soil texture classification (5 models), and these methods using original data and log ratio transformed data for indirect soil texture classification (20 models) and soil PSF interpolation (20 models) (Table 1). The data were randomly divided into two sets: 448 soil samples (70 %) for training and 192 soil samples (30 %) for validation. This process was repeated 30 times.

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

Transformation methods	Soil texture classification (direct)	Soil texture classification (indirect) & soil PSF interpolation
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

We used the overall accuracy, kappa coefficients, area under the precision-recall curve (AUPRC) and abundance index to validate the performance of different models. The first two indicators were selected to evaluate the overall prediction performance of soil texture types, and the last two were applied to evaluate the performance of each soil texture type.

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 overall accuracy is defined as follows (Brus et al., 2011):

$$Overall\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (11)$$

where TP, TN, FP, FN were true positive, true negative, false positive and false negative, respectively. Kappa coefficient demonstrates the agreement of observed classes and measured classes, which is calculated based on the confusion matrix, the equation is defined as:

$$kappa = \frac{p_o - p_e}{1 - p_e}, \quad (12)$$

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). 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 precision and recall, which indicated the extent of identifying positive cases:

$$Recall = \frac{TP}{TP+FN}, \quad (13)$$

$$Precison = \frac{TP}{TP+FP}, \quad (14)$$

Soil texture are a class-imbalanced data set of positive and negative with 62.5% silt loam types, 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). PRCs are informative in dealing with class-imbalanced data (Fu et al., 2017). The R package "precrec" (Saito and Rehmsmeier, 2017) can generate PRCs and compute AUPRC for each soil texture type. This process was repeated 30 times and eventually, the average PRCs and AUPRCs were obtained.

Similarly, confusion index (COI) based on prediction probability was calculated to evaluate the uncertainties of machine-learning models of classification (Burrough et al., 1997). The equation was as follows:

$$COI = \frac{\sum_{i=1}^{n}[1-(P_{max,i}-P_{secmax,i})]}{n} \quad (15)$$

where $P_{max,i}$ refers to the maximum value of probability of soil sampling point i, and $P_{secmax,i}$ represents the second highest value of probability of soil sampling point i. A lower COI indicates better performance of model.

Abundance index was applied to describe the proportion of all soil texture types and well-classified soil texture types in prediction maps, which was defined as follows:

$$Abundance\ index = p/t, \quad (16)$$

where p is all soil texture types in prediction maps and t is all soil texture type(s) of soil samples. All of the soil texture types were involved to ensure the balance of the soil texture types, including 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).

2.6.3 Validation indicators for soil PSF interpolation

Five statistical indicators, including Spearman rank correlation coefficient (RCC), 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), were used to validate the methods of soil PSF interpolation. The equations for the validation indicators RCC, RMSE, MAE, AD and STRESS are as follows:

$$RCC = \rho_{xy}(rank) = \frac{\sigma_{xy(rank)}}{\sigma_{x(rank)}\sigma_{y(rank)}}, \quad (17)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{i,m} - Y_{i,e})^2}, \quad (18)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_{i,m} - Y_{i,e}\right|, \quad (19)$$

where $Y_{i,m}, Y_{i,e}, \bar{Y}_{i,m}$ are measured, estimated and the mean of measured soil PSFs, and n is the number of observations (soil sampling points for validation). $\sigma_{x(rank)}$ and $\sigma_{y(rank)}$ are variances for measured and estimated data, respectively. $\sigma_{xy(rank)}$ is covariance. Rank refers to assigning rank as 1 to the smallest value, rank as 2 to the next highest value, and so on (Mishra and Datta-Gupta, 2018). Higher RCC and lower RMSE and MAE show better performance of models.

$$AD = \left[\sum_{i=1}^{D}\left[log\frac{x_i}{g(x)} - log\frac{X_i}{g(X)}\right]^2\right]^{1/2}, \quad (20)$$

$$STRESS = \left[\frac{\sum_{i<j}(AD_{x,ij} - AD_{X,ij})^2}{\sum_{i<j}(AD_{x,ij})^2}\right]^{1/2}, \quad (21)$$

where x is the observed value; X is the predicted value; D is the number of dimensions (for soil PSFs are 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 PSFs and the predicted soil PSFs at sites i and j. Both of them present that model performances are better when the values are lower. The standard deviation (SD) of prediction values and the ranges of 95 % confidence interval (CI) (Streiner, 1996) of indicators derived from 30 times model running to assess model uncertainty.

2.7 Statistical analysis for the original and log ratio transformed data

The standard deviation (SD), coefficient of variation (CV), mean, minimum (Min), maximum (Max), median absolute deviation (MAD), skewness (Skew), kurtosis and Kolmogorov-Smirnov (k-s) test ($p > 0.05$) were employed for descriptive statistical analysis of the original and log ratio transformed data. The means of log ratio transformed data are calculated as follows: (1) transform the data using a log-ratio method, (2) calculate the mean values of transformed values (ALRs, CLRs or ILRs), (3) back-transform the calculated mean values to the initial closed space. Furthermore, multivariate median based on depth measures (Bedall and Zimmermann, 1979; Small, 1990) were used because of the sum-constraint of compositional soil PSF data. The arithmetic mean of log ratio transformation data should be back-transformed to the original space. For $\mathbf{X} = [X_1, \ldots, X_n]$, the MAD can be calculated according to the Eq. (22) as below:

$$MAD(\mathbf{X}) = median(|X_i - median(\mathbf{X})|). \quad (22)$$

3 Results

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

For the original data of sand content, the mean (30.64 %) was much higher than that of median center (26.06 %). In contrast, silt and clay contents were the opposite, with lower means (silt, 55.79 %; clay, 13.57 %) than median centers (silt, 59.51 %; clay, 14.43 %). For the log ratio transformed data, different log ratio methods delivered the same means for sand, silt and clay. Additionally, the means of sand (28.69 %) and silt (60.54 %) were closer to the median centers of the original data except for clay (10.78 %). For SD and CV, soil PSFs data in the log ratio geometry had more stability and less variability compared with the original data. ILR and CLR had the lowest MAD for the first component (0.66) and the second component (0.43), respectively (Fig. 2). Although the *p* values of the original and different log ratio transformed data were not significant, log ratios made the data more symmetric according to the skews (Fig. 2). All log ratio transformed data had lower skews (ALR: 0.77, CLR: 0.88, ILR: -1.20) than those of the original data (1.24) of the first component. All the kurtosis of log ratio transformed data were much higher compared with original data.

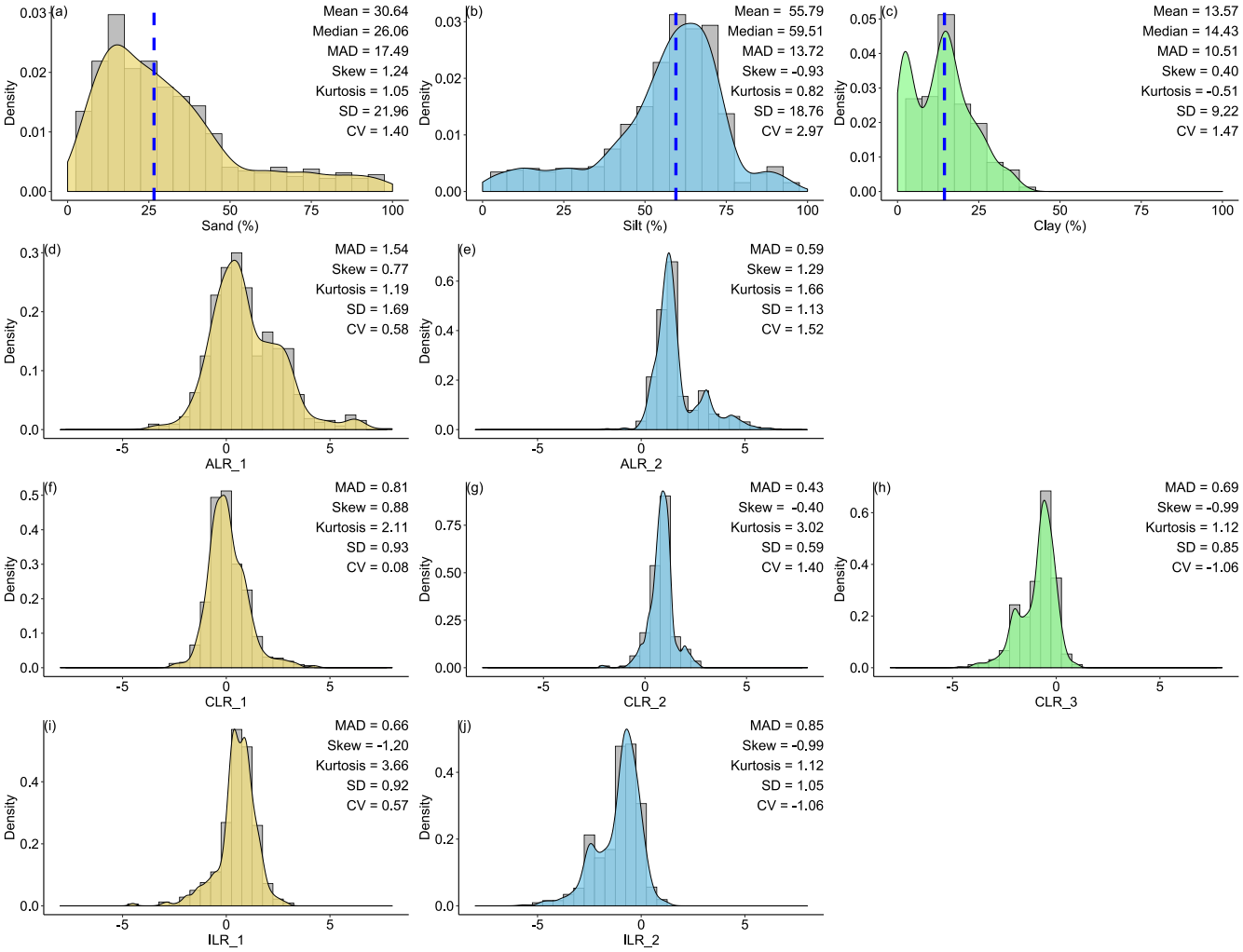


Figure 2. Descriptive statistical analysis for the original and log ratio transformed data of (a) sand, (b) silt, (c) clay, (d) ALR_1, (e) ALR_2, (f) CLR_1, (g) CLR_2,(h) CLR_3, (i) ILR_1 and (j) ILR_2. SD is standard deviation; CV is the coefficient of variation; the Median is multivariate median based on depth measures. ALR and ILR transformed S^3 (the simplex) to R^2 (the real space), and CLR transformed S^3 to R^3. Blue dashed lines showed the multivariate medians of original data.

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 all models ranged from 0.613 to 0.636. (Fig. 3a). RF had the highest overall accuracy (0.636) among the five models, followed closely by the accuracy of KNN (0.630) and MLP (0.627). In addition, SVM (0.618) and XGB (0.613) were relatively lower than other models. The highest kappa coefficient was generated from MLP (0.242), followed by

RF (0.238), XGB (0.229), KNN (0.213) and SVM (0.213) (Fig. 3b). For confusion indices (COIs), XGB (0.278) delivered the best performance, and RF (0.501) demonstrated the highest confusion of models (Fig. 3c).

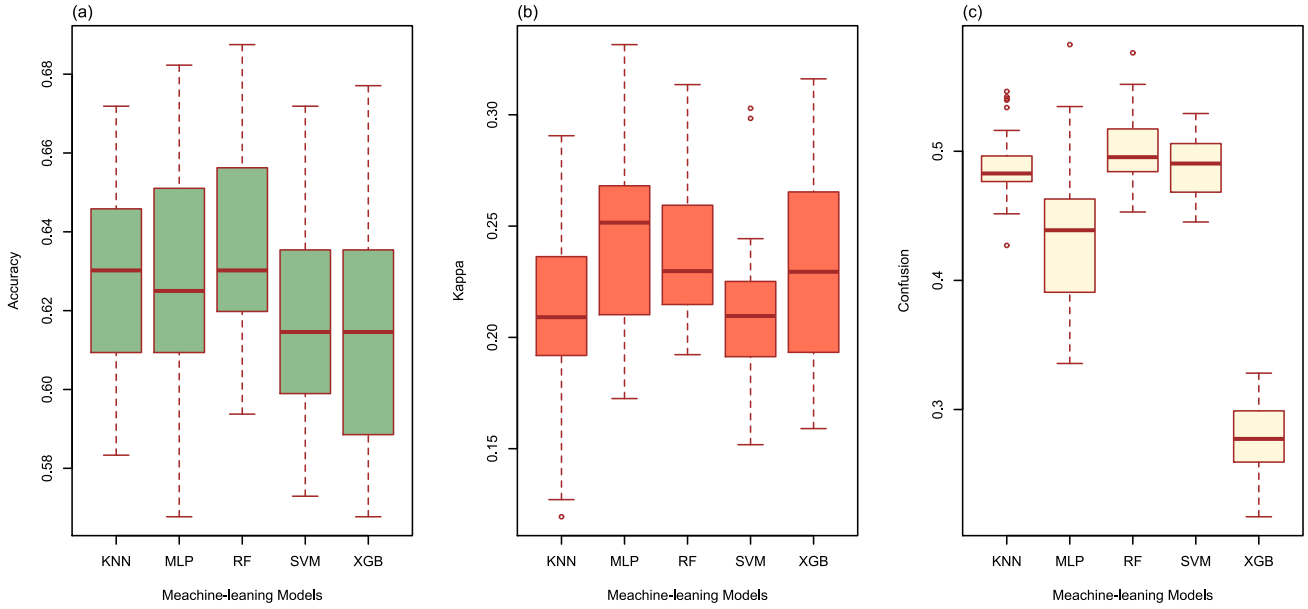


Figure 3. (a) The overall accuracy, (b) kappa coefficients and (c) confusion indices (COIs) for KNN, MLP, RF, SVM and XGB.

We combined the PRCs of the five machine-learning models to evaluate the performance of predicting each soil texture type using imbalanced data with different samples of each type (Fig. 4). The AUPRCs of the types with fewer positive examples were typically small, especially for SaClLo (only 4 samples), which delivered unsatisfying results. This was because the lack of soil sampling points made models learn poorly during the training process. In contrast, 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. For soil texture types with more samples, RF and XGB performed better. For soil texture classes with less samples, RF and SVM had better performance according to the AUPRCs.

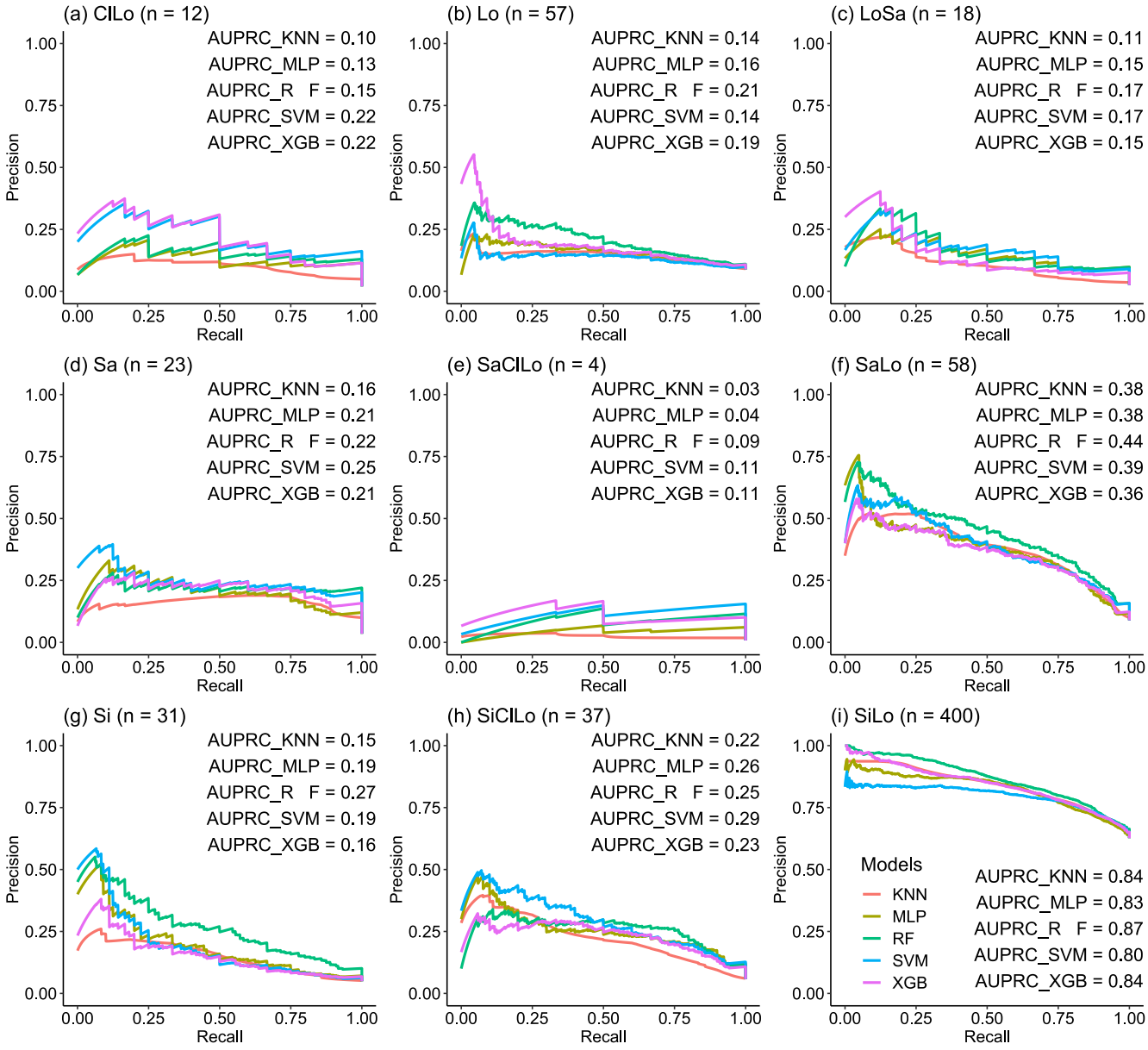


Figure 4. The AUPRCs for different machine-learning models in the prediction of each soil texture type (a) ClLo (b) Lo (c) LoSa (d) Sa (e) SaClLo (f) SaLo (g) Si (h) SiClLo (i) SiLo. n was the sampling points of different soil texture types.

3.2.2 Comparison of the prediction maps for soil texture classification

Prediction maps of soil texture types delivered quite different spatial distributions in the overall performance of different models (Fig. 5). The abundance indices pointed out that SVM can predicted all of 9 types, KNN and XGB predicted 8 of 9

types, followed closely by RF (7 of 9 types) and MLP (6 of 9 types). The maps predicted by RF, SVM and XGB illustrated that the main soil texture types in the northwest of the lower reaches of the 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 (Fig. 1).

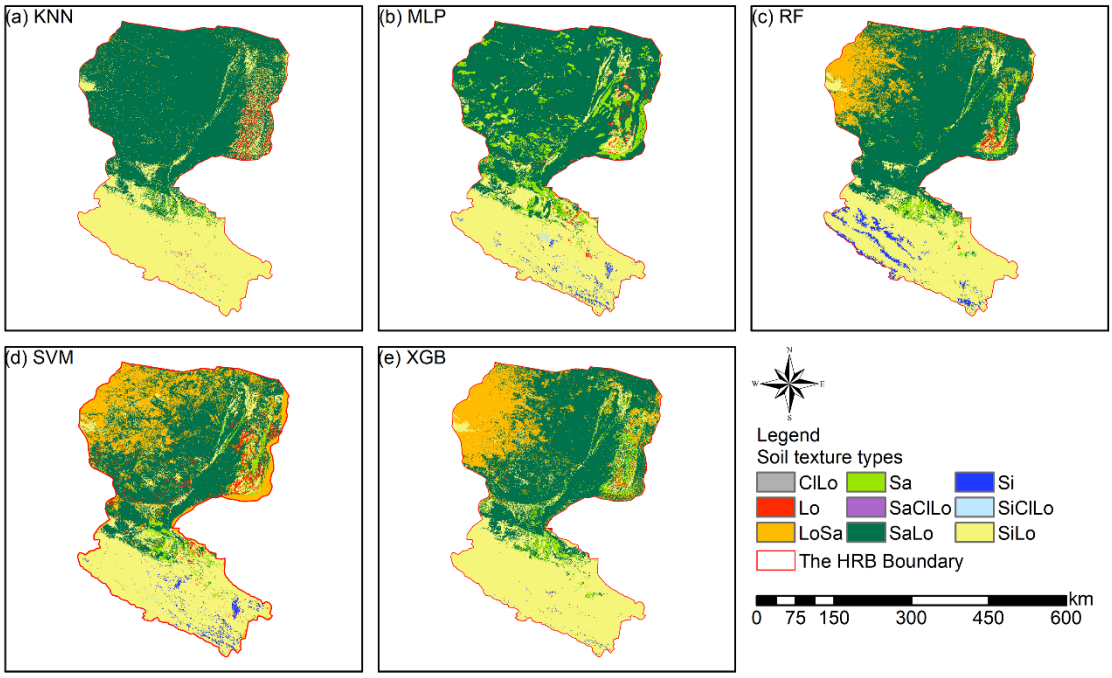


Figure 5. Soil texture classification prediction maps of different soil texture types of (a) KNN, (b) MLP, (c) RF, (d) SVM and (e) XGB.

3.3 Comparison of the machine-learning methods combined with log ratio transformed data in soil PSF interpolation

3.3.1 Comparison of the validation indicators for interpolation of soil PSFs

We compared the performance of each machine-learning model using the original and log ratio transformed data. The results indicated that the STRESS of the methods using log ratio transformed data were superior to these methods using original data (Table 2). The RMSE, MAE, RCC and AD generated from KNN, MLP, RF and XGB using original data outperformed the results using log ratio transformed data. By comparison among different log ratio transformed data of the same machine-learning model, ILR and CLR outperformed ALR. KNN_CLR demonstrated the most remarkable performance with the highest RCC and the lowest RMSE and MAE among KNN using the three log ratio transformation methods. Furthermore, RF and SVM using CLR and ILR transformed data generated relatively similar results. XGB_ILR showed the best performance with most of the indicators except for RMSE (6.75 %) and MAE (5.36 %) of clay, and STRESS (0.63). RF had the lowest RMSE

and MAE, the highest RCC, and the lowest AD and STRESS for ALR, CLR and ILR. For original data, RF also outperformed other models.

Table 2. The comparisons of accuracy of different machine-learning models combined with original and transformed data.

	RMSE (%)			MAE (%)			RCC			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	0.65	0.60	0.63	0.90	0.62
KNN_CLR	15.82	14.77	7.09	11.21	10.74	5.58	0.66	0.61	0.63	0.88	0.62
KNN_ILR	15.82	14.82	7.14	11.22	10.84	5.60	0.66	0.61	0.63	0.88	0.64
KNN_ORI	15.51	14.47	7.05	11.12	10.51	5.49	0.67	0.62	0.63	**0.84**	0.66
MLP_ALR	15.83	15.07	7.43	11.42	11.06	5.97	0.64	0.57	0.64	0.92	0.66
MLP_CLR	15.84	15.07	7.41	11.45	11.05	5.96	0.64	0.57	0.64	0.92	0.66
MLP_ILR	15.84	15.07	7.40	11.46	11.04	5.95	0.64	0.57	0.64	0.92	0.66
MLP_ORI	15.80	14.72	6.96	11.50	10.85	5.52	0.65	0.58	0.65	0.90	0.68
RF_ALR	15.50	14.43	6.62	10.90	10.52	5.24	**0.69**	0.65	0.68	0.86	**0.61**
RF_CLR	15.28	14.22	6.61	10.70	10.25	5.21	**0.69**	0.66	0.68	0.86	**0.61**
RF_ILR	15.27	14.25	6.66	10.66	10.26	5.26	**0.69**	0.66	0.68	0.86	**0.61**
RF_ORI	**15.09**	**13.86**	**6.31**	**10.65**	**9.99**	**5.00**	**0.69**	**0.67**	**0.69**	**0.84**	0.66
SVM_ALR	15.66	14.59	6.76	11.66	10.88	5.34	0.66	0.57	0.66	0.88	0.66
SVM_CLR	15.27	14.36	6.87	11.01	10.41	5.41	0.66	0.60	0.65	0.87	0.65
SVM_ILR	15.29	14.37	6.84	10.92	10.43	5.42	0.67	0.61	0.65	0.87	0.65
SVM_ORI	15.30	14.38	6.92	10.94	10.32	5.43	0.67	0.61	0.66	0.87	0.67
XGB_ALR	15.82	14.92	6.72	11.32	11.01	5.35	0.67	0.62	0.67	0.88	0.64
XGB_CLR	15.70	14.80	6.75	10.96	10.67	5.39	0.67	0.63	0.67	0.88	0.62
XGB_ILR	15.45	14.57	6.75	10.91	10.52	5.36	0.67	0.62	0.66	0.88	0.63
XGB_ORI	15.15	14.05	6.47	10.88	10.15	5.15	0.67	0.66	0.67	0.86	0.68

3.3.2 Comparison of the interpolation prediction maps of soil PSFs

Interpolation prediction maps of soil PSFs using log ratio transformed data (ILR) and original data were represented in Figs. 6, S4.1 and S4.2. The maps generated from ILR transformed data showed closer ranges to the original soil sampling data in terms of the ranges of sand (0.98 – 99.66 %), silt (0.17 – 95.87 %) and clay (0.03 – 39.77 %), and the texture features were more consistent with the distributions of the real environment (Figs. 6, S4.1 and S4.2). With respect to different machine-learning models, RF and XGB delivered prediction maps that were closer to the range of the distribution of original data than KNN, SVM and MLP.

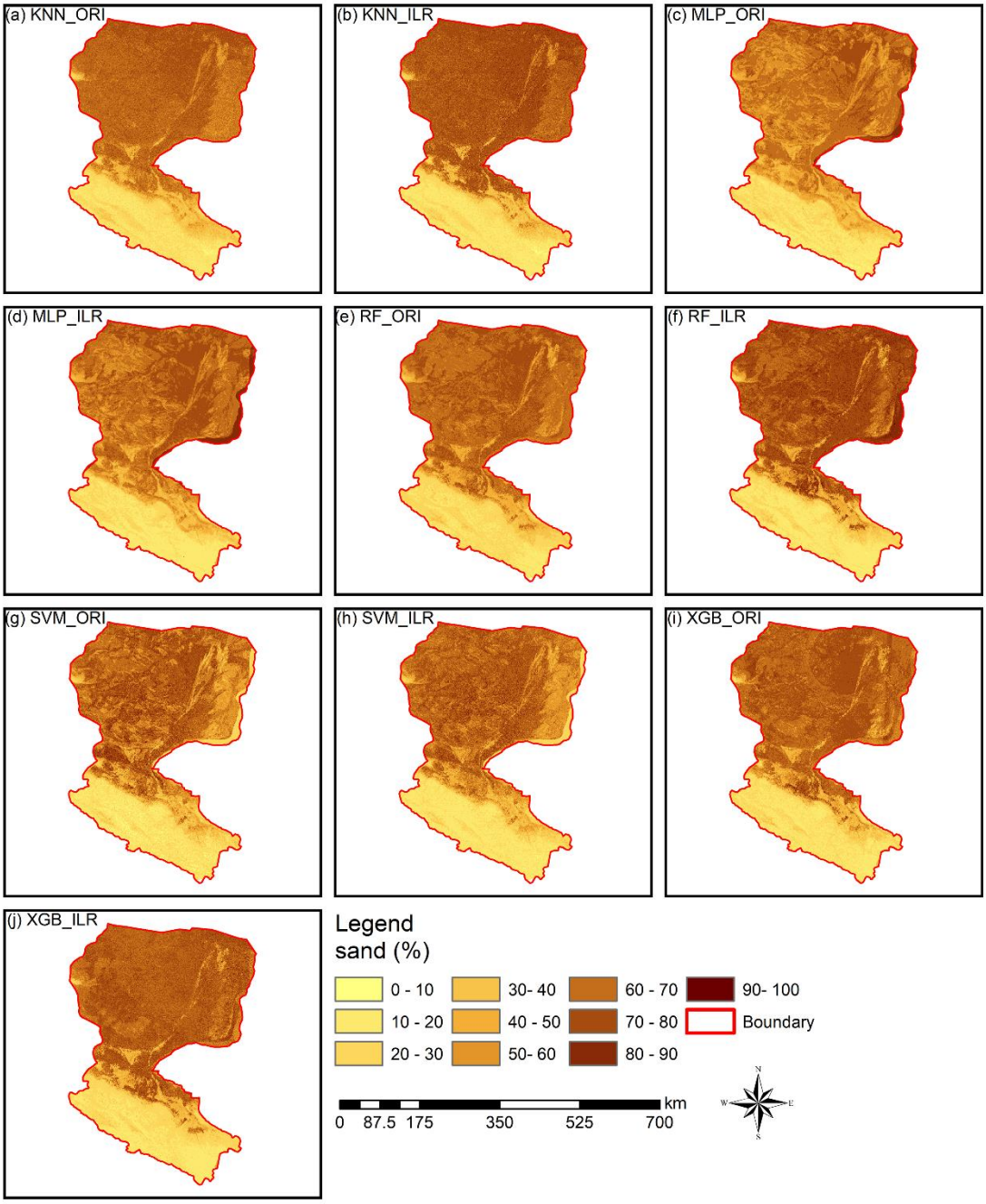


Figure 6. The prediction maps of sand fraction. All the ranges of prediction maps of sand (approximately 9.0 – 90.0 %) were within the range of original data (0.98 – 99.66 %). RF_ILR (7.9 – 94.7 %) and XGB_ORI (1.8 – 92.4 %) generated wider output distributions and were relatively closer to the range of the distribution of original data than other prediction maps, such

as KNN_ILR (7.3 – 88.6 %), KNN_ORI (7.8 – 80.8 %), MLP_ILR (8.8 – 90.8 %), MLP_ORI (9.0 – 90.3 %), RF_ORI (9.0 – 81.0 %), SVM_ILR (6.5 – 85.6 %), SVM_ORI (7.3 – 90.0 %) and XGB_ILR (5.0 – 88.5 %).

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

The overall accuracy and kappa coefficients of indirect classification were improved by using log ratio transformed data, especially for RF and XGB (Fig. 7). ILR showed the highest overall accuracy among three log ratio transformations, which also demonstrated the best performance in terms of kappa coefficients, except for MLP. We also compared direct classification with indirect classification and found that the differences of overall accuracy of direct and indirect classification were negligible. However, the kappa coefficients were greatly modified using indirect classification compared with direct classification except for MLP; peculiarly, RF_ILR increased the kappa coefficient to 0.291 (21.3 % improvement) and kept the accuracy stable.

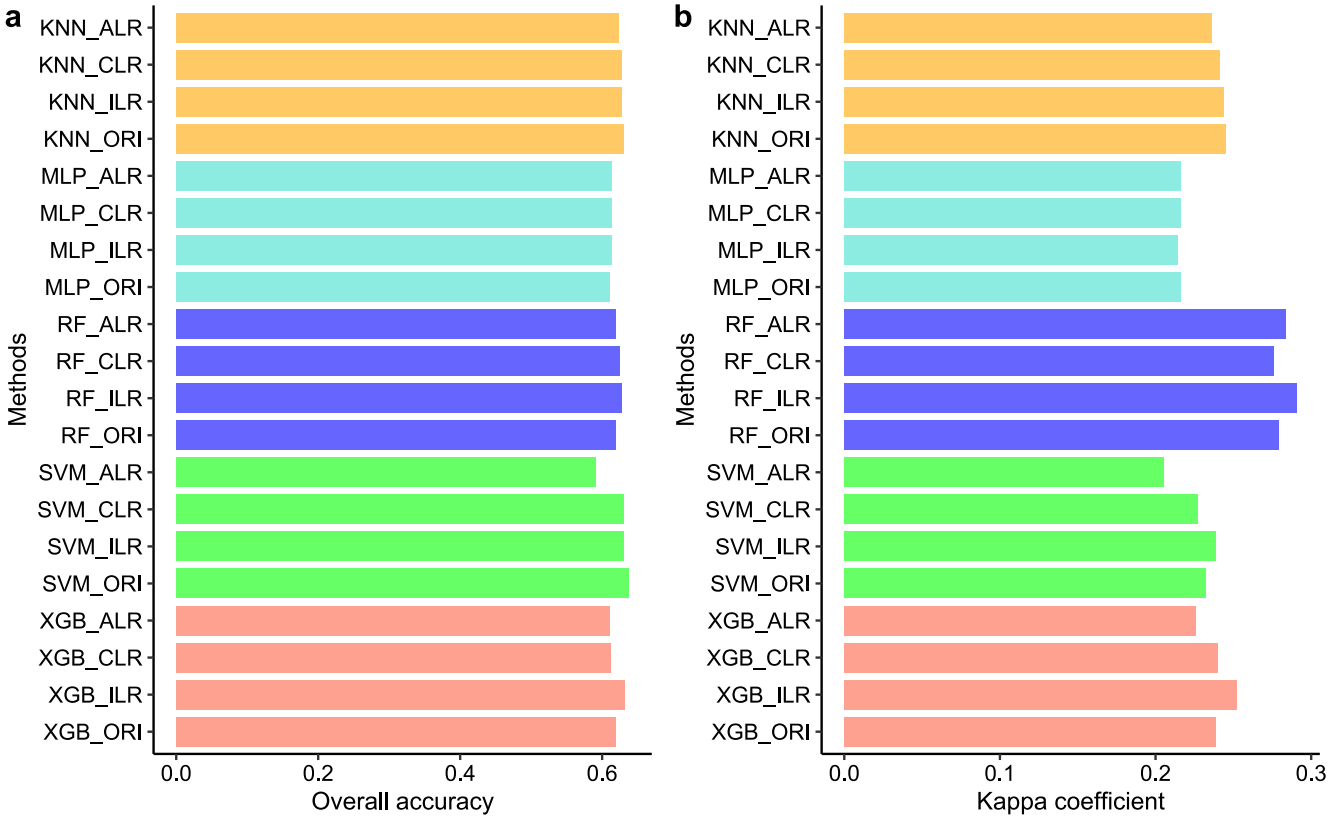
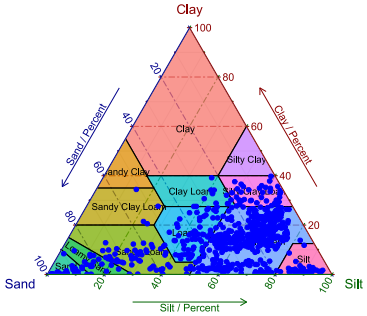


Figure 7. Overall accuracy and kappa coefficients calculated from soil texture classification by soil PSF interpolation using five machine-learning models combined with original data and log ratio transformed data.

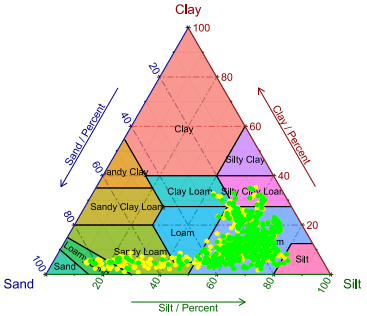
3.4.2 The prediction performance of soil texture types from different methods

The distributions of soil texture types using original and ILR transformed data were illustrated in the United States Department of Agriculture (USDA) soil texture triangle (Fig. 8). The triangle of the original data of soil PSFs (Fig. 8a) demonstrated wider ranges of spatial dispersion than the interpolated data using machine-learning methods. These predictions revealed the properties of aggregating from the sides to the center of triangles. With respect to the machine-learning models, RF showed the most dispersed feature in accordance with the original soil PSF data. The predictions from models combined with ILR transformed data were more discrete and more associated with the original soil PSF data than those resulting from ORI methods. The prediction results represented significant differences in the error ratio (yellow color) of soil sampling points in types of between the left part (LoSa, SaLo and Lo) and right part of the triangles (SiLo and Si) for most of the models, especially for KNN and MLP. The log ratio methods over-calculated the mean value of silt in the process of transformation (Fig. 2), so 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 turning to 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 method (83.9%) compared with ORI (81.7%). In the case of other models, the differences between ORI and ILR 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 for RF and XGB. However, we removed this evaluation indicator because the same data sets were employed in the processes of training and predicting.

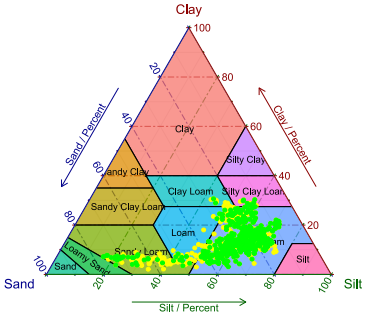
(a) Soil PSFs sampling data



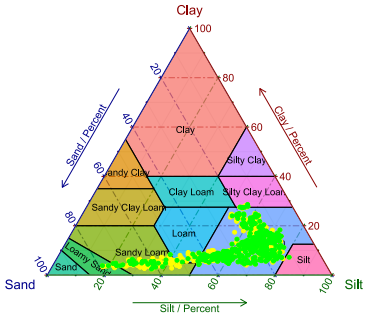
(b) KNN_ILR (65.0%)



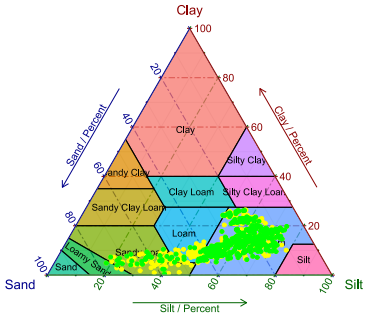
(c) KNN_ORI (65.9%)



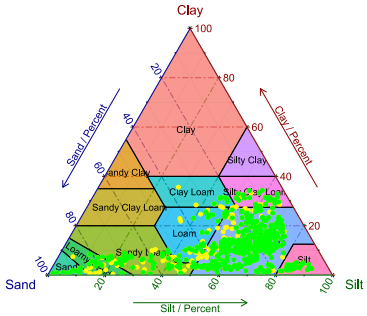
(d) MLP_ILR (63.3%)



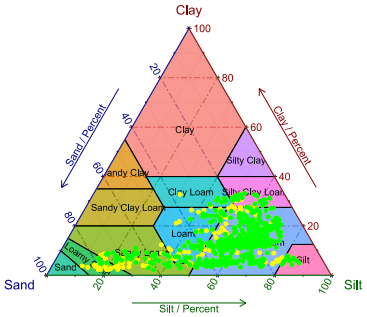
(e) MLP_ORI (63.6%)



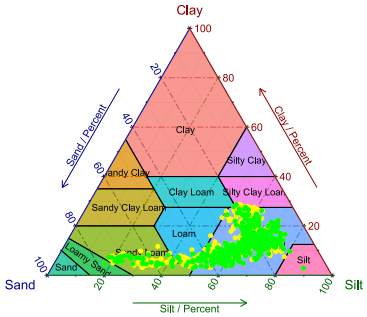
(f) RF_ILR (83.9%)



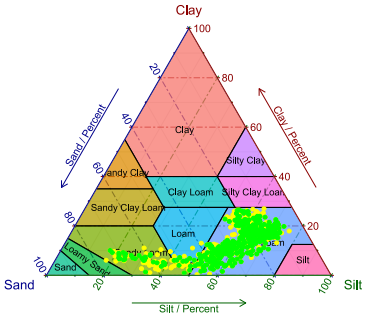
(g) RF_ORI (81.7%)



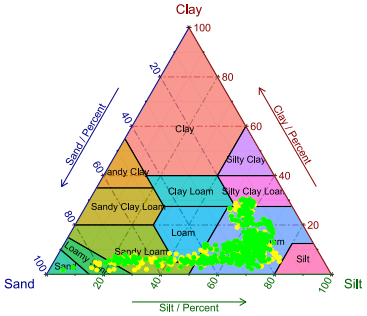
(h) SVM_ILR (66.1%)



(i) SVM_ORI (66.4%)



(j) XGB_ILR (67.8%)



(k) XGB_ORI (68.0%)

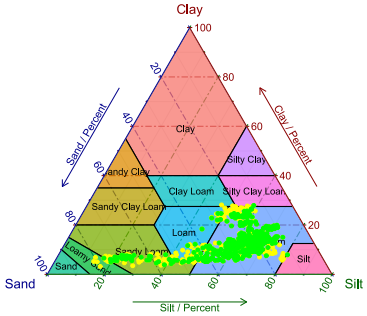


Figure 8. Soil texture types of 640 soil samples shown in the USDA texture triangle. The results of soil PSFs were generated from (a) soil PSF samples (b) KNN_ILR, (c) KNN_ORI, (d) MLP_ILR, (e) MLP_ORI, (f) RF_ILR, (g) RF_ORI, (h) SVM_ILR, (i) SVM_ORI, (j) XGB_ILR, and (k) XGB_ORI. Note that right points (green) means that the predicted and original soil texture classes are the same, and the rest of points (yellow) are those related to misclassification of the soil texture classes. The predicted right ratios (RRs) of the soil texture classes were in the bracket after interpolators in plots.

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

The soil texture maps predicted using original data were different from that generated using log ratio transformed data, and classification maps of the machine-learning models combined with the log ratio transformed data had more detailed information (Figs. 9 and S5.1). The results of machine-learning models using three log ratio transformed data were similar to the number of predicted types; however, there were significant differences between using original data and log ratio transformed data. All machine-learning models combined with original data predicted more Lo and SaLo, and fewer LoSa and Si (Fig. 9). We also compared the prediction of soil texture types by direct classification (Fig. 5) with those generated from indirect classification using the same machine-learning models, revealing different distributions of LoSa existed among them on the lower reaches of Heihe River Basin. For the upper reaches, prediction maps of ILR methods generated more Si and less Lo than ORI method. Si were mainly distributed on the middle and southeast of the upper reaches of the HRB in the predictions combined with ILR methods. For the middle reaches, ILR prediction maps were recommended and more in line with the real environment than the ORI methods, because more SaLo and less Lo on the middle reaches of the HRB were predicted. Furthermore, predicted soil texture of indirect methods were more abundant than the direct one on the middle reaches (Fig .5).

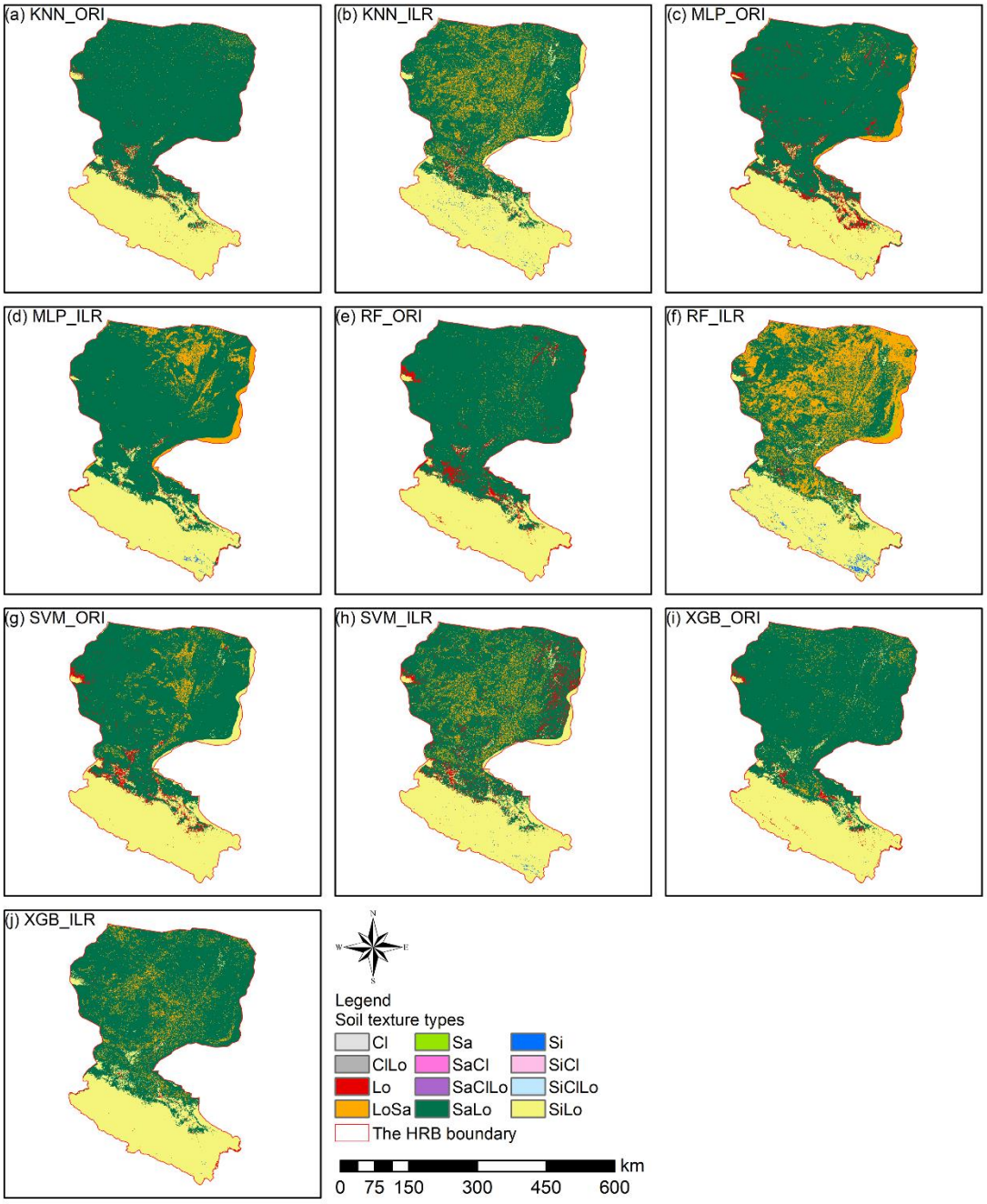


Figure 9. The prediction maps of soil texture classification by indirect methods using KNN, MLP, RF, SVM and XGB using ILR transformed data or original data.

3.4.4 Comparison of total computing time for each model in soil texture classification and soil PSF interpolation

Running time of the models were computed and compared for different machine-learning models in soil texture classification and soil PSF interpolation (Fig. 10). Because the time spent of ORI and log ratio methods were similar, time spent of ILR was selected for soil PSF interpolation. For different models, RFs required the longest time for both classification (453.73 s) and interpolation (188.87 s), which may cause it to lose advantages in processing large data sets. KNN (classification, 4.2 s; interpolation, 23.6 s) and SVM (classification, 4.15 s; interpolation, 12.4 s) showed shorter time in both classification and interpolation. XGB (classification, 21.6 s; interpolation, 17.13 s) was much more stable and spent less time, and the data processes were simpler compared with MLP (classification: 47.28 s, interpolation: 152.31 s). Moreover, XGB delivered better performance than KNN and SVM in prediction maps, demonstrating that it is an effective way of dealing with large data sets.

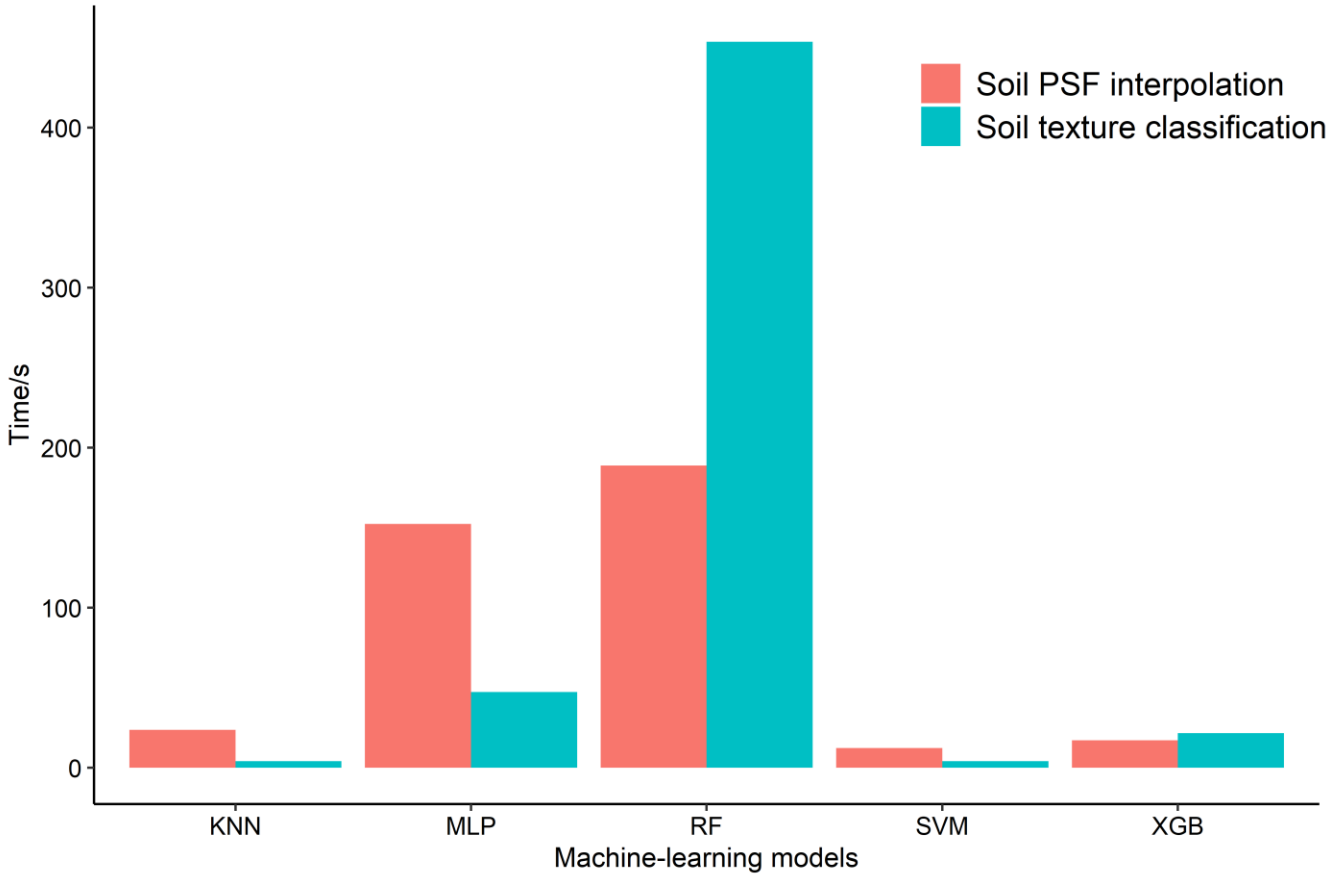


Figure 10. Average time spent running 30 times of KNN, MLP, RF, SVM and XGB of soil texture classification and soil PSF interpolation.

4 Discussion

4.1 The systematic comparison of the five machine-learning models

The range of applicability of the study is limited to independent modeling, i.e., the component-wise approaches. However, joint fractions modeling could lead to different results. We found that tree-based machine-learning models – RF and XGB delivered better performance than KNN, MLP and SVM, which also has concluded by Heung et al. (2016). For the total computing time, RF revealed the longest time with respect to both classification and interpolation mode. In addition, for trade-offs between the total computing time of model and accuracy, XGB was superior to the other four models, reducing the computing time significantly while maintaining the acceptable accuracy. In fact, 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 models (Chen and Guestrin, 2016). Therefore, XGB was recommended with sub-optimal accuracy but fast at the expense of a loss in accuracy, which can be selected when researchers deal with large data sets in study areas. Moreover, some joint fractions approaches – compositional kriging (Wang and Shi, 2017), High Accuracy Surface Modeling (HASM) (Yue et al., 2015; Yue et al., 2016) and the Dirichlet regression (Hijazi and Jernigan, 2009) – can consider the multivariate treatment for soil PSFs using a joint model, but machine-learning models are more convenient to combine environmental covariables. For the machine-learning models in our study, KNN, MLP, RF and SVM also can be applied to multivariate vectors combined with log ratio methods. For example, the Multivariate Random Forest (MRF) method, which is the extended version of RF, calculates predictions of all output features using single model (Segal and Xiao, 2011).

4.2 The systematic comparison of the models using log ratio transformed data and original data

Log ratio transformation methods can open the data and remove the “closure effect”, which induce spurious correlation. The opened data can be interpolated into the mapping area and then the results can be back-transformed using inverse equations. However, in the process of parameters optimization, the optimal parameters of different machine-learning models were obtained using log ratio transformed data, which cannot guarantee the most accurate back-transformed results. This is because that the values of assessment indicators (e.g. MAEs, RMSEs, etc.) will remain stable with limited differences due to the small value range of log ratio transformed data. Therefore, when the prediction values of log ratio methods were back-transformed to the real space, these values of indicators will be enlarged.

Due to the contraction of the predicted values (Fig. 8), there were small numbers of predictions beyond the range of original data value, including the negative predictions using ORI data. Although these few negative predictions can be eliminated by parameter adjustment in our study, there is still a drawback using ORI data. Among the three log ratio methods, ILR and CLR were superior to ALR, which can be explained as ILR and CLR were isometric transformations and they could preserve distances (Filzmoser and Hron, 2009). Moreover, ALR has been criticized because the results were affected by the subjective

choice of the denominator. In addition, slightly better performance of ILR than CLR was obtained, because in CLR, the geometric mean composed of all compositions of soil PSFs is the denominator, and one-to-one mapping of equations and soil PSFs could be implemented. Nevertheless, the sum of the dimensions of CLR is 0, the problem of collinear is still present. ILR transformed all the information into D-1 orthogonal log contrasts (so-called balances) (Egozcue et al. 2003) and overcame the data collinearity and sub-compositional incoherence in CLR by using an appropriate choice of the basis (Egozcue and Pawlowsky-Glahn, 2005). Moreover, in ILR method, multiple sets of ILR transformed data can be generated by permutations of components (different sequential binary partitions, SBPs) in compositional data, and different choices of ILR balances influenced the model accuracy. The choice of a specific SBP for compositions is crucial for the intended interpretation of coordinates (Fiserova and Hron, 2011). The choice of SBPs can be applied blindly (Fiserova and Hron, 2011), or based on priori expert knowledge or using a compositional biplot (Lloyd et al., 2012), and the best ILR balance also can be chosen using variograms and cross-variograms (Molayemat et al., 2018). All three SBPs were demonstrated in Supplementary Section S6 (Table S6.1). The ILR balance chose in our study was SBP1, because the ILR transformed data using SBP1 were more symmetric than other two SBPs. However, there will be different results and prediction maps when different SBPs are used, which need further research. Furthermore, each component of log ratio or original soil PSF data were independently modeled using component-wise approaches (machine-learning methods), that may be sub-optimal compared with joint fractions approach under the circumstances dealing with the multivariate treatment. For example, CLR transformed data are still characterized by collinearity, but there is no guarantee that the sum of three components of CLR is 0 due to the use of independent modeling. Although the final predictions were not influenced (still sum to 100 %) since the inverse equations for CLR, collinear constraints reduced the prediction accuracy. By contrast, the ILR method is more meaningful and appropriate than the other log ratio methods because it indeed removes the data constraints. Therefore, ILR is more recommended as a combination method with machine-learning models for component-wise modeling, unless multivariate extensions of the methods (e.g., functional compositions) are considered.

4.3 The systematic comparison of the direct and indirect soil texture classification

Compared with the real soil texture distribution and environment of the HRB, SiLo overlaid the upper reaches of the HRB, and SaLo and Lo were in the south of the upper reaches of the HRB, showing a strip distribution. Moreover, an uncovered area was detected in the northwest of the lower reaches of the HRB, where it cannot be predicted accurately due to a lack of information input in the process of model training. The main soil texture types on the lower reaches of the HRB were SiLo, LoSa and small areas of SaLo and Lo, which distributed in the uncovered area. The main soil texture types predicted from 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 the HRB was SaLo, which was not fitted for the real environment (LoSa). In fact, LoSa and SaLo were obviously the most confusing. However, they are fairly similar to each other (Fig. 8). In addition, because of the limitation of the training subsets, direct

classification can only predict types which contained in training subsets. In contrast, indirect classification broke such limitations, and new prediction types arose due to the transformation from soil PSFs to soil texture types. Moreover, more suitable matching performance with the real environment should be considered such as the log ratio methods of MLP and RF models, KNN_ ALR, KNN_ ILR and XGB_CLR.

5 Conclusion

We systematically compared five machine-learning models using original data and three log ratio transformed data in the HRB for direct and indirect soil texture classification and soil PSF interpolation. As flexible and stable models, tree learners – RF models delivered powerful performance in both classification and interpolation and were superior to other machine-learning models mentioned above. As a new and sub-optimal machine-learning method in soil science, XGB appeared to be more computationally efficient in processing large data sets. RF and XGB were recommended to evaluate classification capacity of imbalanced data. In addition, the log ratio methods especially ILR had advantages of modifying STRESS in soil PSF interpolation. Moreover, the indirect methods for soil texture classification outperformed the direct ones, especially when combined with the log ratio transformations. The indirect methods for soil texture classification generated preferable results in both cases of accuracy indicators and prediction maps. The keys to improve the interpolator accuracy are using more appropriate interpolation techniques with environmental covariates, transforming soil PSF data by more efficient transformed methods, using compositional data analysis in the multivariate studies, and using systematic parameter adjustment algorithms for compositional data.

Data availability. The 640 soil sampling data of the HRB (http://westdc.westgis.ac.cn/DOI:10.3972/heihe.009.2013.db; http://westdc.westgis.ac.cn/DOI:10.3972/heihe.009.2013.db; http://westdc.westgis.ac.cn/DOI:10.3972/heihe.00135.2016.db; http://westdc.westgis.ac.cn/DOI:10.3972/hiwater.147.2013.db; http://westdc.westgis.ac.cn/DOI:10.3972/heihe.037.2014.db; http://westdc.westgis.ac.cn/DOI:10.3972/heihe.0034.2013.db; http://westdc.westgis.ac.cn/DOI:10.3972/heihe.093.2013.db) and part of environmental covariates data can be accessed through http://westdc.westgis.ac.cn/ (last access: 14 March 2020). The meteorological data can be accessed through http://data.cma.cn/ (last access: 14 March 2020).

Author contributions. Wenjiao Shi contributed to soil data sampling, oversaw the design of the entire project. Mo Zhang performed the analysis and wrote the manuscript, and Ziwei Xu collected and analyzed data. All 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 Key Research and Development Program of China (No. 2017YFA0604703), the National Natural Science Foundation of China (Grant No. 41771364 and 41771111), Fund for Excellent Young Talents in Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (2016RC201), the Youth Innovation Promotion Association, CAS (No. 2018071), Investigation and monitoring project of Ministry of Natural Resources (JCQQ191504-06) and a grant from State Key Laboratory of Resources and Environmental Information System.

References

Abdi, D., Cade-Menun, B. J., Ziadi, N., and Parent, L. E.: Compositional statistical analysis of soil 31P-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, Chapman and Hall, 139-177 pp., 1982.

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

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.

Bedall, F. K., and Zimmermann, H.: Algorithm as 143: The mediancentre, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 325-328, https://doi.org/10.2307/2347218, 1979.

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, https://doi.org/10.1016/S0166-2481(06)31025-2, 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, https://doi.org/10.18637/jss.v046.i07, 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.

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.

Burrough, P. A., van Gaans, P. F. M., and Hootsmans, R.: Continuous classification in soil survey: Spatial correlation, confusion and boundaries, Geoderma, 77, 115-135, https://doi.org/10.1016/S0016-7061(97)00018-9, 1997.

Butler, J. C.: Effects of closure on the moments of a distribution, Journal of the International Association for Mathematical Geology, 11, 75-84, https://doi.org/10.1007/bf01043247, 1979.

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, https://doi.org/10.1145/2939672.2939785, 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: 14 March 2020), 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) version 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.

Egozcue, J. J., and Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis, Mathematical Geology, 37, 795-828, https://doi.org/10.1007/s11004-005-7381-9, 2005.

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.

Fiserova, E., and Hron, K.: On the interpretation of orthonormal coordinates for compositional data, Mathematical Geosciences, 43, 455-468, https://doi.org/10.1007/s11004-011-9333-x, 2011.

Filzmoser, P., and Hron, K.: Correlation analysis for compositional data, Mathematical Geosciences, 41, 905-919, https://doi.org/10.1007/s11004-008-9196-y, 2009.

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., Xu, F., Zhang, B., and Yi, L.: 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.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., and Graeler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, Peerj, 6, https://doi.org/10.7717/peerj.5518, 2018.

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.

Hijazi, R., and Jernigan, R.: Modelling compositional data using Dirichlet regression models, Journal of Applied Probability and Statistics, 4, 77-91, 2009.

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.

Kuhn, M.: Caret: Classification and regression training, R package version 6.0-80, available at: https://CRAN.R-project.org/package=caret (last access: 14 March 2020), 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.

Lloyd, C. D., Pawlowsky-Glahn, V., and Jose Egozcue, J.: Compositional data analysis in population studies, Annals of the Association of American Geographers, 102, 1251-1266, https://doi.org/10.1080/00045608.2011.652855, 2012.

Liaw, A., and Wiener, M.: Classification and regression by randomforest, R News, 2, 18-22, https://CRAN.R-project.org/doc/Rnews/, 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.

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.

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, https://doi.org/10.1002/2015wr018369, 2016b.

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, 2016a.

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: 14 March 2020), 2017.

Mishra, S., and Datta-Gupta, A.: Chapter 2 - Exploratory data analysis, in: Applied Statistical Modeling and Data Analytics, edited by: Mishra, S., and Datta-Gupta, A., Elsevier, 15-29, https://doi.org/10.1016/B978-0-12-803279-4.00002-X, 2018.

Molayemat, H., Torab, F. M., Pawlowsky-Glahn, V., Morshedy, A. H., and Jose Egozcue, J.: The impact of the compositional nature of data on coal reserve evaluation, a case study in Parvadeh IV coal deposit, Central Iran, International Journal of Coal Geology, 188, 94-111, https://doi.org/10.1016/j.coal.2018.02.003, 2018.

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.

Reimann, C., and Filzmoser, P.: Normal and lognormal data distribution in geochemistry: Death of a myth. Consequences for the statistical treatment of geochemical and environmental data, Environmental Geology, 39, 1001-1014, https://doi.org/10.1007/s002549900081, 2000

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: 14 March 2020), 2016.

Segal, M., and Xiao, Y. Y.: Multivariate random forests, Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 1, 80-87, https://doi.org/10.1002/widm.12, 2011.

Small, C. G.: A survey of multidimensional medians, International Statistical Review, 58, 263-277, https://doi.org/10.2307/1403809, 1990.

Song, X., Brus, D. J., Liu, F., Li, D., Zhao, Y., Yang, J., and Zhang, G.: 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.

Streiner, D. L.: Maintaining standards: Differences between the standard deviation and standard error, and when to use each, Canadian Journal of Psychiatry-Revue Canadienne De Psychiatrie, 41, 498-502, https://doi.org/10.1177/070674379604100805, 1996.

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.

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, https://doi.org/10.1016/B978-0-12-386941-8.00021-6, 2012.

Tolosana-Delgado, R., Mueller, U., and van den Boogaart, K. G.: Geostatistics for compositional data: An overview, Mathematical Geosciences, 51, 485-526, https://doi.org/10.1007/s11004-018-9769-3, 2019.

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., https://doi.org/10.1007/978-1-4615-5703-6_3, 1998.

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., He, X., Ma, R., Liu, H., and Lv, J.: 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.

Xu, T., He, X., Bateni, S. M., Auligne, T., Liu, S., Xu, Z., Zhou, J., and Mao, K.: Mapping regional turbulent heat fluxes via variational assimilation of land surface temperature data from polar orbiting satellites, Remote Sensing of Environment, 221, 444-461, https://doi.org/10.1016/j.rse.2018.11.023, 2019.

Yang, R., Zhang, G., Liu, F., Lu, Y., Yang, F., Yang, F., Yang, M., Zhao, Y., and Li, D.: 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, https://doi.org/10.11766/trxb201402180069, 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.

Yue, T., Zhang, L., Zhao, N., Zhao, M., Chen, C., Du, Z., Song, D., Fan, Z., Shi, W., Wang, S., Yan, C., Li, Q., Sun, X., Yang, H., Wilson, J., and Xu, B.: A review of recent developments in HASM, Environmental Earth Sciences, 74, 6541-6549, https://doi.org/10.1007/s12665-015-4489-1, 2015.

Yue, T., Liu, Y., Zhao, M., Du, Z., and Zhao, N.: A fundamental theorem of Earth's surface modelling, Environmental Earth Sciences, 75, 751, https://doi.org/10.1007/s12665-016-5310-5 , 2016.

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.

Zhang, S., Shen, C., Chen, X., Ye, H., Huang, Y., 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.

Zhang, X., Liu, H., Zhang, X., Yu, S., Dou, X., Xie, Y., and Wang, N.: Allocate soil individuals to soil classes with topsoil spectral characteristics and decision trees, Geoderma, 320, 12-22, https://doi.org/10.1016/j.geoderma.2018.01.023, 2018.