



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

Mo Zhang<sup>1,2</sup>, Wenjiao Shi<sup>1,3</sup>, and Ziwei Xu<sup>4</sup>

<sup>1</sup>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

<sup>2</sup>School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China

<sup>3</sup>College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

<sup>4</sup>State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

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

Received: 5 December 2019 – Discussion started: 6 January 2020

Revised: 14 March 2020 – Accepted: 26 March 2020 – Published: 14 May 2020

**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 have systematically compared their performance with respect to both classification and interpolation. Here, five machine-learning models – K-nearest neighbour (KNN), multilayer perceptron neural network (MLP), random forest (RF), support vector machines (SVM), and extreme gradient boosting (XGB) – combined with the original data and three log-ratio transformation methods – additive log ratio (ALR), centred log ratio (CLR), and isometric log ratio (ILR) – were applied to evaluate soil texture and PSFs using both raw and log-ratio-transformed data from 640 soil samples in the Heihe River basin (HRB) in China. The results demonstrated that the log-ratio transformations decreased the skewness of soil PSF data. For soil texture classification, RF and XGB showed better performance with a higher overall accuracy and kappa coefficient. They were also recommended to evaluate the classification capacity of imbalanced data according to the area under the precision–recall curve (AUPRC). For soil PSF interpolation, RF delivered the best performance among five machine-learning models with the lowest root-mean-

square error (RMSE; sand had a RMSE of 15.09 %, silt was 13.86 %, and clay was 6.31 %), mean absolute error (MAE; sand had a MAD of 10.65 %, silt was 9.99 %, and clay was 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 was 0.69, silt was 0.67, and clay was 0.69). STRESS was improved by using log-ratio methods, especially for CLR and ILR. Prediction maps from both direct and indirect classification were similar in the middle and upper reaches of the HRB. However, indirect classification maps using log-ratio-transformed data provided more detailed information in the lower reaches of the HRB. There was a pronounced improvement of 21.3 % in the kappa coefficient when using indirect methods for soil texture classification compared with direct methods. RF was recommended as the best strategy among the five machine-learning models, based on the accuracy evaluation of the 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 over other models when the trade-off between the accuracy and runtime was considered. Our findings provide a reference for future works with respect to the spatial prediction

of soil PSFs and texture using machine-learning models with skewed distributions of soil PSF data over 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, and so on. 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 the estimation of the 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 over 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 have shown 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 a low bias when dealing with large data sets (Hengl et al., 2015, 2017). Among machine-learning models, artificial neural networks and “tree learners” (e.g. decision trees) have been preferred due to their relatively high overall accuracy and kappa coefficients, the interpretability of the results, and the speed of the parameterization in the prediction of soil classes (Taghizadeh-Mehrjardi et al., 2015; Heung et al., 2016). Most previous studies have used machine-learning algorithms to simulate soil category or continuous properties for classification or regression problems. However, few studies have systematically analysed 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 are also compositional data – thus, the constant sum (1 % or 100 %) should be guaranteed. Soil PSF data are typical compositional data with three components that are not independent of each other but are

rather expressed as a percentage (Filzmoser et al., 2009). Because of the spurious correlations between components, different results 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 have been 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 has been a combination of log-ratio transformation methods, including the additive log ratio (ALR; Aitchison, 1982), the centred log ratio (CLR; Aitchison, 1982), and the isometric log ratio (ILR; Egozcue et al., 2003). Soil PSFs have been 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), as 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 for 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), which can be jointly applied to the fractions to fully exploit the richness of information. Menafoglio et al. (2016b) applied such functional-compositional data for the stochastic simulation of PSCs based on a geostatistical Monte Carlo and Bayes space approach combined with a CLR transformation method in heterogeneous aquifer systems in hydrogeology, demonstrating a remarkable improvement of the characterization of the spatial variability and uncertainty compared with traditional methods. However, most soil PSF data used in studies are discrete (i.e. sand, silt, and clay), and few studies have conducted a systematic comparison of the accuracy, strengths, and weaknesses of 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 can also 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 (ANNs; 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 modelled PSFs at a 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 have systematically compared the different machine-learning models for both direct and indirect soil texture classification.

In our study, five machine-learning models – K-nearest neighbour (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, including Gansu and Qinghai provinces, and is the second largest inland river basin in China with an area of 146 700 km<sup>2</sup> (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 ranges from −5 to 4 °C, and 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), and the mean annual temperature ranges 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 an annual precipitation of under 50 mm and an annual average evaporation of above 3500 mm; the mean annual temperature ranges 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 the northwest. The main vegetation types are alpine vegetation (4000–5000 m), the alpine meadow vegetation belt (3000–4000 m), alpine shrub meadow (3200–3800 m), the mountain forest meadow belt (2400–3200 m), the mountain grassland belt (1800–2400 m), and the desert base belt (less than 1800 m). The main vegetation types in 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 the 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), grey cinnamon soil and Chernozem (3200–3600 m), Sierozem and grey cinnamon soil (2600–3200 m), grey cinnamon soil (2300–2600 m), and Sierozem (1900–2300 m) in the upper reaches of the HRB. The main soil

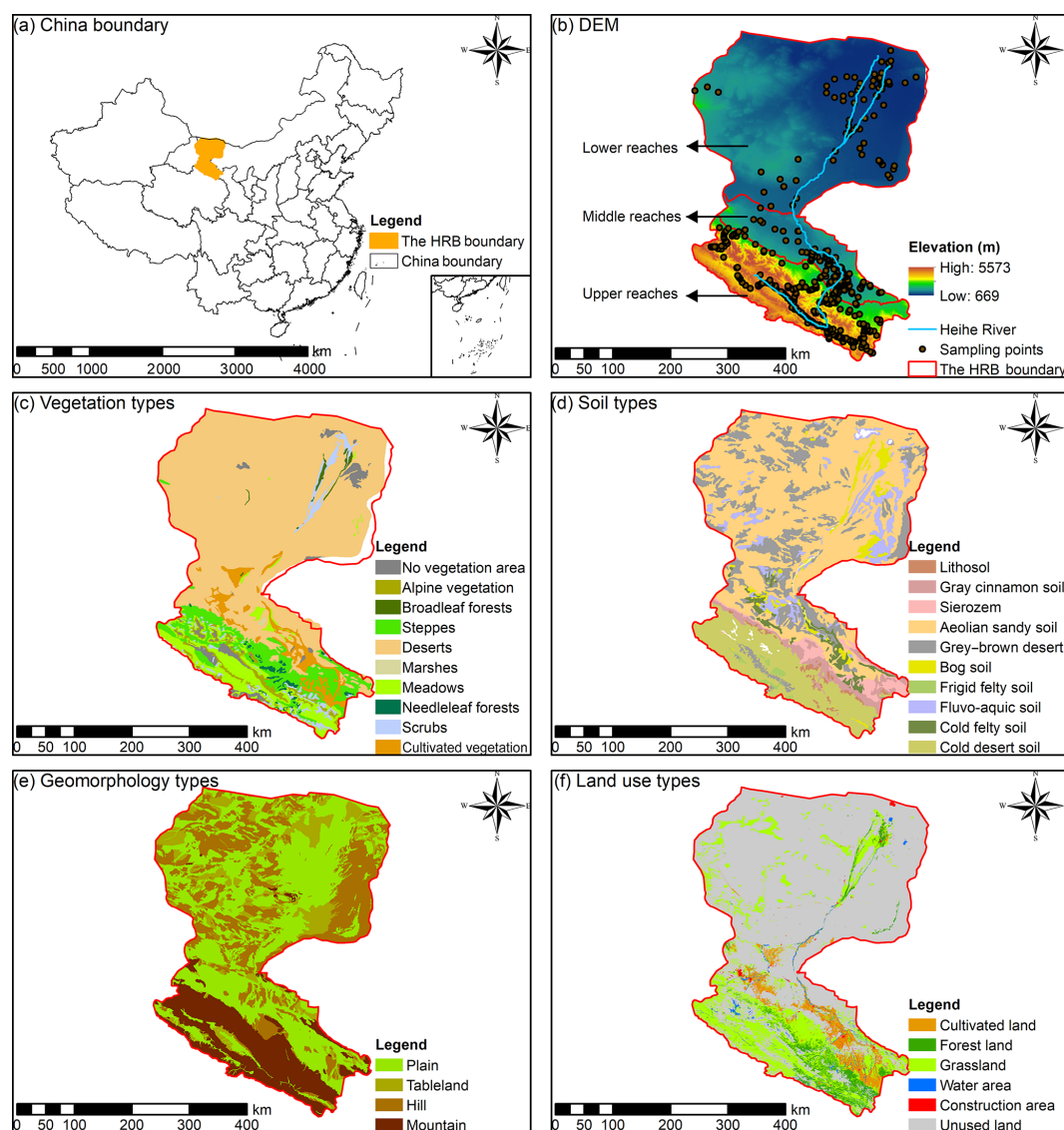
types in the middle reaches of the HRB are aeolian sandy soil, frigid frozen soil, and grey brown desert soil. The main soil types in the lower reaches of the HRB are aeolian sandy soil, grey–brown desert soil (northwest), and Lithosol (northeast).

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

The main land use type in the upper reaches, middle reaches, and lower reaches of the Heihe River were forest land and grassland, cultivated land, and unused land respectively (Fig. 1f). The water area and construction area were mainly distributed near the river in the middle reaches of the HRB.

### 2.2 Soil sampling

A total of 640 soil sampling points was collected in the HRB from the National Tibetan Plateau Data Center (NT-PDC) in China (<http://data.tpdc.ac.cn/zh-hans/>, last access: 5 May 2020), including 392 samples from the upper reaches and 248 samples from the middle and lower reaches of the HRB (Fig. 1b). The distribution of the soil types, vegetation types, the digital elevation model (DEM), and the geomorphology types of the HRB were considered in the soil sample collection, in terms of their location and proportion, in order to obtain samples that were 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, and so on. Due to the complicated soil types and vegetation types in the middle and upper reaches of the HRB, there were more soil sampling points in these areas. In contrast, fewer samples were collected in the lower reaches because of the relatively similar vegetation types. To reduce the noise effect of soil samples, the average of three to five mixed 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, analysed, and measured for soil PSFs (approximately 30 g of each sample). Soil PSFs were analysed using a Malvern Panalytical Mastersizer 2000 laser diffraction particle size analyser (the average measurement error is less than 3 %).



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

### 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 the soil PSFs. System for Automated Geoscientific Analysis (SAGA) GIS (Conrad et al., 2015) was used to compute the topographic variables from the DEM, including the 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/>, last access: 29 April 2020). Furthermore, latitude and longitude were also considered because of the large area of the HRB. Mean annual surface evapotranspiration data (Wu et al., 2012) were gathered from the NTPDC (<http://data.tpd.ac.cn/zh-hans/>, last access: 5 May 2020) 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, 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 parameter optimization

### 2.4.1 K-nearest neighbour

K-nearest neighbour (KNN) is a simple and nonparametric classifier that is based on using the known instance to label the unknown instance (Cover and Hart, 1967). For the test set, K-nearest training set vectors ( $k$ ) were found based on distance, and the maximum summed kernel densities were computed for classification. Moreover, continuous variables can also be predicted for regression with the average values of K-nearest neighbours. The parameters of KNN contain the maximum value of  $k$  (kmax), the distances of the nearest neighbours (distance), and different kernel functions (kernel). The KNN model is available in the “knn” R package (Schliep and Hechenbichler, 2016).

### 2.4.2 Multilayer perceptron neural network

Multilayer perceptron neural network (MLP), which is one of the most common multilayer feed-forward back-propagation networks (Zhang et al., 2018), was selected to train the artificial neural network (ANN) models due to the rapid operation, the small set of training requirements, and the 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: the input layer, the hidden layer, and the output layer. The resilient back-propagation 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 a scale from zero to one. MLP can be run using the “RSNNS” R package (Bergmeir and Benítez, 2012).

### 2.4.3 Random forest

Random forest (RF) was developed by Breiman (2001), combining the bagging method (Breiman, 1996) with 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. The benefits of using RFs are that the ensembles of trees are used without pruning. In addition, RF is relatively robust to overfitting. 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 “randomForest” R package (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 to classification and regression problems in soil science (Burgess, 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, which is known as support vector regression. The SVM is more effective in high dimensional spaces. A linear function was selected for SVM as the kernel function in our study. Additionally, cost and gamma are two other parameters that needed to be tuned, as these parameters control the trade-off between the classification accuracy and complexity and the ranges of the radial effect respectively. The SVM model is available in the “e1071” R package (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 “xgboost” R package (Chen et al., 2018).

### 2.4.6 Parameter optimization

The equation description of five machine-learning models can be found in the Supplement (Sect. S1). The “caret” R package (Kuhn, 2018) for MLP, SVM, and XGB; the “randomForest” R package for RF; and the “knn” R package 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 are selected as the best parameters. There are 11 dependent variables (i.e. “sand”, “silt”, “clay”, “ILR1”, “ILR2”, “ALR1”, “ALR2”, “CLR1”, “CLR2”, and “CLR3” for regression and

“class” for classification) trained with environmental covariates (independent variables). All methods were applied on these 11 components independently, and all of the adjusted parameters for the different models are listed in Table S1. More details about parameter optimization and independent modelling are given in Sect. S2.

## 2.5 Log-ratio transformation methods

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

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

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

$$\mathbf{z} = (z_1, \dots, z_{D-1}) = \text{ILR}(\mathbf{x}), \quad (3)$$

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

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) and are defined as follows:

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

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

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

$$\text{ILR}(x_0) = \text{ILR}(x_D) = 0, \quad (8)$$

$$\overline{\text{ILR}}(x_j) = \frac{\exp(Y(x_j))}{\sum_{j=1}^D \exp(Y(x_j))}. \quad (9)$$

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

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

where  $\text{sand}_s$  is the content of sand after standardization; this is the same for the 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, and ILR) in this study, including five machine-learning models

for direct soil texture classification (five models); we also use the above-mentioned methods with 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.

### 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 all 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):

$$\text{overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (11)$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative respectively. The kappa coefficient demonstrates the agreement between the observed classes and the measured classes, which is calculated based on the confusion matrix; the equation is defined as follows:

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

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 represents slight, 0.21–0.40 represents fair, 0.41–0.60 represents moderate, 0.61–0.80 represents substantial, and 0.81–1.00 represents 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 indicate the extent of identifying positive cases:

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

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (14)$$

Soil texture is a class-imbalanced data set of positive and negative with 62.5 % silt loam types; the negative classifier would be overvalued under these circumstances because of the overabundance of majority (negative) examples, additionally revealing overly optimistic findings (Davis and

**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, and XGB	KNN_ORI, MLP_ORI, RF_ORI, SVM_ORI, and XGB_ORI
Log-ratio-transformed data (ALR, CLR, and 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, and XGB_ILR

Goadrich, 2006). PRCs are informative in dealing with class-imbalanced data (Fu et al., 2017). The “precrc” R package (Saito and Rehmsmeier, 2017) can generate PRCs and compute AUPRC for each soil texture type. This process was repeated 30 times, and the average PRCs and AUPRCs were eventually obtained.

Similarly, the 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:

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

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

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

$$\text{abundance index} = p/t, \quad (16)$$

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 (CiLo: 12), loam (Lo: 57), loamy sand (LoSa: 18), sand (Sa: 23), sandy clay loam (SaCiLo: 4), sandy loam (SaLo: 58), silt (Si: 31), silty clay loam (SiCiLo: 37), and silt loam (SiLo: 400).

### 2.6.3 Validation indicators for soil PSF interpolation

Five statistical indicators, including the 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:

$$\text{RCC} = \rho_{xy}(\text{rank}) = \frac{\sigma_{xy}(\text{rank})}{\sigma_x(\text{rank})\sigma_y(\text{rank})}, \quad (17)$$

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

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

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

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

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

where  $x$  is the observed value,  $X$  is the predicted value,  $D$  is the number of dimensions (for soil PSFs are three),  $g(x)$  denotes the geometric mean  $(x_1 \dots x_D)^{1/D}$ , and  $\text{AD}_{x,ij}$  and  $\text{AD}_{X,ij}$  are the ADs between the observed soil PSFs and the predicted soil PSFs at sites  $i$  and  $j$ . Both of these parameters show that model performance is better when the values are lower. The standard deviation (SD) of prediction values and the ranges of the 95 % confidence interval (CI; Streiner, 1996) of the indicators were derived from 30 model runs to assess the model uncertainty.

### 2.7 Statistical analysis for the original and log-ratio-transformed data

The standard deviation (SD), coefficient of variation (CV), mean value, minimum value (Min), maximum value (Max),

median absolute deviation (MAD), skewness (Skew), kurtosis, and the 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 the 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), and (3) back-transform the calculated mean values to the initial closed space. Furthermore, the multivariate median values 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-transformed data should be back-transformed to the original space. For  $X = [X_1, \dots, X_n]$ , the MAD can be calculated according to Eq. (22):

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

### 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 the median centre (26.06 %). In contrast, silt and clay contents were the opposite, with lower means (silt, 55.79 %; clay, 13.57 %) than median centres (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 centres of the original data, except for clay (10.78 %). With respect to the SD and CV, soil PSF data in the log-ratio geometry had more stability and less variability than 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 the various 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 of the kurtosis values for log-ratio-transformed data were much higher than for the 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 KNN (0.630) and MLP (0.627). In addition, SVM (0.618) and XGB (0.613) had relatively lower accuracy than the 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). With respect to the confusion indices (COIs), XGB (0.278) delivered the best performance, and RF (0.501) demonstrated the highest confusion of models (Fig. 3c).

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 SaCiLo (only four samples), and 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, and SiCiLo) 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 the nine types. For soil texture types with more samples, RF and XGB performed better. For soil texture classes with less samples, RF and SVM showed better performance according to the AUPRCs.

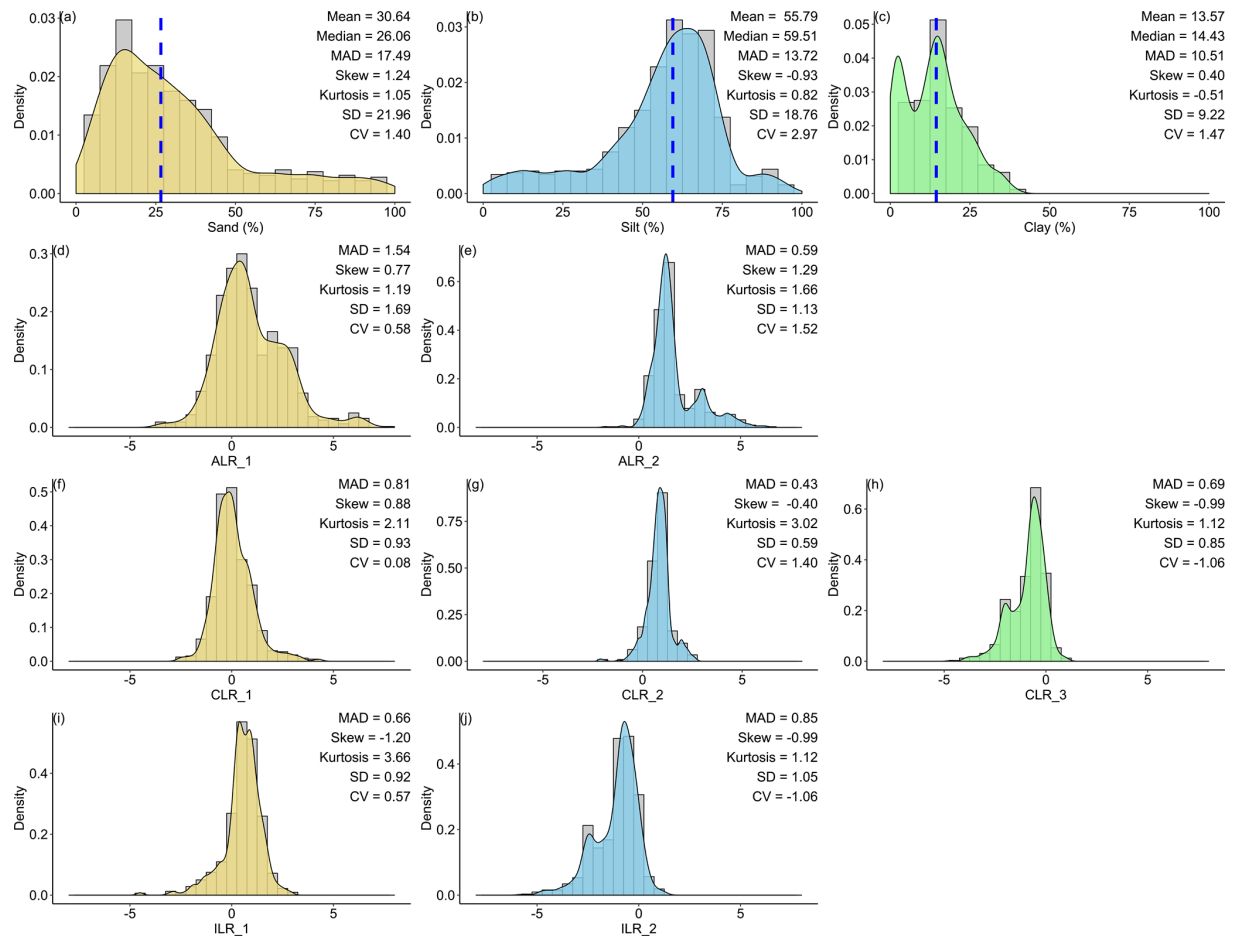
##### 3.2.2 Comparison of the prediction maps for soil texture classification

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 could predict all nine types, KNN and XGB predicted eight of nine types, followed closely by RF (seven of nine types) and MLP (six of nine 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. In 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).

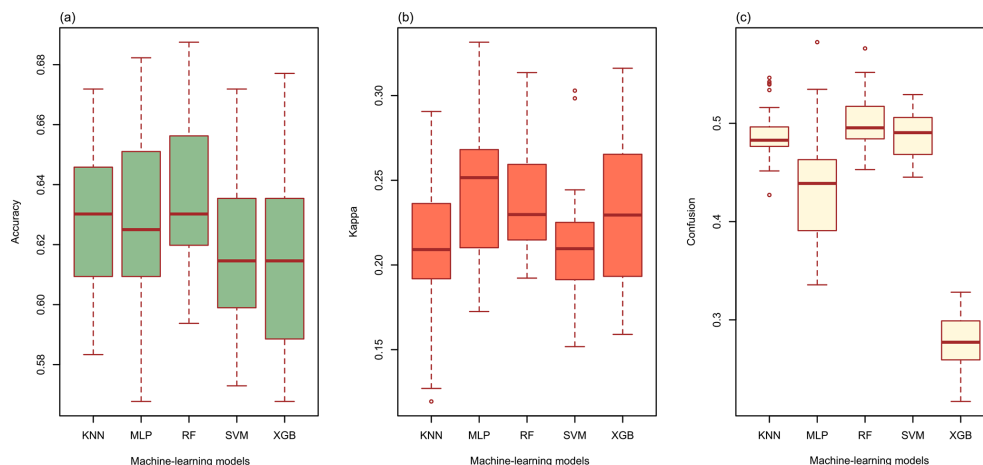
#### 3.3 Comparison of the machine-learning methods combined with log-ratio-transformed data in soil PSF interpolation

##### 3.3.1 Comparison of the validation indicators for the interpolation of soil PSFs

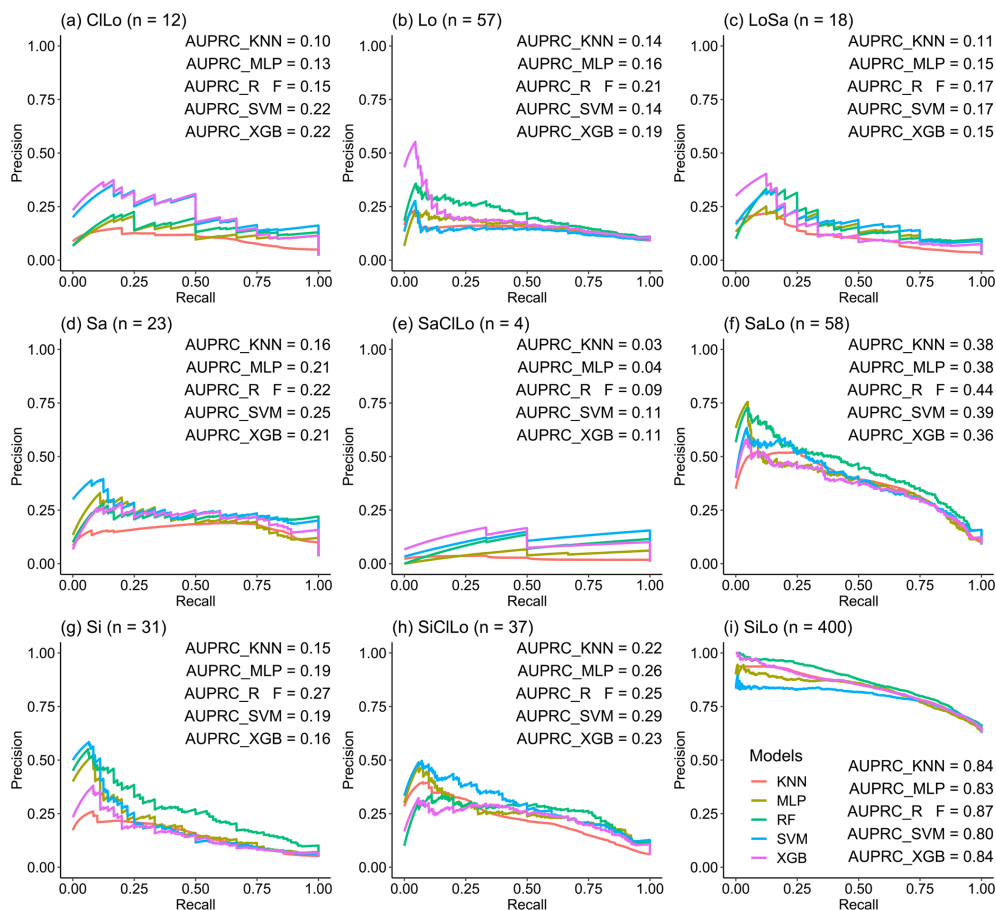
We compared the performance of each machine-learning model using the original and log-ratio-transformed data. The results indicated that the STRESS of the methods using log-ratio-transformed data were superior to the methods using original data (Table 2). The RMSE, MAE, RCC, and AD generated from KNN, MLP, RF, and XGB using original data outperformed the results using log-ratio-transformed data. By comparison, among different log-ratio-transformed data



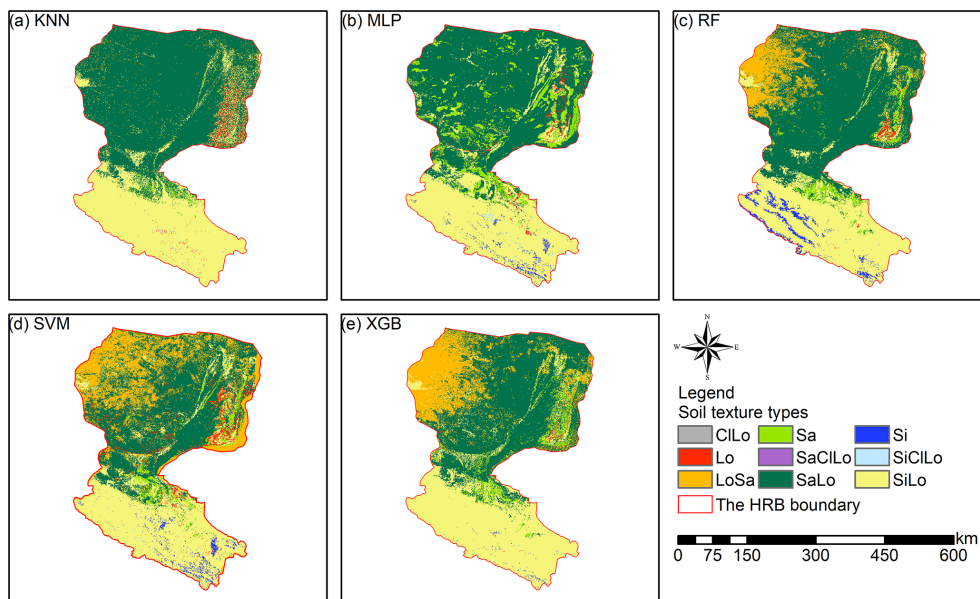
**Figure 2.** Descriptive statistical analysis for the original and log-ratio-transformed data for (a) sand, (b) silt, (c) clay, (d) ALR\_1, (e) ALR\_2, (f) CLR\_1, (g) CLR\_2, (h) CLR\_3, (i) ILR\_1, and (j) ILR\_2. “MAD” refers to median absolute deviation, “SD” refers to standard deviation, “CV” refers to the coefficient of variation, and “Median” is the 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 the original data.



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



**Figure 4.** The AUPRCs for different machine-learning models in the prediction of each soil texture type: (a) CiLo, (b) Lo, (c) LoSa, (d) Sa, (e) SaCiLo, (f) SaLo, (g) Si, (h) SiCiLo, and (i) SiLo. “n” denotes the number of sampling points for different soil texture types.



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



**Table 2.** Comparisons of the accuracy of different machine-learning models combined with original and transformed data. Bold values denote the best model performance for different indicators.

	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	<b>0.84</b>	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	<b>0.69</b>	0.65	0.68	0.86	<b>0.61</b>
RF_CLR	15.28	14.22	6.61	10.70	10.25	5.21	<b>0.69</b>	0.66	0.68	0.86	<b>0.61</b>
RF_ILR	15.27	14.25	6.66	10.66	10.26	5.26	<b>0.69</b>	0.66	0.68	0.86	<b>0.61</b>
RF_ORI	<b>15.09</b>	<b>13.86</b>	<b>6.31</b>	<b>10.65</b>	<b>9.99</b>	<b>5.00</b>	<b>0.69</b>	<b>0.67</b>	<b>0.69</b>	<b>0.84</b>	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

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 for KNN using the three log-ratio transformation methods. Furthermore, RF and SVM generated relatively similar results using CLR- and ILR-transformed data. 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.

### 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 are represented in Figs. 6, S1, and S2. 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, S1, S2). With respect to different machine-learning models, RF and XGB delivered prediction maps that were closer to the range of the distribution of the original data than KNN, SVM, or MLP.

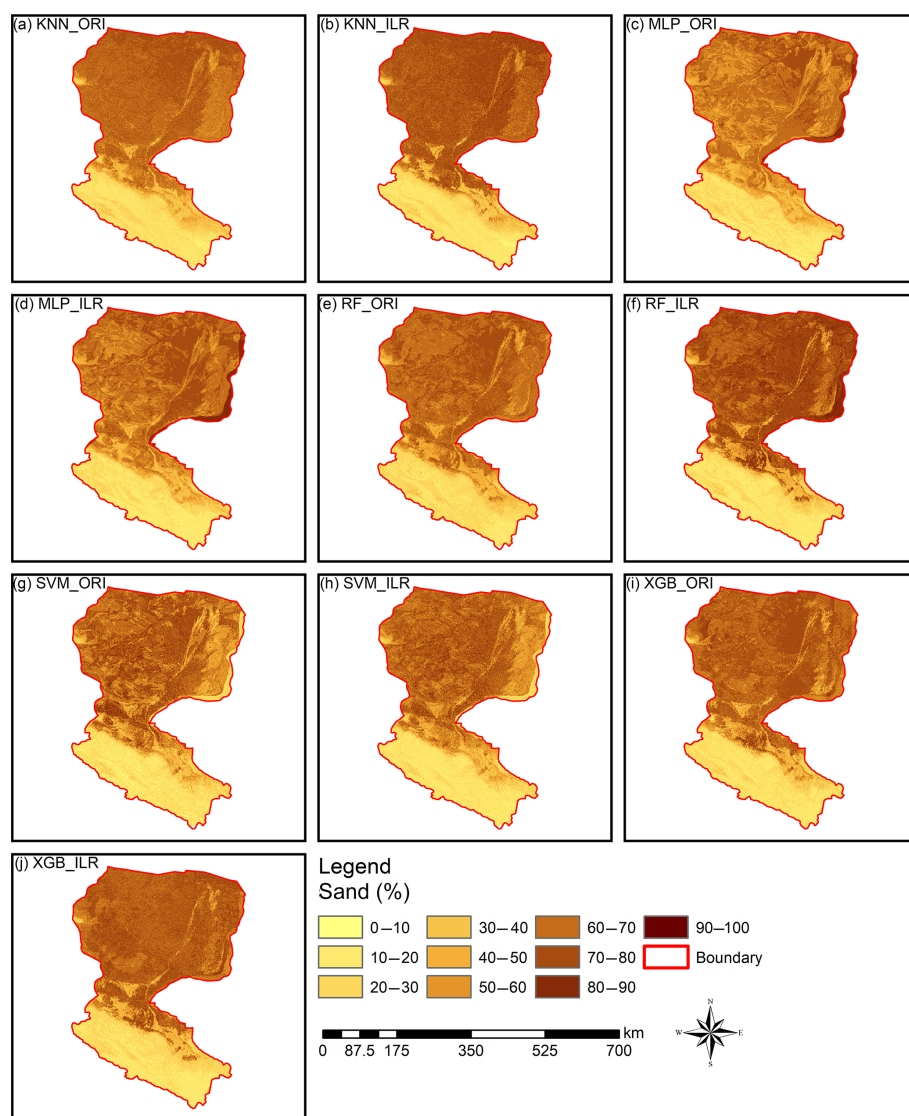
## 3.4 Comparison of direct and indirect soil texture classification

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

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

### 3.4.2 The prediction performance of soil texture types from different methods

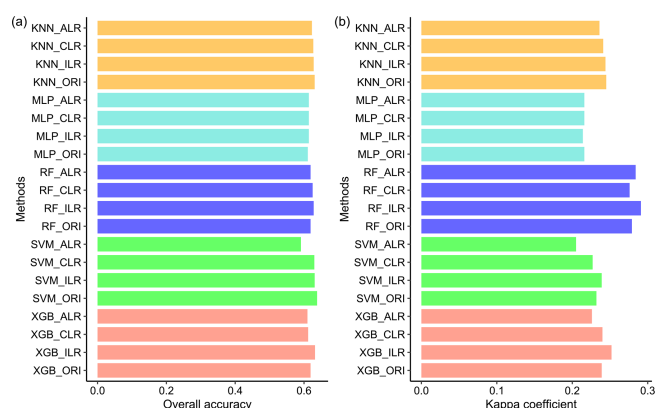
The distributions of soil texture types using original and ILR-transformed data are illustrated in Fig. 8 using the United States Department of Agriculture (USDA) soil texture triangle. The triangle of the original data of soil PSFs (Fig. 8a) demonstrates wider ranges of spatial dispersion than the interpolated data using machine-learning methods. These pre-



**Figure 6.** Prediction maps of the sand fraction. All of the ranges of the prediction maps of sand (approximately 9.0 %–90.0 %) were within the range of the 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 the 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 %).

dictions reveal the properties of aggregating from the sides to the centre of triangles. With respect to the machine-learning models, RF shows the most dispersed feature in accordance with the original soil PSF data. The predictions from models combined with ILR-transformed data are more discrete and more associated with the original soil PSF data than those resulting from ORI methods. The prediction results represent significant differences in the error ratio (yellow symbols, Fig. 8) of the soil sampling points with respect to soil types 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-calculate the

mean value of silt in the process of transformation (Fig. 2), so these points are 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) delivers the highest right ratio (RR) among these models, and the classification accuracy is enhanced using the ILR method (83.9 %) compared with ORI (81.7 %). In the case of other models, the differences between ORI and ILR are 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,



**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.

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.

### 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 the map 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, S3). The results of machine-learning models using three log-ratio-transformed data sets were similar to the number of predicted types; however, there were significant differences between the results using original data and log-ratio-transformed data. All machine-learning models combined with original data predicted more Lo and SaLo soil texture types and fewer LoSa and Si types (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, which revealed that different distributions of LoSa existed among them in the lower reaches of Heihe River basin. For the upper reaches, prediction maps of the ILR methods generated more Si and less Lo than the ORI method. Si soil texture types were mainly distributed in 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 were more in line with the real environment than the ORI methods, because more SaLo and less Lo soil texture types were predicted in the middle reaches of the HRB. Furthermore, the predicted soil texture using indirect methods was more abundant than the directly predicted soil texture in the middle reaches (Fig. 5).

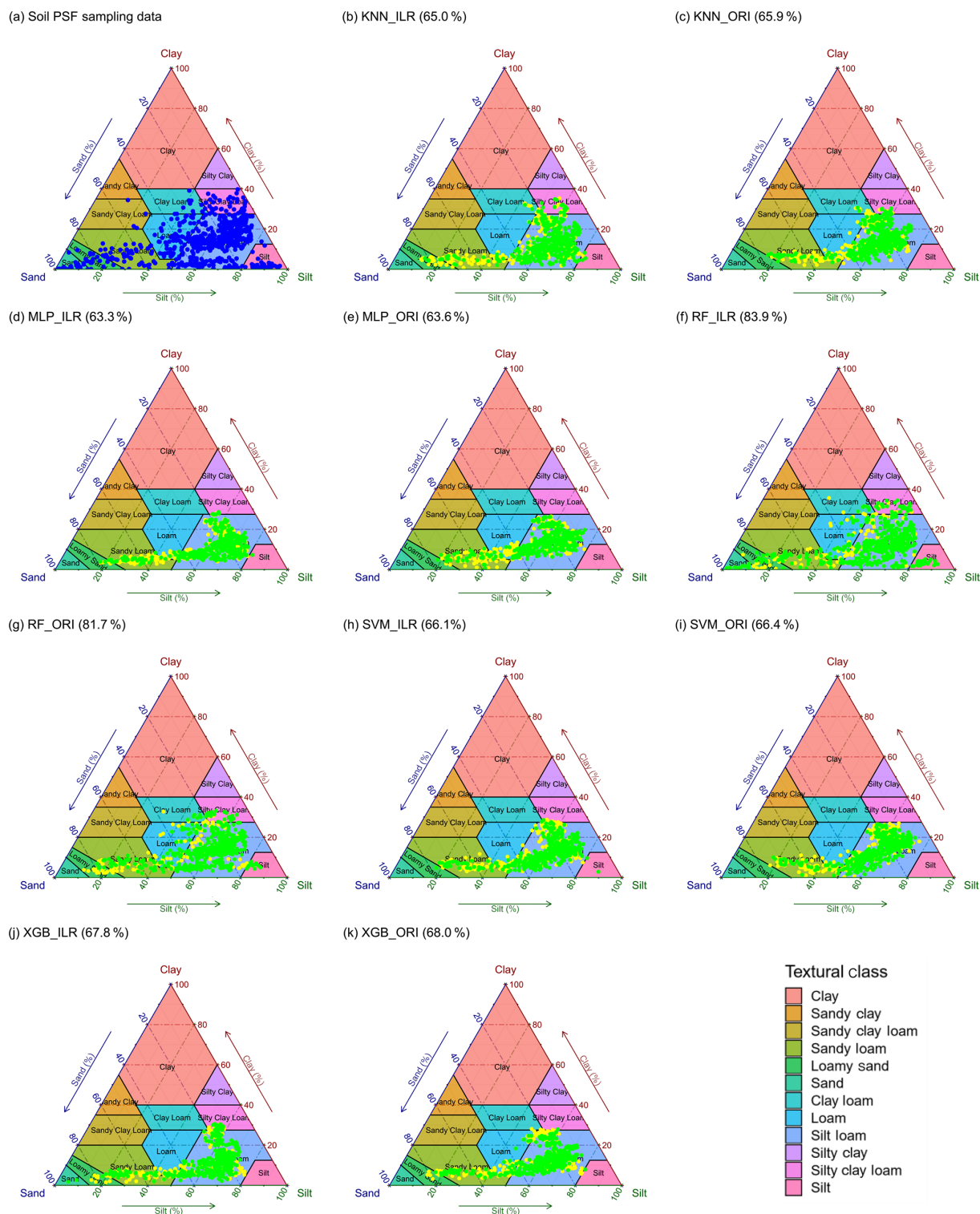
### 3.4.4 Comparison of total computing time for each model in soil texture classification and soil PSF interpolation

The run times of the models were computed and compared for different machine-learning models in soil texture classification and soil PSF interpolation (Fig. 10). Because the run times of the ORI and log-ratio methods were similar, the ILR was selected for soil PSF interpolation. With respect to the different models, RFs required the longest time for both classification (453.73 s) and interpolation (188.87 s), which may cause it to lose its advantage over the other models when processing large data sets. KNN (classification, 4.2 s; interpolation, 23.6 s) and SVM (classification, 4.15 s; interpolation, 12.4 s) had shorter run times with respect to both classification and interpolation. XGB (classification, 21.6 s; interpolation, 17.13 s) was much more stable and required less time; the data processes were also 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.

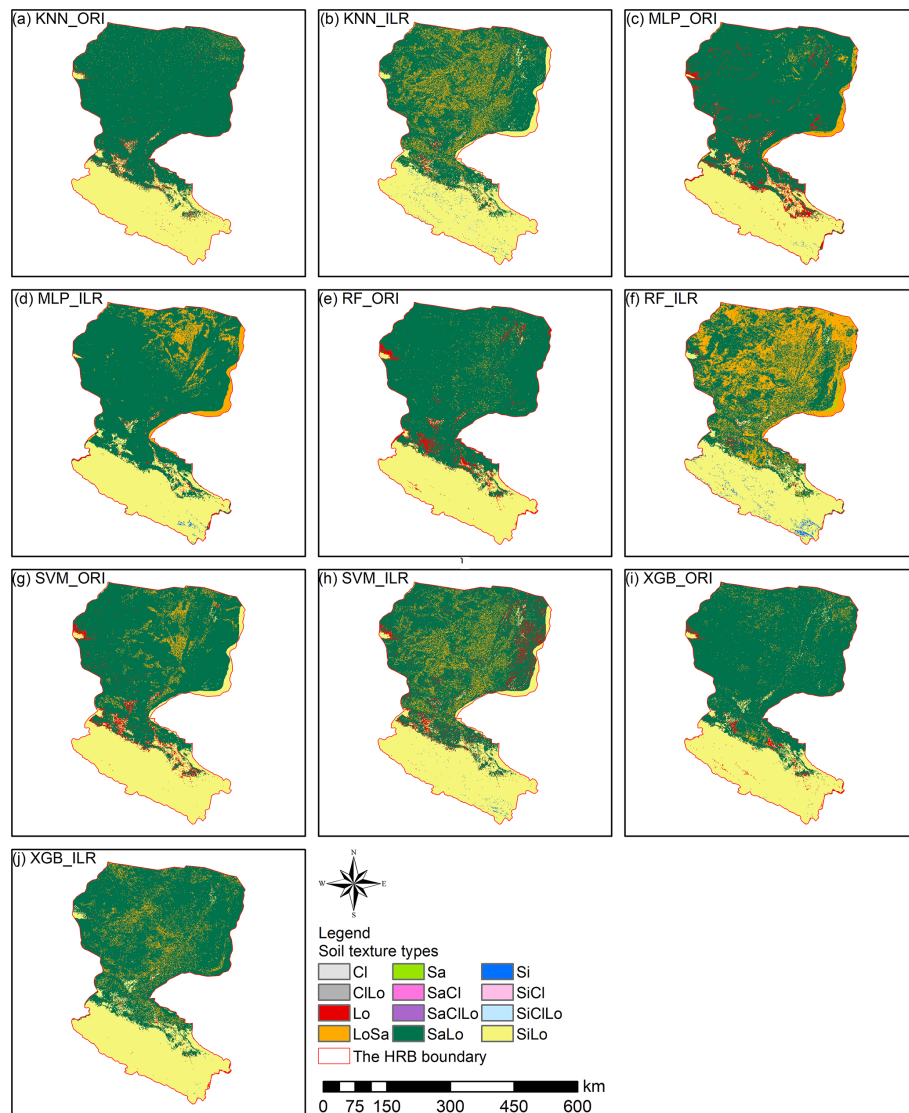
## 4 Discussion

### 4.1 The systematic comparison of the five machine-learning models

The range of applicability of the study is limited to independent modelling, i.e. the component-wise approaches. However, joint fractions modelling 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 was also concluded by Heung et al. (2016). With respect to the total computing time, RF revealed the longest run time with respect to both the classification and interpolation mode. In addition, regarding trade-offs between the total computing time of the model and the accuracy, XGB was superior to the other four models, reducing the computing time significantly while maintaining acceptable accuracy. In fact, parallel calculations can be automatically executed during the training phase of the XGB model: this is a great advantage when working with large data sets, as the XGB can be more than 10 times faster than the existing gradient boosting models (Chen and Guestrin, 2016). Therefore, XGB is recommended due to its speed (although this is at the expense of suboptimal accuracy) when researchers are dealing with large data sets in study areas. Moreover, some joint fractions approaches – compositional kriging (Wang and Shi, 2017), high accuracy surface modelling (HASM; Yue et al., 2015, 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 for combining environmental covariables.



**Figure 8.** Soil texture types of 640 soil samples shown using 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 the green symbols represent that the predicted and original soil texture classes were the same, whereas the yellow symbols represent the misclassification of the soil texture classes. The predicted right ratios (RRs) of the soil texture classes are given in parentheses after the interpolators above the plots.



**Figure 9.** The prediction maps of soil texture classification by indirect methods using KNN, MLP, RF, SVM, and XGB with either ILR-transformed data (ILR) or original data (ORI).

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 a single model (Segal and Xiao, 2011).

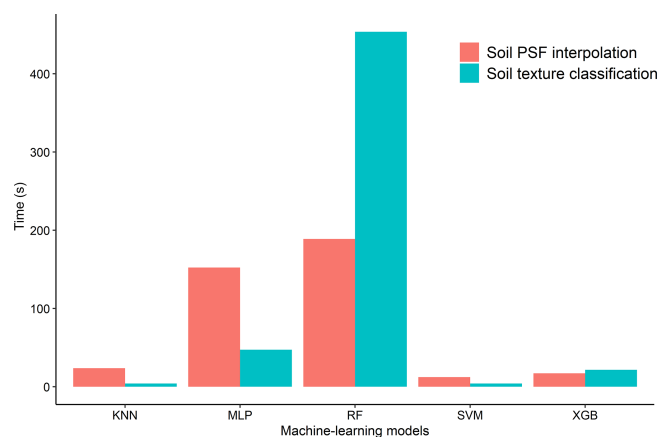
#### 4.2 The systematic comparison of the models using log-ratio-transformed data and original data

Log-ratio transformation methods can open the data and remove the “closure effect”, which induces spurious correlation. The opened data can be interpolated into the mapping area, and the results can then be back-transformed using inverse equations. However, in the process of pa-

rameter optimization, the optimal parameters of different machine-learning models are obtained using log-ratio-transformed data, which cannot guarantee the most accurate back-transformed results. This is because the values of assessment indicators (e.g. MAEs and RMSEs) 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 are back-transformed to the real space, these indicator values will be enlarged.

Due to the contraction of the predicted values (Fig. 8), there were small numbers of predictions beyond the range of the original data values, 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 to using ORI data. Among the three log-





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

ratio methods, ILR and CLR were superior to ALR, which can be explained by the fact that ILR and CLR are 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, ILR showed slightly better performance than CLR, because the geometric mean composed of all compositions of soil PSFs is the denominator in CLR, and one-to-one mapping of equations and soil PSFs could be implemented. Nevertheless, the sum of the dimensions of CLR is zero, and the problem of collinearity is still present. ILR transformed all of 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 the 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), can be based on a priori expert knowledge, or can be based on 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 are demonstrated in Sect. S6 (Table S3). The ILR balance chosen 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 requires further research. Furthermore, each component of log-ratio or original soil PSF data was independently modelled using component-wise approaches (machine-learning methods), which may be suboptimal com-

pared with the joint fractions approach under the circumstances (when dealing with the multivariate treatment). For example, CLR-transformed data are still characterized by collinearity, but there is no guarantee that the sum of the three components of CLR is zero due to the use of independent modelling. Although the final predictions were not influenced (still sum to 100 %) due to 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 recommended as a combination method with machine-learning models for component-wise modelling 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 present 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 input information in the model training process. The main soil texture types in the lower reaches of the HRB were SiLo, LoSa, and small areas of SaLo and Lo, which were 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 of these models predicted that the main soil type in the lower reaches of the HRB was SaLo, which did not fit with 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, due to the limitation of the training subsets, direct classification can only predict types that are 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 respect to the real environment should be considered such as the log-ratio methods of the 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 the other



machine-learning models mentioned above. As a new and suboptimal 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 the classification capacity of imbalanced data. In addition, the log-ratio methods, especially ILR, had the advantage of modifying STRESS in soil PSF interpolation. Moreover, the indirect methods for soil texture classification outperformed the direct methods, especially when combined with log-ratio transformations. The indirect methods for soil texture classification generated preferable results with respect to both the accuracy indicators and the prediction maps. The keys to improving the interpolator accuracy are using more appropriate interpolation techniques with environmental covariates, transforming soil PSF data using more efficient transformation methods, utilizing compositional data analysis in the multivariate studies, and using systematic parameter adjustment algorithms for compositional data.

## Appendix A: Abbreviations

PSFs	Particle size fractions
HRB	Heihe River basin
KNN	K-nearest neighbour
MLP	Multilayer perceptron neural network
RF	Random forest
SVM	Support vector machines
XGB	Extreme gradient boosting
ALR	Additive log ratio
CLR	Centred log ratio
ILR	Isometric log ratio
ORI	Original data
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, and KNN_ORI	KNN combined with ALR, CLR, ILR, and ORI respectively
MLP_ALR, MLP_CLR, MLP_ILR, and MLP_ORI	MLP combined with ALR, CLR, ILR, and ORI respectively
RF_ALR, RF_CLR, RF_ILR, and RF_ORI	RF combined with ALR, CLR, ILR, and ORI respectively
SVM_ALR, SVM_CLR, SVM_ILR, SVM_ORI	SVM combined with ALR, CLR, ILR, and ORI respectively
XGB_ALR, XGB_CLR, XGB_ILR, XGB_ORI	XGB combined with ALR, CLR, ILR, and ORI respectively
CI <sub>Lo</sub>	Clay loam
Lo	Loam
LoSa	Loamy sand
Sa	Sand
SaCI <sub>Lo</sub>	Sandy clay loam
SaLo	Sandy loam
Si	Silt
SiCI <sub>Lo</sub>	Silty clay loam
SiLo	Silt loam

**Data availability.** The 640 soil sampling data for the HRB, <http://data.tpdac.ac.cn/zh-hans/data/b5835154-1e3c-41a4-ba6c-a6ec5c968949/> (Zhang, 2020), <http://data.tpdac.ac.cn/zh-hans/data/2e9cbc1a-5972-4e29-945d-99a1902cadb7/> (Huang and Jiang, 2020), <http://data.tpdac.ac.cn/zh-hans/data/737e4d01-c5f8-4940-98d2-3bda306784ad/> (Yue and Zhao, 2020a), <http://data.tpdac.ac.cn/zh-hans/data/7f91d36d-8bbd-40d5-8eaf-7c035e742f40/> (Yue and Zhao, 2020b), <http://data.tpdac.ac.cn/zh-hans/data/371ce545-e8d0-4e96-81e1-e862dbfc3b50/> (Ma, 2020), <http://data.tpdac.ac.cn/zh-hans/data/b8bfbb8b-97e4-4622-acbd-06b5ac466403/> (Zhao and Ma, 2020), and <http://data.tpdac.ac.cn/zh-hans/data/438fc689-ad9e-4370-8961-5b2de53d8b87/> (Si, 2020); the saturated water content, field water holding capacity, wilt water content, and saturated hydraulic conductivity data, <http://data.tpdac.ac.cn/zh-hans/data/e977f5e8-972b-42a5-bffe-cd0195f3b42b/> (Zhang and Song, 2020a); and the soil thickness data, <http://data.tpdac.ac.cn/zh-hans/data/fc84083e-8c66-4a42-b729-4f19334d0d67/> (Zhang and Song, 2020b), can be accessed through <http://data.tpdac.ac.cn/zh-hans/> (last access: 5 May 2020). The meteorological data can be accessed through <http://data.cma.cn/> (last access: 14 March 2020).

**Supplement.** The supplement related to this article is available online at: <https://doi.org/10.5194/hess-24-2505-2020-supplement>.

**Author contributions.** WS contributed to soil data sampling and oversaw the design of the entire project. MZ performed the analysis and wrote the paper. ZX collected and analysed data. All authors contributed to writing the paper and interpreting data.

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

**Acknowledgements.** We acknowledge the comments from the editor, Alberto Guadagnini; the reviewers, Tom Hengl and Alfred Stein; and the anonymous referees that helped us improve the quality of the paper. Thanks are also due to the National Meteorological Information Center for providing the meteorological data and the National Tibetan Plateau Data Center for the soil particle size fractions data.

**Financial support.** This study was supported by the National Key Research and Development Program of China (grant no. 2017YFA0604703), the National Natural Science Foundation of China (grant nos. 41771364 and 41771111), the Fund for Excellent Young Talents in Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (CAS; grant no. 2016RC201), the Youth Innovation Promotion Association, CAS (grant no. 2018071), the Investigation and Monitoring project of Ministry of Natural Resources (grant no. JCQQ191504-

06) and a grant from the State Key Laboratory of Resources and Environmental Information System.

**Review statement.** This paper was edited by Alberto Guadagnini and reviewed by two anonymous referees.

## References

- Abdi, D., Cade-Menun, B. J., Ziadi, N., and Parent, L. E.: Compositional statistical analysis of soil <sup>31</sup>P-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, 1982.
- Aitchison, J.: On criteria for measures of compositional difference, *Math. Geol.*, 24, 365–379, <https://doi.org/10.1007/bf00891269>, 1992.
- Bagheri Bodaghabadi, M., Antonio Martinez-Casasnovas, J., Salehi, M. H., Mohammadi, J., Esfandiarpour Borujeni, I., Toomanian, N., and Gandomkar, A.: Digital soil mapping using artificial neural networks and terrain-related attributes, *Pedosphere*, 25, 580–591, 2015.
- Bationo, A., Kihara, J., Vanlauwe, B., Waswa, B., and Kimetu, J.: Soil organic carbon dynamics, functions and management in west african agro-ecosystems, *Agr. Syst.*, 94, 13–25, <https://doi.org/10.1016/j.agsy.2005.08.011>, 2007.
- Bedall, F. K. and Zimmermann, H.: Algorithm as 143: The mediant centre, *J. Roy. Stat. Soc. C-Appl.*, 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](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, *J. Stat. Softw.*, 46, 1–26, <https://doi.org/10.18637/jss.v046.i07>, 2012.
- Breiman, L.: Bagging predictors, *Mach. Learn.*, 24, 123–140, <https://doi.org/10.1023/a:1018054314350>, 1996.
- Breiman, L.: Random forests, *Mach. Learn.*, 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, *Eur. J. Soil Sci.*, 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 Min. Knowl. Disc.*, 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](https://doi.org/10.1016/S0016-7061(97)00018-9), 1997.
- Butler, J. C.: Effects of closure on the moments of a distribution, *J. Int. Ass. Math. Geol.*, 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) 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, *Mach. Learn.*, 20, 273–297, <https://doi.org/10.1023/a:1022627411411>, 1995.
- Cover, T. M. and Hart, P. E.: Nearest neighbor pattern classification, *IEEE T. Inform. Theory*, 13, 21–27, <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, *Math. Geol.*, 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, *Math. Geol.*, 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, *J. Anim. Ecol.*, 77, 802–813, <https://doi.org/10.1111/j.1365-2656.2008.01390.x>, 2008.
- Filzmoser, P. and Hron, K.: Correlation analysis for compositional data, *Math. Geosci.*, 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, *Sci. Total Environ.*, 407, 6100–6108, <https://doi.org/10.1016/j.scitotenv.2009.08.008>, 2009.
- Fiserova, E. and Hron, K.: On the interpretation of orthonormal coordinates for compositional data, *Math. Geosci.*, 43, 455–468, <https://doi.org/10.1007/s11004-011-9333-x>, 2011.
- 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, *Chemometr. Intell. Lab.*, 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, *Phys. Chem. Earth Pt. B*, 26, 41–45, [https://doi.org/10.1016/s1464-1909\(01\)85012-7](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, *J. Arid Environ.*, 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, e0125814, <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, e0169748, <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, e5518, <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, G. and Jiang, Y.: Soil texture of soil sampling points in Yingke Irrigation District, available at: <http://data.tpdac.ac.cn/zh-hans/data/2e9cbc1a-5972-4e29-945d-99a1902cadb7/>, last access: 11 May 2020.
- Huang, J., Subasinghe, R., and Triantafyllis, J.: Mapping particle-size fractions as a composition using additive log-ratio transformation and ancillary data, *Soil Sci. Soc. Am. J.*, 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 Sens. Environ.*, 83, 195–213, [https://doi.org/10.1016/s0034-4257\(02\)00096-2](https://doi.org/10.1016/s0034-4257(02)00096-2), 2002.
- Huete, A. R.: A soil-adjusted vegetation index (SAVI), *Remote Sens. Environ.*, 25, 295–309, [https://doi.org/10.1016/0034-4257\(88\)90106-x](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.
- Liaw, A., and Wiener, M.: Classification and regression by random-forest, *R News*, 2, 18–22, available at: <https://CRAN.R-project.org/doc/Rnews/> (last access: 29 April 2020), 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.
- Lloyd, C. D., Pawlowsky-Glahn, V., and Jose Egozcue, J.: Compositional data analysis in population studies, *Ann. Assoc. Am. Geogr.*, 102, 1251–1266, <https://doi.org/10.1080/00045608.2011.652855>, 2012.
- Ma, M.: HiWATER: Dataset of soil parameters in the midstream of the Heihe River Basin (2012), available at: <http://data.tpd.ac.cn/zh-hans/data/371ce545-e8d0-4e96-81e1-e862dbfc3b50/>, last access: 11 May 2020.
- Martin-Fernandez, J. A., Olea-Meneses, R. A., and Pawlowsky-Glahn, V.: Criteria to compare estimation methods of regionalized compositions, *Math. Geol.*, 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 semiarid, snowmelt-driven catchment, *Hydrol. Process.*, 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., 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.
- Menafoglio, A., Guadagnini, A., and Secchi, P.: Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a Bayes space approach, *Water Resour. Res.*, 52, 5708–5726, <https://doi.org/10.1002/2015wr018369>, 2016b.
- Metternicht, G. I. and Zinck, J. A.: Remote sensing of soil salinity: Potentials and constraints, *Remote Sens. Environ.*, 85, 1–20, [https://doi.org/10.1016/s0034-4257\(02\)00188-8](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.: Exploratory data analysis, in: *Applied Statistical Modeling and Data Analytics*, chap. 2, 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, *Int. J. Coal Geol.*, 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, *Environ. Geol.*, 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, *Stoch. Environ. Res. Risk A.*, 29, 1737–1749, <https://doi.org/10.1007/s00477-014-1010-4>, 2015.
- Schliep, K. and Hechenbichler, K.: kkn: Weighted K-nearest neighbors, R package version 1.3.1, available at: <https://CRAN.R-project.org/package=kkn> (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.
- Si, J.: Data set of soil moisture in the lower reaches of Heihe River (2012), available at: <http://data.tpd.ac.cn/zh-hans/data/438fc689-ad9e-4370-8961-5b2de53d8b87/>, last access: 12 May 2020.
- Small, C. G.: A survey of multidimensional medians, *Int. Stat. Rev.*, 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, *Can. J. Psychiat.*, 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 Syst. Appl.*, 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 Triantafyllis, 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.: Digital soil mapping: Interactions with and applications for hydrogeology, chap. 21, in: *Hydrogeology*, 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, *Math.*

- Geosci., 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, *Comput. Geosci.*, 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, [https://doi.org/10.1007/978-1-4615-5703-6\\_3](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, *J. Hydrol.*, 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, *J. Hydrol.*, 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, *Comput. Electron. Agr.*, 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 Sens. Environ.*, 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, *Ecol. Indic.*, 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. and Zhao, N.: Digital soil mapping dataset of soil texture (soil particle-size fractions) in the Tianlaochi basin (2012–2014), available at: <http://data.tpdac.ac.cn/zh-hans/data/737e4d01-c5f8-4940-98d2-3bda306784ad/>, last access: 11 May 2020a.
- Yue, T. and Zhao, N.: Digital soil mapping dataset of soil texture (soil particle-size fractions) in the upstream of the Heihe river basin (2012–2016), available at: <http://data.tpdac.ac.cn/zh-hans/data/7f91d36d-8bbd-40d5-8eaf-7c035e742f40/>, last access: 11 May 2020b.
- 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, *Environ. Earth Sci.*, 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, *Environ. Earth Sci.*, 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, G.: Soil texture of representative samples in the Heihe River Basin, available at: <http://data.tpdac.ac.cn/zh-hans/data/b5835154-1e3c-41a4-ba6c-a6ec5c968949/>, last access: 11 May 2020.
- Zhang, G. and Song, X.: Digital soil mapping dataset of hydrological parameters in the Heihe River Basin (2012), available at: <http://data.tpdac.ac.cn/zh-hans/data/e977f5e8-972b-42a5-bffe-cd0195f3b42b/>, last access: 11 May 2020a.
- Zhang, G. and Song, X.: Digital soil mapping dataset of soil depth in the Heihe River Basin (2012–2014), available at: <http://data.tpdac.ac.cn/zh-hans/data/fc84083e-8c66-4a42-b729-4f19334d0d67/>, last access: 11 May 2020b.
- 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, *J. Integr. Agr.*, 12, 1673–1683, [https://doi.org/10.1016/s2095-3119\(13\)60395-0](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.
- Zhao, C. and Ma, W.: Soil physical properties-soil bulk density and mechanical composition dataset of Tianlaochi Watershed in Qilian Mountains, available at: <http://data.tpdac.ac.cn/zh-hans/data/b8bfbb8b-97e4-4622-acbd-06b5ac466403/>, last access: 12 May 2020.