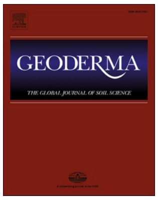
Geoderma 384 (2021) 114809

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma



# A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data

Lei Zhang [a,d], Lin Yang [a,b,*], Tianwu Ma [c,d,e], Feixue Shen [a], Yanyan Cai [a], Chenghu Zhou [a,b]

[a] *School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China*
[b] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China*
[c] *School of Geography, Nanjing Normal University, Nanjing 210023, China*
[d] *Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China*
[e] *Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing 210023, China*

ARTICLE INFO

Handling Editor: Kristin Piikki

*Keywords:*
Digital soil sampling
Machine learning
Semi-supervised learning
Self-training
Predictive mapping

ABSTRACT

Numerous machine learning models have been developed for constructing the relationship between soil classes or properties and its environmental covariates in digital soil mapping (DSM). Most machine learning models are trained with a supervised learning (SL) method based on training samples. However, the collected sample data is often limited in practice due to that field sampling is expensive and time-consuming. The insufficient samples may limit the learning ability of the model to a large extent. Semi-supervised machine learning, a new machine learning paradigm that makes use of both unsampled data and a small amount of sampled data in the learning process, can be a potential effective method for DSM. In this study, we present a self-training semi-supervised learning (SSL) method for DSM. Different with the SL method for machine learning models, the SSL method not only utilizes the sampled locations but also the abundant environmental covariate information at the unvisited locations. Its basic idea is to iteratively enlarge the training data set by adding the unsampled points with high prediction confidence from the unvisited locations until a stopping criterion reached. The proposed SSL method was applied in machine learning models for predicting soil classes in Heshan Farm of Nenjiang County in Heilongjiang Province, China. Three machine learning models, including multinomial logistic regression (MLR), k-nearest neighbor (KNN) and random forest (RF), were selected to evaluate the efficiency of the SSL method. The entropy threshold was an important parameter in the SSL method, and a sensitivity analysis on this parameter was conducted with using a series of entropy thresholds. The SSL method was compared with the SL method for the three machine learning models for soil prediction. A cross-validation was employed to evaluate the accuracy of the predicted soil class maps generated based on each method. The results showed that the prediction accuracies (the proportion of the correctly predicted samples over the total number of validation samples) of the SSL method were higher than those of the SL method for MLR, KNN, and RF by 5.9%, 12.2%, and 6.0%, respectively. RF-SSL was the most accurate model in the study area, followed by KNN-SSL. Meanwhile, the self-training SSL method for the KNN model had the largest improvement comparing with the other two models. Furthermore, the predicted soil maps using the SSL method showed a more reasonable spatial variation pattern of soil classes. In the study area, a suitable value of the entropy threshold was 0.8 ~ 1.0. We concluded that the SSL method improved the soil prediction accuracy compared with the SL method when applying machine learning models for DSM, and thus is a potential efficient method for DSM with limit sample data.

## 1. Introduction

Soil class maps provide essential information for natural resource management, environmental and ecological modeling (Arrouays et al., 2020; Lagacherie, 2008; Lamichhane et al., 2019; McBratney et al., 2003; Sanchez et al., 2009; Scull et al., 2003). Digital soil mapping (DSM) techniques formalize a generic framework for exploring the quantitative relationship linking soil to its environmental covariates.

* Corresponding author at: School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China.
*E-mail address:* yanglin@nju.edu.cn (L. Yang).

https://doi.org/10.1016/j.geoderma.2020.114809
Received 20 May 2020; Received in revised form 23 October 2020; Accepted 23 October 2020
Available online 21 November 2020

This relationship is used to calibrate a model to predict at locations where soil information is not available (Lagacherie, 2008; McBratney et al., 2003). The quality of the constructed soil-environment relationship and the mapping result are greatly impacted by field samples and the predictive methods (de Gruijter et al., 2006). In this paper we mainly focus on the predictive methods.

Numerous methods have been developed for predicting the spatial distribution of soil over the past decades. Geostatistical interpolation and machine learning are the two major methods. Geostatistical methods follow the principle of geostatistics (Matheron, 1963) to model the spatial dependence structure of the observed samples with (out) accounting for the deterministic trend. Various types of kriging algorithms have been developed (Heuvelink and Webster, 2001), such as ordinary kriging (OK) (Rawlins et al., 2011), co-kriging (Goovaerts, 1997) and regression co-kriging (Heuvelink et al., 2016; Malone et al., 2014; Mishra et al., 2012; Sun et al., 2012). Machine learning is another widely-used method for predicting the spatial variation of soil classes and soil properties (Brungard et al., 2015; Heung et al., 2016; Keskin et al., 2019; Lamichhane et al., 2019; Scull et al., 2003). Machine learning methods are mainly data-driven and they build soil-environment relationships based on training samples. Compared with geostatistical methods, machine learning does not require rigorous statistical assumptions about the distribution of soil samples. It can handle correlated environmental covariates, which can be either continuous or categorical variables. Numerous machine learning models have been developed and applied in DSM, including linear models (e.g. Chen et al., 2020; Kempen et al., 2009; Meersmans et al., 2008; Vasques et al., 2014), k-nearest neighbor (KNN) (e.g. Mansuy et al., 2014), support vector machine (e.g. Kovačević et al., 2010; Were et al., 2015), artificial neural networks (e.g. Behrens et al., 2005; Priori et al., 2014; Zhu, 2000), and ensemble models such as random forest (RF) (e.g. Grimm et al., 2008; Hengl et al., 2018; Heung et al., 2014; Li et al., 2021; Ma et al., 2020; Stoorvogel et al., 2009; Wiesmeier et al., 2014).

Most of the current machine learning models employ supervised learning (SL) (Russell and Norvig, 1995), of which the basic concept is to learn the relationships between the target variable and its independent variables based on training samples with their environmental covariates. The supervised learning methods usually require a large number of training samples. However, the number of the collected samples is often limited in practice due to the labor-intensive field sampling campaign (de Gruijter et al., 2006; Webster and Oliver, 1990; Zhu et al., 2015, 2008). The insufficient sample data may limit the learning ability of the supervised learning methods.

In supervised learning, only the training sample, which is called labeled data, is used for model calibration. The environmental information of the substantial unsampled locations, which can be called unlabeled data, is wasted. One way to improve the machine learning model performance is to make use of the unlabeled data. Semi-supervised learning (SSL), a new machine learning paradigm that utilizes both the labeled data and a large amount of unlabeled data, is a potential effective method with improved learning ability (Chapelle et al., 2006; Du et al., 2020; Levatić et al., 2017; Triguero et al., 2015; Zhu, 2006). To better understand the difference between supervised and semi-supervised learning, we showed an intuitional figure (Fig. 1). The figure represents the classification results considering only label data and considering both labeled and unlabeled data, respectively. It can be seen that both the classification boundaries based on SL and SSL are well split the labeled data points, however, the boundary generated by considering both the labeled and unlabeled data is better in differentiating the two groups of points than the boundary obtained by only considering the labeled data. This illustration shows a better classification ability of the semi-supervised learning method.

Several semi-supervised learning methods have been developed in machine learning field, such as self-training (Yarowsky, 1995), co-training (Blum and Mitchell, 1998), graph-based algorithms (Blum and Chawla, 2001), and semi-supervised support vector machine (Zhu and Goldberg, 2009). Self-training or co-training has recently been applied in remote sensing studies to improve the image recognition capacity (Han et al., 2015, 2018; Romaszewski et al., 2016; Yao et al., 2016). A semi-supervised weighted support vector machine learning method has been proposed by Silva et al. (2017) for land cover class mapping. However, there has been very few studies on using semi-supervised learning for soil prediction. Du et al. (2020) presented a graph-based semi-supervised learning approach for predictive soil class mapping with small field sample data, and proved its efficiency in improving the prediction accuracy. Yet, a complex graph model describing the environmental similarities between sample locations and unvisited locations was required. By comparison, self-training is adaptable and simple, which can use any machine learning model as the base model. It can improve the model by repeatedly enlarging the labeled data set from the unlabeled data set. Compared with other SSL methods, the self-training method is more concise. Therefore, it is desirable to research how to utilize the self-training semi-supervised learning to improve the accuracy of predictive soil mapping by engaging a large amount of unlabeled data.

This paper presents a self-training semi-supervised learning method

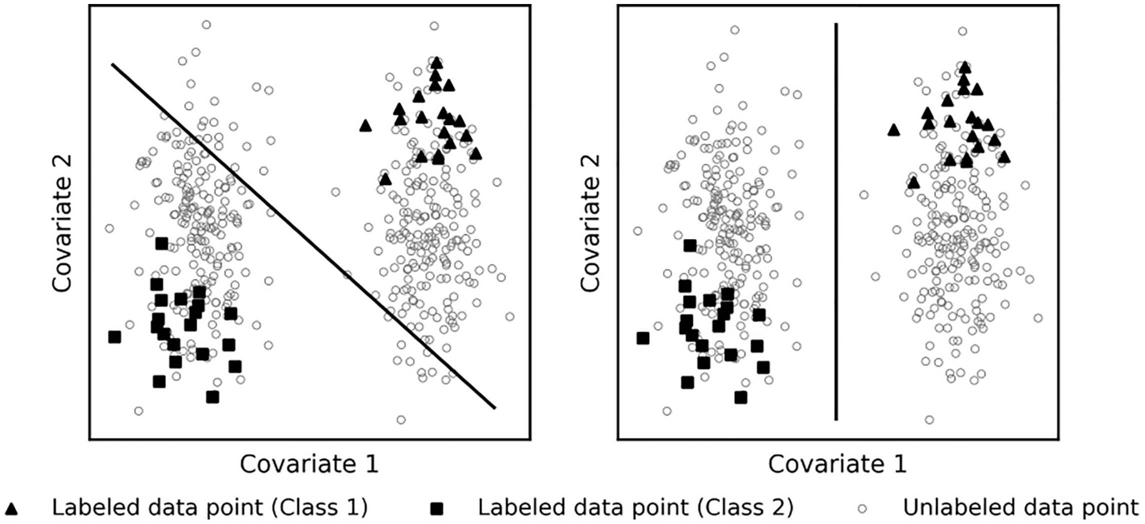


**Fig. 1.** An illustration of the usefulness of unlabeled data. Left one shows the optimal classification boundary only based on labeled data, right one shows the optimal boundary with considering both labeled and unlabeled data.

for soil class prediction with limited sample data. The next section of this paper presents a detailed introduction to the methodology and a case study was conducted in Heilongjiang province, China. The self-training semi-supervised learning method was compared with the supervised learning method for three machine learning models for DSM in the study area.

## 2. Materials and methodology

### 2.1. Study area and data

The study area is located in Heshan farm of Nenjiang County in Heilongjiang Province, China (Fig. 2a). The size of this area is approximately 60 km$^2$. Elevation within the area ranges from 276 to 363 m with slope gradient generally under 4°. The annual temperature is the area ranges from −38 to 36 °C, and the annual accumulated temperature above 10 °C is about 2000–2300 °C day. The average annual precipitation is 500–600 mm. Most soils in the area were formed on deposits of silt loam loess, except for the valley, where the underlying parent material is fluvial deposits. Crops in the area are generally limited to soybean (*Glycine* Willd.) and wheat (*Triticum* L.). The land use of this area is mainly croplands under mechanized tillage and no organic fertilizer has been applied to this area because of the naturally high contents of organic matter in the soil of this area. (Miao et al., 2011; Zhu et al., 2010).

According to the previous studies on the characteristics of the environmental conditions and the soil-environment relationships in this area (Yang et al., 2007; Zhu et al., 2010, 2018), some commonly used covariates for soil, such as macroclimate, parent materials and vegetation conditions are overall uniform over this small area. Topographic covariates were selected as indicators for the spatial variation of soil in the study area. We selected six topographic covariates for digital soil mapping. The selected environmental covariates are elevation (m), slope gradient (%), plan curvature, profile curvature, relative position index (%), and topographic wetness index. A 10-m resolution Digital Elevation Model (DEM) (Fig. 2b) was created from a 1:10,000 topographic map (published by Chinese Bureau of Surveying and Mapping (1987)) using the TOPOGRID and TINLATTICE in Arc/Info (Yang et al., 2007). Slope gradient, contour curvature and profile curvature were derived from this DEM. The TWI was calculated according to the method of Beven and Kirkby (1979). A multiple flow direction strategy (MFD-md) was used to estimate the upslope drainage area in the computation of TWI (Qin et al., 2011, 2007). The relative position index was calculated with the algorithm proposed by Qin et al., 2009.

One hundred and twenty-nine field samples were collected in the study area as shown in Fig. 2b. At each location, a pit was dug and the soil expert classified its soil type using a Chinese soil taxonomy system (Chinese Soil Taxonomy Research Group, 2001). There are six soil classes in this area: Mollic Bori-Udic Cambosols, Typic Hapli-Udic Isohumosols, Typic Bori-Udic Cambosols, Lithic Udi-Orthic Primosols, Pachic Stagni-Udic Isohumosols, and Fibric Histic-Typic Haplic Stagnic Gleyosols.

### 2.2. The semi-supervised learning (SSL) problem

In this section, we present the definition of semi-supervised learning (SSL) basic concept and its difference from supervised learning (SL). Given the environmental covariates data set $X$ with the size of $N \times k$ ($N$ is the number of raster cells in a certain study area, $k$ is the number of the environmental covariates), and $n$ sample data with their soil classes $Y$. The $n$ samples form the labeled data set $D_l = \{x_i, y_i | i = 1, 2, \cdots, n\}$, where $x_i$ is a feature vector describing the environmental covariates at the $i$th sample point, $y_i \in \{1, 2, \cdots, c\}$ is the corresponding target class label and $c$ is the number of the classes. The labeled data set $D_l$ consists of two parts, $X_l = (x_1, x_2, \cdots, x_n)$ and $Y_l = (y_1, y_2, \cdots, y_n)$. The unlabeled data set is defined as $D_u = \{x_{n+j} | j = 1, 2, \cdots, m\}$, where $m$ is the number of the unlabeled data and equals to $N - n$, generally $m \gg n$. $D_u$ can also be regarded as $X_u (X_u = X - X_l)$.

Formally, the task of SL is to learn a functional relationship $f : \mathscr{X} \mapsto \mathscr{Y}$ from a training data set $D_l$. Different from the SL, SSL is intended to learn a better model by including unlabeled data set $D_u$ rather than using only the labeled set $D_l$. Thus, the training data set in SSL is $D_l \cup D_u$.

### 2.3. The self-training SSL method

Self-training is a typical self-labeled strategy to tackle the semi-supervised learning problem (Triguero et al., 2015). It attempts to iteratively enlarge the labeled data set by adding those unlabeled points with high prediction confidence based on the trained model, and retrain the model with the enlarged labeled data set until a stopping criterion is reached. The overview framework of the self-training method is shown in Fig. 3.

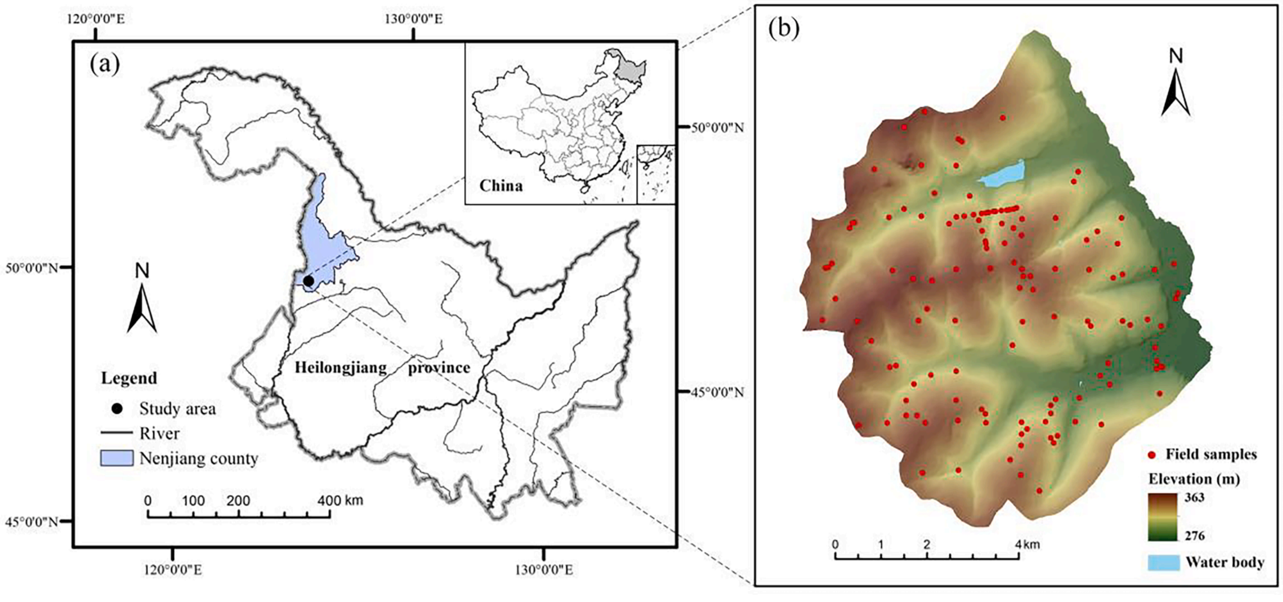


**Fig. 2.** (a) The location of the study area; (b) Digital elevation model (DEM) and field sample points of the study area.

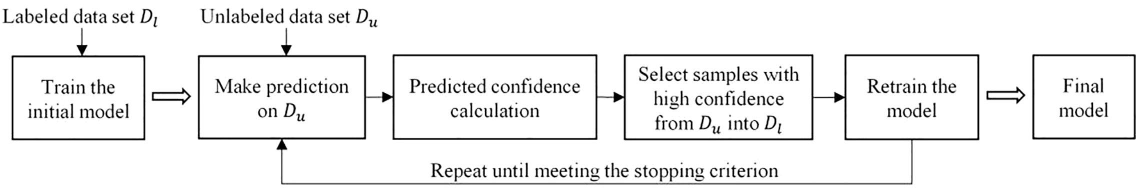

Labeled data set $D_l$
Unlabeled data set $D_u$
Train the initial model
Make prediction on $D_u$
Predicted confidence calculation
Select samples with high confidence from $D_u$ into $D_l$
Retrain the model
Final model
Repeat until meeting the stopping criterion

**Fig. 3.** The overall framework of the self-training semi-supervised learning method.

The specific procedure of the self-training SSL method was shown as follows:

(1) Use the labeled data set $D_l$ to train the model.
(2) Use the trained model to predict labels for the unlabeled data set $D_u$.
(3) Calculate the prediction confidence of each unlabeled sample point. The prediction confidence represents the level of the probability of the correct prediction of an unvisited point. The trained model determines the probability of each unlabeled sample point belonging to each class. We used the information entropy (Shannon, 1948) of the probabilities of all the classes to quantitatively determine the prediction confidence of each unlabeled point. The entropy calculation was as follow:

$$\mathrm{Ent}(p) = -\sum_{i=1}^{c} p_i log_2 p_i \tag{1}$$

where $c$ is the number of classes, $p$ represents the vector of predicted probabilities on all classes that calculated by the trained model. $p_i$ is the predicted probability of the $i$th class, and is determined according to the selected base model. For example, when using a tree-based model, the predicted probability is estimated by the proportion of a certain class at the leaf of a decision tree. When using a logistic regression model, it is determined by a logistic function of the linear combination of the input feature vector (James et al., 2013a; 2013b). The more distinguishing the predicted probabilities of the different classes, the smaller the entropy value for vector $p$, indicating a more confident prediction. Thus, the unlabeled data with the entropy smaller than a certain user-defined threshold can be assigned pseudo-labels. The pseudo-labeled data set is denoted as $D_p$.

(4) Remove the selected pseudo-labeled data from unlabeled data. $D_u \leftarrow D_u - D_p$.
(5) Combine the pseudo-labeled data into the labeled data set. $D_l \leftarrow D_l \cup D_p$.
(6) Repeat steps (1)-(5) until $D_u$ is empty or no example in $D_u$ has the entropy value derived from the retrained model higher than the threshold.

### 2.4. *Experimental design*

#### 2.4.1. *Data pre-processing*

According to the self-training SSL method mentioned in Section 2.3, the total dataset consisted of the labeled data set $D_l$ and unlabeled data set $D_u$. $D_l$ was the 129 sample points $Y_l$ with environmental covariates values $X_l$ and their soil class labels in our case study. In order to reduce the computational complexity, we selected a random sample set of 10, 000 raster cells out of the total 596, 158 raster cells in the study area (except the cells at the sample locations) as the unlabeled points. The unsampled data $D_u$ was the unlabeled points with their environmental covariates. Table 1 shows the descriptive statistics of the environmental covariates at soil sample points ($X_l$), the 10,000 randomly selected unvisited points ($X_u$) and the entire study area. The values of each covariate were scaled to lie between 0 and 100. It can be observed that the coverage of the environmental covariate space by $X_l$ and $X_u$ generally matched that of the entire study area. Due to the large number of points, $X_u$ had a better coverage in environmental covariate space than $X_l$, and the environmental characteristics of $X_u$ was highly consistent with the environmental characteristics of the entire study area.

#### 2.4.2. *The base machine learning models and parameter setting*

A base model is needed when applying the self-training SSL. In this study, we selected three commonly used models in DSM as base models to test the performance of SSL, including multinomial logistic regression (MLR), k-nearest neighbor (KNN), and random forest (RF). The three models were directly conducted as supervised learning without using unlabeled data to compare with the semi-supervised learning models. Therefore, six models were trained: MLR-SL, KNN-SL, RF-SL as supervised learning models and MLR-SSL, KNN-SSL, RF-SSL as semi-supervised learning method. The scikit-learn (Pedregosa et al., 2011) and SciPy (Virtanen et al., 2020) packages in Python programming language (Pérez et al., 2011) were used to apply models. The following is a brief introduction to the three base models and the parameter setting.

MLR is a classification method that generalizes binomial logistic regression to multiclass problems. A logistic function is used that gives outputs between 0 and 1 for a linear combination of the input covariates. MLR is the most frequently used linear approach for classification purposes of soil (Debella-Gilo and Etzelmüller, 2009; Jafari et al., 2012; Kempen et al., 2009). As the simplicity of the MLR model, it does not require parameter tuning.

The second model KNN is a neighbors-based learning method and widely used for the agricultural land cover classification (Samaniego and Schulz, 2009), forest inventory (Beaudoin et al., 2014; Bernier et al., 2010) and digital soil mapping (Subburayalu and Slater, 2013). The basic principle of KNN is to predict the class of an unobserved point based on the predefined number of training samples that have the closest

**Table 1**
Statistics of the standardized values of the environmental covariates.

| | $X_l$ | | | | $X_u$ | | | | The entire study area | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Min. | Median | Mean | Max. | Min. | Median | Mean | Max. | Min. | Median | Mean | Max. |
| Elevation | 0.3 | 58.4 | 53.8 | 96.2 | 0.0 | 55.0 | 53.2 | 99.2 | 0.0 | 55.1 | 53.0 | 100.0 |
| Slope gradient | 0.0 | 17.1 | 16.9 | 58.5 | 0.0 | 16.9 | 16.5 | 98.6 | 0.0 | 16.9 | 16.5 | 100.0 |
| Planform curvature | 39.8 | 46.6 | 46.8 | 76.5 | 1.3 | 46.6 | 46.8 | 97.6 | 0.0 | 46.6 | 46.9 | 100.0 |
| Profile curvature | 35.6 | 51.5 | 51.3 | 63.0 | 0.7 | 51.4 | 51.5 | 98.4 | 0.0 | 51.4 | 51.5 | 100.0 |
| Relative position index | 0.0 | 44.6 | 45.9 | 100.0 | 0.0 | 38.4 | 42.7 | 100.0 | 0.0 | 38.1 | 42.4 | 100.0 |
| Topographic wetness index | 16.1 | 28.9 | 33.8 | 95.0 | 0.1 | 29.0 | 35.0 | 99.9 | 0.0 | 29.0 | 35.1 | 100.0 |

$X_l$: environmental covariates of labeled data (sample points); $X_u$: environmental covariates of unlabeled data.

distance in covariate space to it. The key parameter for the KNN is the number of neighbors. It was set to the default value of 5.

RF is a representative of the state-of-the-art ensemble models. It is an extension of Bagging (Breiman, 1996). Bagging applies bootstrap sampling (Efron and Tibshirani, 1993) to obtain the data subsets for training the base learners and adopts the voting or averaging strategy for aggregating the base learners. The major improvement of RF over Bagging is the incorporation of randomized feature selection (Breiman, 2001). This strategy can effectively reduce the risk of overfitting and lead to a better generalization ability. RF has been proved as outperformed than other machine learning models in many studies (Brungard et al., 2015; Heung et al., 2016; Keskin et al., 2019; Zeraatpisheh et al., 2019). There are two important user-defined parameters in the RF. The first is the number of covariates that randomly selected for each tree building process. We used the rounded down square root of the total number of covariates as this parameter value by default (Breiman, 2001). The second parameter is *ntree*, which is defined as the number of trees to be learned in the forest. We set *ntree* = 200, for the previous studies showed that it is sufficient to obtain stable results when the number of trees is larger than 150 (Lopes, 2015; Wadoux et al., 2019).

In the self-training SSL models, the threshold of entropy needed to be set. A sensitivity analysis of the impact of the entropy threshold parameter was conducted by setting the entropy threshold from 0.2 to 2.0 with an interval of 0.2.

### 2.4.3. Evaluation of the predicted soil maps based on different models

Soil class maps could be predicted based on different machine learning models. The performance of the models was evaluated by the soil prediction accuracy (ACC) based on the validation sample set. It is defined as the proportion of the correctly predicted samples over the total number of validation samples. The following equation was adopted:

$$\mathrm{ACC} = \frac{1}{N_v}\sum_{i=1}^{N_v}\mathbb{I}\left(\widehat{y}_i, y_i\right) \tag{2}$$

where $N_v$ is the total number of the validation samples, $y_i$ and $\widehat{y}_i$ are the true and predicted soil class for the $i$th validation sample respectively, $\mathbb{I}(\cdot)$ is an indicator function that $\mathbb{I}(\cdot) = 1$ if $\widehat{y}_i = y_i$ and $\mathbb{I}(\cdot) = 0$ otherwise.

As there is not enough sample data available to create a separate validation set, the validation method used in this study was cross-validation (Brus et al., 2011). A 5-fold cross-validation was employed. The total sample data set was partitioned by stratified split into five equal-size disjoint subsets, and 80% of the data (four of the partitions) were used to train the model and the remaining 20% data were used for

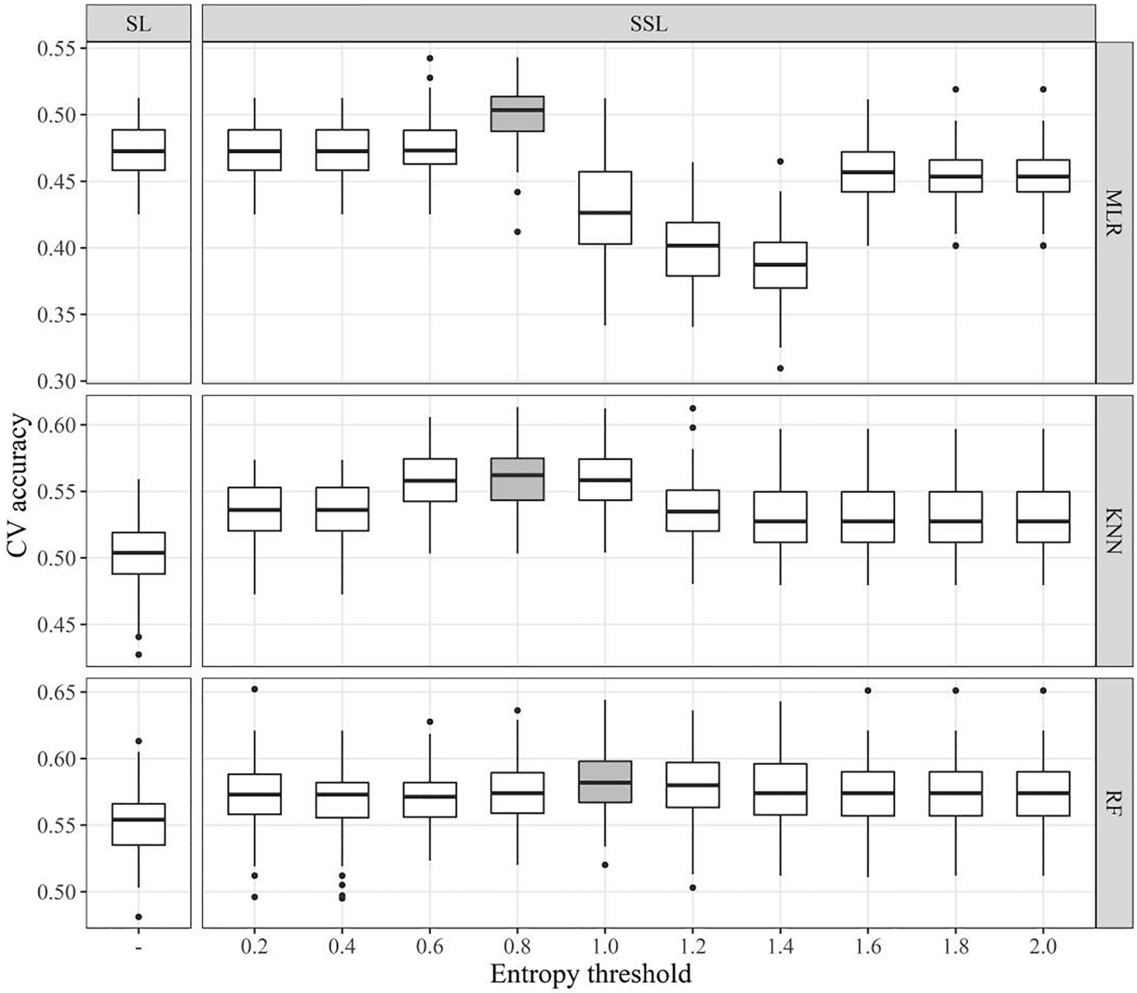


**Fig. 4.** Boxplots of cross-validation (CV) accuracies based on three models (multinomial logistic regression (MLR), k-nearest neighbor (KNN) and random forest (RF)) with the supervised learning (SL) method and the semi-supervised learning (SSL) method over different entropy thresholds, boxes marked in gray color for SSL represent the highest average CV accuracy at that threshold value. The midline represents the median value. The bottom line of the box represents the 25th percentile (Q1) of the values. The top line of the box represents the 75th percentile (Q3) of the values. The bottom whisker represents the range from Q1 – 1.5 * IQR (interquartile range) to Q1. The top whisker represents the range from Q3 to Q3 + 1.5 * IQR. The points represent outliers.“

validation. Then 5 runs of this process were performed by using each fold for validation once. To reduce the influence of randomness introduced by data split, the 5-fold cross-validation was repeated 100 times. The average cross-validation accuracy of the 100 times were taken as the results of the validation. Similar validation procedures were adopted in many previous studies (Brungard et al., 2015; Heung et al., 2016; 2014; Schmidt et al., 2008; Yang et al., 2019). The degree of the accuracy improvement of the SSL model compared with the SL model was calculated by the improvement ratio of the average cross-validation accuracy of the SSL model to the average cross-validation accuracy of the SL model.

## 3. Results

### 3.1. Comparisons of the prediction accuracies

The boxplots of the cross-validation accuracies for the three models with SL and SSL methods are presented in Fig. 4. These results for SSL were with a series of entropy thresholds ranging from 0.2 to 2.0 with an interval of 0.2. Boxes marked in gray color represent the highest average cross-validation accuracy at that threshold value. The average accuracy for SSL with KNN and RF was higher than that for SL at all threshold values. Yet, the average accuracy for SSL with MLR was not improved compared with SL with MLR for the threshold larger than 0.8. The value of 0.8 was the optimal threshold for MLR-SSL and KNN-SSL generating the highest average accuracy, and 1.0 was the optimal for RF-SSL. Under the optimal entropy threshold, the average cross-validation accuracies of SSL for MLR, KNN and RF were 0.50, 0.56, and 0.58, respectively, all of which were higher than the average accuracies of SL for the three machine learning models (0.47 for MLR, 0.50 for KNN, and 0.55 for RF). In this case, the self-training SSL method outperformed the SL method with an increase of accuracy by 5.9%, 12.2% and 6.0% for MLR, KNN and RF models, respectively. Although KNN generated the largest improvement, RF was the most accurate model for both SL and SSL. The results indicated that taking use of the unlabeled data improved the prediction accuracies. As shown in Table 2, a paired *t* test further shows that the accuracy improvement of the self-training SSL method is statistically significant at the 99% confidence (p-value $< 0.01$) (except for the MLR-SSL with the entropy threshold $\leq 0.6$ or $\geq 1.0$).

### 3.2. The predicted mapping results

Fig. 5 shows the maps of the predicted soil classes with the six models. It can be seen that the generated soil class maps with SL or SSL method showed generally similar spatial distribution patterns of soil classes for each machine learning model. The difference of the maps between model types were larger than between training method (SL or SSL). The soil maps with MLR models showed more difference with the maps with KNN or RF models. Compared to the predicted map generated in the previous study in this study area (Zhu et al., 2010), the soil maps based on RF models showed a more reasonable spatial distribution of soil classes. As for the comparison between predicted maps generated by the SL method and the SSL method, it can be seen that the predicted maps generated by the SSL method showed more reasonable than those generated by the SL method. For example, the soil type called Lithic Udi-Orthic Primosols can be more mapped from MLR-SSL comparing with MLR-SL, which is more similar to the previous map generated by the expert knowledge (Zhu et al., 2010). The maps generated with the SSL method showed a smoother effect and soil spatial distribution patterns become less speckled in appearance, which is an expected behavior consequence of using unlabeled data.

**Table 2**
Results of *t* test for the accuracy improvement of the models with semi-supervised learning (SSL) method over various entropy threshold values to the models with supervised learning (SL) method.

| Entropy threshold | p-value | | |
|---|---|---|---|
| | **MLR** | **KNN** | **RF** |
| 0.2 | 0.50 | $2.14 \times 10^{-23}$ | $6.43 \times 10^{-8}$ |
| 0.4 | 0.46 | $2.14 \times 10^{-23}$ | $1.21 \times 10^{-6}$ |
| 0.6 | 0.16 | $1.12 \times 10^{-41}$ | $3.54 \times 10^{-10}$ |
| 0.8 | $9.77 \times 10^{-17}$ | $4.58 \times 10^{-44}$ | $2.99 \times 10^{-12}$ |
| 1.0 | 1.00 | $4.14 \times 10^{-41}$ | $7.00 \times 10^{-18}$ |
| 1.2 | 1.00 | $1.80 \times 10^{-20}$ | $8.42 \times 10^{-13}$ |
| 1.4 | 1.00 | $2.14 \times 10^{-15}$ | $2.62 \times 10^{-11}$ |
| 1.6 | 1.00 | $2.14 \times 10^{-15}$ | $2.82 \times 10^{-10}$ |
| 1.8 | 1.00 | $2.14 \times 10^{-15}$ | $1.89 \times 10^{-10}$ |
| 2.0 | 1.00 | $2.14 \times 10^{-15}$ | $1.89 \times 10^{-10}$ |

## 4. Discussion

### 4.1. Impact of the entropy threshold for SSL

The entropy threshold is an important parameter in the proposed SSL method. Fig. 6 shows the accuracy improvement of the self-training SSL relative to SL for the three models over different entropy threshold values. It shows that the entropy threshold had an important impact on prediction accuracy. Yet the impact on different machine learning models was different. The accuracy improvement with the change of the entropy threshold had a similar trend for the three models. The three improvement curves all had an upward trend from the beginning, then achieved the highest point around a threshold of 0.8 ~ 1.0, and then the improvement decreased till stable under a threshold larger than 1.4 ~ 1.6. Both KNN and RF achieved a positive average improvement (larger than zero) over different entropy thresholds. The improvement using KNN was larger than that using RF at each threshold. For the RF model, the overall improvement with different entropy thresholds was relatively more stable with a range from 3.1 to 6.0. However, the accuracy improvement for MLR had a sharp decline with the threshold larger than 0.8. The accuracy improvement for MLR stayed negative starting from the threshold equal to 1.0. This was probably because that more wrong pseudo-label data were added when the entropy threshold became larger with the poor predictive performance of MLR-SL.

Table 3 shows percentage numbers of the increased (PN-I), decreased (PN-D) and equal (PN-E) accuracy with SSL to SL as evaluated based on 100 times cross-validation for three models over different thresholds of entropy. Similar to Fig. 6, KNN-SSL had the highest PN-I, followed by RF, and MLR had the lowest. It also can be seen that KNN-SSL guaranteed that the PN-I values were all greater than 90% with all entropy thresholds, and that of RF-SSL were greater than 85% when the entropy threshold was larger than or equaled to 0.8. The PN-I value of MLR-SSL was greater than its PN-D value when the threshold was less than or equaled to 0.8.

These results revealed that MLR-SSL was sensitive to the entropy threshold, KNN and RF was more robust. It seemed that a moderate value of the threshold was suitable for the model. This was consistent with a study of (Levatić et al., 2017), on the effect of the threshold on the performance of self-training. Their results indicated that a too permissive entropy threshold may include a large number of incorrect pseudo-labeled data, thus may lead to a worse performance, while a too strict threshold does not allow self-training to benefit from the unlabeled data. In our case study, 0.8 was an appropriate threshold value for MLR and KNN, while 1.0 was the most appropriate for RF.

### 4.2. Applicability and limitation of the self-training method

This paper employed a self-training semi-supervised learning method for the first time in soil prediction and mapping. The results showed a better performance of the self-training SSL method compared with the SL method in a case study with limited training samples. The

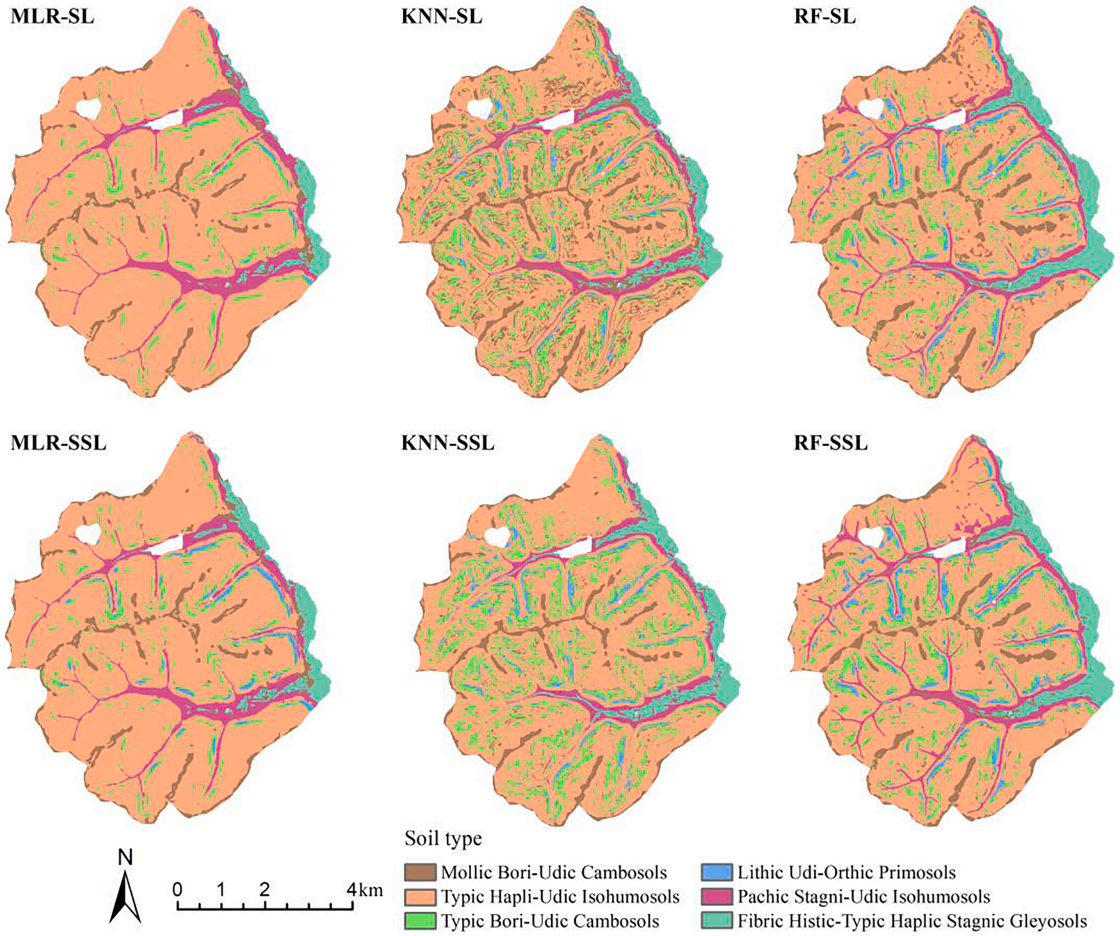

MLR-SL
KNN-SL
RF-SL
MLR-SSL
KNN-SSL
RF-SSL
N
0 1 2 4km
Soil type
Mollic Bori-Udic Cambosols
Typic Hapli-Udic Isohumosols
Typic Bori-Udic Cambosols
Lithic Udi-Orthic Primosols
Pachic Stagni-Udic Isohumosols
Fibric Histic-Typic Haplic Stagnic Gleyosols

**Fig. 5.** Predictive mapping of soil classes by using three models (multinomial logistic regression (MLR), k-nearest neighbor (KNN) and random forest (RF)) with supervised learning (SL) and semi-supervised learning (SSL) methods.

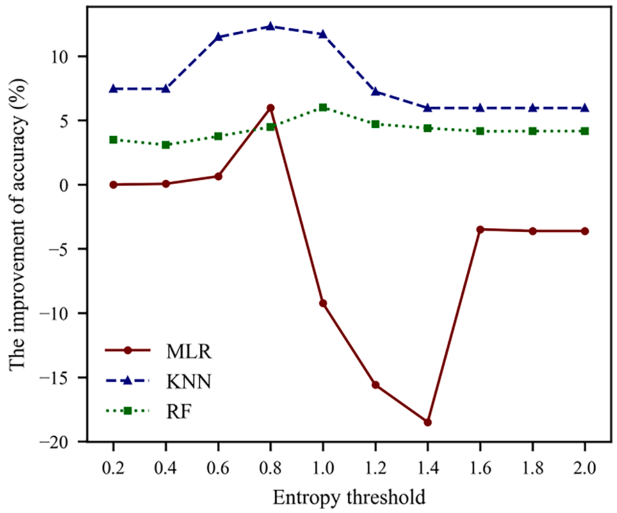


**Fig. 6.** The accuracy improvement of three models (multinomial logistic regression (MLR), k-nearest neighbor (KNN) and random forest (RF)) with semi-supervised learning (SSL) to supervised learning (SL) over different thresholds of entropy.

accuracy of the three models (MLR, KNN and RF) were all improved with the self-training method, however, the improvement was different for the three models. This may due to the following reasons. As MLR is a simple linear model, the correctness of the pseudo-label set generated by MLR might be relatively low, leading to a low improvement of MLR-SSL. Our result showed that KNN was a better model than MLR, which was consistent with the previous studies (Brungard et al., 2015; Heung et al., 2016). This led to the correctness of the pseudo-label determined by KNN could be relatively larger than that by MLR, which contributed to the higher performance gain of the KNN model. For the RF model, it achieved the highest performance with SL, and generated a high correctness of the pseudo-label set leading to the most stable improvement of RF. However, the improvement of RF was less than that of KNN. As many previous comparative studies have shown that RF was consistently the most accurate model across different study areas, maybe it was relatively more difficult to have a larger improvement for RF with a high accuracy. However, with the self-training SSL method, the predictive performance of RF can still be effectively further improved. From the result of our case study, RF-SSL is the most accurate model. it is recommended to adopt RF as the base model when using the self-training SSL method. KNN is also an alternative base model with the SSL method. For MLR, as its low accuracy with SL, it may have a risk of accuracy reduction when using self-training with large entropy thresholds.

Although the self-training SSL method could improve the model performance compared to the SL method in most cases, there were still cases in which it did not improve the performance, especially for MLR. To better understand the potential reason for this, we generated the scatter plots between the values of the cross-validation (CV) accuracy of the SL method and the values of the improvement of SSL over SL for the three models, as shown in Fig. 7. The entropy threshold was set to 0.8 for

**Table 3**
Percentage numbers of the increased, decreased and equal accuracy with semi-supervised learning (SSL) to supervised learning (SL) in the 100 repeats as evaluated based on cross-validation for three models (multinomial logistic regression (MLR), k-nearest neighbor (KNN) and random forest (RF)) over different thresholds of entropy.

| Model | Entropy threshold | PN-I (%) | PN-D (%) | PN-E (%) |
|---|---|---|---|---|
| MLR-SSL | 0.2 | 0 | 0 | 100 |
| | 0.4 | 2 | 0 | 98 |
| | 0.6 | 33 | 15 | 52 |
| | 0.8 | 87 | 12 | 1 |
| | 1.0 | 10 | 90 | 0 |
| | 1.2 | 0 | 100 | 0 |
| | 1.4 | 0 | 100 | 0 |
| | 1.6 | 10 | 84 | 6 |
| | 1.8 | 10 | 84 | 6 |
| | 2.0 | 10 | 84 | 6 |
| KNN-SSL | 0.2 | 90 | 10 | 0 |
| | 0.4 | 90 | 10 | 0 |
| | 0.6 | 97 | 2 | 1 |
| | 0.8 | 99 | 1 | 0 |
| | 1.0 | 99 | 1 | 0 |
| | 1.2 | 95 | 5 | 0 |
| | 1.4 | 91 | 7 | 2 |
| | 1.6 | 91 | 7 | 2 |
| | 1.8 | 91 | 7 | 2 |
| | 2.0 | 91 | 7 | 2 |
| RF-SSL | 0.2 | 75 | 19 | 6 |
| | 0.4 | 74 | 22 | 4 |
| | 0.6 | 74 | 26 | 0 |
| | 0.8 | 85 | 9 | 6 |
| | 1.0 | 95 | 3 | 2 |
| | 1.2 | 88 | 8 | 4 |
| | 1.4 | 87 | 7 | 6 |
| | 1.6 | 90 | 7 | 3 |
| | 1.8 | 91 | 7 | 2 |
| | 2.0 | 91 | 7 | 2 |

PN-I: percentage number of the increased accuracy with SSL to SL in the 100 repeats; PN-D: percentage number of the decreased accuracy with SSL to SL in the 100 repeats; PN-E: percentage number of the equal accuracy with SSL to SL in the 100 repeats.

both MLR and KNN, and the threshold was set to 1.0 for RF. It is worth noting that there was a clear negative relationship between the SL accuracy and the improvement of accuracy. The situations that the accuracy increased significantly mostly occurs when the SL accuracy was relatively low. Therefore, it indicated that the proposed self-training SSL method was more applicable to the cases with the low SL model performance. That is to say, when using a set of training samples that make it difficult to train a model with good performance, self-training is an effective method for improving accuracy. Meanwhile, it is also can be seen that the self-training method could still have the possibility of improving the accuracy when the accuracy was relatively high under supervised learning, though the improvement may not be significant.

The entropy threshold is an important user-defined parameter in the proposed method and had an important impact on the result. This threshold is a tuning parameter and needs to be adjusted when using the self-training method. According to the result of the sensitive analysis about the threshold, we suggest a reasonable solution to select the appropriate threshold is to start with a small value (e.g. 0.1), and then increase the threshold step by step with a short interval until the model performance cannot be improved.

In addition, the proposed self-training SSL method is suitable in situations where field samples are limited. It can be considered as an alternative way to further improve the model performance. The self-training method has the advantage of being able to be “wrapped” around any existing supervised machine learning method as the base predictive model and can improve the performance of the base supervised learning model by using the abundant unlabeled data.

## 5. Conclusions

This paper developed a self-training semi-supervised learning method that used both labeled and unlabeled data for soil class prediction. MLR, KNN, and RF were used as the base machine learning models for comparing semi-supervised and supervised learning methods. The results showed that the self-training SSL method improved the soil prediction accuracy compared with the SL method. RF-SSL was the most accurate model in the study area with an average accuracy of 0.58 at an entropy threshold of 1.0, followed by KNN-SSL. Meanwhile, the self-training method for the KNN model had the largest improvement comparing with the other two models. Yet the performance of MLR can also be improved with the self-training method, but it needs to be more careful with the parameter tuning to avoid a risk of accuracy reduction. In our future study, we will focus on improving the algorithm

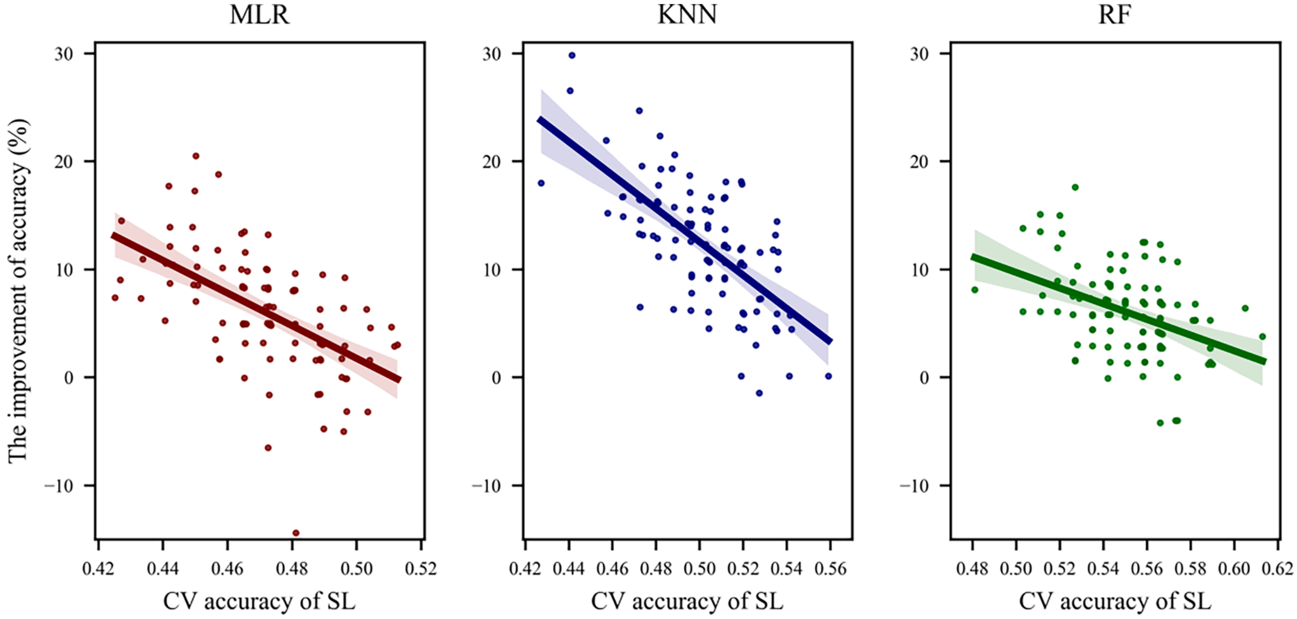


**Fig. 7.** The relationship between cross-validation (CV) accuracy and the improvement ratio of semi-supervised learning (SSL) over supervised learning (SL) by three models.

to reduce the possibility of accuracy reduction, for example, by designing a strategy to pre-select the representative unlabeled data to avoid the addition of the noise into the model. As for the important parameter, entropy threshold, the results showed that a moderate threshold, neither too permissive nor strict, could lead to a good performance. The suitable value of the entropy threshold was 0.8 ~ 1.0 in the case study. In the future, how to automatically extract the entropy threshold value, rather than setting by subjective experience, will be an important research question. Finally, we concluded that the self-training SSL method is a potential effective and accurate method for DSM with limit training sample data.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This study is supported by the National Natural Science Foundation of China (Project No. 41971054; 41530749; 41871300). We would like to express our sincere gratitude to Editor Kristin Piikki and the two anonymous reviewers, whose valuable comments and suggestions have greatly improved the quality of the paper.

## References

Arrouays, D., Poggio, L., Salazar Guerrero, O.A., Mulder, V.L., 2020. Digital soil mapping and GlobalSoilMap. Main advances and ways forward. Geoderma Reg. 21, e00265. https://doi.org/10.1016/j.geodrs.2020.e00265.

Beaudoin, A., Bernier, P.Y., Guindon, L., Villemaire, P., Guo, X.J., Stinson, G., Bergeron, T., Magnussen, S., Hall, R.J., 2014. Mapping attributes of Canada's forests at moderate resolution through k NN and MODIS imagery. Can. J. For. Res. 44 (5), 521–532. https://doi.org/10.1139/cjfr-2013-0401.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Z. Pflanzenernähr. Bodenk. 168 (1), 21–33. https://doi.org/10.1002/jpln.200421414.

Bernier, P.Y., Daigle, G., Rivest, L.P., Ung, C.H., Labbé, F., Bergeron, C., Patry, A., 2010. From plots to landscape: A k-NN-based method for estimating stand-level merchantable volume in the Province of Québec. Can. For. Chron. 86, 461–468. https://doi.org/10.5558/tfc86461-4.

Beven, K.J., Kirkby, N.J., 1979. A physically based variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69.

Blum, A., Chawla, S., 2001. Learning from Labeled and Unlabeled Data using Graph Mincuts. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp. 19–26. https://doi.org/10.1184/R1/6606860.v1.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, WI. ACM, New York, pp. 92–100. https://doi.org/10.1145/279943.279962.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., 1996. Bagging predictors. Mach Learn 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr., T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. 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.

Chapelle, O., Scholkopf, B., Zien, A., 2006. Semi-Supervised Learning. MIT Press, Cambridge.

Chen, Z., Chen, D., Zhao, C., Kwan, M.P., Cai, J., Zhuang, Y., Zhao, B., Wang, X., Chen, B., Yang, J., Li, R., He, B., Gao, B., Wang, K., Xu, B., 2020. Influence of meteorological conditions on $PM_{2.5}$ concentrations across China: A review of methodology and mechanism. Environ. Int. 139, 105558. https://doi.org/10.1016/j.envint.2020.105558.

Chinese Soil Taxonomy Research Group, 2001. Keys to Chinese Soil Taxonomy, 3rd edition. University of Science and Technology of China Press, Hefei.

de Gruijter, J.J., Bierkens, M.F.P., Brus, D.J., Knotters, M. (Eds.), 2006. Sampling for Natural Resource Monitoring. Springer Berlin Heidelberg, Berlin, Heidelberg.

Debella-Gilo, M., Etzelmüller, B., 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. Catena 77 (1), 8–18. https://doi.org/10.1016/j.catena.2008.12.001.

Du, F., Zhu, A.X., Liu, J., Yang, L., 2020. Predictive mapping with small field sample data using semi-supervised machine learning. Transactions in GIS 24 (2), 315–331. https://doi.org/10.1111/tgis.12598.

Efron, B., Tibshirani, R.J. (Eds.), 1993. An Introduction to the Bootstrap. Springer US, Boston, MA.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146 (1-2), 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008.

Han, J., Zhang, D., Cheng, G., Guo, L., Ren, J., 2015. Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning. IEEE Trans. Geosci. Remote Sensing 53 (6), 3325–3337. https://doi.org/10.1109/TGRS.2014.2374218.

Han, W., Feng, R., Wang, L., Cheng, Y., 2018. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. ISPRS J. Photogramm. Remote Sens. 145, 23–43. https://doi.org/10.1016/j.isprsjprs.2017.11.004.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. 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.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma 214-215, 141–154. https://doi.org/10.1016/j.geoderma.2013.09.016.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. 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.

Heuvelink, G.B.M., Kros, J., Reinds, G.J., De Vries, W., 2016. Geostatistical prediction and simulation of European soil property maps. Geoderma Regional 7 (2), 201–215. https://doi.org/10.1016/j.geodrs.2016.04.002.

Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100 (3-4), 269–301. https://doi.org/10.1016/S0016-7061(01)00025-8.

Jafari, A., Finke, P.A., Vande Wauw, J., Ayoubi, S., Khademi, H., 2012. Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. Eur. J. Soil Sci. 63, 284–298. https://doi.org/10.1111/j.1365-2389.2012.01425.x.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013a. Tree-Based Methods. In: An Introduction to Statistical Learning: With Applications in R. Springer, New York, New York, NY, pp. 303–335. https://doi.org/10.1007/978-1-4614-7138-7_8.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013b. Classification. In: An Introduction to Statistical Learning: With Applications in R. Springer, New York, New York, NY, pp. 127–173. https://doi.org/10.1007/978-1-4614-7138-7_4.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151 (3-4), 311–326. https://doi.org/10.1016/j.geoderma.2009.04.023.

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58. https://doi.org/10.1016/j.geoderma.2018.12.037.

Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154 (3-4), 340–347. https://doi.org/10.1016/j.geoderma.2009.11.005.

Lagacherie, P., 2008. Digital Soil Mapping: A State of the Art. In: Hartemink, A.E., McBratney, A., Mendonça-Santosde, M.L. (Eds.), Digital soil mapping with limited data. Springer Netherlands, Dordrecht, pp. 3–14. https://doi.org/10.1007/978-1-4020-8592-5_1.

Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma 352, 395–413. https://doi.org/10.1016/j.geoderma.2019.05.031.

Levatić, J., Ceci, M., Kocev, D., Džeroski, S., 2017. Self-training for multi-target regression with tree ensembles. Knowl.-Based Syst. 123, 41–60. https://doi.org/10.1016/j.knosys.2017.02.014.

Li, R., Xu, M., Chen, Z., Gao, B., Cai, J., Shen, F., He, X., Zhuang, Y., Chen, D., 2021. Phenology-based classification of crop species and rotation types using fused MODIS and Landsat data: The comparison of a random-forest-based model and a decision-rule-based model. Soil Tillage Res. 206, 104838. https://doi.org/10.1016/j.still.2020.104838.

Lopes, M.E., 2015. Measuring the algorithmic convergence of random forests via bootstrap extrapolation. Technical Report. Department of Statistics. University of California, Davis CA.

Ma, T., Brus, D.J., Zhu, A.X., Zhang, L., Scholten, T., 2020. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. Geoderma 370, 114366. https://doi.org/10.1016/j.geoderma.2020.114366.

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232-234, 34–44. https://doi.org/10.1016/j.geoderma.2014.04.033.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method. Geoderma 235-236, 59–73. https://doi.org/10.1016/j.geoderma.2014.06.032.

Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58, 1246–1266.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1-2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., Van Molle, M., 2008. A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 143 (1-2), 1–13. https://doi.org/10.1016/j.geoderma.2007.08.025.

Miao, C., Yang, L., Liu, B., Gao, Y., Li, S., 2011. Streamflow changes and its influencing factors in the mainstream of the Songhua River basin, Northeast China over the past 50 years. Environ. Earth Sci. 63, 489–499. https://doi.org/10.1007/s12665-010-0717-x.

Mishra, U., Torn, M.S., Masanet, E., Ogle, S.M., 2012. Improving regional soil carbon inventories: Combining the IPCC carbon inventory method with regression kriging. Geoderma 189-190, 288–295. https://doi.org/10.1016/j.geoderma.2012.06.022.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perez, F., Granger, B.E., Hunter, J.D., 2011. Python: An Ecosystem for Scientific Computing. Comput. Sci. Eng. 13 (2), 13–21. https://doi.org/10.1109/MCSE.2010.119.

Priori, S., Bianconi, N., Costantini, E.A.C., 2014. Can γ-radiometrics predict soil textural data and stoniness in different parent materials? A comparison of two machine-learning methods. Geoderma 226-227, 354–364. https://doi.org/10.1016/j.geoderma.2014.03.012.

Qin, C.Z., Zhu, A.X., Pei, T., Li, B.-L., Scholten, T., Behrens, T., Zhou, C.H., 2011. An approach to computing topographic wetness index based on maximum downslope gradient. Precision Agric 12 (1), 32–43. https://doi.org/10.1007/s11119-009-9152-y.

Qin, C.Z., Zhu, A.X., Shi, X., Li, B.L., Pei, T., Zhou, C.-H., 2009. Quantification of spatial gradation of slope positions. Geomorphology 110 (3–4), 152–161. https://doi.org/10.1016/j.geomorph.2009.04.003.

Qin, C.Z., Zhu, A.X., Pei, T., Li, B., Zhou, C., Yang, L., 2007. An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. International Journal of Geographical Information Science 21 (4), 443–458. https://doi.org/10.1080/13658810601073240.

Rawlins, B.G., Henrys, P., Breward, N., Robinson, D.A., Keith, A.M., Garcia-Bajo, M., 2011. The importance of inorganic carbon in soil carbon databases and stock estimates: a case study from England. Soil Use Manag. 27, 312–320. https://doi.org/10.1111/j.1475-2743.2011.00348.x.

Romaszewski, M., Głomb, P., Cholewa, M., 2016. Semi-supervised hyperspectral classification from a small number of training samples using a co-training approach. ISPRS J. Photogramm. Remote Sens. 121, 60–76. https://doi.org/10.1016/j.isprsjprs.2016.08.011.

Russell, S.J., Norvig, P., 1995. Artificial Intelligence: A Modern Approach. Prentice-Hall Inc, USA.

Samaniego, L., Schulz, K., 2009. Supervised classification of agricultural land cover using a modified k-NN technique (MNN) and Landsat remote sensing imagery. Remote Sens 1, 875–895. https://doi.org/10.3390/rs1040875.

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.D.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009. Digital Soil Map of the World. Science 325 (5941), 680–681. https://doi.org/10.1126/science:1175084.

Schmidt, K., Behrens, T., Scholten, T., 2008. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146 (1-2), 138–146. https://doi.org/10.1016/j.geoderma.2008.05.010.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography: Earth and Environment 27 (2), 171–197. https://doi.org/10.1191/0309133303pp366ra.

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Silva, J., Bacao, F., Caetano, M., 2017. Specific land cover class mapping by semi-supervised weighted support vector machines. Remote Sens 9. https://doi.org/10.3390/rs9020181.

Stoorvogel, J.J., Kempen, B., Heuvelink, G.B.M., de Bruin, S., 2009. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. Geoderma 149 (1-2), 161–170. https://doi.org/10.1016/j.geoderma.2008.11.039.

Subburayalu, S.K., Slater, B.K., 2013. Soil series mapping by knowledge discovery from an ohio county soil map. Soil Sci. Soc. Am. J. 77 (4), 1254–1268. https://doi.org/10.2136/sssaj2012.0321.

Sun, W., Minasny, B., McBratney, A., 2012. Analysis and prediction of soil properties using local regression-kriging. Geoderma 171-172, 16–23. https://doi.org/10.1016/j.geoderma.2011.02.010.

Triguero, I., García, S., Herrera, F., 2015. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowl. Inf. Syst. 42 (2), 245–284. https://doi.org/10.1007/s10115-013-0706-y.

Vasques, G.M., Demattê, J.A.M., Viscarra Rossel, R.A., Ramírez-López, L., Terra, F.S., 2014. Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. Geoderma 223-225, 73–78. https://doi.org/10.1016/j.geoderma.2014.01.019.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A.P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.-L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y., Contributors, S., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Wadoux, A.M.J.C., Brus, D.J., Heuvelink, G.B.M., 2019. Sampling design optimization for soil mapping with random forest. Geoderma 355, 113913. https://doi.org/10.1016/j.geoderma.2019.113913.

Webster, R., Oliver, M.A., 1990. Statistical methods in soil and land resource survey. Oxford University Press, New-York, USA.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Ind. 52, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028.

Wiesmeier, M., Barthold, F., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., Schilling, B., Angst, G., von Lützow, M., Kögel-Knabner, I., 2014. Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany). Geoderma Regional 1, 67–78. https://doi.org/10.1016/j.geodrs.2014.09.001.

Yang, L., Song, M., Zhu, A.-X., Qin, C., Zhou, C., Qi, F., Li, X., Chen, Z., Gao, B., 2019. Predicting soil organic carbon content in croplands using crop rotation and Fourier transform decomposed variables. Geoderma 340, 289–302. https://doi.org/10.1016/j.geoderma.2019.01.015.

Yang, L., Zhu, A.X., Li, B.L., Qin, C.Z., Pei, T., Liu, B.Y., Li, R.K., Cai, Q.G., 2007. Extraction of knowledge about soil-environment relationship for soil mapping using fuzzy c-means (FCM) clustering. Acta Pedol. Sin. 44, 16–23.

Yao, X., Han, J., Cheng, G., Qian, X., Guo, L., 2016. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. IEEE Trans. Geosci. Remote Sensing 54 (6), 3660–3671. https://doi.org/10.1109/TGRS.2016.2523563.

Yarowsky, D., 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, ACL 1995. Association for Computational Linguistics, USA, pp. 189–196. doi:10.3115/981658.981684.

Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P., 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338, 445–452. https://doi.org/10.1016/j.geoderma.2018.09.006.

Zhu, A.X., 2000. Mapping soil landscape as spatial continua: The Neural Network Approach. Water Resour. Res. 36 (3), 663–677. https://doi.org/10.1029/1999WR900315.

Zhu, A.X., Liu, J., Du, F., Zhang, S.J., Qin, C.Z., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. Eur. J. Soil Sci. 66, 535–547. https://doi.org/10.1111/ejss.12244.

Zhu, A.X., Lu, G., Liu, J., Qin, C., Zhou, C., 2018. Spatial prediction based on Third Law of Geography. Ann. Gis 24 (4), 225–240. https://doi.org/10.1080/19475683.2018.1534890.

Zhu, A.X., Yang, L., Li, B.L., Qin, C.Z., English, E., Burt, J.E., Zhou, C.H., 2008. Purposive Sampling for Digital Soil Mapping for Areas with Limited Data. In: Hartemink, A.E., McBratney, A., Mendonça-Santosde, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer Netherlands, Dordrecht, pp. 233–245. https://doi.org/10.1007/978-1-4020-8592-5_20.

Zhu, A.X., Yang, L., Li, B., Qin, C., Pei, T., Liu, B., 2010. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma 155 (3–4), 164–174. https://doi.org/10.1016/j.geoderma.2009.05.024.

Zhu, X., 2006. Semi-Supervised Learning Literature Survey. Technical Report 1530. Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI.

Zhu, X., Goldberg, A.B., 2009. Introduction to Semi-Supervised Learning. Synth. Lect. Artif. Intell. Mach. Learn. 3 (1), 1–130. https://doi.org/10.2200/S00196ED1V01Y200906AIM006.