Geoderma 406 (2022) 115531

Contents lists available at ScienceDirect

Geoderma

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

A multiple soil properties oriented representative sampling strategy for digital soil mapping

Lei Zhang [a], Lin Yang [a,b,*], Yanyan Cai [a], Haili Huang [a], Jingjing Shi [b], Chenghu Zhou [a,b]

[a] School of Geography and Ocean Science, Nanjing University, Nanjing 210023, PR 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, PR China

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Soil sampling
Multiple soil properties
Clustering
Environmental variables
Digital soil mapping

ABSTRACT

Sampling design plays a key role in digital soil mapping (DSM). Efficient sampling design for multiple soil properties is increasingly needed for multivariate soil survey and mapping. However, most of the present sampling methods are not developed for multiple soil properties. Different soil properties have different influential covariates, but usually only one set of covariates is used in designing samples for multiple soil properties which makes simultaneously mapping multiple soil properties accurately difficult. This paper proposed a multiple soil properties oriented representative sampling strategy (MPRS) by considering the influential environmental covariates for each soil property. The method first selects the most influential set of environmental covariates for each soil property, then uses fuzzy *c*-means (FCM) clustering to generate environmental clusters relating to spatial variation patterns for each soil property, and the selected samples are representative of as many typical locations of environmental clusters for multiple soil properties as possible. The proposed sampling method was applied for mapping soil sand content and soil organic matter content at surface (0–20 cm) and subsurface (20–40 cm) layers in a study area with 5900 km^2 located in Anhui Province, China, and compared with two methods, the purposive sampling (PS) method and integrative hierarchical stepwise sampling (IHS) method. The results showed that the proposed sampling method achieved the most accurate prediction for most of the four soil properties over different sample sizes. The proposed sampling method also has an advantage to extract representative samples which can better cover multiple soil properties with a limit of a small sample size. On average, the improvement of prediction accuracy by using the MPRS method was 38.1% and 36.3% compared with PS and IHS in terms of R^2, 4.8% and 4.6% in terms of RMSE, and 11.7% and 13.7% in terms of CCC, respectively. Our case study confirmed the necessity to consider the difference of the influential environmental variable combinations for the multiple soil properties oriented sampling design. We conclude that MPRS is a potential effective method for supporting DSM for multiple soil properties.

1. Introduction

The spatial distribution of soil types/properties is basic data for land management and ecological modeling (Band and Moore, 1995; McBratney et al., 2003). Digital soil mapping (DSM) has become the most common approach to obtain soil maps over the past decades (Lagacherie, 2008; Lamichhane et al., 2019; McBratney et al., 2003; Zhu et al., 2010; Zhu et al., 2015; Heung et al., 2016; Arrouays et al., 2020; Yang et al., 2021a). DSM uses sample data and environmental variables (covariates) to predict the spatial distribution of soil types or properties. Sampling methods to determine where to collect samples profoundly impact the quality of the resulting soil maps (Brus and de Gruijter, 1997; de Gruijter et al., 2006; Carter and Gregorich, 2007; Gregoire and Valentine, 2008; Brus, 2015). However, field sampling is costly and labor-intensive (Webster and Oliver, 1990; Zhu et al., 2008; Yang et al., 2018). Therefore, how to design a set of limited sample points to generate accurate soil maps is a key research topic in DSM.

Numerous studies have been devoted to sampling optimization for soil mapping. When only considering the spatial distribution of samples, the spatial coverage sampling, such as sampling based on *k*-means clustering in geographical space, has been developed to design an even distribution of sample set in geographic space (Brus et al., 2006; Royle

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

https://doi.org/10.1016/j.geoderma.2021.115531
Received 28 June 2021; Received in revised form 7 October 2021; Accepted 7 October 2021
0016-7061/© 2021 Elsevier B.V. All rights reserved.

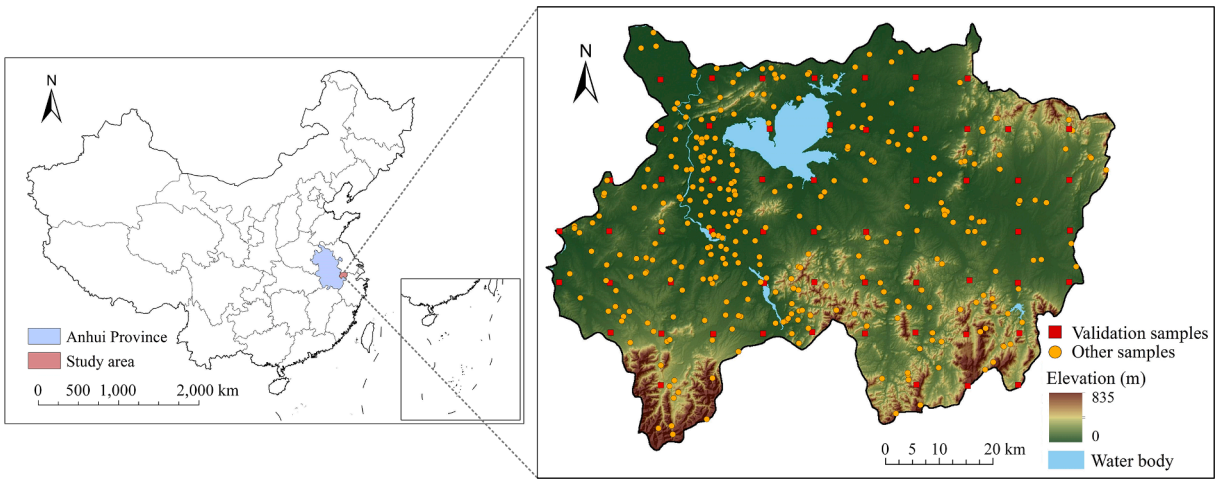

Fig. 1. The location, the digital elevation model and field samples points of the study area.

and Nychka, 1998; Walvoort et al., 2010). Sampling based on geostatistics has also been applied in many studies, in which a sample design is optimized based on a prior semivariogram of the target soil property (Goovaerts, 1999; Hengl et al., 2004). According to Jenny's (1941) theory of soil forming factors for linking the soil development with multiple environmental factors, many methods have been developed to use ancillary environmental covariates to optimize the sample design in the environmental feature space (Isaaks and Srivastava, 1989; Goovaerts, 1999; Minasny and McBratney, 2006; Brus and Heuvelink, 2007; Ma et al., 2020a). For example, a popular sampling method, the conditional Latin hypercube sampling (cLHS), selects samples that fully cover the range of each ancillary environmental variable by maximally stratifying the marginal distribution of environmental variables (Minasny and McBratney, 2006). In addition to the way of selecting samples by fully covering the feature space, researchers designed samples by capturing the typical types of soil variations with assistance of environmental covariates. Zhu et al. (2008) proposed a purposive sampling (PS) approach based on the concept that unique soil conditions (soil types or soil properties) can be associated with unique combination (configuration) of environmental conditions. This method used fuzzy *c*-means (FCM) clustering to identify those unique combinations and take locations on cluster centroids as typical sample points. Furthermore, Yang et al. (2013) proposed an integrative hierarchical stepwise sampling (IHS) strategy by selecting typical sample points representative of large-scale spatial patterns and local patterns of soil variations successively. It was tested to generate soil maps with higher accuracy than the stratified random sampling (SRS) or cLHS with limited samples at both watershed and regional scales (Yang et al., 2016; Yang et al., 2017).

In practice, it is often that the purpose of a soil survey is to map multiple soil properties. In this case, most of the present sampling methods may be not applicable or efficient (Vašát et al., 2010). It is, therefore, increasingly significant to design sampling methods optimized for multivariate soil mapping. Vašát et al. (2010) developed a sampling method for multiple soil variables by minimizing the average kriging variance for multiple soil properties simultaneously. In their method, a linear model of coregionalization with prior sample data was employed to fit a variogram model for adapting the mutual spatial dependence of different soil properties. Szatmári et al. (2016) used the regression kriging variance to optimize the sampling design for multiple soil properties by using an extended spatial simulated annealing (SSA) method. Szatmári et al. (2019) further proposed a two-phase sampling method based on regression kriging for multivariate soil mapping considering the economic efficiency. The above sampling methods generally adopted geostatistical models such as ordinary kriging or regression kriging. However, it is often difficult to verify the stationary assumption for the geostatistical models in many complicated field conditions. Besides, the influential environmental covariates for different soil properties are not explicitly accounted for in these methods.

Those strategies exploring clusters in environmental feature space to relating to combinations of environmental covariates to design samples can be a potential effective solution for mapping multiple soil properties. The spatial variation type of each soil property can be associated with one combination of environmental variables (Qi et al., 2006; Zhu et al., 2008; 2010). It is effective to sample at locations typical of soil variation types. Brus (2019) mentioned in a review that further studies into efficient sampling designs based on fuzzy memberships are needed. In the aforementioned sampling methods of Zhu et al. (2008) and Yang et al. (2013), only one set of environmental variables were employed to design samples based on environmental clustering. However, the most influential set of environmental variables for mapping different soil properties are different (Behrens et al., 2014; Miller et al., 2015; Shi et al., 2018), because the environmental variables have various impacts on different soil properties under different formation mechanisms (Gessler et al., 1995; Xiong et al., 2012; Luo et al., 2017; Ma et al., 2019; Xu et al., 2019; Yang et al., 2021b). Considering the different influential environmental variables for multiple soil properties can be used to select typical samples representative of spatial variations for multiple soil properties.

In this paper, we propose a multiple soil properties oriented representative sampling strategy (MPRS). This strategy designs one sample set that can be simultaneously representative toward different soil properties based on an environmental clustering analysis after knowing the most influential set of environmental variables to each target soil property. The influential environmental variables to each soil property was determined using machine learning based on prior sample data in this study. The next section presents a detailed introduction to the methodology and a case study was conducted in Xuancheng, Anhui province, China. The proposed MPRS method was evaluated and compared with the purposive sampling (PS) and integrative hierarchical stepwise sampling (IHS) for mapping soil sand content and soil organic matter content of both the surface layer (0–20 cm) and the subsurface layer (20–40 cm).

2. Materials and methodology

2.1. Study area and data

2.1.1. Study area

The study area is Xuanzhou City, Guangde County, and Langxi County in Anhui Province, China (Fig. 1). Its area is about 5900 km^2. The climate of this area is warm and humid in summer and relatively cool

Table 1
Description of the environmental variables used in the study area.

Variable type	Variable name	Abbreviation	Reference or data source
Topography	Elevation	ELEV	Shuttle Radar Topography Mission (SRTM)
	Slope (cosine)	SLP	Zevenbergen and Thorne (1987)
	Aspect	ASP	Zevenbergen and Thorne (1987)
	Profile curvature	PROFC	Zevenbergen and Thorne (1987)
	Plan curvature	PLANC	Zevenbergen and Thorne (1987)
	Relief degree	REF	Park and van de Giesen (2004)
	Terrain ruggedness index	TRI	Riley et al. (1999)
	Total curvature	TC	Wilson and Gallant (2000)
	Topographic position index	TPI	Qin et al. (2009)
	Terrain characterization index	TCI	Park and van de Giesen (2004)
	Surface curvature	CS	Park and van de Giesen (2004)
	Specific catchment area	SCA	Beven and Kirkby, (1979)
	Topographic wetness index	TWI	Quinn et al. (1991); Qin et al. (2011)
	Stream power index	SPI	Moore et al. (1993)
Climate	Annual mean temperature	TEMP	National meteorological dataset of China (http://www.resdc.cn)
	Annual precipitation	PRECI	National meteorological dataset of China (http://www.resdc.cn)
Vegetation	NDVI in the spring (the average from 2009 to 2011)	NDVI-1	http://earthexplorer.usgs.gov
	NDVI in the summer (the average from 2009 to 2011)	NDVI-2	http://earthexplorer.usgs.gov
	NDVI in the autumn (the average from 2009 to 2011)	NDVI-3	http://earthexplorer.usgs.gov
Parent materials	Lithology type	PM	Chinese ecological environmental database (1:500,000)

and dry in winter. Its annual average temperature ranges from 11 to 16℃ and the average precipitation is 1200–1800 mm. The elevation ranges from 0 to 1039 m with flat plains in the northwest and hills in the northeast and south. The main soil types in this area are red soil and paddy soil in the Chinese genetic soil classification system (Chinese National Soil Survey Office, 1992). Land use is mainly cultivated land with rice as the dominant crop, and forest land covered by bamboo, fir, shrub, and other evergreen broad-leaved trees.

The soil sand content of the surface layer at a depth of 0–20 cm (SandA) and the subsurface layer at a depth of 20–40 cm (SandB), and the soil organic matter content of the surface layer (SOMA) and the subsurface layer (SOMB) are the target soil properties.

2.1.2. Environmental covariate data

According to the previous studies on the soil-environment relationships in this area (An et al., 2018; Yang et al., 2019), we used the following environmental covariates, including parent material, vegetation, and topographic variables such as elevation, slope gradient, slope aspect, planform curvature, profile curvature, and topographic wetness index. The raw form of slope aspect was transformed into a measure of northness vs. southness using the cosine function. The transformed values are close to 1 if the aspect is generally northward, close to −1 if the aspect is southward, and close to 0 if the aspect is either east or west (Roberts, 1986). The information of these covariates are described in Table 1. The parent material map was generated from the 1: 500,000 geological map of China. There are eight parent lithology types in the study area (Fig. 2). All topographic variables were derived from the digital elevation model (DEM) obtained from the Shuttle Radar Topographic Mission (SRTM) dataset (Farr and Kobrick, 2000). The Normalized Difference Vegetation Index (NDVI) was used as the vegetation index. We averaged the Landsat-5 remote sensing image data with cloud cover less than 10% of the three seasons separately (spring: March to May, summer: June to August, and autumn: September to November) from 2009 to 2011 (cover the sampling time). All covariate maps were resampled into the resolution of 90 m which was consistent with the resolution of the DEM dataset.

2.1.3. Sample data

There were 393 sampling points collected in 2011, 2015 and 2016 through several investigations in the study area (Fig. 1). Among these samples, 57 points collected based on a systematic sampling strategy of a 10 km by 10 km grid arrangement were used as an independent validation set in this study. The remaining 336 samples included 58 samples selected using an integrative hierarchical stepwise sampling (Yang et al., 2013), 57 samples collected by a heuristic uncertainty directed sampling (Zhang et al., 2016), 30 samples collected by environmental similarity-based recommendation (Ma et al, 2020b, Zhu et al., 2018), and other samples were collected by using stratified random sampling with the parent lithology types as strata.

At each sampling location, soils were sampled at surface (0–20 cm) and subsurface (20–40 cm). The soil sand content (Sand) (%) and the soil organic matter content (SOM) (g/kg) at these two depths were measured. The soil samples were air-dried, sieved through a 2-mm mesh sieve, and analyzed using a Mastersizer 2000 laser particle size analyzer (Malvern Instruments, England). The dichromate oxidation method (external heat applied) was used to measure the SOM concentration (Liu et al., 1996; Nelson and Sommers, 1983).

The 58 samples selected using the integrative hierarchical stepwise sampling (Yang et al., 2013) were taken as the IHS samples in this study directly. The proposed sampling method and PS method selected their samples from the 336 samples, on account of those samples were generally uniformly distributed over the geographic space (Fig. 1) and represented the distribution of the environmental variables well (Fig. S1).

2.2. The multiple soil properties oriented representative sampling design method

The basic idea of the proposed MPRS method is to first select the most influential sets of environmental features for each soil property, and generate environmental clusters relating to spatial variation patterns for each soil property, then determine the sample locations representative of as many typical locations of environmental clusters for soil properties as possible. The procedure of the method consists of three steps: 1) to select the set of influential environmental variables for each soil property; 2) to generate the environmental clusters for representing the spatial variation of each soil property and determine the typical locations of each environmental clusters; 3) to select samples for multiple soil properties based on the typical locations for each soil property. Fig. 3 shows the overall framework of the sampling method.

2.2.1. Selecting influential environmental variables for each soil property

There are three commonly-used approaches to select the influential environmental variables for a soil property. The first is based on the expert knowledge of local soil pedogenesis. This approach relies on the experience of local experts. Another approach to select variables is based on literature review, or namely *meta*-analysis techniques. The third way to select the influential variables using machine learning techniques based on sample data. Studies showed that variables selected using

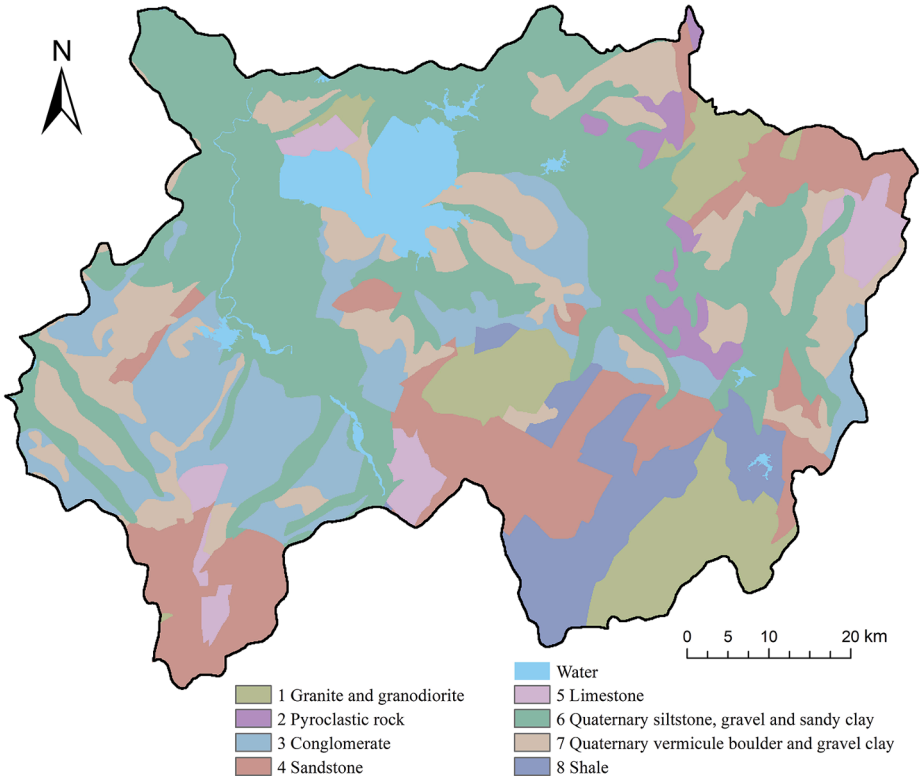

Fig. 2. The map of parent lithology types in the study area.

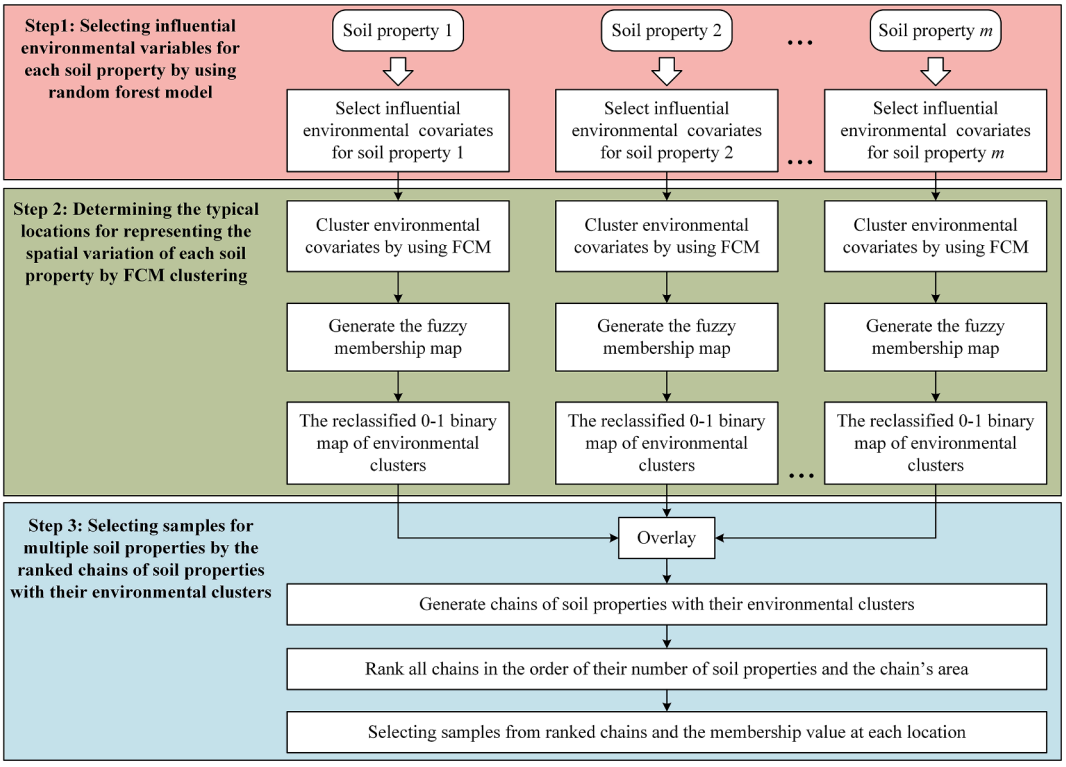

Fig. 3. The framework of the multiple soil properties oriented representative sampling method, FCM: fuzzy *c*-means.

machine learning outperformed those selected based on expert knowledge in soil mapping (Brungard et al., 2015; Shi et al., 2018). In this study, we selected the influential variables based on the calculated variable importance using a random forest (RF) model (Breiman, 2001).

Random forest is an ensemble model that applies bootstrap sampling to obtain the data subsets for training the base learners, and uses the voting or averaging strategy for aggregating the base learners. RF has been widely adopted to calculate the variable importance in predicting

soil properties or classes (e.g. Brungard et al., 2015; Hengl et al., 2017; Yang et al., 2020; He et al., 2020). In the RF model, bootstrap sampling is used to obtain the data subsets for training each base learner, and the rest of data are used as the out-of-bag (OOB) data for validation. RF provides two measures of variable importance: mean decrease in accuracy (MDA) and mean decrease in Gini (MDG). MDA is usually considered to be a more reliable measure than MDG (Bureau et al., 2003). In this study, we adopted the MDA as the variable importance index. The mean decrease in accuracy is assessed by replacing each environmental variable in turn with random noise and observing the average decrease in the prediction accuracies for all trees based on the out of bag (OOB) validation. The larger value of the mean decrease accuracy, the stronger the capability of interpreting the predicted soil property. This approach has been shown to be effective for variable reduction and improving model accuracy (Svetnik et al., 2003; Xiong et al., 2012; Heung et al., 2014). This means that the selection of influential environmental variables for different soil property is necessary. The specific procedure to select influential variables for each soil property based on MDA is as follows:

(1) All the environmental variables are adopted as the candidate variables to fit the initial RF model.
(2) Rank all the candidate variables by their variable importance calculated by the mean decrease in accuracy.
(3) Remove the environmental variables with the importance value less than a threshold. This threshold was set as zero in our case study.

For each soil property, the above procedure was performed based on all the samples. The representation of the environmental feature space ensured a good chance of determining the influential environmental variables for each soil property. The 'randomForest' package (Liaw and Wiener, 2002) in R (R Development Core Team, 2012) was used to calculate the variable importance.

2.2.2. Determining the typical locations for representing the spatial variation of each soil property

Each type of soil property variation can be associated with one combination of environmental variables (Qi et al., 2006; Zhu et al., 2008; 2010). In this study, fuzzy c-means clustering (FCM) (Bezdek, 1981) was performed on the selected environmental covariates for each soil property to generate clusters in the environmental feature space. These environmental clusters were assumed to be related to the soil variation types. The FCM clustering algorithm is an unsupervised learning method, developed by Dunn (1973) and improved by Bezdek (1981). It can optimally partition a dataset, such as the selected environmental covariates data, into a given number of classes. It has been successfully used in soil classification and terrain analysis (de Bruin and Stein, 1998; Burrough, et al., 2000; English, 2001; Hanesch et al., 2001; McBratney et al., 1992; Yang et al., 2013; Zhu et al., 2008).

Compared to the well-known k-means clustering algorithm (MacQueen, 1967), FCM is a "soft" clustering algorithm that can compute the membership of each data element associated with each class (Bezdek, 1981). Thus the generated fuzzy memberships to each class can better recognize the spatial continuous distribution of geographical variables (such as soils). Let the environmental covariates dataset be a finite collection of n data vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$, and a list of c cluster centroid vectors $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_c\}$, the fuzzy clustering criterion which recognizes clusters can be based on distances from data points to cluster centroids.

The objective function in FCM is the total weighted sum of squared distances between each point and each class centroid, which is quantified as:

$$J_m(\boldsymbol{X}, \boldsymbol{V}) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m (D_{ik})^2 \quad (1)$$

where J_m is the fuzzy partition error, it decreases as all data points tend to be overall closer to their cluster centroids. D_{ik} is the Euclidean distance between the feature vector of the kth data point $\boldsymbol{x}_k$ and the vector of the ith cluster centroid $\boldsymbol{v}_i$. m is the parameter which defines the fuzziness of the resulting clusters. The larger value of m determines the higher level of cluster fuzziness. In this study, we set m as 2 as default. u_{ik} is the membership of the kth point belonging to the ith cluster, it is defined as Eq. (2):

$$u_{ik} = \left[\sum_{j}^{c} \frac{(D_{ik})^2}{(D_{jk})^2} \right]^{-\frac{2}{m-1}} \quad (2)$$

The centroids of the clusters are defined as follow:

$$\text{centroid}_i = \frac{\sum_{k=1}^{n} {u_{ik}}^m \boldsymbol{x}_k}{\sum_{k=1}^{n} {u_{ik}}^m} \quad (3)$$

It is often difficult to determine the optimal number of clustering. In this study, we used normalized entropy (H) (Bezdek et al., 1984) in our study to determine the optimal number of clusters. When the number of clusters is c, H is determined as follow:

$$H_c = -\sum_{k=1}^{n} \sum_{i=1}^{c} [u_{ik} \log_a(u_{ik})] \Big/ \mathrm{n} \quad (4)$$

H_c is a scalar measure of the level of fuzziness in a given fuzzy partition (Bezdek et al., 1984). Generally, H increases with the numbers of clusters increased. We can examine the improvement of the entropy value over two adjacent numbers of cluster results to determine the optimal number of clusters (English, 2001; Yang et al., 2013). When the increment of H with cluster number changing from $c^{'}-1$ to c (i.e., $H_{c^{'}} - H_{c^{'}-1}$) is smaller than the increment with cluster number changing from $c^{'}$ to $c^{'}+1$ (i.e., $H_{c^{'}+1} - H_{c^{'}}$), the current cluster number can be considered as a satisfying partition of the dataset and $c^{'}$ is the possible optimal cluster number.

The clustering results include cluster centroids and fuzzy membership maps of all environmental clusters. The cluster centroids are described by the values of their environmental feature vectors. The fuzzy membership map for a given environmental cluster contains the membership at each pixel to that cluster.

To identify the typical locations of environmental clusters that could indicate typical soil variation, the values of a fuzzy membership maps were reclassified and transformed into a binary map that only includes 0 and 1 by applying a membership threshold. A pixel with the value of 1 in the binary map indicates that it is a typical location of a certain environmental cluster, and a pixel with the value of 0 indicates that it is not typical of that environmental cluster. A higher threshold may lead to a narrower selection of typical samples. We set the threshold to 0.7, according to a previous study of An et al. (2018) in this area.

In this step, we generated all the binary maps of environmental clusters for each soil property. Those pixels with a value of 1 in those binary maps indicated the typical locations of environmental clusters for each soil property.

If categorical variables (such as parent materials) are included in the environmental database, these variables will be used to divide the whole study area into strata. In this study, parent materials were taken as strata. FCM clustering was applied in each stratum, and samples were designed for each stratum according to the next step.

2.2.3. Selecting representative samples for multiple soil properties

For multiple soil properties, all the hardened binary maps were overlaid to produce an overall frequency map of typical locations in each stratum. For example, if there are four target soil properties, a location with a value of three in the frequency map means that three soil properties can be represented at this location. The higher frequency at a location, the more representative there for multiple soil properties, and the higher priority to sampling at this location.

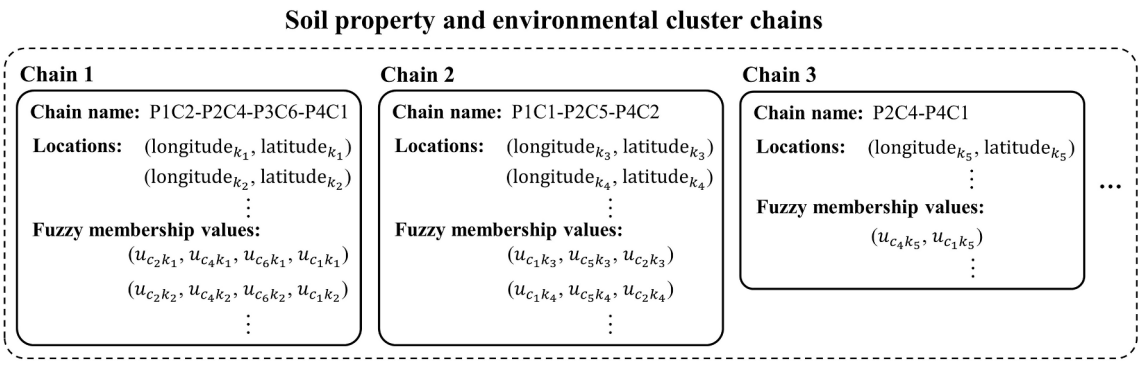

Fig. 4. An example diagram for illustrating the record of soil property and environmental cluster chains. Each chain contains three parts: the chain name (expressed by connecting multiple combinations of soil property and cluster), locations of all pixels that belong to the chain, and the corresponding fuzzy membership values, u_{ik} is the membership of the jth point belonging to the ith cluster.

Table 2
The differences in the use of the purposive sampling (PS), integrative hierarchical stepwise sampling (IHS), and multiple soil properties oriented sampling (MPRS) methods in this study.

	PS	IHS	MPRS
Selection of environmental variables	The union of the selected environmental variables for each soil property in MPRS.	The union of the selected environmental variables for each soil property in MPRS.	Selecting the most influential set of variables for each soil property using a machine learning approach.
Determination of Environmental Clusters	Environmental clusters were generated under an optimal cluster number for all the soil properties.	Multiple groups of environmental clusters were generated under different numbers of clusters.	The optimal number of clusters for each soil property was determined, and environmental clusters were generated for each soil property.
Selection of the representative samples	Samples were selected at the cluster centroids.	Samples were selected based on the ranked "environmental cluster chain" derived from multiple clustering maps.	Samples were selected based on the ranked "chain of soil properties with their environmental cluster" derived from multiple clustering maps for each soil property.

Table 3
Sample sizes of four groups on each parent material (MP) and the total number of samples for each group.

Group	Sample size								
	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	Sum
G1	3	2	4	5	2	8	5	3	32
G2	5	7	6	5	5	8	5	5	46
G3	6	7	7	6	6	10	10	6	58
G4	12	10	11	8	11	11	11	12	86

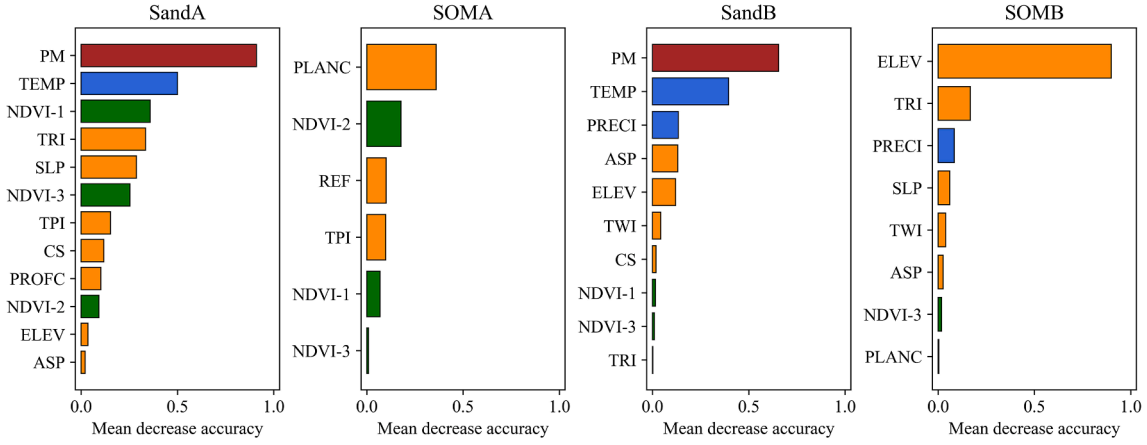

Fig. 5. The sorted importance of the selected environmental covariates based on mean decrease in accuracy for four different soil properties. SandA and SandB are the soil sand content of the surface layer (0–20 cm) and the subsurface layer (20–40 cm), respectively, and SOMA and SOMB are the soil organic matter content of the surface layer and the subsurface layer, respectively. The topographical variables are in orange, the climate variables are in blue, the vegetation variables are in green and the parent material is in brown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

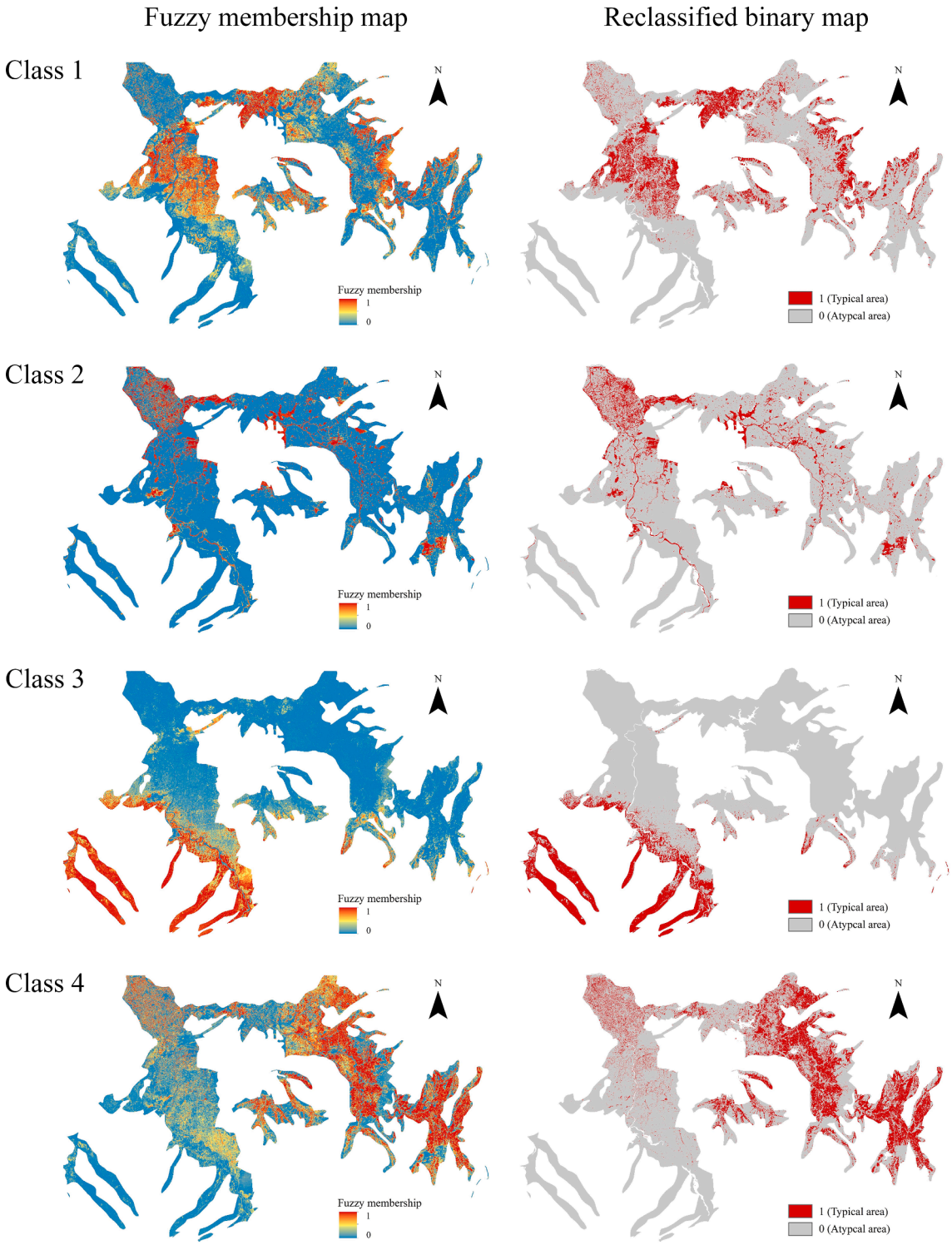

Fig. 6. The fuzzy membership maps of environmental clusters (left) and their reclassified binary maps (right) for the surface soil organic matter content on one stratum (the 6th parent material).

The locations with the same frequency may represent different typical positions of different environmental clusters for multiple soil properties. To examine the class configurations for different soil properties under each frequency, the concept of the "chain of soil properties with their environmental clusters" is introduced to record the specific list of clusters and the associated soil properties. This concept, additionally takes into account of different soil properties, is an extension of the "environmental cluster chain" proposed by Yang et al. (2013). For example, if one location with a frequency of 3 is represented by a chain: "P1C1-P2C5-P4C2", this means that this location can be identified as a typical position of "P1C1" (i.e., the 1st environmental cluster of the 1st soil property), the 5th environmental cluster (C5) for the 2nd soil property (P2), and the 2nd environmental cluster for the 4th soil property. Each location is thus labeled with a chain name except for those locations with the frequency of zero. Selecting samples on those locations with high frequencies means that it is able to capture several

Table 4
The ranked "chain of soil properties with their environmental clusters" list on one stratum (6th parent material as an example), the top 24 chains were listed. The chains in italic were discarded and the chains in bold were remained for sample design. Note that all the chains represented four soil properties in this stratum, but chains represented different numbers of soil properties in other strata.

Chain name	Number of soil properties in the chain	Area of the chain (number of pixels)
SandAClass2-SOMAClass3-SandBClass2-SOMBClass2	4	51,783
SandAClass4-SOMAClass3-SandBClass2-SOMBClass2	4	37,950
SandAClass3-SOMAClass4-SandBClass6-SOMBClass3	4	25,591
SandAClass6-SOMAClass1-SandBClass5-SOMBClass3	4	19,408
SandAClass5-SOMAClass1-SandBClass5-SOMBClass3	4	17,911
SandAClass8-SOMAClass4-SandBClass6-SOMBClass3	4	14,467
SandAClass1-SOMAClass4-SandBClass6-SOMBClass3	4	13,237
SandAClass7-SOMAClass4-SandBClass6-SOMBClass3	4	11,837
SandAClass5-SOMAClass4-SandBClass6-SOMBClass3	4	11,696
SandAClass3-SOMAClass2-SandBClass1-SOMBClass4	4	11,478
SandAClass6-SOMAClass4-SandBClass6-SOMBClass3	4	11,145
SandAClass6-SOMAClass1-SandBClass4-SOMBClass3	4	10,697
SandAClass3-SOMAClass2-SandBClass6-SOMBClass4	4	10,422
SandAClass7-SOMAClass2-SandBClass6-SOMBClass4	4	9702
SandAClass5-SOMAClass1-SandBClass4-SOMBClass3	4	9469
SandAClass3-SOMAClass4-SandBClass4-SOMBClass3	4	7844
SandAClass3-SOMAClass4-SandBClass6-SOMBClass1	4	5921
SandAClass1-SOMAClass4-SandBClass4-SOMBClass3	4	5074
SandAClass8-SOMAClass4-SandBClass4-SOMBClass3	4	5037
SandAClass3-SOMAClass2-SandBClass6-SOMBClass1	4	4375
SandAClass5-SOMAClass1-SandBClass4-SOMBClass1	4	4222
SandAClass1-SOMAClass1-SandBClass5-SOMBClass3	4	4149
SandAClass7-SOMAClass4-SandBClass4-SOMBClass3	4	3917
SandAClass5-SOMAClass1-SandBClass3-SOMBClass1	4	3753

typical locations for multiple soil properties once a time. As an example illustrated in Fig. 4, for each chain, it contains information about all the point locations and fuzzy membership values at those locations.

All chains need to be ranked for designing samples because they represent different numbers of soil properties and have different pixel numbers (areas). We thus ranked these chains based on two criteria successively: 1) the number of soil properties represented in a chain; 2) the area (the number of pixels) of each chain. All chains were ranked by these two criteria in order from largest to smallest. Note, if a chain with a lower ranking order was completely contained in a chain with a higher ranking order, this lower ranked chain will be discarded. This is because that the higher ranked chains have already covered those typical locations of environmental clusters for soil properties represented by the lower ranked chains. Then, with the same number of occurred soil properties, the chains were ranked by their areas. The chain with the area of coverage smaller than a pre-defined threshold was considered atypical for soil mapping and excluded from the sampling (Yang et al., 2013). After the ordering and removing of the redundant or small chains, all the remaining chains were unique patterns.

Finally, we successively selected the sample locations based on the ranked chains. For every pixel in a chain, we calculated its mean fuzzy membership value to all the clusters in the chain. Take a location labeled with the chain "P1C1-P2C5-P4C2" as an example, assuming that the fuzzy membership values to the three clusters are 0.9 (C1 to P1), 0.85 (C5 to P2) and 0.95 (C2 to P4) at this location, the mean membership value of this chain at this location is 0.9. The samples were selected at those locations with the highest mean membership value in each chains. Due to the order of the chains, the selected samples had correspondingly representativeness.

2.3. Two sampling methods for comparison

To evaluate the proposed sampling method, we adopted the other two sampling methods, PS (Zhu et al., 2008) and IHS (Yang et al., 2013, 2016), for mapping multiple soil properties. The reasons of choosing these two methods for comparison are that they both use FCM to generate fuzzy membership maps of clusters for sample design. Furthermore, IHS has been validated to be able to collect less sampling points with a higher mapping accuracy than stratified random sampling or cLHS in two case studies (Yang et al., 2016).

The PS method used FCM to generate environmental clusters based on all the available environmental variables (Zhu et al., 2008). Stratification of the study area was first implemented by using soil parent materials, and the clustering is performed in each stratum. Only one optimal cluster number was determined for each stratum. The samples generated by PS method were selected from the 336 collected samples. Particularly, the sample point with the highest fuzzy membership to a cluster was determined as the sample point for the environmental cluster. A total of 86 environmental clusters were generated after clustering for all the strata, thus, 86 points were selected as PS samples.

IHS also used FCM to generate clusters based on all the available environmental variables together (Yang et al., 2016). Stratification by the parent materials was implemented first, and the FCM clustering was performed in each stratum. For each stratum, a range of cluster numbers was determined to generate multiple groups of environmental clusters, which is used to approximate soil spatial variations at different scales. A few clusters are assumed to represent the major soil variations types, and increasing the number of clusters leads to the emergence of minor types. In other words, the dominant large-scale soil variation types can always be detected regardless of the number of clusters, but small-scale local types are only revealed when the number of clusters is greater (Yang et al., 2017). Based on this, all the environmental clusters were reclassified to binary maps to discriminate their typical locations. The binary maps under a series of cluster numbers were overlaid to generate the so-called "environmental cluster chain", recording the specific list of classes and the associated clustering number for each pixel. The more classes occurred in a chain, the larger spatial variation type the chain represented. Samples were selected based on the environmental cluster chains after ranking by the occurrence number of environmental clusters in environmental cluster chains and the areas of environmental cluster chains. In this study, we adopted the 58 samples selected by using IHS in the same study area which is detailed described in Yang et al. (2016). These samples represented a total of 58 environmental cluster chains in the area, and the sample sets were with three representativeness grades. The representativeness grade was determined by the occurrence number of environmental clusters in the environmental cluster chains where the samples designed. The detail of samples generated by IHS can refer to Yang et al. (2016; 2017).

Differences in the use of these three sampling methods in this study were illustrated in Table 2.

Table 5
Descriptive statistics of the selected samples by three sampling methods (purposive sampling (PS), integrative hierarchical stepwise sampling (IHS) and multiple soil properties oriented representative sampling (MPRS)) with four sample sizes. SandA and SandB are the soil sand content of the surface layer (0–20 cm) and the subsurface layer (20–40 cm), respectively, and SOMA and SOMB are the soil organic matter content of the surface layer and the subsurface layer, respectively.

Sample size	Method	SandA				SandB				SOMA				SOMB			
		Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
32	PS	29.0	18.9	16.4	56.8	29.2	21.6	11.3	69.0	21.4	12.2	8.8	43.7	10.0	5.5	4.8	43.8
	IHS	30.1	15.9	11.9	65.1	30.1	20.6	11.0	71.3	20.0	10.3	8.8	67.0	9.0	6.4	3.7	46.2
	MPRS	28.8	16.6	11.5	69.2	28.0	19.1	12.8	74.1	20.8	13.5	6.1	69.5	9.8	6.3	3.7	46.2
46	PS	31.2	19.2	10.5	65.1	31.6	24.5	11.3	74.1	22.4	11.5	8.8	67.0	11.0	5.9	4.3	47.8
	IHS	29.3	16.5	7.9	72.4	32.2	21.9	11.0	76.0	19.2	9.5	8.3	71.0	9.2	4.8	3.7	47.8
	MPRS	26.8	18.4	7.9	73.0	29.6	18.6	7.3	78.2	21.1	11.3	8.3	71.0	10.3	4.6	3.5	52.8
58	PS	28.2	18.3	5.4	79.5	32.1	23.7	6.3	76.0	22.0	11.0	8.3	68.7	10.7	4.0	4.3	47.8
	IHS	28.7	15.7	7.9	78.5	27.7	20.8	7.8	78.2	19.0	8.9	8.3	71.0	9.0	4.6	3.5	49.2
	MPRS	27.0	17.0	6.5	79.8	27.0	18.3	7.3	78.2	22.3	11.5	6.1	71.0	10.3	3.8	3.5	52.8
86	PS	27.7	16.5	5.4	78.5	26.3	19.3	6.3	78.2	23.5	10.1	5.6	73.7	11.3	3.5	3.5	52.8
	MPRS	26.5	15.3	5.4	79.8	25.3	21.7	6.3	79.8	24.1	9.6	5.6	73.7	10.9	3.6	3.5	52.8

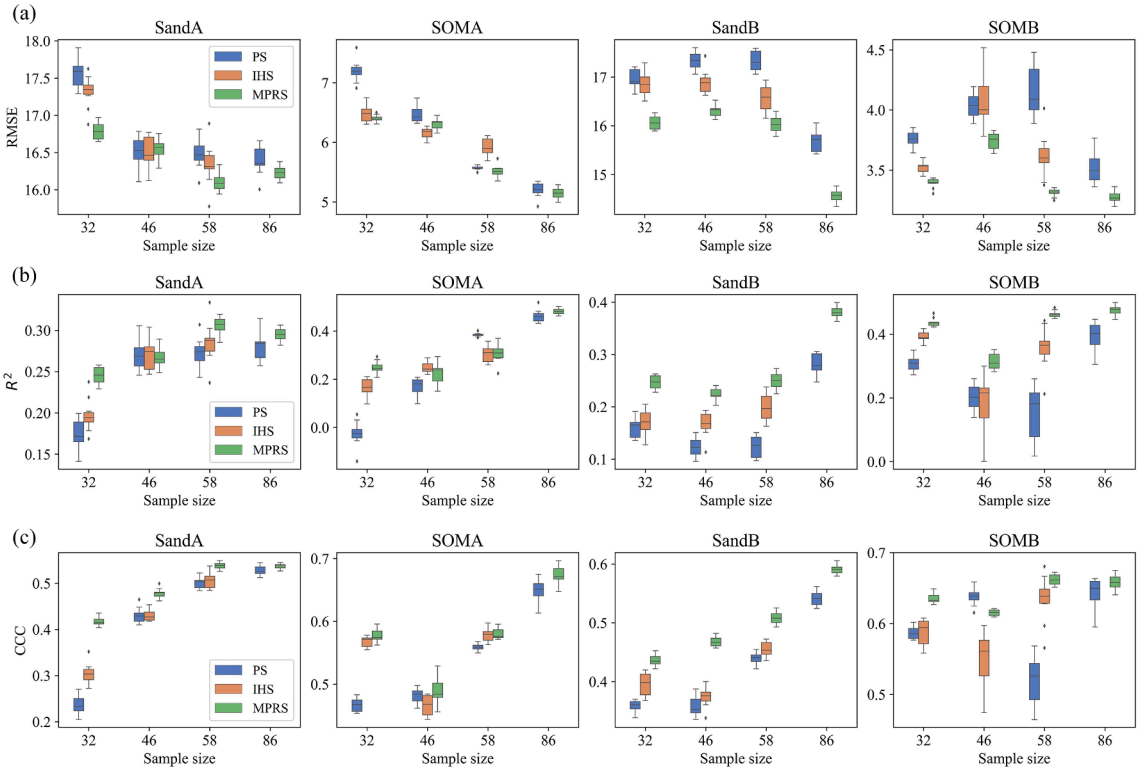

Fig. 7. The boxplots of RMSE (a), R^2 (b) and CCC (c) for the purposive sampling (PS), integrative hierarchical stepwise sampling (IHS) and multiple soil properties oriented representative sampling (MPRS) methods with different sample sizes. SandA and SandB are the soil sand content of the surface layer (0–20 cm) and the subsurface layer (20–40 cm), respectively, and SOMA and SOMB are the soil organic matter content of the surface layer and the subsurface layer, respectively. RMSE: root mean square error; R^2: coefficient of determination; CCC: concordance correlation coefficient.

2.4. The groups of samples for the three sampling methods based on representativeness

In order to evaluate the representativeness of samples and compare the performance of the proposed method with the other two sampling methods under different sample sizes, all samples generated by three sampling methods were selected into four groups. The sample sizes of each group were determined as follows. First, according to the 58 samples collected using IHS by Yang et al. (2016), the first group was 32 samples with the highest representativeness grade. The second group was samples with the highest and second representativeness grades, with a sample size of 46. The third group with a sample size of 58 was samples with the highest, second and third representativeness grades. These three sample sizes and the sample size selected in each stratum were taken as the number of samples of first three groups for the other two sampling methods (Table 3).

Because there are 86 PS samples in the study area, 86 was taken as the sample size of the fourth group for PS and MPRS. To determine the PS samples in the first three groups, the selected samples in each stratum were sorted by the area (number of pixels) of the corresponding clusters from highest to lowest. The sample size for each stratum for the first, second, and third groups were determined as shown in Table 3. The samples with the corresponding numbers were selected according to their orders in each stratum to obtain the samples of the first, second and third group, successively.

To be consistent with IHS and PS, four groups of samples were selected using MPRS. The FCM clustering was performed with different influential environmental covariates for each soil property in each stratum. From the existing 336 samples, the sample point with the highest fuzzy membership to a "chain of soil properties with their

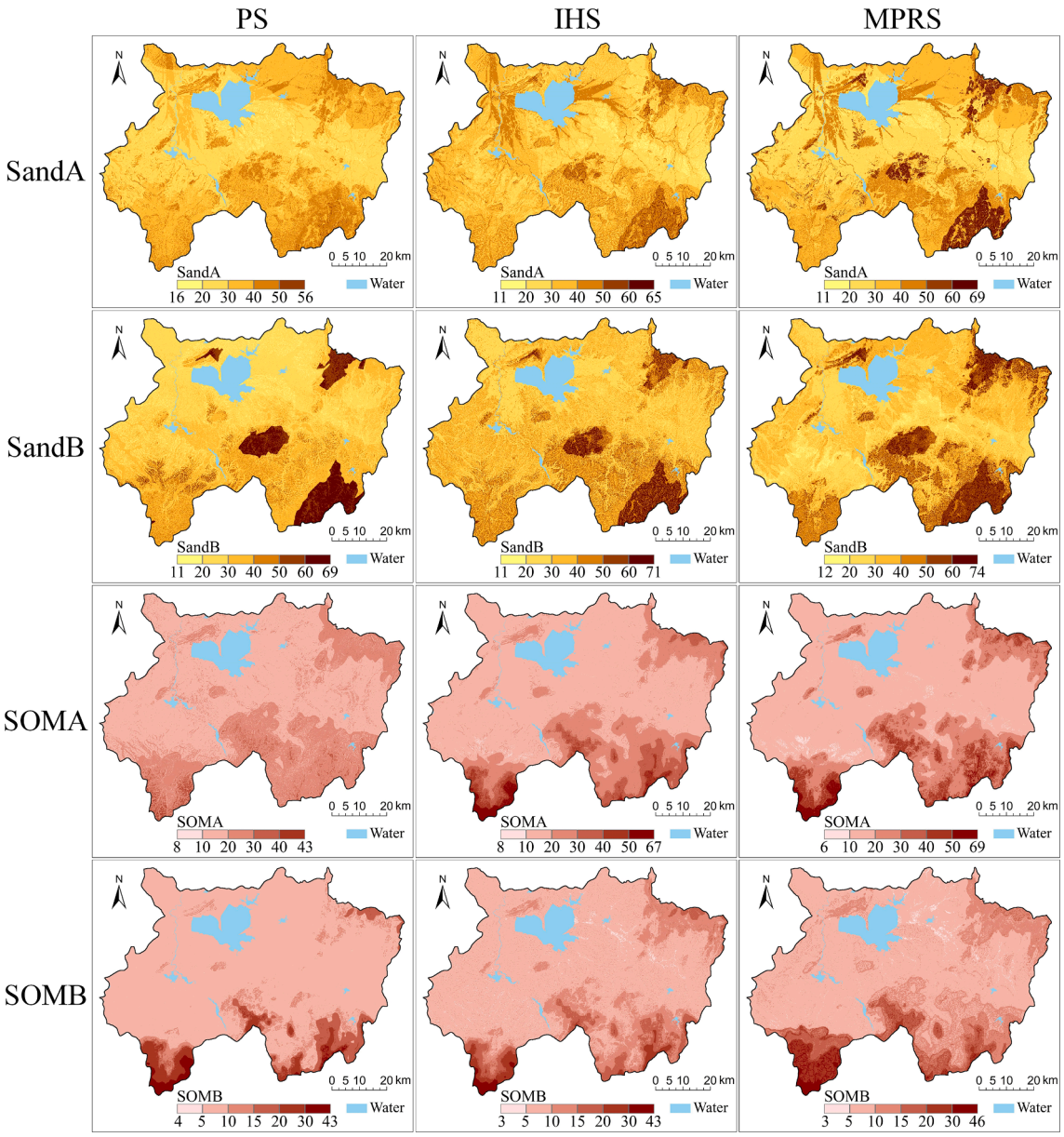

Fig. 8. The predicted maps of the four soil properties (the soil sand content at surface layer (SandA) and subsurface layer (SandB), and the soil organic matter content at surface layer (SOMA) and subsurface layer (SOMB)) with 32 samples generated by the purposive sampling (PS), integrative hierarchical stepwise sampling (IHS) and multiple soil properties oriented representative sampling (MPRS) method.

environmental clusters" was determined as the sample point for this chain. In each stratum, according to the ranking order of the "chain of soil properties with their environmental clusters", samples were ranked. The sample size in each stratum was the same as the above two methods for the four sample groups. The samples with the corresponding numbers were selected according to their orders in each stratum to obtain the samples of the first, second, third and fourth group, successively. Note that only the fourth sample group of PS and MPRS were compared due to there were only 58 samples of IHS method (Yang et al., 2016).

2.5. Soil mapping and validation

Soil maps based on different groups of samples designed by the three sampling methods were generated and validated to examine how the mapping accuracy differs over the three methods. Random forest was used as the soil prediction model. RF is a popular ensemble machine learning model, and has been widely used in DSM (Heung et al., 2014; Hengl et al., 2017; Zhang et al., 2021). The major advantage of RF is the incorporation of randomized feature selection and sample selection in the multiple trees building process (Breiman, 2001). Thus it can effectively reduce the risk of overfitting and lead to a good generalization ability. Two important user-defined parameters need to be set in 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 the number of trees to be learned in the forest. We set it to be 200.

The predicted soil property maps were evaluated based on the 57 independent validation points in Fig. 1. Three indices were used for the validation: root mean square error (RMSE), coefficient of determination (R^2) and Lin's concordance correlation coefficient (CCC; Lin, 1989), which were calculated as follows:

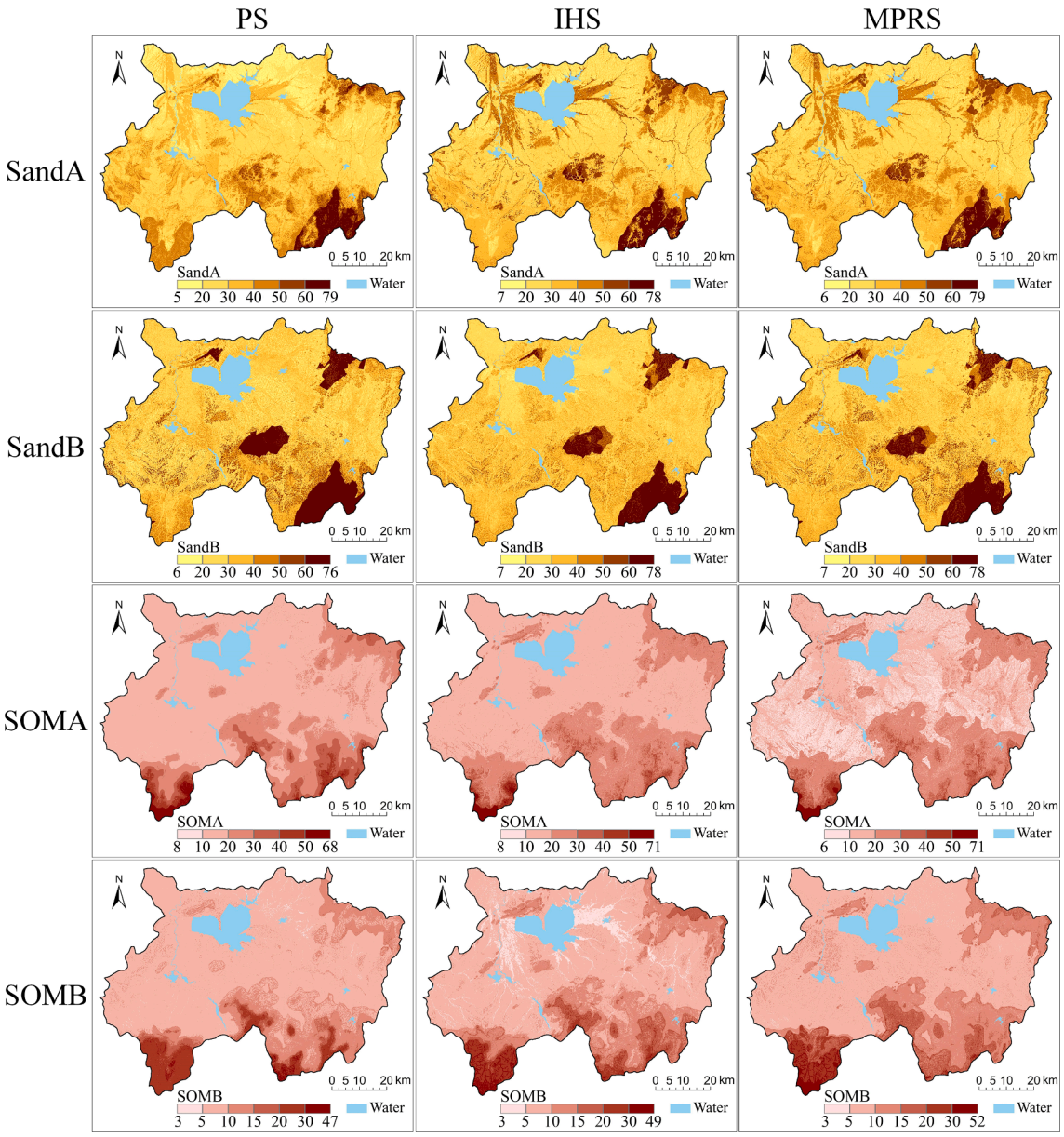

Fig. 9. The predicted maps of the four soil properties (the soil sand content at surface layer (SandA) and subsurface layer (SandB), and the soil organic matter content at surface layer (SOMA) and subsurface layer (SOMB)) with 58 samples generated by the purposive sampling (PS), integrative hierarchical stepwise sampling (IHS) and multiple soil properties oriented representative sampling (MPRS) method.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i - P_i)^2} \tag{5}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Z_i - P_i)^2}{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2} \tag{6}$$

$$\mathrm{CCC} = \frac{2r\sigma_z\sigma_p}{\sigma_z^2 + \sigma_p^2 + \left(\overline{Z} - \overline{P}\right)^2} \tag{7}$$

where Z_i and P_i are observed and predicted values respectively; $\overline{Z}$ and $\overline{P}$ are the averages of n observed and predicted values; σ_z and σ_p are the corresponding standard deviations; and r is the correlation coefficient value between the predicted and observed values. A smaller value of RMSE or a larger value of R^2 and CCC means a higher prediction accuracy. We repeated the training of the RF model 100 times using different random seeds under each set of training samples. The average accuracy over repeated experiments was used for evaluate the model performance.

3. Results

3.1. The selected environmental covariates for each soil property

The sorted importance of the selected environmental covariates for each soil property is shown in Fig. 5. Variables with the importance value smaller than zero were removed since the predictive model could not be further improved when adding those variables. It can be seen that the selected sets of important covariates for each soil property were different. Parent materials, temperature and topography impacted the sand content at both layers, while topography, vegetation impacted SOM at surface soil and topography mainly impacted SOM at the subsurface soil. The number of influential variables for Sand was generally larger than that for SOM. In addition, the influence of vegetation

variables on the two surface soil properties was relatively larger than that at the subsurface layer which was more influenced by topographical variables.

3.2. The selected samples by the MPRS method

With MPRS method, we first clustered the selected environmental covariates for each soil property on each stratum (parent material). Fig. 6 shows the clustering results of fuzzy membership maps and reclassified binary maps for SOMA (soil organic matter content at surface layer) on the largest stratum (No. 6 parent material as shown in Fig. 2) in the study area. All binary maps for four soil properties were overlaid, and the list of "chain of soil properties with their environmental clusters" was generated. The chains for this stratum are listed in Table 4. Following the ranking criteria described in 2.2.3, these chains were ranked. The chain names in bold were selected and the chain names in italic were removed because these chains were completely contained in the chains with the higher ranking order. According to the sample sizes designed in section 2.4, four groups of samples were successively selected from the ranked chains. The descriptive statistics of four groups of samples generated by MPRS and other two sampling methods (PS and IHS) was shown in Table 5.

3.3. Comparison of the soil prediction accuracy and maps for the three sampling methods

3.3.1. The prediction accuracy

The boxplots of RMSE, R^2 and CCC of predicting multiple soil properties with different sample sizes are illustrated in Fig. 7. It showed that the samples generated by the MPRS method achieved the best prediction accuracy in most cases, and IHS generally performed better than PS. Averaging the validation indices of the four target soil properties based on the four sizes of samples, the improvements of accuracy by using the MPRS method were 38.1% and 36.3% in terms of R^2 compared with PS and IHS, respectively, 11.7% and 13.7% in terms of CCC, and 4.8% % and 4.6% in terms of RMSE. It also showed that the variation of prediction accuracy over repeated experiments for MPRS was smaller than those for other two sampling methods generally, which indicated that samples generated by MPRS was more representative and had a more stable prediction performance. Particularly, MPRS obtained the highest accuracy at every sample size for every subsurface soil property. For the first sample group with the smallest sample size of 32, the RMSE (or R^2 and CCC) by using the MPRS method were all lower (higher) than those using the other two sampling methods. This indicates that the proposed sampling method can effectively generate a set of highly representative samples for multiple soil properties with a small sample size. At the sample size of 46 and 58, although some results of the MPRS method were not performed best in predicting the surface soil properties, the differences between MPRS and other two methods were very small. For predicting the two subsurface soil properties, it showed that the prediction accuracies of MPRS were largely higher than the other two methods. It was also worth noting that the accuracies of subsurface soil properties were higher than those at surface in most cases from our results, although the selected environmental covariates (such as topographic variables and vegetation index) were supposed to be more representative of surface conditions than of deeper soil layers. The possible reason for the difference between the accuracies of surface and subsurface soil properties is that surface soil is more interfered by human activities and those variables selected for the subsurface soil properties well described their spatial variations. At the sample size of 86, MPRS also achieved the best results in all four soil properties prediction. In general, over the different sample sizes, the MPRS method can guarantee that most of the soil properties can achieve the most accurate prediction result, while the other two sampling methods cannot.

The prediction accuracies improved with the increment of sample size in general. However, the trends in accuracies changing with the increase of the sample size were different for surface and subsurface soil properties. The rates of accuracy changing for the surface layer were relatively fast, but the rates of the accuracy changing for the subsurface fluctuated with the increase of the sample size. One possible reason is that the surface soil is more disturbed by human activity than that at subsurface layer, and capturing the variation of top soils may be more dependent on the sufficient number of samples.

3.3.2. The predicted soil property maps

The predicted soil property maps based on the 32 and 58 samples generated by three sampling methods are shown in Fig. 8 and Fig. 9 as examples, respectively. It can be seen that the predicted maps with three sampling methods show a generally similar spatial distribution pattern for each soil property, and more similar with a larger number of 58 samples. The soil property maps generated with the MPRS method showed more details of soil spatial variation in most cases. It also can be seen that the ranges of soil properties predicted by MPRS samples are usually larger than those predicted by the other two sampling methods. For example, the range of the predicted SandA and SOMA based on the 32 MPRS samples was 11 ~ 69% and 6 ~ 69 g/kg, respectively, while those based on PS samples were 16 ~ 56%, and 8 ~ 43 g/kg, respectively. This indicates that MPRS selected samples representing more local minimum or maximum for multiple soil properties with the same sample size.

4. Discussion

To use their influential environmental variables for each soil property for soil mapping is increasingly adopted in DSM (Brungard et al., 2015; Yang et al., 2020; He et al., 2021). However, sampling design considering the influential environmental variables for each soil property has not been paid much attention. The results of our study showed a better soil prediction performance of the MPRS method compared with other two sampling methods (PS and IHS) in the study area. One main difference among the three methods is that PS and IHS use all the available environmental variables together for clustering, while MPRS uses the influential environmental variables for each soil property. This reveals the importance of using the influential environmental covariates for each target soil property in sample design. IHS generally generated higher mapping accuracies than PS. This is probably because IHS discerned spatial variation types of soil properties at different scales. In the same study area, IHS was compared with SRS and cLHS with different sample sizes and generated more accurate soil property or soil type maps (Yang et al., 2016). Based on the results of our study, MPRS is supposed to be more accurate than SRS or cLHS for mapping multiple soil properties. More studies on comparisons between MPRS can be conducted in the future.

The influential environmental variables for different soil properties were different. Various environmental variables of climate, topography and vegetation had impact on soil sand content, while topography and vegetation variables were the most influential for SOM in our study. The number of influential variables for Sand was larger than that for SOM. Similar findings were also found in previous studies (Heung et al., 2014; Wang et al., 2020). In the study of Hengl et al. (2017), they found climatic variables (especially precipitation) and surface reflectance seemed to be the most important for predicting soil chemical properties, while the combination of topography, vegetation dynamics, and parent material was more important for predicting soil physical properties. However, different from their findings, climate variables did not appear to be of high importance for predicting SOM in our study. This is probably because our study area is much smaller than their global soil mapping research. In addition, the influence of vegetation variables on these two surface soil properties was relatively larger than those at the subsurface layer which were more influenced by topographical variables. This was consistent with the study of Hong et al. (2020).

To determine the influential environmental variables for each soil property is needed when applying our method. In this study, we used machine learning to select the influential environmental variables based on sample data. The selected influential variables for each soil property by using RF and prior samples can be transferred in other study area with near distance or similar environmental conditions. For example, we found that the influence of vegetation variables on the surface soil properties was larger than that at the deeper layer, which can be a useful prior knowledge for guiding the sampling strategy in other nearby areas or regions with similar altitude, land cover and climatic conditions.

Similar to the previous studies on the sampling method for multiple soil properties (Vašát et al., 2010; Szatmári et al., 2016; 2019), prior sample data is also needed when using the proposed MPRS method. However, there are also some other options to avoid this requirement. One is to take advantage of expert knowledge in a study area. The knowledge between soil and environmental covariates can be acquired if soil experts or expert knowledge are available in the target area (e.g. Zhu, 1999). Although it has been found that the mapping accuracy by using knowledge-based environmental variables may not higher than that by using variables selected by machine learning approaches (Shi et al., 2018), it is still an effective way that can be developed for multivariate sampling. Additionally, the *meta*-analytical method (quantitative analysis of research results by literature review) (e.g. Guo and Gifford, 2002) can also be another potential option for pre-selecting the influential covariates for each soil property. Investigations of such two approaches would deserve more attentions in the multiple soil properties oriented sampling in future.

In this study, we used existing samples with a good coverage in environmental feature space to select samples for the proposed method and PS, which is an economic way. Some points were selected as samples for two or three sampling methods. This is because some of the extracted typical locations representative of variation types for soil properties by different sampling methods were overlap over space.

As previous studies on sampling design for multiple soil properties by Vasat et al. (2010) and Szatmari et al. (2016), their proposed sampling methods were 'fully optimal' for one dominant soil property and 'sub-optimal' for the others. In our study, we cannot guarantee the designed samples are fully optimal for all soil properties either. However, the results showed that the prediction accuracy were highest by using samples generated by MPRS in most cases. The objective of our method is to obtain a set of samples that can represent the relationships between different soil properties and corresponding influential environment covariates as much as possible in a way based on clustering analysis and machine learning. Supported by the proposed concept of the "chain of soil properties with their environmental clusters", the finally selected samples from the top ranked chains could better cover typical locations of environmental clusters for multiple soil properties. From this aspect, our study provided an alternative heuristic approach to design a multiple soil properties oriented sampling method, which was a potential effort for improving the soil prediction accuracy.

5. Conclusions

This paper proposed a multiple soil properties oriented representative sampling method. This sampling strategy first generated environmental clusters for each soil property based on their influential environmental covariates, and designed samples that locate on as many typical locations of environmental clusters for multiple soil properties simultaneously as possible. From the results of the case study, the sets of important covariates for different soil properties were different. Parent materials, temperature and topography impacted the sand content for both surface and subsurface soil, while topography, vegetation impacted the SOM at surface soil and topography mainly impacted the SOM at the subsurface soil. This confirmed the necessity to consider the difference of the influential environmental variable combinations for the multiple soil properties oriented sampling design. Compared with the other two sampling methods, the proposed sampling method achieved the most accurate prediction for four soil properties over different sample sizes in most cases. We conclude that the proposed MPRS method is a potential effective sampling method for DSM with multiple soil properties.

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), and the Leading Funds for the First Class Universities (0209-14912203 and 0209-14380079).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2021.115531.

References

An, Y., Yang, L., Zhu, A.X., Qin, C., Shi, J., 2018. Identification of representative samples from existing samples for digital soil mapping. Geoderma 311, 109–119. https://doi.org/10.1016/j.geoderma.2017.03.014.

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 https://doi.org/10.1016/j.geodrs.2020.e00265.

Band, L.E., Moore, I.D., 1995. Scale: Landscape attributes and geographical information systems. Hydrol. Process. 9, 401–422. https://doi.org/10.1002/hyp.3360090312.

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.X., Scholten, T., 2014. Hyper-scale digital soil mapping and soil formation analysis. Geoderma 213 (1), 578–588. https://doi.org/10.1016/j.geoderma.2013.07.031.

Bezdek, J.C., 1981. Models for Pattern Recognition. In: Bezdek, J.C. (Ed.), Pattern Recognition with Fuzzy Objective Function Algorithms, Advanced Applications in Pattern Recognition. Springer, US, Boston, MA, pp. 1–13. https://doi.org/10.1007/978-1-4757-0450-1_1.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. 10 (2–3), 191–203. https://doi.org/10.1080/01969727308546046.

Beven, K.J., Kirkby, N.J., 1979. A physically based variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69. https://doi.org/10.1080/02626667909491834.

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

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., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80 (1–2), 1–44. https://doi.org/10.1016/S0016-7061(97)00072-4.

Brus, D.J., 2015. Balanced sampling: A versatile sampling approach for statistical soil surveys. Geoderma 253–254, 111–121. https://doi.org/10.1016/j.geoderma.2015.04.009.

Brus, D.J., Heuvelink, G.B., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138 (1–2), 86–95. https://doi.org/10.1016/j.geoderma.2006.10.016.

Brus, D.J., De Gruijter, J.J., Van Groenigen, J.W., 2006. Developments in Soil Science. In: Chapter 14 Designing Spatial Coverage Samples Using the k-means Clustering Algorithm. Elsevier, pp. 183–192. https://doi.org/10.1016/S0166-2481(06)31014-8.

Brus, D.J., 2019. Sampling for digital soil mapping: a tutorial supported by R scripts. Geoderma 338, 464–480. https://doi.org/10.1016/j.geoderma.2018.07.036.

Bureau, A., Dupuis, J., Hayward, B., Falls, K., Van Eerdewegh, P., 2003. Mapping complex traits using Random Forest. BMC Genet. 4, S64. https://doi.org/10.1186/1471-2156-4-S1-S64.

Burrough, P.A., van Gaans, P.F., MacMillan, R.A., 2000. High-resolution landform classification using fuzzy k-means. Fuzzy Sets Syst. 113 (1), 37–52. https://doi.org/10.1016/S0165-0114(99)00011-1.

Carter, M. R., Gregorich, E. G. (Eds.)., 2007. Soil sampling and methods of analysis. CRC Press.

Chinese National Soil Survey Office, 1992. Chinese soil taxonomy system. Agriculture Press, Beijing.

de Bruin, S., Stein, A., 1998. Soil-landscape modeling using fuzzy c-means clustering of attribute data derived from a Digital Elevation Model (DEM). Geoderma 83 (1–2), 17–33. https://doi.org/10.1016/S0016-7061(97)00143-2.

de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, Berlin.
Dunn, J.C., 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics 3, 32–57. https://doi.org/10.1080/01969727308546046.
English, E.M., 2001. Assisting knowledge-based inferential soil mapping: The application of fuzzy c-means clustering to expose environmental niches. University of Wisconsin-Madison.
Farr, T.G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. Eos, Transactions American Geophysical Union 81 (48), 583–585. https://doi.org/10.1029/EO081i048p00583.
Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil-landscape modelling and spatial prediction of soil attributes. Int. J. Geogr. Inf. Sci. 9 (4), 421–432. https://doi.org/10.1080/02693799508902047.
Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1–45. https://doi.org/10.1016/S0016-7061(98)00078-0.
Gregoire, T.G., Valentine, H.T., 2008. Sampling strategies for natural resources and the environment. Chapman & Hall/CRC, Boca Raton, Florida.
Guo, L.B., Gifford, R.M., 2002. Soil carbon stocks and land use change: a meta analysis. Glob. Change Biol. 8, 345–360. https://doi.org/10.1046/j.1354-1013.2002.00486.x.
Hanesch, M., Scholger, R., Dekkers, M.J., 2001. The application of fuzzy c-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites. Phw Chem. Earth(A) 26 (11–12), 885–891. https://doi.org/10.1016/S1464-1895(01)00137-5.
Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120 (1–2), 75–93. https://doi.org/10.1016/j.geoderma.2003.08.018.
Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, 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., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12. https://doi.org/10.1371/journal.pone.0169748.
He, X., Yang, L., Li, A., Zhang, L., Shen, F., Cai, Y., Zhou, C., 2021. Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. Catena 205, 105442. https://doi.org/10.1016/j.catena.2021.105442.
Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214, 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.
Hong, S., Yin, G., Piao, S., Dybzinski, R., Cong, N., Li, X., Wang, K., Peñuelas, J., Zeng, H., Chen, A., 2020. Divergent responses of soil organic carbon to afforestation. Nat. Sustain. 3, 694–700. https://doi.org/10.1038/s41893-020-0557-y.
Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, p. 561.
Jenny, H., 1941. Factors of Soil Formation. A System of Quantitative Pedology, McGraw-Hill, NewYork.
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, Dordrecht, The Netherlands, pp. 3–14. https://dx.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.
Liaw, A., Wiener, M., 2002. Classification and Regression by random Forest. R News 2, 18–22.
Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268.
Liu, G.S., Jiang, N.H., Zhang, L.D., Liu, Z.L., 1996. Soil Physical and chemical analysis and description of soil profile. China Standardization Publishing House, Beijing, pp. 131–134 (In Chinese).
Luo, Z., Feng, W., Luo, Y., Baldock, J., Wang, E., 2017. Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. Glob. Change Biol. 23 (10), 4430–4439. https://doi.org/10.1111/gcb.13767.
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.
Ma, T., Wei, T., Qin, C.Z., Zhu, A.X., Qi, F., Liu, J., Zhao, F., Pan, H., 2020. In-situ recommendation of alternative soil samples during field sampling based on environmental similarity. Earth Sci. Informatics. 13, 39–53. https://doi.org/10.1007/s12145-019-00407-x.
Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur. J. Soil Sci. 70, 216–235. https://doi.org/10.1111/ejss.12790.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berke-ley symposium on mathematical statistics and probability 1 (14), 281–297.
McBratney, A.B., De Gruijter, J.J., Brus, D.J., 1992. Spacial prediction and mapping of continuous soil classes. Geoderma 54, 39–64. https://doi.org/10.1016/0016-7061(92)90097-Q.
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.
Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M., 2015. Impact of multi-scale predictor selection for modeling soil properties. Geoderma 239, 97–106. https://doi.org/10.1016/j.geoderma.2014.09.018.
Minasny, B., McBratney, A.B., 2006. Mechanistic soil–landscape modelling as an approach to developing pedogenetic classifications. Geoderma 133 (1–2), 138–149. https://doi.org/10.1016/j.geoderma.2006.03.042.
Moore, I.D., Gessler, P.E., Nielsen, G.A.E., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57 (2), 443–452. https://doi.org/10.2136/sssaj1993.03615995005700020026x.
Nelson, D.W., Sommers, L.E., 1983. Total Carbon, Organic Carbon, and Organic Matter, in: Methods of Soil Analysis. John Wiley & Sons, Ltd, pp. 539–579. https://doi.org/10.2134/agronmonogr9.2.2ed.c29.
Park, S.J., van de Giesen, N., 2004. Soil–landscape delineation to define spatial sampling domains for hillslope hydrology. J. Hydrol. 295, 28–46. https://doi.org/10.1016/j.jhydrol.2004.02.022.
Qi, F., Zhu, A.X., Harrower, M., Burt, J.E., 2006. Fuzzy soil mapping based on prototype category theory. Geoderma 136, 774–787. https://doi.org/10.1016/j.geoderma.2006.06.001.
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.
Quinn, P.F.B.J., Beven, K., Chevallier, P., Planchon, O., 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrol. Process. 5 (1), 59–79. https://doi.org/10.1002/hyp.3360050106.
R Development Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project. org.
Riley, S.J., DeGloria, S.D., Elliot, R., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain J. Sci. 5, 23–27.
Roberts, D.W., 1986. Ordination on the basis of fuzzy set theory. Vegetatio 66, 123–131. https://doi.org/10.1007/BF00039905.
Royle, J.A., Nychka, D., 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Comput. Geosci. 24 (5), 479–488. https://doi.org/10.1016/S0098-3004(98)00020-X.
Szatmári, G., Barta, K., Pásztor, L., 2016. Multivariate sampling design optimization for digital soil mapping. In: Zhang, G.-L., Brus, D., Liu, F., Song, X.-D., Lagacherie, P. (Eds.), Digital Soil Mapping Across Paradigms, Scales and Boundaries, Springer Environmental Science and Engineering. Springer, Singapore, pp. 77–87. https://doi.org/10.1007/978-981-10-0415-5_7.
Szatmári, G., László, P., Takács, K., Szabó, J., Bakacsi, Z., Koós, S., Pásztor, L., 2019. Optimization of second-phase sampling for multivariate soil mapping purposes: Case study from a wine region, Hungary. Geoderma 352, 373–384. https://doi.org/10.1016/j.geoderma.2018.02.030.
Shi, J., Yang, L., Zhu, A.X., Qin, C., Liang, P., Zeng, C., Pei, T., 2018. Machine-learning variables at different scales vs. knowledge-based variables for mapping multiple soil properties. Soil Sci. Soc. Am. J. 82 (3), 645–656. https://doi.org/10.2136/sssaj2017.11.0392.
Svetnik, V., Liaw, A., Tong, C., Culberson, C., Sheridan, R.P., Feuston, B.P., 2003. Random Forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958. https://doi.org/10.1021/ci034160g.
Vašát, R., Heuvelink, G.B.M., Borůvka, L., 2010. Sampling design optimization for multivariate soil mapping. Geoderma 155, 147–153. https://doi.org/10.1016/j.geoderma.2009.07.005.
Walvoort, D.J.J., Brus, D.J., de Gruijter, J.J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Comput. Geosci. 36 (10), 1261–1267. https://doi.org/10.1016/j.cageo.2010.04.005.
Wang, S., Zhuang, Q., Jin, X., Yang, Z., Liu, H., 2020. Predicting soil organic carbon and soil nitrogen stocks in topsoil of forest ecosystems in northeastern China using remote sensing data. Remote Sens. 12, 1115. https://doi.org/10.3390/rs12071115.
Webster, R., Oliver, M.A., 1990. Statistical methods in soil and land resource survey. Oxford University Press, New-York, USA.
Wilson, J.P., Gallant, J.C. (Eds.), 2000. Terrain analysis: principles and applications. Wiley, New York Chichester.
Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2012. Which Covariates Are Needed for Soil Carbon Models in Florida. In: Digital Soil Assessment and Beyond. CRC Press, pp. 109–113. https://doi.org/10.1201/b12728-23.
Xu, W., Yuan, W., Cui, L., Ma, M., Zhang, F., 2019. Responses of soil organic carbon decomposition to warming depend on the natural warming gradient. Geoderma 343, 10–18. https://doi.org/10.1016/j.geoderma.2019.02.017.
Yang, L., Zhu, A.X., Qi, F., Qin, C.Z., Li, B.L., Pei, T., 2013. An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping. Int. J. Geogr. Inf. Sci. 27 (1), 1–23. https://doi.org/10.1080/13658816.2012.658053.
Yang, L., Brus, D.J., Zhu, A.X., Li, X.M., Shi, J.J., 2018. Accounting for access costs in validation of soil maps: A comparison of design-based sampling strategies. Geoderma 315, 160–169. https://doi.org/10.1016/j.geoderma.2017.11.028.
Yang, L., Qi, F., Zhu, A.X., Shi, J.J., An, Y.M., 2016. Evaluation of integrative hierarchical stepwise sampling for digital soil mapping.Soil Sci. Soc. Am. J. 80 (3), 637–651. https://doi.org/10.2136/sssaj2015.08.0285.

Yang, L., Zhu, A.X., Zhao, Y., Li, D., Zhang, G., Zhang, S., Band, L.E., 2017. Regional soil mapping using multi-grade representative sampling and a fuzzy membership-based mapping approach. Pedosphere 27, 344–357. https://doi.org/10.1016/S1002-0160(17)60322-9.

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., He, X., Shen, F., Zhou, C., Zhu, A.X., Gao, B., Chen, Z., Li, M., 2020. Improving prediction of soil organic carbon content in croplands using phenological parameters extracted from NDVI time series data. Soil Tillage Res. 196 https://doi.org/10.1016/j.still.2019.104465.

Yang, L., Cai, Y., Zhang, L., Guo, M., Li, A., Zhou, C., 2021a. A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology variables. Int. J. Appl. Earth Obs. Geoinf. 102 https://doi.org/10.1016/j.jag.2021.102428.

Yang, L., Shen, F., Zhang, L., Cai, Y., Yi, F., Zhou, C., 2021b. Quantifying influences of natural and anthropogenic factors on vegetation changes using structural equation modeling: A case study in Jiangsu Province, China. Journal of Cleaner Production 280. https://doi.org/10.1016/j.jclepro.2020.124330.

Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surf. Proc. Land. 12, 47–56. https://doi.org/10.1002/esp.3290120107.

Zhang, S.J., Zhu, A.X., Liu, J., Yang, L., Qin, C.Z., An, Y.M., 2016. An heuristic uncertainty directed field sampling design for digital soil mapping. Geoderma 267, 123–136. https://doi.org/10.1016/j.geoderma.2015.12.009.

Zhang, L., Yang, L., Ma, T., Shen, F., Cai, Y., Zhou, C., 2021. A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data. Geoderma 384. https://doi.org/10.1016/j.geoderma.2020.114809.

Zhu, A.X., 1999. A personal construct-based knowledge acquisition process for natural resource mapping. International Journal of Geographical Information Science 13, 119–141. https://doi.org/10.1080/136588199241382.

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, 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., Lu, G., Liu, J., Qin, C., Zhou, C., 2018. Spatial prediction based on Third Law of Geography. Ann. Gis 24, 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.