Geoderma 351 (2019) 130–143

Contents lists available at ScienceDirect

Geoderma

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

A representativeness heuristic for mitigating spatial bias in existing soil samples for digital soil mapping

Guiming Zhang[a,*], A-Xing Zhu[b,c,d,e]

[a] Department of Geography & the Environment, University of Denver, Denver 80208, USA
[b] Department of Geography, University of Wisconsin-Madison, Madison 53706, USA
[c] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
[d] Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing 210023, China
[e] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China

ARTICLE INFO

Handling Editor: Alex McBratney

Keywords:
Sample representativeness
Existing soil samples
Spatial bias
Digital soil mapping (DSM)

ABSTRACT

Digital soil mapping (DSM) often relies on existing soil samples obtained from various sources. However, the spatial distribution of such soil samples can be biased, for example, towards areas of better accessibility. Such biased coverage over the geographic space (i.e., spatial bias) often leads to biased coverage of the soil samples over the environmental covariate space. As a result, spatial bias degrades the correlation or statistical relationship between samples and covariates in the study area and impedes DSM accuracy. This paper presents a representativeness heuristic for mitigating spatial bias in existing soil samples for improving DSM accuracy. The key idea of the heuristic was to define and quantify sample representativeness as the *goodness-of-coverage* of the soil samples over the environmental covariate space. Spatial bias was then mitigated by weighting the samples towards maximizing their representativeness. Determination of the sample weights was conceived as an optimization problem and accordingly the optimal weights were determined using a genetic algorithm. To evaluate the effectiveness of the representativeness heuristic, a case study of mapping soil organic matter (SOM) content using existing soil samples was conducted in Heshan study area, northeastern China. Results showed that weighting soil samples using the optimal weights determined from the representativeness heuristic improved SOM content mapping accuracy. Moreover, a positive relationship between sample representativeness and mapping accuracy was observed, suggesting sample representativeness is an effective indicator of mapping accuracy. Additionally, the determined optimal weights were informative of individual sample importance and thus can be used as guidance to filter existing soil samples to improve DSM accuracy.

1. Introduction

Information on the spatial distribution of soil is a crucial ingredient for environmental monitoring, modeling and management (Arrouays et al., 2014). For example, soil spatial information is a key input to land surface processes modeling such as hydrological models (Singh and Woolhiser, 2002; Zhu and Mackay, 2001). Digital soil mapping (DSM) is widely used to predict soil spatial information from environmental covariate layers based on the soil-environment covariation relationship derived from soil samples (McBratney et al., 2003; Minasny and McBratney, 2016; Scull et al., 2003). Therefore, soil samples used for DSM need to be representative of the soil-environment relationship over the geographic mapping area, in order to achieve satisfactory mapping accuracy (An et al., 2018; Vaysse and Lagacherie, 2015).

Representative soil samples are usually obtained through sampling campaigns following well-designed sampling schemes such as stratified sampling, systematic sampling, and purposive sampling. Under these sampling schemes, sample locations are allocated to well cover the geographic and/or covariate space to be mapped (De Gruijter et al., 2006; Yang et al., 2013; S. Zhang et al., 2016). For example, under stratified sampling, when a mapping area contains sub-areas of different land cover types, soil samples are taken within each of these sub-areas to ensure that all of the variation present in a sub-area is measured in the samples (Jensen and Shumway, 2010). However, conducting such field sampling campaigns to collect soil samples can be very expensive, labor intensive, and time consuming. Therefore, obtaining soil samples for DSM through additional sampling may not always be feasible due to possible budgetary and/or time constraints.

* Corresponding author at: 2050 E. Iliff Ave., Denver, CO 80208, USA.
E-mail address: guiming.zhang@du.edu (G. Zhang).

https://doi.org/10.1016/j.geoderma.2019.05.024
Received 1 November 2018; Received in revised form 10 May 2019; Accepted 15 May 2019

In such cases, existing soil samples available from various sources are often pooled and utilized for DSM (Liu, 2017). Existing samples also provide a basis for making decisions regarding where to place future sample locations should additional soil sampling be conducted (Carré et al., 2007; S. Zhang et al., 2016). Examples of existing soil samples are legacy samples from past soil surveys (Vaysse and Lagacherie, 2015), samples collected by various groups of researchers (An et al., 2018), and samples contributed by volunteer citizen scientists (Rossiter et al., 2015).

Nevertheless, existing soil samples are susceptible to spatial bias (Liu, 2017). That is, the spatial distribution of existing soil samples may be biased towards certain geographic area. For instance, soil samples contributed by volunteers are more concentrated in areas of better accessibility (Kadmon et al., 2004). Soil samples from past soil surveys may contain spatial bias as sampling methods adopted by surveyors are generally empirical and lack statistical criteria (Carré et al., 2007). Even soil samples collected following well-designed sampling schemes can be subject to spatial bias. For example, some designed sample locations may not be sampled due to field condition constraints (e.g., inaccessibility). Under such circumstances, spare soil samples shall be taken at locations chosen at sampler's discretion (S. Zhang et al. 2016). Such deviations from the original sampling design may lead to spatial bias in the samples. Besides, pooling existing soil samples from different sources may also result in spatial bias in the pooled samples because samples collected by different research groups or agencies may cover different parts of the mapping area at imbalanced sample densities (An et al., 2018).

Such spatial bias of soil samples over the geographic space can result in biased coverage of the samples over the environmental covariate space (De Gruijter et al., 2006; Minasny and McBratney, 2006), although bias in covariate space does not necessarily equate to spatial bias. Spatial bias thus would degrade representativeness of the samples, which is the degree to which the samples capture the variabilities of the environmental covariates over the mapping area. The spatial bias in soil samples, if not properly accounted for, would result in relatively low DSM accuracy compared with approaches that take spatial bias into account (An et al., 2018; Carré et al., 2007).

There are few studies related to assessing or improving representativeness of existing soil samples for DSM. Carré et al. (2007) proposed a method for estimating and improving the representativeness of existing legacy soil samples. In this method, the principle of Latin hypercube sampling proposed by Minasny and McBratney (2006) was used to assess the representativeness of existing soil samples through the HELS (Hypercube Evaluation of a Legacy Sample) algorithm and to guide additional sampling efforts through the HISQ algorithm. The two algorithms are detailed in Carré et al. (2007) and the basic ideas of the algorithms were summarized here as follows. The representativeness of existing soil samples was estimated by calculating the relative densities of the existing sampling units and the environmental covariate data. The relative densities indicate areas where there is over- or under-observation in the existing soil samples relative to the covariates. Additional sampling should then be prioritized to areas with high degree of under-observation (Carré et al., 2007). This method improves the representativeness of existing soil samples by supplementing additional samples; it does not provide a way to improve the accuracy DSM utilizing all existing samples.

An et al. (2018) developed an approach for identifying representative samples from existing soil samples based on representativeness of individual samples. In this approach, the representativeness of an individual soil sample is measured as the fuzzy membership of the sample location to an environmental cluster (i.e., distance between the sample location and an environmental cluster centroid in the environmental covariate space) (Yang et al., 2013). A set of representative samples were then identified by selecting soil samples from existing samples based on their individual sample representativeness (e.g., selecting sample locations whose fuzzy membership values exceed a user-defined threshold). The reported case study of this approach showed that the accuracy of DSM achieved using the representative samples identified from all existing samples was generally higher than the accuracy achieved using non-representative samples (i.e., all samples minus representative samples), but it was only comparable to the accuracy achieved using all samples (An et al., 2018). This approach is useful for identifying representative samples from existing soil samples. However, using the identified representative samples for DSM did not achieve a clear improvement in prediction accuracy over using all existing samples.

This paper presents a novel representativeness heuristic for mitigating spatial bias in existing soil samples to improve the accuracy of DSM. The basic idea underlying this heuristic follows. DSM utilizing representative soil samples can achieve high accuracy because the samples sufficiently capture the variation of the environmental covariates (Minasny and McBratney, 2006; Qi and Zhu, 2003; Yang et al., 2013). In light of this, representativeness of soil samples was defined and quantified in this study as the goodness-of-coverage of the sample locations over the covariate space (Kruskal and Mosteller, 1979). Representativeness of the soil samples was then improved through weighting the samples towards maximal goodness-of-coverage. The sample weights determined based on this representativeness heuristic were then used to weight the soil samples in deriving the soil-environment covariation relationship for DSM. The hypothesis was that weighting existing soil samples using the weights determined from the heuristic can mitigate spatial bias in the samples to increase their representativeness and thus improve DSM accuracy (Zhang et al., 2018).

The main objective of this study was to present the representativeness heuristic and evaluate its effectiveness in mitigating spatial bias in existing soil samples for improving DSM accuracy. A case study of mapping A-horizon soil organic matter (SOM) content using existing soil samples in Heshan study area, northeastern China was conducted to thoroughly evaluate the effectiveness of the proposed heuristic using two soil mapping methods: the individual soil mapping method (Zhu et al., 2015) and multiple linear regression.

2. Material and methods

2.1. Study area and data

2.1.1. Study area

The 60 km^2 study area is located at Heshan farm (116°12′E, 48°57′N) in Heilongjiang province, northeastern China (Fig. 1). It has a maximum terrain relief of about ninety meters and is generally flat with a gentle slope gradient of less than four degrees. The soils in this area are mostly formed on deposits of silt loam loess except the valley where the underlying parent material is fluvial deposits. The farm has been cultivated for over forty years to grow soybeans and wheat. There is a thick *A*-horizon (top-layer of soil) with high organic matter content. The land use and soil management have been uniform throughout the area and no organic fertilizer has been applied to these soils to maintain agricultural productivity because of the naturally high organic matter content (Zhu et al., 2010).

2.1.2. Soil samples

2.1.2.1. Existing soil samples. There are 59 existing soil samples in the study area obtained through sampling campaigns for various purposes in previous studies (Yang et al., 2013; Zeng et al., 2016; Zhu et al., 2010) (Fig. 1). Among these soil samples, 29 samples were collected through integrative hierarchical stepwise sampling (Yang et al., 2013), 10 samples through subjective sampling, and 20 samples through transect sampling (Zhu et al., 2010). Interested readers can refer to the above references for details. An extensive discussion of the sampling methods is not necessary here because the representativeness heuristic proposed in this study imposes no requirements or assumptions on the methods used to collect existing soil samples. The soil organic matter

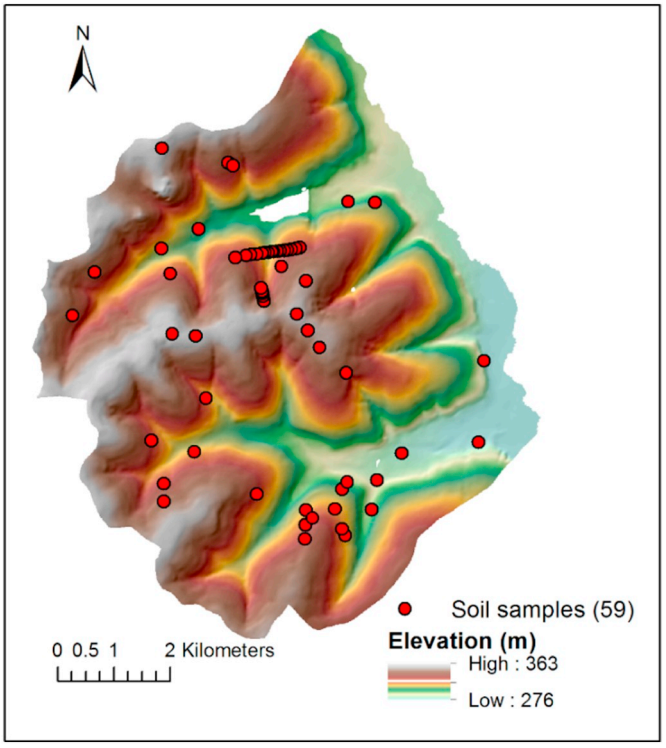

Fig. 1. Study area and soil samples.

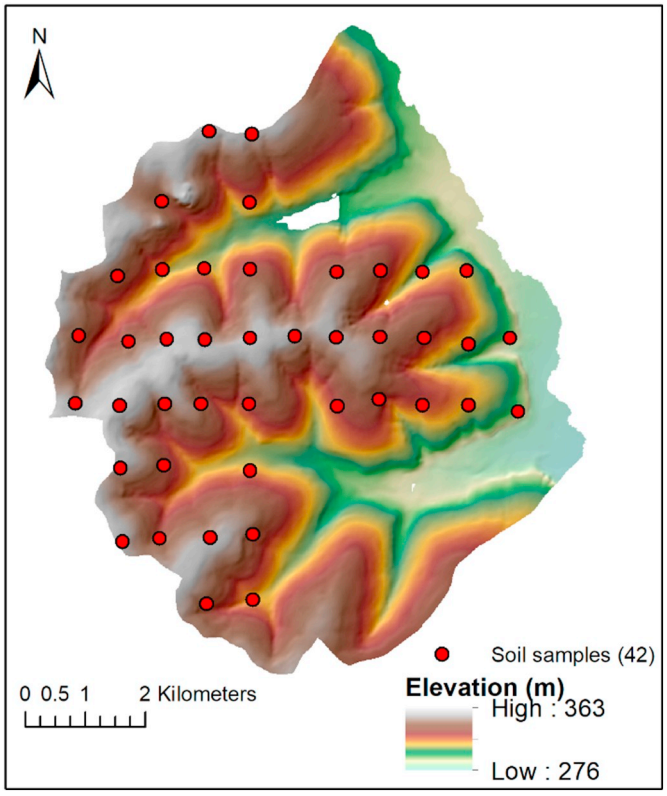

Fig. 2. Validation soil samples in the study area.

(SOM) content (%) in *A*-horizon soil was measured for each of the soil samples. The mean SOM content over the sample locations was 4.454%, with a standard deviation of 1.638%.

The existing soil samples were subject to spatial bias. As can be seen from their spatial distribution, there are areas of clusters of samples where soil samples are more concentrated than other areas. When using these existing soil samples to train DSM models for mapping SOM content in the study area, the proposed representativeness heuristic was applied to mitigate the spatial bias in these soil samples to improve mapping accuracy (Section 2.2).

2.1.2.2. Validation soil samples. Accuracy of the predicted SOM content maps were assessed based on 42 soil samples collected through systematic sampling on a 1100 m × 740 m grid in the study area (Fig. 2) (Zhu et al., 2010). The *A*-horizon SOM content (%) was measured for each of the soil samples. The mean SOM content over the sample locations was 4.319%, with a standard deviation of 0.806%. These regularly spaced soil samples cover most parts of the study area and spread across various slope positions (footslope, backslope, slope shoulder, ridge) (Qin et al., 2009), except the lowest parts of the landscape which are floodplains that were not sampled. The representativeness of the samples was 0.858, evaluated using the method for quantifying sample representativeness as detailed in Section 2.2.1. These soil samples were used as validation samples (Zhu et al., 2015) to evaluate accuracies of the SOM content maps predicted from the 59 existing soil samples (Section 2.4).

2.1.3. Environmental covariates

Environmental covariates were selected based on soil forming factors (Dokuchayev, 1883; Jenny, 1994). There are five categories of soil forming factors: climate, organisms, terrain, parent materials, and time. In this small study area, parent materials and macro-climatic conditions are relatively uniform and micro-climatic variations can be reflected by topographic conditions. The spatial variation of the *A*-horizon SOM content in the study area is mostly influenced by topographic and vegetation conditions (Yang et al., 2013; Zhu et al., 2015, 2010). Thus, six topographic covariates including elevation, slope gradient, contour curvature, profile curvature, relative slope position and topographic wetness index (TWI), and one vegetation covariate normalize difference vegetation index (NDVI) were selected as environmental covariates for soil mapping in the study area (Fig. 3) (Zhu et al., 2015). A digital elevation model (DEM) of the study area at 10-m spatial resolution was created from the 1:10,000 topographic map of the area. Elevation, slope gradient, contour curvature, profile curvature, relative slope position (Qin et al., 2012, 2009), and TWI (Pei et al., 2010; Qin et al., 2007) were then derived from the DEM. NDVI was derived from a Landsat ETM+ image of the area obtained on September 25, 2000 (Zhu et al., 2015).

Principal component analysis (PCA) (Jolliffe, 2002) was adopted to eliminate collinearity among the covariates and reduce the number of covariates. Prior to PCA transformation, outliers in the covariate data were removed and the covariates were linearly stretched to range 0 to 100 (elevation, slope gradient, relative slope position, TWI, NDVI) or range −50 to 50 (contour curvature, profile curvature) (Yang et al., 2013). PCA was then adopted for transforming the covariate data to derive linearly independent principal components (PCs). The first three PCs retaining 91.7% (66.6%, 17.7%, and 7.4%, respectively) of the total variance were used as new environmental covariates for mapping SOM content in the study area (Fig. 4).

2.2. The representativeness heuristic for mitigating spatial bias

2.2.1. Quantifying sample representativeness

Based upon the basic ideas presented in the Introduction, *representativeness* of soil samples in this study is defined as the goodness-of-coverage of the samples in the covariate space (Kruskal and Mosteller, 1979). It was quantified as the similarity between two probability density distributions in the covariate space: the distribution over soil sample locations (*sample distribution* hereafter) and the

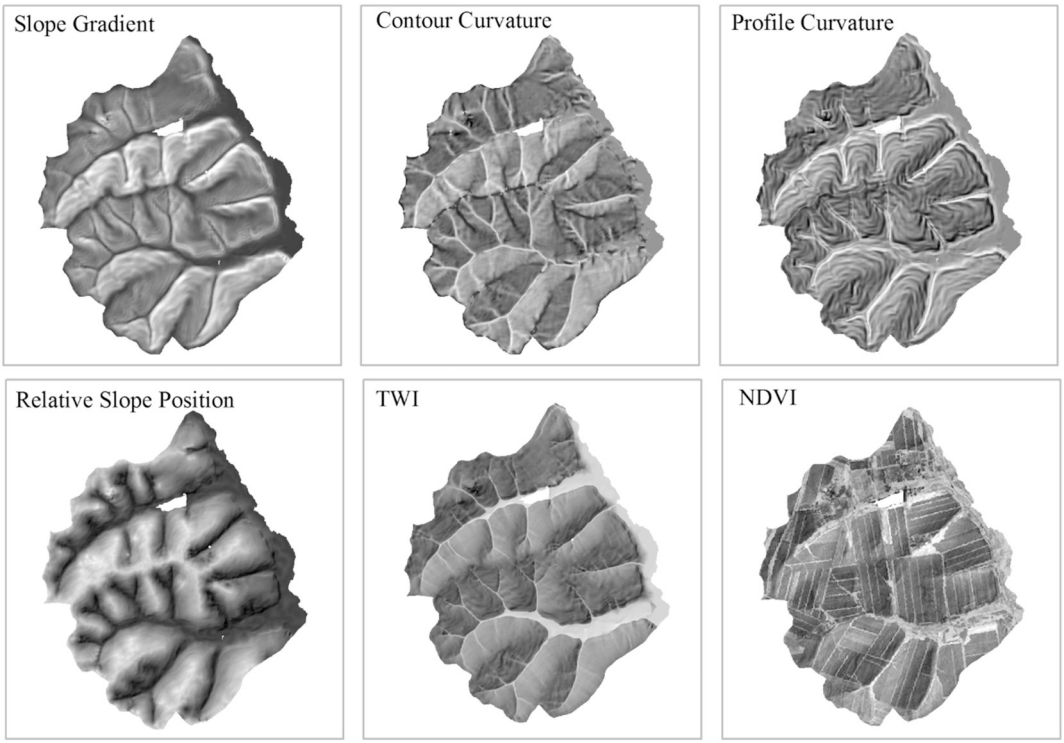

Fig. 3. Environmental covariates selected for mapping *A*-horizon soil organic matter (SOM) content in the study area. Lighter color indicates high covariate values.

distribution over the study area (*population distribution* hereafter) (Zhang, 2018).

Kernel density estimation (KDE) was adopted to estimate the two distributions in the covariate space. KDE is a nonparametric method that can estimate continuous probability density functions (PDF) from discrete sample values (Silverman, 1986) using equation

$$f(v) = \sum_{i=1}^{n} w_i \frac{1}{h} K\left(\frac{v - V_i}{h}\right) \quad (1)$$

where $f(v)$ is the estimated PDF over variable v, V_i the ith sample value of v, w_i the weight for the ith sample value, and n the total number of sample values. The Gaussian kernel was adopted for the kernel function K in this study (Silverman, 1986):

$$K\left(\frac{v - V_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(v-V_i)^2}{2h^2}} \quad (2)$$

where h is a smoothing parameter called bandwidth. Bandwidth is a crucial parameter for KDE. A too-large bandwidth would result in a flat PDF that fails to reflect variability in the sample values. A too-small bandwidth would result in a spiky PDF that contains too much noise. The rule-of-thumb algorithm is often used to determine the bandwidth h for KDE when sample size (i.e., n) is large (Silverman, 1986):

$$h = 1.06 \cdot \sigma_v \cdot n^{-1/5} \quad (3)$$

in which σ_v is the standard deviation of the n sample values of v. When the sample size is small, the "golden section search" procedure (Brunsdon, 1995) can be used to determine the optimal bandwidth for KDE based on the maximum likelihood estimation principle (Zhang et al., 2017). Interested readers are referred to Brunsdon (1995) for details of the procedure.

Representativeness of the existing soil samples was computed following three steps. *First*, the sample distribution and the population distribution regarding each environmental covariate (i.e., each of the three principal components) were estimated using KDE (Eq. (1)). When estimating the sample distribution using KDE, covariate values at soil sample locations were taken as the sample values. Each soil sample may carry a different sample weight (Section 2.2.2). Given the small sample size (i.e., number of existing soil samples), the bandwidth was determined using the "golden section search" procedure. When estimating the population distribution using KDE, covariate values at all raster

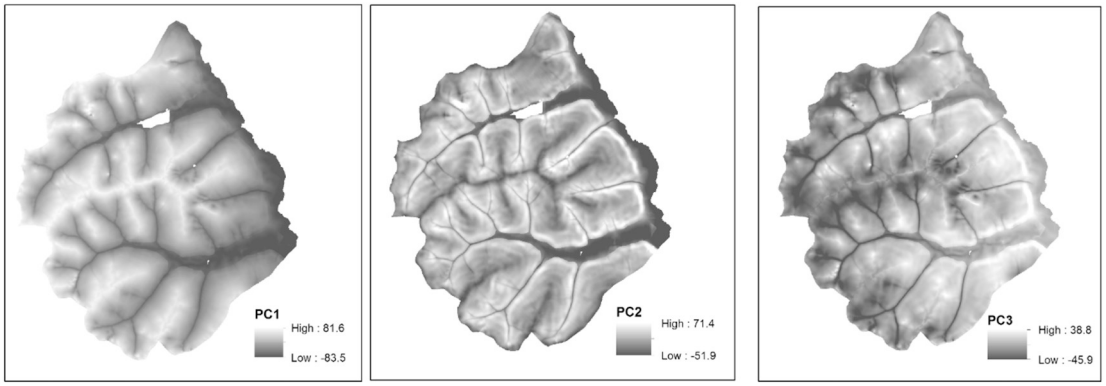

Fig. 4. Selected principal components used as environmental covariates for mapping soil organic matter content in the study area.

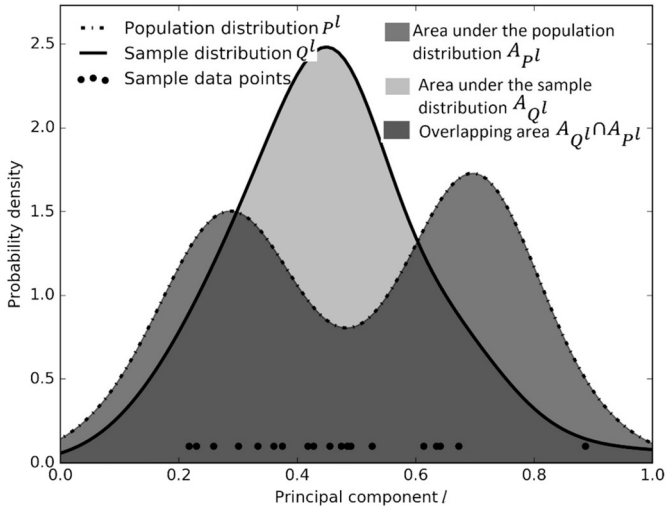

Fig. 5. Schematic example of the overlapping area between the sample distribution and the population distribution.

cells in the study area were taken as sample values of equal weight. Given the large sample size (i.e., number of cells in the study area), the bandwidth was determined using the rule-of-thumb algorithm.

Second, the similarity between the sample distribution and the population distribution regarding each environmental covariate was computed as the overlapping area under the two probability distribution curves (Zhu, 1999):

$$SIM^l = \frac{2 \times A_{Q^l} \cap A_{P^l}}{A_{Q^l} + A_{P^l}}. \quad (4)$$

In the above equation, the sample distribution and population distribution regarding the l^{th} covariate estimated in the first step are denoted as Q^l and P^l, respectively. Areas under the two distribution curves are denoted as A_{Q^l} and A_{P^l}, respectively. The similarity between the two distributions is denoted as SIM^l, which equals the area of the overlapping part of the two distribution curves denoted as $A_{Q^l} \cap A_{P^l}$ (Fig. 5). The similarity SIM^l has a value range of [0, 1.0] with a higher value reflecting better goodness-of-coverage of the soil samples regarding the lth covariate.

The Kulback-Liebler (KL) divergence is a commonly used measure of (dis)similarity between two probability distributions. However, we found the KL divergence measure to be easily saturated (i.e., KL divergence is very close to zero on two distributions that are not so similar). This makes it an inappropriate objective function for the optimization algorithm to optimize sample weights (see Section 2.2.2). Thus, the overlapping area under the two probability density curves (Zhu, 1999) was adopted as a similarity measure in this study.

Third, sample representativeness was then computed as a weighted average of the similarities regarding individual covariates (Eq. (5)). The weight was proportional to the proportion of the variance each principal component retains. The rationale is that a principal component retaining larger variance is more prominent for the soil samples to capture covariate variations, and thus has a larger contribution to sample representativeness.

$$R = SIM^{overall} = \sum_{l=1}^{L} \frac{\lambda^l}{\sum_{j=1}^{L} \lambda^j} SIM^l. \quad (5)$$

In the above equation, sample representativeness R equals the overall similarity $SIM^{overall}$. The similarity regarding the lth covariate is SIM^l. The eigenvalue of the lth principal component λ^l indicates the percentage of variance it retains. The number of selected principal components is L. Sample representativeness R has a value range of [0, 1.0], with a higher value indicating better sample representativeness.

2.2.2. Improving sample representativeness

The spatial bias in existing soil samples was mitigated by improving sample representativeness, i.e., increasing the overall similarity between the sample distribution and the population distribution. Recall that sample weight plays a role in estimating the sample distribution (Eq. (1)). Thus, representativeness improvement was accomplished by adjusting the weight of individual soil samples towards increasing the similarity between the two distributions.

Here the key is to determine the sample weights. Determination of the sample weights in this study was conceived as an optimization problem, where the objective is to find a set of optimal sample weights that maximize the representativeness of the soil samples. A genetic algorithm (GA) implemented in the Distributed Evolutionary Algorithms in Python (DEAP) package (Rainville et al., 2012) was adopted to determine the optimal sample weights. GA works as follows. In essence, GA represents sample weights as an "*individual*" that consists of a list of ordered "*genes*", where each "*gene*" is one sample weight value. At first, an initial pool ("*population*") of candidate sample weight lists ("*individuals*") were generated, and each list was filled with sample weight values randomly drawn from a uniform distribution in the interval of [1.0, W_{max}] (W_{max} is the maximum possible sample weight; W_{max} was set to 10.0 by default). GA then evaluates sample representativeness given each sample weight list in the pool. Sample weight lists were then selected with selection probabilities proportional to the representativeness to remain in the pool. Pairs of sample weight lists were also selected with the selection probabilities to apply the crossover operator to produce new weight lists. After the new weight lists were added to the pool, a portion (e.g., 5%) of the weight lists in the pool were selected with a uniform probability and each selected weight list was mutated by adjusting the weight value at randomly selected positions by a small random amount drawn from a Gaussian distribution (mean = 0; standard deviation = 0.5). In this way, the pool was updated and subsequently, another iteration of evaluation, selection, crossover, and mutation were repeated. GA terminates after going through a prescribed number of iterations ("*generations*", e.g., 200) or the highest sample representativeness exceeds a predefined threshold (e.g. 0.9). The sample weight list in the current pool corresponding to the highest representativeness value is then returned as the optimal sample weights.

2.3. Soil mapping methods

Two methods were adopted for mapping the *A*-horizon SOM content in the study area. One is the individual predictive soil mapping method proposed by Zhu et al. (2015) for mapping soil properties. The other is the multiple linear regression method, a general approach for modeling multivariate linear relationships. Soil mapping using these two methods allows examination of the effectiveness of the proposed representativeness heuristic on a domain-specific soil mapping method and a general predictive mapping method.

2.3.1. Individual predictive soil mapping (iPSM)

iPSM is a method specially designed for digital soil mapping (Zhu et al., 2015) and has been used in a wide range of studies (An et al., 2018; Liu, 2017; Zeng et al., 2016; S. Zhang et al. 2016). An overview of the iPSM method is provided as follows. Interested readers may refer to (Zhu et al., 2015) for full details of the method. iPSM uses the soil-environment relationship at each individual soil sample location to predict soil properties at unsampled locations. Based on the assumption that locations of similar environment conditions shall have similar soil property values, iPSM predicts soil property value at an unsampled location as a weighted average of soil property values observed at sample locations, where the environmental similarities between the unsampled location and soil sample locations are used as weights. iPSM imposes no requirements on sample size and does not require the set of sample locations being representative. It is an effective alternative for

soil mapping when existing soil samples are limited in terms of representing the study area (Zhu et al., 2015).

The iPSM method for mapping soil properties includes two main operational steps. The *first step* is to calculate environmental similarity. Environmental similarity between an unsampled location j and sample location i is first evaluated at the individual environmental variable (i.e., principal component) level, and then similarities based on all environmental variables are integrated to represent the overall similarity between unsampled location j and sample location i.

The environmental similarity between unsampled location j and sample location i w.r.t. the lth principal component, $S_{i,j}{}^{l}$, is calculated as:

$$S_{i,j}^{l} = \exp\left[-\frac{(V_i^l - V_j^l)^2}{2 \times \left(\frac{SD^l}{SD_j^l} \times SD^l\right)^2}\right]. \tag{6}$$

In the above equation, the value of the lth principal component at sample location i and unsampled location j are denoted by V_i^l and V_j^l, respectively. The standard deviation of the lth principal component is SD^l. The standard deviation of the lth principal component from V_j^l (instead of from the mean), SD_j^l, is computed by:

$$SD_j^l = \sqrt{\frac{\sum_{p=1}^{m}(V_p^l - V_j^l)^2}{m}} \tag{7}$$

where $V_p{}^l$ is the value of the lth principal component at raster cell p, and m is the total number of raster cells in the study area.

The overall environmental similarity between unsampled location j and sample location i considering all L selected principal components, $S_{i,j}$, is then determined following a limiting factor approach based on the simplistic assumption that the least similar environmental factor determines the overall environmental similarity between two locations. A minimum operator was applied to take the minimum of the environmental similarities to individual principal components (i.e., $S_{i,j}{}^{1}$, $S_{i,j}{}^{2}$, … , $S_{i,j}{}^{L}$) as the overall environmental similarity $S_{i,j}$ (Zhu et al., 2015; Zhu and Band, 1994):

$$S_{i,j} = \min(S_{i,j}^1, S_{i,j}^2, ..., S_{i,j}^L) \tag{8}$$

The environmental similarity between unsampled location j and each of the n sample locations can be computed following Eq. (6)–(8).

The *second step* of iPSM is to compute the soil property value at unsampled location j based on its environmental similarities to the n sample locations. A weighted average approach is adopted for this purpose:

$$\hat{T}_j = \frac{\sum_{i=1}^{n} S_{i,j} \times T_i}{\sum_{i=1}^{n} S_{i,j}} \tag{9}$$

where $\hat{T}_j$ is the predicted value of the soil property (i.e., A-horizon SOM content) at unsampled location j, and T_i the observed value of the soil property at sample location i.

The original iPSM method does not account for sample weight when predicting soil property values. This study extends the iPSM method on weighted soil samples. Soil samples can be weighted with the optimal sample weights determined from the proposed representativeness heuristic in predicting the soil property value at unsampled location j:

$$\hat{T}_j = \frac{\sum_{i=1}^{n} w_i \times S_{i,j} \times T_i}{\sum_{i=1}^{n} w_i \times S_{i,j}}. \tag{10}$$

Here w_i is the weight of the soil sample at location i. Everything else being equal, a soil sample of a larger weight has larger contribution to the estimation at a unsampled location.

2.3.2. Multiple linear regression (MLR)

MLR is a general method for modeling multivariate linear relationships between a dependent variable and independent variables. It has been widely used for mapping soil properties from environmental covariates (Brus et al., 2019; Grunwald, 2009; Zeng et al., 2016; Zhu et al., 2015). Unlike iPSM, MLR has stricter requirements on sample size and the representativeness of the sample set for building a statistically robust model. An MLR model takes the following form:

$$\hat{T}_j = \beta^0 + \sum_{l=1}^{L} \beta^l \times V_j^l \tag{11}$$

where $\hat{T}_j$ is the predicted value of the soil property (i.e., A-horizon SOM content) at unsampled location j, V_j^l the value of the lth principal component at location j, β^0 the intercept, and β^l the coefficient for the lth principal component.

The intercept β^0 and coefficients β^ls are determined by fitting the model on training data (i.e., soil property values and values of the principal component at the n sample locations) based on the ordinary least squares (OLS) criterion, i.e., finding the values of β^0 and β^ls that minimize the sum of squared residuals between predicted soil property values and observed soil property values at the n sample locations:

$$\beta^0, \beta^1, ..., \beta^L = argmin_{\beta^0,\beta^1,...,\beta^L}\left\{\sum_{i=1}^{n}\left[T_i - \left(\beta^0 + \sum_{l=1}^{L}\beta^l \times V_i^l\right)\right]^2\right\} \tag{12}$$

The OLS procedures implemented in the Scikit-learn package (Pedregosa et al., 2012) were adopted to train an MLR model using (unweighted) soil samples.

This study extends the MLR method on weighted soil samples. In training an MLR model, soil samples can be weighted with the optimal sample weights determined from the proposed representativeness heuristic. Specifically, sample weights are used to weight individual squared residuals in fitting the model parameters using OLS:

$$\beta^0, \beta^1, ..., \beta^L = argmin_{\beta^0,\beta^1,...,\beta^L}\left\{\sum_{i=1}^{n} w_i\left[T_i - \left(\beta^0 + \sum_{l=1}^{L}\beta^l \times V_i^l\right)\right]^2\right\} \tag{13}$$

where w_i is the weight of the soil sample at location i. Samples with larger weights are treated as more important in this model fitting process. The OLS procedures implemented in the Scikit-learn package (Pedregosa et al., 2012), capable of accounting for sample weights, were adopted to train an MLR model using weighted soil samples.

The MLR model, trained with either unweighted soil samples or weighted soil samples, was then applied to the environmental condition at every location (raster cell) in the study area to predict SOM content values.

2.4. Accuracy assessment

Three indices, the mean error (ME), root mean square error (RMSE), and mean absolute error (MAE), were used to measure accuracy of the predicted SOM content maps. All three indices are computed based on the differences between the predicted and observed SOM content values at the validation soil sample locations (Section 2.1.2.2):

$$ME = \frac{1}{k}\sum_{i=1}^{k}(T_i - \hat{T}_i), \tag{14}$$

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(T_i - \hat{T}_i)^2}, \tag{15}$$

and

$$MAE = \frac{1}{k}\sum_{i=1}^{k}|T_i - \hat{T}_i|. \tag{16}$$

In the above equations, k is the number of validation soil samples, and $\hat{T}_i$ and T_i the predicted and observed SOM content values at validation sample location i, respectively.

ME was adopted as a measure of prediction bias, with values closer to zero indicating lower prediction bias. RMSE and MAE express

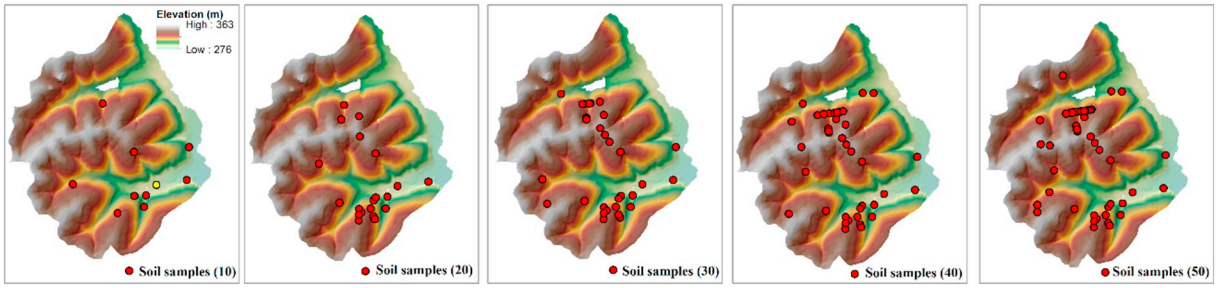

Fig. 6. Soil samples at varying sample sizes selected from the 59 soil samples in the study area.

average model prediction error, with lower values indicate higher prediction accuracy. RMSE is more indicative of large prediction errors than MAE as it gives a relatively high weight to large prediction errors (the errors are squared before being averaged) (Chai and Draxler, 2014).

In addition, the explained variance score (EVS) was adopted to indicate the overall goodness-of-fit of the soil mapping models. EVS was computed using the following equation:

$$EVS = 1 - \frac{Var\{T - \hat{T}\}}{Var\{T\}}, \quad (17)$$

where $Var\{T\}$ and $Var\{T - \hat{T}\}$ are variance of the observed SOM content values and variance of the residuals across the validation soil sample locations, respectively. EVS reflects the proportion of the spatial variability in SOM content as represented by the validation samples explained by the models. The best possible EVS is 1.0 indicating perfect model fitting and lower values indicating worse model fitting.

2.5. Experiment design

Experiments were designed to evaluate the effectiveness of the representativeness heuristic in improving SOM content mapping accuracy and to examine the impact of sample size. Additionally, the optimal weights determined from the heuristic were also used as guidance to filter (instead of weighting) existing soil samples for improving SOM content mapping accuracy.

2.5.1. Effectiveness of the representativeness heuristic

Effectiveness of the representativeness heuristic was evaluated from three aspects. *First*, all 59 existing soil samples (Fig. 1) were used as training samples in the two soil mapping methods (iPSM and MLR) to map SOM content. Accuracies of SOM content maps predicted using models trained with unweighted soil samples were compared to those predicted using models trained with soil samples weighted by the optimal weights determined from the representativeness heuristic. If weighting the soil samples resulted in more accurate SOM content maps, the heuristic can be proved effective.

Second, two tests were performed to examine the statistical significance of the effects of weighting soil samples by the optimal sample weights. For the first test, prediction accuracy achieved under the optimal sample weights was compared to prediction accuracy achieved under randomly assigned weights. For this purpose, one hundred sets of random sample weights were generated, where each weight value was randomly drawn from a uniform distribution over the range [1, W_{max}]. The soil samples weighted by each set of the random weights were used to train the models and to map SOM content. One sample *t*-test was then applied to test if the accuracy achieved under the optimal weights is statistically significantly higher than the average accuracy achieved under the random sample weights. For the second test, one hundred sets of weights were generated by randomly shuffling the order of the optimal weights. The soil samples weighted by each set of the shuffled weights were used to train the models and to map SOM content. One sample t-test was applied to test if the accuracy achieved under the optimal weights is statistically significantly higher than the average accuracy achieved under the randomly shuffled optimal weights. If weighting the soil samples by the optimal weights resulted in statistically significantly higher SOM content prediction accuracies than weighting by random weights or randomly shuffled optimal weights, it can be proved that the improvements in SOM content mapping accuracy achieved through weighting soil samples by the optimal weights did not happen by random chance (i.e., random weight values or optimal weight values in random order), which would in turn suggest that the heuristic is effective.

Third, the relationship between prediction accuracy and sample representativeness was examined. In the GA used to optimize sample weights (Section 2.2.2), sample weights and sample representativeness gradually improves over the generations. At each generation, the best sample weights corresponding to the highest sample representativeness were used to weight the soil samples to train models to map SOM content. The relationship between accuracy of the predicted SOM content map and sample representativeness was examined by plotting prediction accuracy against sample representativeness over the generations of the GA. If the relationship was positive, it can be proved that higher sample representativeness as quantified by the heuristic can effectively indicate higher prediction accuracy.

2.5.2. Impact of sample size

The representativeness heuristic was applied on samples of varying sample sizes (sample size = 10, 20, 30, 40, 50) to investigate the impact of sample size on effectiveness of the heuristic. Soil sample sets at various sample sizes were subjectively selected from the 59 soil samples in a way such that the sample sets maintain certain characteristics of spatial bias (Fig. 6). The procedures for selecting these subjective sample sets were as follows. A sample location on the flood plain in the south part of the study area was chosen as the seed sample. Then samples were drawn randomly from the remaining 58 samples (without replacement) at selection probabilities being inversely related to their distances to the seed sample (i.e., a sample closer to the seed sample have a higher probability of being selected). For each sample size, multiple sets of samples were generated following the above procedures. Spatial distribution patterns of the sample sets were then visually examined. One sample set was subjectively chosen following the principle that the samples should have a relatively wide spatial coverage while concentrating in some areas more than others. For example, the selected sample set of size 10 spreads across most part of the study area but most of the samples are on the floodplain and foot-slopes. Similarly, the sample set of size 20 has a wide spread but most samples are clustered on the foot-slopes and hill-slopes in the south part of the study area.

The representativeness heuristic was applied to determine the optimal sample weights for each set of the subjective samples for soil mapping. For each set of soil samples, accuracy of the SOM content map predicted using models trained with unweighted soil samples were compared to that predicted using models trained with soil samples

weighted by the optimal weights. The impact of sample size on performance of the heuristic was then be examined.

2.5.3. Optimal sample weights as guidance to filter soil samples

Filtering existing samples (i.e., selecting a subset of samples) is another commonly adopted strategy to reduce spatial bias in the samples (An et al., 2018; Boria et al., 2014; Varela et al., 2014). The optimal weights of the 59 existing soil samples determined from the representativeness heuristic were used as guidance to filter soil samples for SOM content mapping. The existing soil samples were first sorted on descending order of their weights. A sample set at size s (s = 20, 21, …, 59) was then obtained by selecting the first s samples (i.e., in descending order of weight). As comparisons, at each sample size, another sample set of the same size s was obtained by selecting the last s samples (i.e., in ascending order of weight). Yet another one hundred sets of samples of size s were obtained, each was constructed by randomly selecting s samples from the 59 samples (i.e., random samples). Sample representativeness was computed for each sample set obtained through the above three filtering strategies; each sample set (unweighted) was also used to train models for SOM content mapping. This experiment allows examining how mapping accuracy responds to sample filtering strategies.

3. Results

3.1. Effectiveness of the representativeness heuristic

By weighting the samples with the optimal weights (Fig. 7) determined from the representativeness heuristic, the representativeness of the 59 soil samples increased from 0.906 to 0.964. The general spatial patterns of the SOM content maps predicted based on the unweighted soil samples and soil samples weighted by the optimal weights (Fig. 8) are similar. Lower-to-toe slopes and floodplain areas were predicted to have high SOM content and upper-to-middle slopes were predicted to have low SOM content. This agrees with our understanding of how the terrain influences SOM content in the study area. On lower-to-toe slopes, gentle depositional processes tend to be the dominant processes that usually lead to higher *A*-horizon SOM content. On upper-to-middle slopes, erosive processes tend to be the dominant processes and that reduce the A-horizon SOM content. SOM content maps predicted with the iPSM method tend to have fewer variations and smaller value ranges than those predicted with the MLR method.

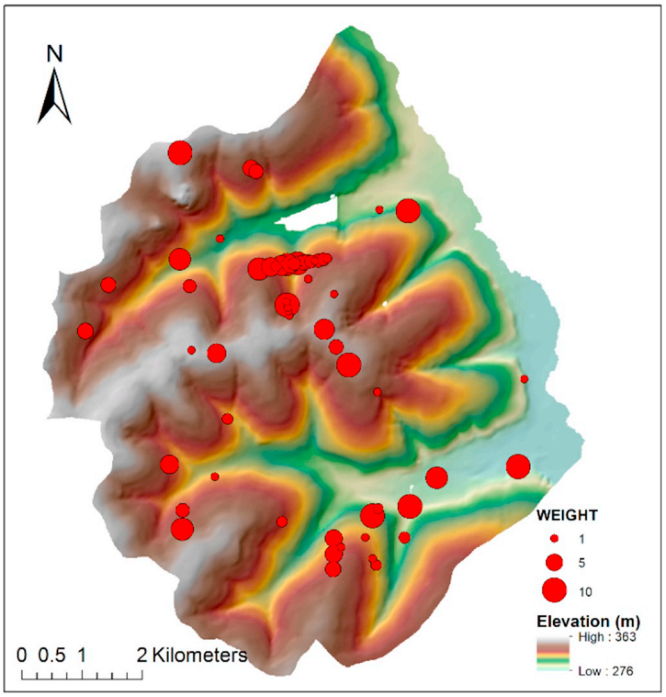

Fig. 7. Optimal weights of the 59 soil samples determined from the representativeness heuristic.

Nonetheless, validation revealed that accuracies of the SOM content maps predicted based on soil samples weighted by the optimal weights were generally higher than those predicted based on unweighted soil samples (Table 1). When the samples were weighted to train an MLR model for mapping SOM content, there was a 14%, 10.5% and 39% decrease in RMSE, MAE and ME, respectively. More outstandingly, the EVS increased from 0.033 to 0.23, suggesting the MLR model trained using weighted samples can explain a much larger proportion of variability in the SOM content over the study area. SOM content maps predicted using the iPSM method were generally more accurate than those predicted using MLR. Yet weighting the samples in iPSM did not improve prediction accuracy much. RMSE slightly increased while MAE and EVS slightly decreased. However, there was still a 59.7% decrease in ME, suggesting that weighting the samples led to less biased predictions.

Moreover, accuracies of the SOM content maps (RMSE, MAE and ME) predicted based on soil samples weighted by the optimal weights were statistically significantly higher than samples weighted by the random assigned weights (Table 2). This observation was consistent for both iPSM and MLR.

In addition, for the MLR method, there were strong negative relationships between RMSE/MAE/ME and sample representativeness, and strong positive relationships between EVS and representativeness (i.e., positive relationship between prediction accuracies and representativeness) (Fig. 9). Results for iPSM were mixed. Although the relationships between RMSE/ME and representativeness were negative, the relationship between MAE and representativeness was positive, and the relationship between EVS and representativeness was negative. Noticeably, the ranges of the accuracy indices for iPSM were narrower than those for MLR, suggesting that the proposed representativeness heuristic had not as much impact on iPSM as on MLR.

3.2. Impact of sample size

Across various sample sizes, weighting soil samples by the optimal weights determined from the representativeness heuristic (Fig. 10) consistently improved SOM content mapping accuracies (Table 3). Generally, the accuracy improvements were less significant on soil samples of larger sample sizes that have higher representativeness (Table 4. Nonetheless, weighting the soil samples still helped decrease ME and increase EVS even at larger sample sizes.

3.3. Effects of filtering soil samples

As more soil samples were selected in descending order of the optimal sample weight, sample representativeness improved dramatically (Fig. 11). It reached a plateau (with fluctuations) starting at sample size 25 and then decreased beyond sample size 48. On the other hand, as more samples were selected at random or in ascending order of the optimal weight, sample representativeness continuously improved. At equal sample sizes, representativeness of the soil samples selected in descending order of the optimal weight was generally higher than representativeness of randomly selected samples, which in turn was higher than representativeness of samples selected in ascending order of the optimal weight. It suggests that the optimal weights determined from the heuristic are informative of individual soil sample's importance for improving sample representativeness.

Overall, at equal sample sizes, the SOM content mapping accuracies achieved based on soil samples selected in descending order of the optimal weight were consistently higher than accuracies achieved on soil samples selected purely at random, which in turn were

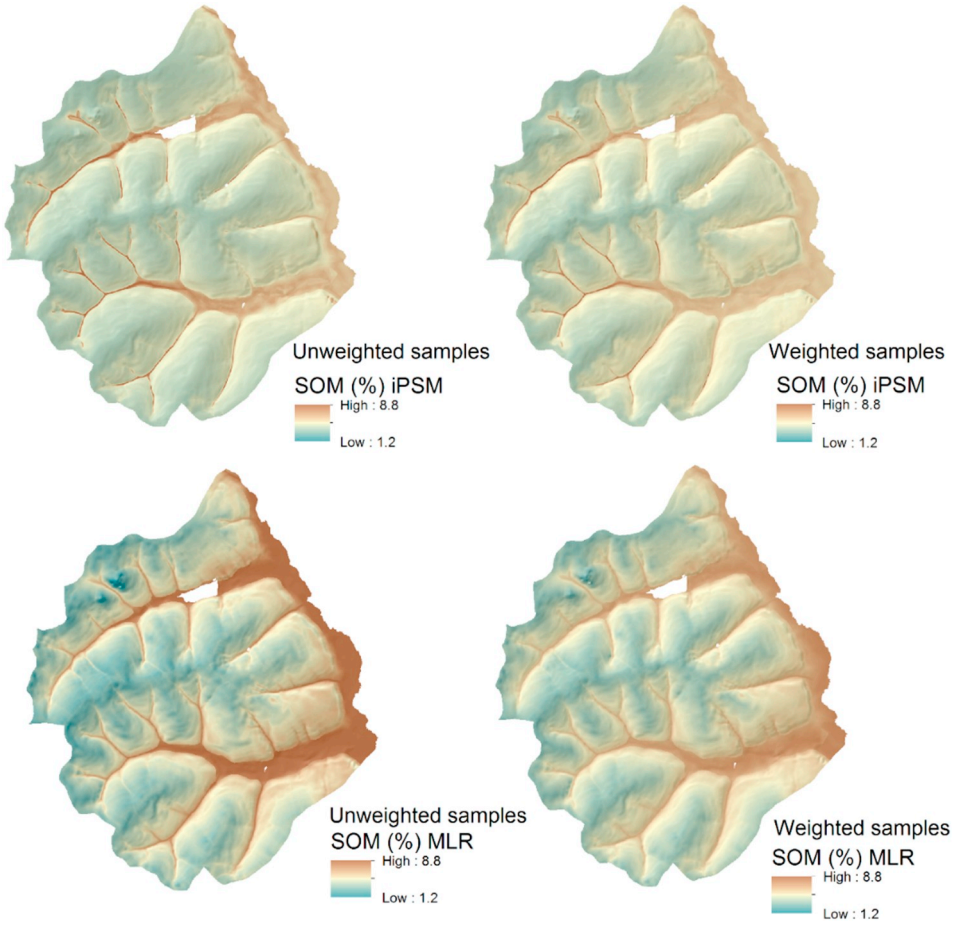

Fig. 8. SOM content maps predicted using the 59 samples.

Table 1
Accuracy of SOM maps predicted using unweighted soil samples and samples weighted by the optimal weights.

Method		Unweighted samples	Weighted samples
iPSM	RMSE	0.684	0.671
	MAE	0.537	0.539
	ME	0.208	0.084
	EVS	0.331	0.300
MLR	RMSE	0.841	0.723
	MAE	0.643	0.575
	ME	0.306	0.187
	EVS	0.033	0.230

considerably higher than accuracies achieved on samples selected at ascending order of the weight (Fig. 12). Prediction accuracies achieved under the three sample filtering strategies converged above certain sample size approaching the full sample size. These observations hold for both iPSM and MLR, again attesting that the optimal weights are informative for filtering samples.

When samples were selected in descending order of the optimal weight, performance of iPSM and MLR responded differently to sample size. SOM content mapping accuracies using iPSM generally improved with increasing sample size (except for ME). When 35 soil samples were selected and used in iPSM, the mapping accuracies (RMSE = 0.674, MAE = 0.543, ME = 0.110 and EVS = 0.302) were comparable to mapping accuracies achieved using all existing samples (RMSE = 0.684, MAE = 0.537, ME = 0.208 and EVS = 0.331; Table 1), or using all existing samples weighted by the optimal weights (RMSE = 0.671, MAE = 0.539, ME = 0.084 and EVS = 0.300; Table 1).

As more soil samples were selected in descending order of the optimal weight, SOM content mapping accuracies using MLR slightly improved until reaching the highest at sample size 26 (RMSE = 0.709, MAE = 0.557, ME = 0.048 and EVS = 0.211). Note that the accuracies

Table 2
One sample *t*-tests to compare the accuracies of SOM maps predicted based on soil samples weighted by the optimal weights and by random weights.

Method		Optimal weights	Random weights				Shuffled optimal weights			
			Mean	Std	t	p	Mean	Std	t	p
iPSM	RMSE	0.671	0.696	0.025	9.718	0.000	0.718	0.039	11.876	0.000
	MAE	0.539	0.548	0.023	4.104	0.000	0.568	0.034	8.680	0.000
	ME	0.084	0.212	0.070	18.333	0.000	0.214	0.112	11.544	0.000
MLR	RMSE	0.723	0.858	0.051	26.226	0.000	0.890	0.093	17.821	0.000
	MAE	0.575	0.656	0.043	18.759	0.000	0.682	0.071	14.901	0.000
	ME	0.187	0.310	0.057	21.338	0.000	0.315	0.094	13.569	0.000

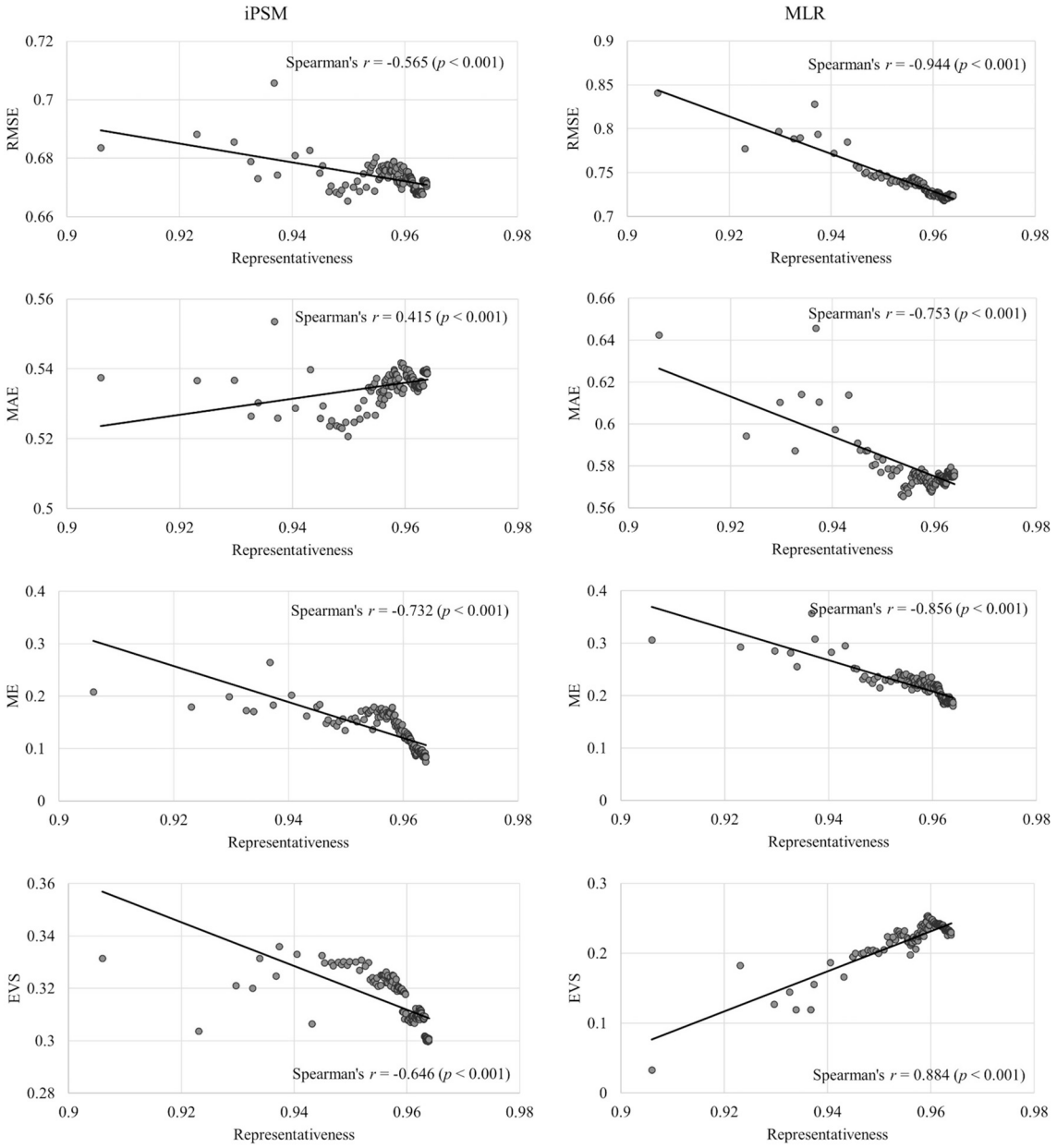

Fig. 9. Relationship between sample representativeness and prediction accuracy over the generations of the genetic algorithm.

were higher than the accuracies achieved using all existing samples (RMSE = 0.841, MAE = 0.643, ME = 0.306 and EVS = 0.033; Table 1) and were even slightly higher than accuracies achieved using all existing samples weighted by the optimal weights (RMSE = 0.723, MAE = 0.575, ME = 0.187 and EVS = 0.230; Table 1). The accuracies then gradually decreased slightly with increasing sample size. This is counter-intuitive at the first glance but could happen under certain scenarios. For instance, if the relationship between SOM content and the environmental covariates (principal components) were not stationary, MLR at best captures the "average" (linear) relationship over

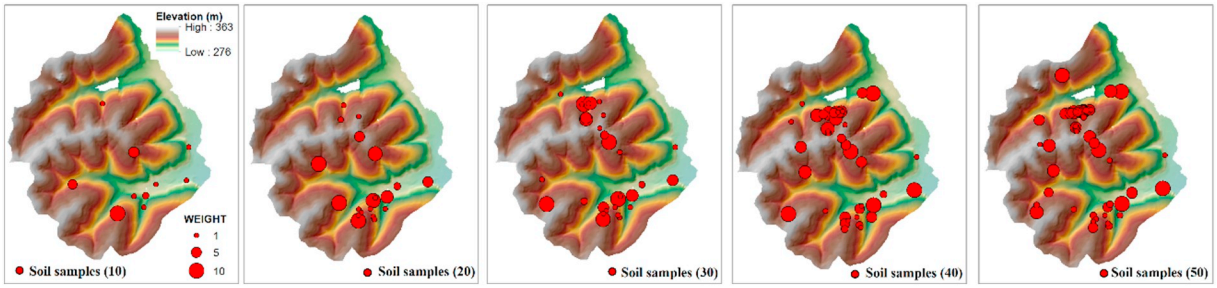

Fig. 10. Optimal weights of the subjective samples determined from the representativeness heuristic.

Table 3
Accuracies of SOM content maps predicted using unweighted (*U.w.*) or weighted (*W.*) soil samples of varying sample sizes.

Method	Sample size	RMSE			MAE			ME		EVS	
		U.w.	W.	Decrease	U.w.	W.	Decrease	U.w.	W.	U.w.	W.
iPSM	10	1.171	0.952	18.7%	0.846	0.724	14.4%	−0.604	−0.509	−0.587	−0.023
	20	0.931	0.705	24.2%	0.652	0.548	16.0%	−0.186	−0.007	−0.311	0.215
	30	0.779	0.726	6.8%	0.615	0.589	4.2%	0.129	0.198	0.070	0.231
	40	0.789	0.772	2.1%	0.619	0.595	3.9%	0.356	0.310	0.219	0.212
	50	0.689	0.669	2.9%	0.541	0.526	2.7%	0.216	0.139	0.324	0.325
MLR	10	1.917	1.066	44.4%	1.607	0.856	46.7%	−0.092	−0.644	−4.785	−0.139
	20	1.187	0.750	36.8%	0.958	0.596	37.9%	−0.548	−0.104	−0.749	0.129
	30	0.894	0.699	21.8%	0.709	0.539	23.9%	−0.138	0.190	−0.230	0.287
	40	0.886	0.844	4.7%	0.708	0.662	6.4%	0.426	0.379	0.048	0.102
	50	0.784	0.710	9.5%	0.606	0.568	6.3%	0.287	0.131	0.159	0.231

Table 4
Representativeness of the soil samples at varying sample sizes.

Sample size	Unweighted samples	Weighted samples
10	0.696	0.864
20	0.884	0.908
30	0.878	0.919
40	0.899	0.958
50	0.903	0.960

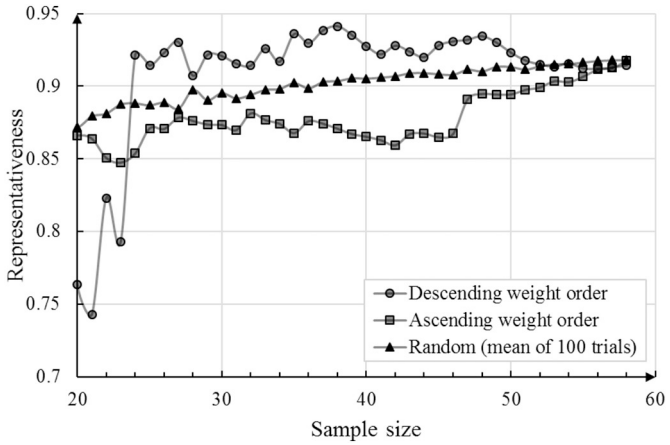

Fig. 11. Impact of the filtering strategies on the representativeness of filtered soil samples.

the study area given representative training soil samples. Adding more soil samples to the already-representative training samples does not necessarily result in better MLR models. For example, if the newly added samples were all from a corner of the study area, the trained MLR model would better fit the corner area but deviate from the "average" model that best fits the whole area.

4. Discussion

4.1. Effectiveness of the representativeness heuristic

Experiment results revealed that weighting soil samples by the optimal weights determined from the proposed representativeness heuristic improved DSM mapping accuracies (especially for the MLR mapping method). Accuracy improvements achieved by weighting soil samples by the optimal weights were less significant on soil samples of larger sample size. This was expected, given that everything else being equal, representativeness of soil samples of larger sample sizes could be good enough and thus there was less space for improvement. On the contrary, representativeness of soil samples of smaller sample sizes would be poorer and thus there is more space for the heuristic to improve sample representativeness to increase DSM accuracy. *Moreover*, statistical significance tests suggest that the accuracy improvements were not achieved by chance, implying that the weight configuration in the optimal weights was statistically meaningful. *Besides*, a clear positive relationship between representativeness of the soil samples and DSM accuracy was observed when using the MLR method for DSM (the relationship was mixed when using the iPSM method). Overall, it suggests sample representativeness as quantified in the heuristic was an effective indicator of DSM accuracy. *Additionally*, the optimal sample weights determined from the representative heuristic were informative for filtering soil samples to improve DSM accuracies (e.g., MLR) or achieve comparable accuracies at a smaller sample size (e.g., iPSM). Similarly, the optimal weights could also inform the selection of appropriate soil samples for validation of DSM models, which is often a problem when working with legacy soil data. *Overall*, the proposed representativeness heuristic offers a novel and effective approach to mitigating spatial bias in existing soil samples for DSM.

4.2. Parameter settings

The upper limit of sample weight W_{max} is a key parameter in the representativeness heuristic as it defines the value range of the sample weights [1.0, W_{max}]. The physical meaning of W_{max} is that a sample with weight W_{max} would be treated as W_{max} times more important than a sample with weight 1.0 in training DSM models. Based on this observation, W_{max} was subjectively set to 10.0 by default in this study (A sample can be at most 10 times more important than another sample). Experiments were run on the 59 existing soil samples with various W_{max} settings to examine the impact of W_{max} on performance of the heuristic. Results (Table 5) showed that applying the heuristic with the default setting (W_{max} = 10.0) achieved the largest accuracy improvements compared to other settings. In studies where data availability allows, W_{max} may be determined objectively through data-driven procedures such as cross-validation, beyond taking its physical meaning into account.

Other less-influential parameters are parameters for the GA. The default GA parameter settings recommended by the DEAP python package (Rainville et al., 2012) were used for this study where appropriate. Readers interested in fine-tuning the GA parameters are referred to relevant references (e.g., Lobo et al., 2007; Smit and Eiben, 2011). Here we offer recommendations for setting *population size* (number of candidate sample weight lists in the pool) and *number of generations* (number of iterations the GA needs to go through before returning the optimal sample weights). A population size that is large enough relative to the problem size (number of sample weights to optimize) is needed to find a good solution (optimal sample weights corresponding to high representativeness). A larger population size allows the GA to evaluate a larger variety of sample weights. Note that a larger population size also implies longer computing time, as the operations of GA are

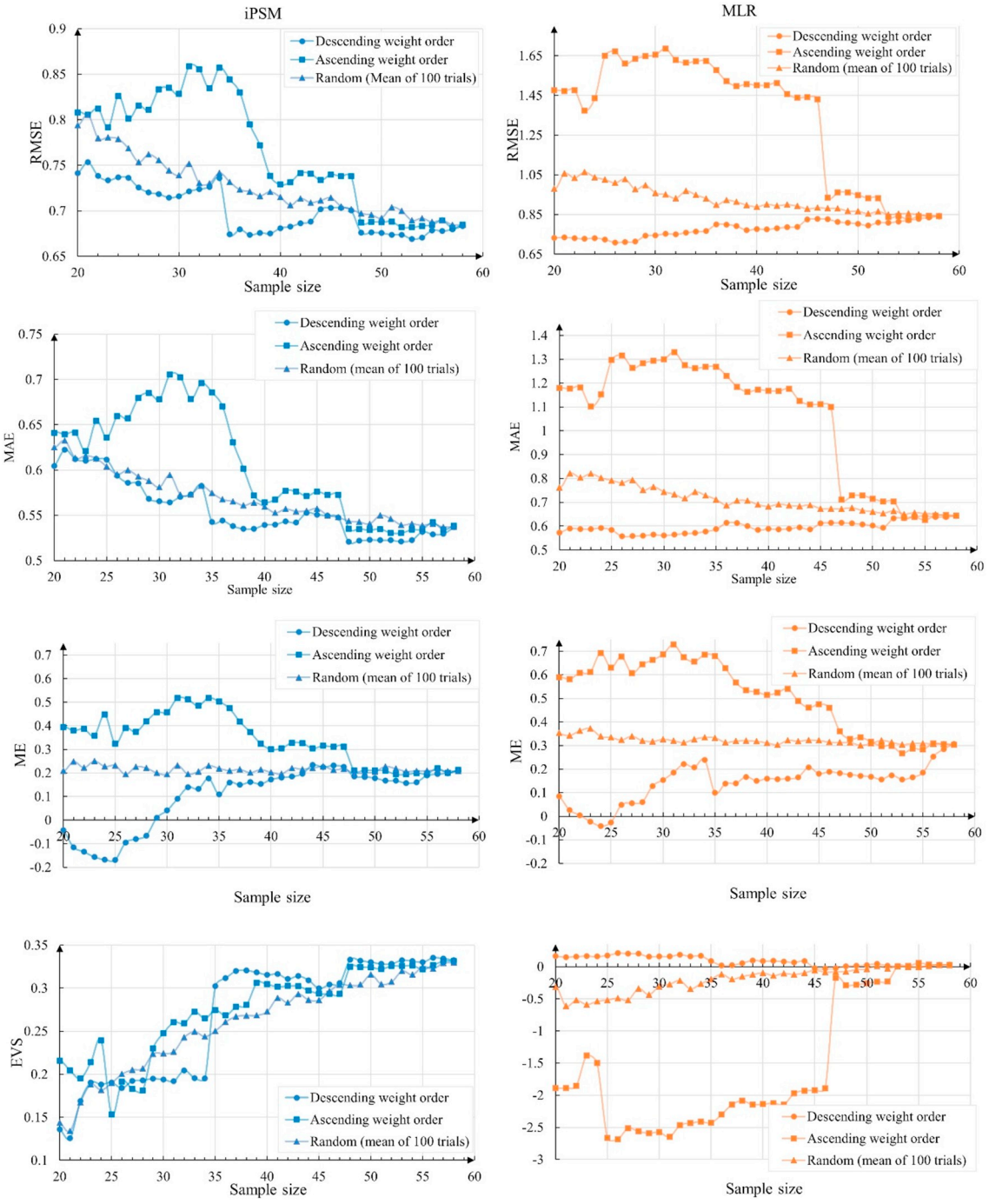

Fig. 12. Impact of the filtering strategies on prediction accuracy using filtered soil samples.

performed on a larger pool of candidate weights. In this study, the population size was set to 200 given there were only 59 sample weights to optimize. This setting as a reasonable balance between GA performance and computing time.

As a "rule-of-thumb", the *number of generations* can be set to the generation at which the representativeness no longer significantly improves. Alternatively, the GA can also terminate and return the optimal weights once the representativeness exceeds a prescribed threshold (e.g., 0.95). Also note that the computing time would be longer if the GA runs through a larger number of generations. In this study, the number of generations was set to 200 based on the observation that sample representativeness reached a plateau in fewer than 200 generations in the GA. When optimizing weights for the 59 soil samples, on average it took 86 s for the GA to complete each iteration (population size was 200).

4.3. Applicability

4.3.1. Mapping methods

The iPSM and MLR methods used for SOM content mapping in this study represent two distinct classes of methods for soil mapping or for spatial prediction in general (Zhu et al., 2018). The underlying premise of applying MLR (and many other regression-based methods; see Grunwald, 2009) is that there exists a statistical (linear) relationship between the target variable to predict (SOM content) and the environmental covariates and the relationship is stationary across the

Table 5
Impact of W_{max} settings on performance of the representativeness heuristic (The 59 existing soil samples were used in the experiments).

Method		Wmax				
		5	10	20	50	100
iPSM	RMSE	0.685	0.671	0.706	0.711	0.689
	MAE	0.542	0.539	0.546	0.552	0.539
	ME	0.151	0.084	0.160	0.188	0.155
	EVS	0.296	0.300	0.254	0.259	0.289
MLR	RMSE	0.754	0.723	0.785	0.773	0.815
	MAE	0.591	0.575	0.613	0.599	0.635
	ME	0.242	0.187	0.288	0.283	0.310
	EVS	0.195	0.230	0.158	0.185	0.104

study area. An MLR model (Section 2.3.2) is first fitted based on values of the target variable and the covariates at sample locations. The model representing the relationship is then applied to unsampled locations to predict values of the target variable based on the in-situ covariate values. In contrast, iPSM (Section 2.3.1) is based on the principle that the more similar geographic environment of two locations, the more similar the values of the target variable at these two locations (Zhu et al., 2015, 2018). For instance, soil scientists studied the formation of soils under certain geographic environment (configuration of climate, geology, topography, vegetation, and time) at some locations and then expected the similar soil formation processes to occur at other locations with similar environment (Jenny, 1994). iPSM does not assume an explicit stationary relationship to exist between the target variable and its environmental covariates and thus no models are trained prior to prediction. It directly predicts value of the target variable at an unsampled location based on environmental similarities between the unsampled location and the sample locations (the environmental similarities are used to weight the values of the target variable at sample locations to predicted value of the target variable at the unsampled location). Nonetheless, a similarity between iPSM and MLR is that both methods rely on covariates to predict the target variable (other methods such as the ordinary kriging rely on spatial distance for prediction).

Weighting soil samples with the optimal weights determined from the representativeness heuristic effectively improved SOM content mapping accuracies using both iPSM and MLR, although the accuracy improvements were more prominent for MLR. In essence, the weighting scheme down-weights samples that disproportionately over-represent the covariate space and up-weights samples that disproportionately under-represent the covariate space. It is thus reasonable to expect the heuristic to be applicable for other DSM methods involving covariates for modeling and prediction, including regression and classification methods (e.g., multivariate nonlinear regression, decision tree) and geostatistical methods with a regression component (e.g., co-kriging) (Grunwald, 2009).

The above-mentioned methods use a “global” model to capture the soil-environment relationship which is assumed to stay the same over the study area. It is also possible to apply the heuristic to “local” modeling methods that account for spatially-varying soil-environment relationships, such as the geographically weighted regression (GWR) (Fotheringham et al., 2003; Zeng et al., 2016). The GWR fits a “local” regression model using samples within the vicinity of a prediction location. The representativeness heuristic could be applied to increase representativeness of the samples in the local area and to improve GWR prediction accuracy.

4.3.2. Sample size

This study used a small data set in a small study area to demonstrate the proposed representativeness heuristic. A total of 59 soil samples were used for SOM content mapping over the 60 km^2 study area (density = 1 sample per km^2). Nonetheless, the heuristic is expected to be applicable for mapping over large areas using a large number of soil samples provided that the samples are subject to spatial bias. A larger number of samples do not always imply higher sample representativeness. For example, if additional samples were all from a small part of the study area, the representativeness of the samples would not increase much. Thus, sample size is not the sole factor that determines sample representativeness (and thus the effectiveness of the heuristic). Another important factor is the spatial distribution pattern of the samples. In the case study, it was observed that SOM content mapping accuracy improvements achieved by weighting the soil samples with the optimal weights were less significant on soil samples of larger sample sizes (Sections 3.2). This was because the representativeness of soil samples of larger sample sizes was already relatively high and thus there was less space for improvement. In general, it is reasonable to hypothesize that the heuristic is more effective on samples with more severe spatial bias.

4.3.3. Computational considerations

The most computationally demanding part of the representativeness heuristic was sample representativeness evaluation (Section 2.2.1) in the GA for determining the optimal sample weights (Section 2.2.2). Sample representativeness was evaluated on each candidate sample weight lists in the pool at each iteration of the GA. For each evaluation, the KDE method estimated the sample distributions (needed for computing representativeness) with a bandwidth parameter determined from the “golden section search” (Brunsdon, 1995), which itself was an iterative and computationally intensive procedure. Initially, a sequential version of the heuristic (running on a single CPU-thread) was implemented using Python 2.7. Using the GA to optimize weights for the 59 soil samples took 4 h and 47 min (each iteration of the GA took 86 s on average) on a Dell Precision 5820 workstation (8-core Intel Xeon CPU 3.7 GHz; 64 GB memory; Windows 10 operating system).

To speed up the computation, computing resources on multi-core CPUs and many-core GPUs (graphics processing units) were exploited to run the computationally critical steps (estimating sample distributions with KDE) in parallel (G. Zhang et al. 2016; Zhang et al., 2017). The parallel version of the heuristic was implemented using the PyOpenCL python programming package (Klöckner et al., 2012). Running the parallel version of the heuristic to optimize weights for the 59 samples on the 8 CPU cores (16 threads) took about 40 min (each GA iteration took about 12 s). It took about 29 min (each GA iteration took about 8 s) to run on an NVDIA Quadro P4000 GPU (5.3 TFLOPS; 8 GB memory). With the support of high-performance computing technologies and resources, the proposed heuristic can be applied to DSM studies involving many soil samples.

5. Conclusions

This paper presents a representativeness heuristic for mitigating spatial bias in existing soil samples for DSM. Experiment results of mapping *A*-horizon SOM content using existing soil samples in the Heshan study area showed that weighting existing soil samples by the optimal sample weights determined from the representativeness heuristic effectively improved DSM accuracy. Moreover, it was observed that the quantified representativeness of soil samples was an effective indicator of DSM accuracy. In addition, the optimal sample weights were informative for identifying representative soil samples and thus can be used as guidance to filter samples for improving DSM accuracy. Overall, the proposed representativeness heuristic can effectively mitigate spatial bias in existing soil samples to improve DSM accuracy using such soil samples.

Soil samples coming from multiple sources are very likely to suffer from spatial bias. The proposed representativeness heuristic is an important contribution to the DSM community as it offers a novel approach to tackling the spatial bias issue in soil samples for DSM. Other than the two DSM methods tested in this study, the heuristic could potentially be applicable to a wide range of “global” regression and

classification approaches for DSM (Grunwald, 2009). It can also be integrated into "local" modeling methods such as the GWR (Fotheringham et al., 2003) to mitigate spatial bias in the soil samples for DSM over large areas.

Acknowledgements

Supports to Guiming Zhang through the Faculty Start-up Funds at the University of Denver and the Whitbeck Graduate Dissertator Award from the Department of Geography, University of Wisconsin-Madison is acknowledged. The work reported here was supported by grants from National Natural Science Foundation of China (Project No.: 41431177, 41871300), National Basic Research Program of China (Project No.: 2015CB954102), PAPD, and Outstanding Innovation Team in Colleges and Universities in Jiangsu Province. Supports to A-Xing Zhu through the Vilas Associate Award, the Hammel Faculty Fellow Award, and the Manasse Chair Professorship from the University of Wisconsin-Madison are greatly appreciated.

Declaration of Competing Interest

The authors have no competing interests to declare.

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.
Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. Adv. Agron. 125, 93–134.
Boria, R.a., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275, 73–77.
Brunsdon, C., 1995. Estimating probability surfaces for geographical point data: an adaptive kernel algorithm. Comput. Geosci. 21, 877–894.
Brus, D.J., Yang, L., Zhu, A.X., 2019. Accounting for differences in costs among sampling locations in optimal stratification. Eur. J. Soil Sci. 70, 200–212.
Carré, F., McBratney, A.B., Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma 141, 1–14.
Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? -arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250.
De Gruijter, J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer Science & Business Media.
Dokuchayev, V.V., 1883. Russkiy Chernozem (the Russian Chernozem). St. Petersburg.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2003. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons (Limited).
Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207.
Jenny, H., 1994. Factors of soil formation: A system of quantitative pedology. Courier Corporation, North Chelmsford, MA.
Jensen, R.R., Shumway, J.M., 2010. Sampling our world, in: Gomez, B., Jones III, J.P. (Eds.), Research Methods in Geography: A Critical Introduction. John Wiley & Sons, pp. 77–90.
Jolliffe, I.T., 2002. Principal component analysis and factor analysis. Princ. Compon. Anal. 150–166.
Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol. Appl. 14, 401–413.
Klöckner, A., Pinto, N., Lee, Y., Catanzaro, B., Ivanov, P., Fasih, A., 2012. PyCUDA and PyOpenCL: a scripting-based approach to GPU run-time code generation. Parallel Comput. 38, 157–174.
Kruskal, W., Mosteller, F., 1979. Representative sampling, III: the current statistical literature. Int. Stat. Rev. 47, 245–265.
Liu, J., 2017. Integration of Samples from Multiple Sources for Predictive Mapping over Large Areas. University of Wisconsin-Madison.
Lobo, F.J., Lima, C.F., Michalewicz, Z., 2007. Parameter Setting in Evolutionary Algorithms. Springer Science & Business Media.
McBratney, A., Mendonça Santos, M., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388.
Minasny, B., McBratney, A.B., 2016. Digital soil mapping: a brief history and some lessons. Geoderma 264, 301–311.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2012. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Pei, T., Qin, C.-Z., Zhu, A.-X., Yang, L., Luo, M., Li, B., Zhou, C., 2010. Mapping soil organic matter using the topographic wetness index: a comparative study based on different flow-direction algorithms and kriging methods. Ecol. Indic. 10, 610–619.
Qi, F., Zhu, A.-X., 2003. Knowledge discovery from soil maps using inductive learning. Int. J. Geogr. Inf. Sci. 17, 771–795.
Qin, C., 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. Int. J. Geogr. Inf. Sci. 21, 443–458.
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, 152–161.
Qin, C.Z., Zhu, A.X., Qiu, W.L., Lu, Y.J., Li, B.L., Pei, T., 2012. Mapping soil organic matter in small low-relief catchments using fuzzy slope position information. Geoderma 171–172, 64–74.
Rainville, D., Fortin, F.-A., Gardner, M.-A., Parizeau, M., Gagné, C., 2012. DEAP: a python framework for evolutionary algorithms. In: Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation. ACM, pp. 85–92.
Rossiter, D.G., Liu, J., Carlisle, S., Zhu, A.-X., 2015. Can citizen science assist digital soil mapping? Geoderma 259–260, 71–80.
Scull, P., Franklin, J., Chadwick, O.a., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27, 171–197.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, UK.
Singh, V.P., Woolhiser, D.A., 2002. Mathematical modeling of watershed hydrology. J. Hydrol. Eng. 7, 270–292.
Smit, S.K., Eiben, A.E., 2011. Parameter tuning for configuring and analyzing evolutionary algorithms. Swarm Evol. Comput. https://doi.org/10.1016/j.swevo.2011.02.001.
Varela, S., Anderson, R.P., García-Valdés, R., Fernández-González, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography (Cop.) 37, 1084–1091.
Vaysse, K., Lagacherie, P., 2015. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Reg 4, 20–30.
Yang, L., Zhu, A., Qi, F., Qin, C., Li, B., 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, 1–23.
Zeng, C., Yang, L., Zhu, A.-X., Rossiter, D.G., Liu, J., Liu, J., Qin, C., Wang, D., 2016. Mapping soil organic matter concentration at different scales using a mixed geographically weighted regression method. Geoderma 281, 69–82.
Zhang, G., 2018. A Representativeness Directed Approach to Spatial Bias Mitigation in VGI for Predictive Mapping. University of Wisconsin-Madison.
Zhang, G., Huang, Q., Zhu, A.-X., Keel, J., 2016. Enabling point pattern analysis on spatial big data using cloud computing: optimizing and accelerating Ripley's K function. Int. J. Geogr. Inf. Sci. 30, 2230–2252.
Zhang, S., 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.
Zhang, G., Zhu, A.-X., Huang, Q., 2017. A GPU-accelerated adaptive kernel density estimation approach for efficient point pattern analysis on spatial big data. Int. J. Geogr. Inf. Sci. 31, 2068–2097.
Zhang, G., Zhu, A.-X., Huang, Z.-P., Ren, G., Qin, C.-Z., Xiao, W., 2018. Validity of historical volunteered geographic information: evaluating citizen data for mapping historical geographic phenomena. Trans. GIS 22, 149–164.
Zhu, A.-X., 1999. A personal construct-based knowledge acquisition process for natural resource mapping. Int. J. Geogr. Inf. Sci. 13, 119–141.
Zhu, A.X., Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. Can. J. Remote. Sens. 20, 208–218.
Zhu, A.X., Mackay, D.S., 2001. Effects of spatial detail of soil information on watershed modeling. J. Hydrol. 248, 54–77.
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, 164–174.
Zhu, A.-X., Liu, J., Du, F., Zhang, S., Qin, C.-Z., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. Eur. J. Soil Sci. 66, 535–547.
Zhu, A., Lu, G., Liu, J., Qin, C., Zhou, C., 2018. Spatial prediction based on third law of geography. Ann. GIS 24, 225–240.