European Journal of Soil Science

European Journal of Soil Science, May 2015, **66**, 535–547 doi: 10.1111/ejss.12244

Predictive soil mapping with limited sample data

A. X. Zhu[a,b,c,d], J. Liu[d], F. Du[d], S. J. Zhang[c], C. Z. Qin[c], J. Burt[d], T. Behrens[e] & T. Scholten[e]

[a]*School of Geography Science, Nanjing Normal University, 1 Wenyuan Road, Nanjing 210023, China,* [b]*Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, 1 Wenyuan Road, Nanjing 210023, China,* [c]*State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A Datun Road, Beijing 100101, China,* [d]*Department of Geography, University of Wisconsin - Madison, 550 North Park Street, Madison, Wisconsin 53706, USA, and* [e]*Institute of Geography, Physical Geography, University of Tübingen, Rümelinstraße 19-23, D-72074 Tübingen, Germany*

Summary

Existing predictive soil mapping (PSM) methods often require soil sample data to be sufficient to represent soil–environment relationships throughout the study area. However, in many parts of the world with only a limited quantity of soil sample data to represent the study area, this is still an issue for PSM application. This paper presents a method, named 'individual predictive soil mapping' (iPSM), which can make use of limited soil sample data for PSM. With the assumption that similar environmental conditions have similar soils, iPSM uses the soil–environment relationship at each individual soil sample location to predict soil properties at unvisited locations and estimate prediction uncertainty. Specifically, the environmental similarities of an unvisited location to a set of soil sample locations are used in a weighted average method to integrate the soil–environment relationships at sample locations for prediction and uncertainty estimation. As a case study, iPSM was applied to map soil organic matter (SOM) content (%) in the topsoil layer using two sets of soil samples. Compared with multiple linear regression (MLR), iPSM produced a more accurate SOM map (root mean squared error (RMSE) 1.43, mean absolute error (MAE) 1.16) than MLR (RMSE 8.54, MAE 7.34) the ability of the sample set to represent the study area is limited and achieved a comparable accuracy (RMSE 1.10, MAE 0.69) with MLR (RMSE 1.01, MAE 0.73) when the sample set could represent the study area better. In addition, the prediction uncertainty estimated by iPSM was positively related to prediction residuals in both scenarios. This study demonstrates that iPSM is an effective alternative when existing soil samples are limited in their ability to represent the study area and the prediction uncertainty in iPSM can be used as an indicator of its prediction accuracy.

Introduction

Soil sample data are often used for calibrating or constructing a predictive soil mapping (PSM) model (McBratney *et al*., 2003; Scull *et al*., 2003; Zhu *et al*., 2010). Generally, it is required that the sample data should be sufficient to represent the soil–environment relationships throughout a study area for reliable model calibration or construction (Ramsey & Hewitt, 2005). From probability theory and geostatistical analysis, sufficient soil sample data usually need to be collected through a well-defined field sampling design, which typically aims to allocate soil samples in geographical space and/or environmental covariate space (feature space) in such a way that the designed sample set is representative of the area (Webster & Oliver, 1990; van Groenigen & Stein, 1998; de Gruijter *et al*., 2006; Minasny & McBratney, 2006; Zhu *et al*., 2008). The fuzzy logic-based techniques (Zhu, 1999, 2001; Qi & Zhu, 2003) also require that the knowledge or the soil sample set needs to be representative of the area for reliable construction of the fuzzy membership function.

In practice, however, it can be difficult to collect a set of soil samples that is sufficiently representative of the study area. Because of the costs of field sampling and laboratory analysis, the number of soil sample points actually collected can be limited. In addition, positioning errors can result in the mismatch between the actual and the design locations, and the limited expertise of soil surveyors can also cause bias in the collected samples (Johnson *et al.,* 2012). Those constraints can easily lead to the collected soil sample set not being representative of the soil-environment relationships throughout the study area.

Correspondence: J. Liu. E-mail: jliu93@wisc.edu

Received 12 June 2014; revised version accepted 20 January 2015

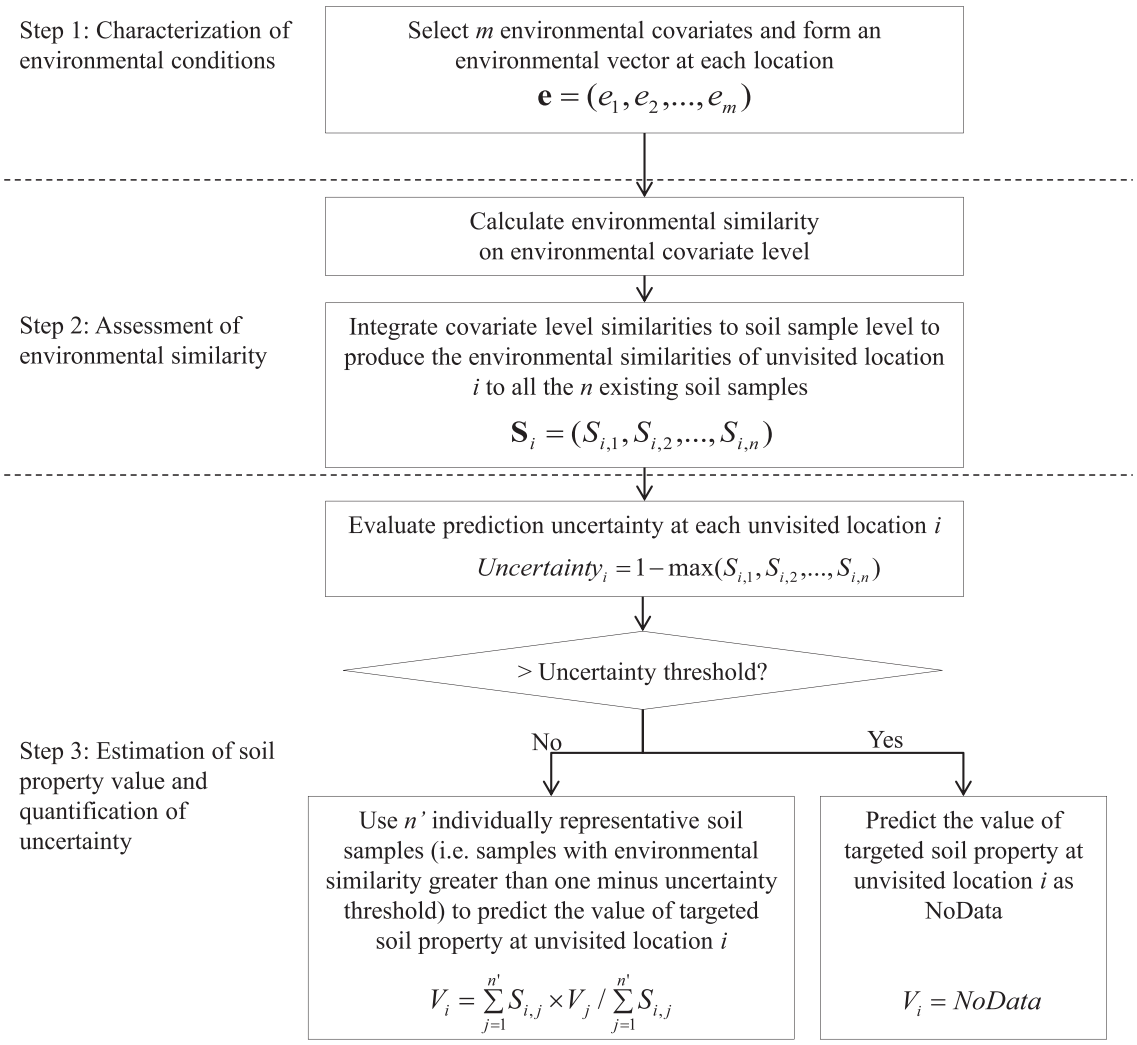

Figure 1 The flowchart of the individual predictive soil mapping (iPSM) method.

The problem of limited soil sample data, which are not adequate to represent soil–environment relationships in a study area, presents challenges for existing PSM methods. One problem is that is the data are not sufficient for developing a reliable soil–environment model with existing PSM methods (such as regression and kriging). As a consequence, the estimation of prediction uncertainty using existing methods, such as the error variance in kriging method, may not be valid.

This paper presents a PSM method, called 'individual predictive soil mapping' (iPSM), which is capable of using limited soil sample data for predictive soil mapping and quantifying prediction uncertainty. The next section of this paper presents the methodologies of iPSM in detail. Then we apply iPSM in a case study for mapping the content of soil organic matter (SOM, %) in the topsoil layer and quantifying prediction uncertainty and compare this with multiple linear regression (MLR). We use two soil sample datasets: one has limitations in representing the study area and the other can represent it better.

Methods

Basic idea and overall design

Soil samples, however limited their ability to represent soil–environment relationships in a study area, still contain valuable information for PSM. For example, each individual soil sample point reflects the *in situ* relationship of soil to the environmental 'niche' that it is in. By assuming that locations with similar environmental conditions have similar soil properties (Hudson, 1992; Mallavan *et al*., 2010), similar values of the targeted soil property can be expected at locations with similar environmental conditions. Therefore, each soil sample point can individually serve as a surrogate representing the locations with similar environmental conditions, and thus can be used to predict the soil conditions for these locations.

The overall design of iPSM contains three major steps (Figure 1): (i) the characterization of environmental conditions at each location, (ii) the assessment of environmental similarities at every unvisited

location and every soil sample location and (iii) the estimation of the value of the targeted soil property and the quantification of prediction uncertainty at each location.

Characterization of environmental conditions

Environmental conditions associated with soil at each location need to be quantitatively characterized. It is important to use the environmental variables (the environmental covariates) that are effective in indicating the spatial variation of a targeted soil property in a given study area. According to the state equation of soil formation (Jenny, 1941), there are five environmental state factors (climate, parent material, topography, biology and time) that need to be considered. Each environmental state factor can be characterized by using various environmental variables (or covariates) (McBratney *et al.,* 2003). The selection of an appropriate environmental covariate depends on its (i) ability to indicate the spatial variation of a targeted soil property and (ii) availability regarding data sources. Commonly used environmental covariates include annual averaged temperature, annual averaged precipitation, parent materials, elevation, slope aspect, slope gradient, profile and planform curvatures, topographic wetness index, distance to streams, land use and cover and vegetation index (McBratney *et al*., 2003; Florinsky, 2011). The advances in earth observation technologies and spatial analysis methods have also provided many new candidate environmental covariates (Mendonça-Santos *et al*., 2006; Kienast-Brown & Boettinger, 2007; Nield *et al*., 2007; Zhu *et al*., 2010; Liu *et al*., 2012). Usually, multiple environmental covariates are needed to characterize the environmental conditions related to the targeted soil properties. At a given location an environmental vector is formed to characterize the environmental conditions:

$$\mathbf{e} = \left(e_1, e_2, \ldots, e_m\right), \tag{1}$$

where m is the number of environmental covariates used.

Assessment of environmental similarity

From the environmental vector, the environmental similarity between an unvisited location i ($i = 1, 2, 3, \ldots, k$; k is the number of locations to be predicted, typically grid cells) and a soil sample point j ($j = 1,2,3, \ldots, n$; n is the number of available soil sample points) needs to be determined at the scale(s) of the environmental covariate and soil sample (Shi *et al.,* 2004):

$$S_{i,j} = P\left(E\left(e_{1i}, e_{1j}\right), E\left(e_{2i}, e_{2j}\right), \ldots, E\left(e_{vi}, e_{vj}\right), \ldots, E\left(e_{mi}, e_{mj}\right)\right), \tag{2}$$

where $S_{i,j}$ represents the environmental similarity between unvisited location i and sample location j, e_{vi} and e_{vj} are the values of the *vth* environmental covariate at the two locations, and $\boldsymbol{E}(\bullet)$ and $\boldsymbol{P}(\bullet)$ are functions for calculating environmental similarity at the environmental covariate scale and soil sample scale, which are explained next.

Function $\boldsymbol{E}(\bullet)$ is for calculating environmental similarity at the environmental covariate scale. Given an environmental covariate, $\boldsymbol{E}(\bullet)$ describes how the environmental similarity between an unvisited location and a sample location changes when the value of the environmental covariate at the unvisited location deviates from the value at the sample location. For iPSM, we define $\boldsymbol{E}(\bullet)$ depending on the measurement scale (Stevens, 1946) of the environmental covariate:

$$E\left(e_{vi}, e_{vj}\right) = \begin{cases} 0 & e_v \text{ is nominal/ordinal and } e_{vi} \neq e_{vj} \\ 1 & e_v \text{ is nominal/ordinal and } e_{vi} = e_{vj} \\ \exp\left(-\dfrac{\left(e_{vi}-e_{vj}\right)^2}{2\left(SD_{e_v} \times \frac{SD_{e_v}}{SD_{e_{vj}}}\right)^2}\right) & v \text{ is interval/ratio} \end{cases}, \tag{3}$$

in which SD_{e_v} is the standard deviation of the *vth* environmental covariate in the study area, and $SD_{e_{vj}}$ is the square root of the mean deviation of the values of the *vth* environmental covariate at all unvisited locations ($i = 1, 2, \ldots, k$) from that at sample location j (i.e. e_{vj}):

$$SD_{e_{vj}} = \sqrt{\frac{\sum_{i=1}^{k}\left(e_{vi} - e_{vj}\right)^2}{k}}. \tag{4}$$

According to Equation (3), if the covariate has been recorded on a nominal or ordinal scale, such as parent material type, a stepped $\boldsymbol{E}(\bullet)$ will be used (Figure 2a). This $\boldsymbol{E}(\bullet)$ produces a value of 1 when the value of the covariate at an unvisited location is the same as that at a sample location and 0 otherwise. For environmental covariates measure on interval or ratio scales such as precipitation or elevation a Gaussian-shaped $\boldsymbol{E}(\bullet)$ is designed to determine the similarity (Figure 2b). The width of the Gaussian-shaped curve, which is controlled by the standard deviation of the vth environmental covariate (SD_{e_v}), is further adjusted by $SD_{e_{vj}}$ (Equation (4)). This adjustment leads to a Gaussian-shaped curve that is spread over the range of the vth environmental covariate when the value at the soil sample location j is similar to many other prediction locations (therefore, smaller $SD_{e_{vj}}$) and *vice versa*. The aim of adjusting the width of the curve is to differentiate soil sample points that represent a common soil-environment relationship in the study area from soil sample points that represent a rare soil-environment relationship in the study area.

Function $\boldsymbol{P}(\bullet)$ is used to integrate environmental similarities for covariates with those for soil samples. The form of $\boldsymbol{P}(\bullet)$ should take into consideration the relative importance of multiple types of environmental covariates in influencing the value of the targeted soil property. The choices of $\boldsymbol{P}(\bullet)$ include weighted average methods, which require the relative weights for every environmental covariate (McBratney *et al*., 2003), decision trees, which allow a hierarchy configuration of the relative importance (Mallavan *et al*., 2010), and

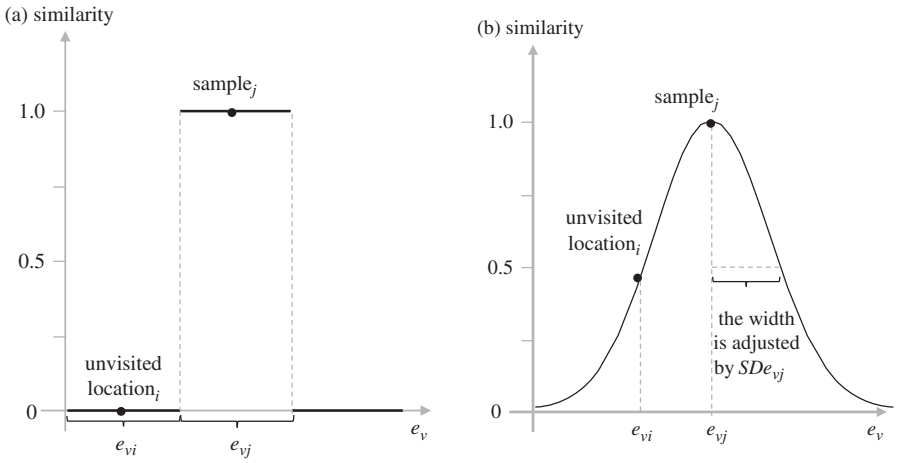

Figure 2 The E function used for estimating environmental similarity at the environmental covariate scale: (a) when the environmental covariate is nominal or ordinal; (b) when the environmental covariate is interval or ratio.

a minimum operator based on the Liebig's law of the minimum (van der Ploeg *et al*., 1999). In our paper, a minimum operator was used: the smallest value among all environmental covariate similarities was used as the environmental similarity at the soil sample level (Zhu *et al*., 1997; Shi *et al*., 2004).

The outputs from the assessment of environmental similarity to all existing soil sample points were organized into a similarity vector at each unvisited location i:

$$\mathbf{S}_i = \left(S_{i,1}, S_{i,2}, \ldots, S_{i,n}\right), \tag{5}$$

in which $\mathbf{S}_i$is the vector containing environmental similarities between unvisited location i and the n soil samples.

Quantification of prediction uncertainty and estimation of soil property value

Prediction uncertainty at each location is inversely related to its environmental similarities to the existing soil samples. The rationale for this is that if the similarities in environmental conditions between the unvisited location and the sample points used are poor, then the existing sample points cannot be used to represent that location. Thus, the use of these samples to predict the soil conditions at that location would lead to much uncertainty. Therefore, the uncertainty measurement in iPSM is basically a measurement of how reliable it is to use existing soil sample points to represent a given unvisited location for prediction. It is given as the following equation:

$$Uncertainty_i = 1 - \max\left(S_{i,1}, S_{i,2}, \ldots, S_{i,n}\right). \tag{6}$$

Equation (6) indicates that prediction uncertainty is expected to be large at unvisited location i when none of the n soil sample points is environmentally similar to this location.

It is necessary to examine whether an unvisited location can be predicted by existing soil samples with an acceptable uncertainty. A user-defined uncertainty threshold is set for this purpose. For a location whose prediction uncertainty is greater than this uncertainty threshold, the value of a targeted soil property at the location is set to '*NoData*' to indicate that prediction at this location using existing soil samples cannot be made with an acceptable range of uncertainty. The choice of uncertainty threshold depends on the user's tolerance regarding the reliability of using existing soil samples to represent the unvisited location. A strict uncertainty threshold such as a small value means that the user is required to use the existing soil samples to predict only the unvisited locations with very similar environmental conditions.

For a location whose prediction uncertainty does not exceed the threshold, the value of the target property at this location is predicted from the soil samples whose environmental similarity exceeds the value of one less the uncertainty threshold. In other words, only those soil samples that are similar enough to the unvisited location are used to compute the value of the targeted property at that site. The estimation of the value of the targeted soil property is calculated with a weighted average method (Zhu *et al.,* 1997):

$$V_i = \frac{\sum_{j=1}^{n'} S_{i,j} \times V_j}{\sum_{j=1}^{n'} S_{i,j}}, \tag{7}$$

in which n' is the number of the selected soil samples whose environmental similarity to the unvisited location i exceeds one minus the uncertainty threshold, $S_{i,j}$ is the environmental similarity of the unvisited location i to the soil sample location j, and V_j is the value of the targeted soil property of soil sample at location j.

Case study

Study area and datasets

To test the effectiveness, iPSM was used for predicting soil organic matter (SOM) content (%) in the topsoil layer of an area of approximately 60 km^2. The study area is located in Heshan Farm, Nenjiang County, Heilongjiang Province of China (Figure 3). The annual temperature ranges from −38 to 36°C, the ≥10°C accumulated temperature over one year is about 2000–2300°C, and the average annual precipitation is 500–600 mm. Most soils

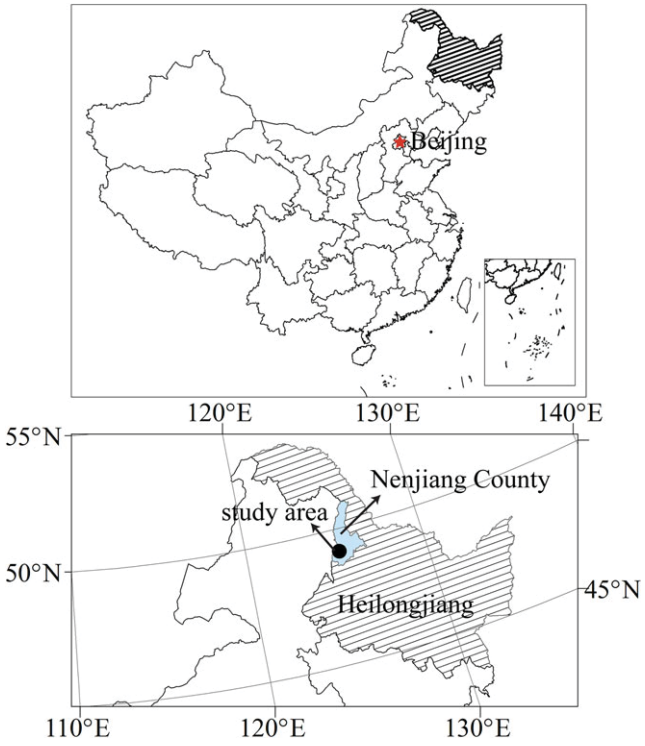

Figure 3 The location of the study area.

in the area were formed on deposits of silt loam loess, with the exception of the valley, where the underlying parent material is fluvial deposits. Elevation within the area ranges from 270 to 360 m and the slope gradient is generally less than 4°. The study area has been cultivated as cropland for over 40 years, with soybean and wheat as the main agricultural products. Because there is a thick top-layer of soil with a naturally large organic matter content it requires little organic fertilizer to maintain productivity.

Given the characteristics of the environmental conditions and the soil-environment relationships in this area (Yang *et al*., 2007; Zhu *et al*., 2008), topographic and vegetation covariates are the most effective indicators of the spatial variation of the SOM content (%) in the topsoil layer. Six topographic covariates (elevation (m), slope gradient (%), contour curvature, profile curvature, relative slope position and topographic wetness index (TWI)) and Normalized Difference Vegetation Index (NDVI) were used to characterize the environmental conditions. A 10-m resolution digital elevation model (DEM) was created from the 1:10 000 topographic map of the area using the TOPOGRID and TINLATTICE in Arc/Info (Yang *et al*., 2007). Slope gradient, contour curvature and profile curvature were derived from this DEM with 3DMapper (www.terrainanalytics.com). The TWI was calculated with the method of Beven & Kirkby (1979). Because of the gentle terrain relief and the wide floodplain in this area, a multiple-flow strategy (Qin *et al*., 2012) was used to estimate the upslope drainage area in the computation of TWI. The relative position index was calculated with the algorithm proposed by Qin *et al.* (2009) and NDVI was calculated with a Landsat ETM+ image of the area obtained on 25 September 2000, which was downloaded from the website of the Global Land Cover Facility (GLCF) served by the University of Maryland (http://glcfapp.glcf.umd.edu:8080/esdi/).

To examine the effectiveness of iPSM when the existing soil sample set is limited in its ability to represent soil–environment relationships throughout the study area (referred to as limited sample scenario hereafter), 10 soil samples collected along a transect line across two slopes at the side of the valleys were used (Figure 4). The locations of those samples were subjectively determined by soil surveyors to reveal the variation of soil properties on slopes only (Zhu *et al*., 2008): ridge areas and wide valley areas were not sampled. This provides a good case study to evaluate the effectiveness of iPSM when the available soil sample data are limited and existing PSM methods may not be applicable.

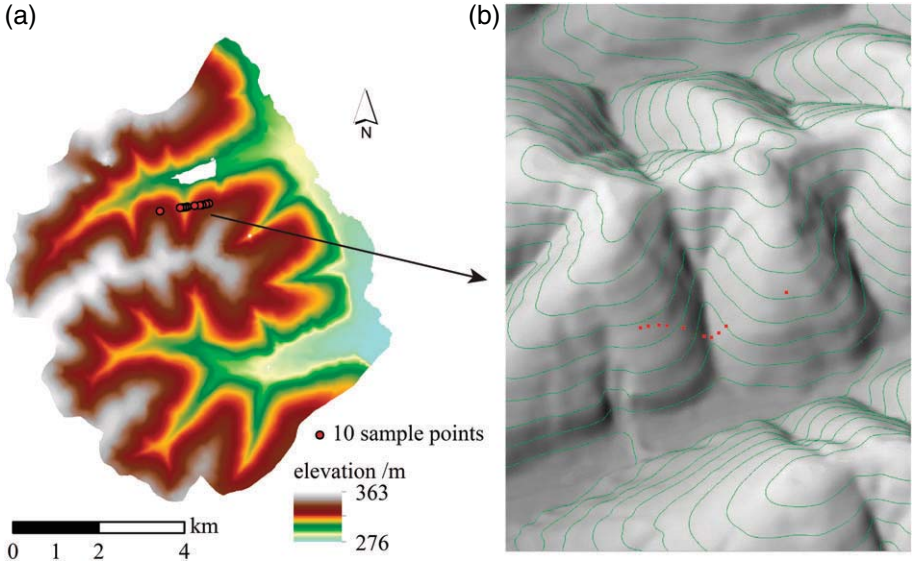

Figure 4 The location of 10 soil samples existing in the study area: (a) a planar view; (b) a 3D view revealing the distribution of the points along the transect line.

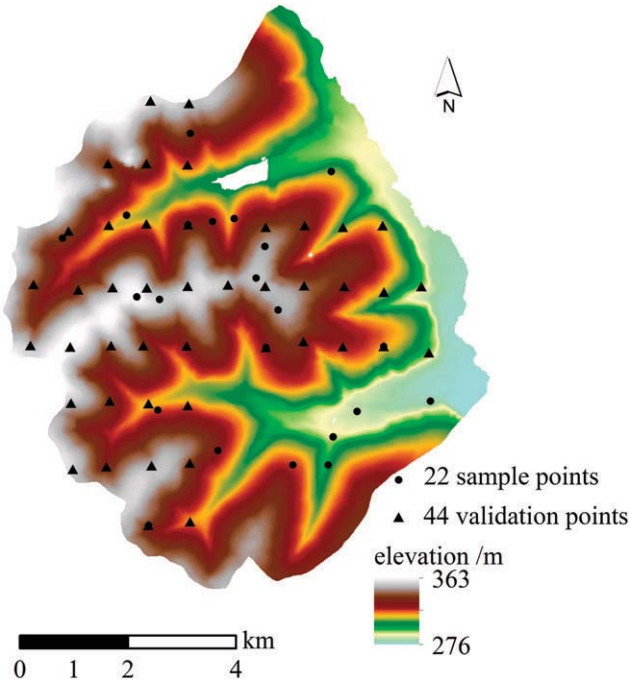

Figure 5 The spatial distribution of the 22 sample points that can represent the study area better, and the spatial distribution of the 44 validation points.

To test whether improving the ability of soil samples to represent the study area can improve the performance of iPSM, and to compare it with existing PSM methods, we applied it to another situation (referred to as the improved sample scenario), which contains 22 sample points collected by using the sampling method proposed by Zhu *et al.* (see Zhu *et al*., 2008, for full details). In brief, the sampling method was designed to sample the locations where the soils are typical of the soil categories in study area, and therefore can better represent the soil–environment relationship throughout the study area. The spatial distribution of the 22 soil sample points is shown in Figure 5.

The validation sample set consisting of 44 sample points was independently collected on a $1100 \times 740\,m^2$ grid (Figure 5). The statistics on the environmental conditions at the 10 soil sample points, the 22 sample points and the 44 validation points are shown in Table 1. Figure 6 uses elevation as an example to show the coverage of the three sample sets in the environmental feature space. It can be observed that the coverage of the environmental feature space by the 10 sample points deviated from that of the study area, while the coverage of the environmental feature space by the 22 sample points and the 44 validation sample points generally matched that of the study area.

Validation using the independent validation sample set

Two indices, the root mean squared error (RMSE) and mean absolute error (MAE), were used to evaluate prediction accuracy. As part of this examination, the effectiveness of quantified prediction uncertainty for indicating prediction accuracy was assessed. Specifically, the prediction uncertainty quantified at each validation point was plotted against the corresponding absolute value of the prediction residual. A positive relationship is expected if the quantified prediction uncertainty is indicative of prediction accuracy.

Comparison with multiple linear regression (MLR) under the limited sample scenario

The iPSM was compared with a multiple linear regression (MLR) method under the limited sample scenario. The MLR was used because with only 10 sample points it is difficult to use a more sophisticated approach such as kriging. In Figure 7, the spatial configuration of the regression residuals of those 10 points was examined but no spatial autocorrelation was found. In addition, before constructing the MLR model, a principal component analysis (PCA) was performed with the standardized values of the seven environmental covariates in the study area to avoid the impact of multi-collinearity on the covariates (Table 2).

A forward stepwise regression was applied to fit the MLR model and using the principal components in sequence. The first three principal components (*PC*1, *PC*2 and *PC*3) captured 91.5% of the variance in the seven environmental covariates and thus were selected to construct the MLR model (Equation (8)):

$$SOM = 5.12 + 0.24 \times PC1 + 0.23 \times PC2 - 0.33 \times PC3, \left(R^2 = 0.91\right), \quad (8)$$

where *SOM* is the soil organic matter content (%) in the topsoil layer, and *PC*1, *PC*2 and *PC*3 are the first three principal components. The summary of the estimation of the coefficients of the MLR model is shown in Table 3.

Comparison with MLR under the improved sample scenario

The iPSM was also compared with MLR under the improved sample scenario. The first three principal components (*PC*1, *PC*2 and *PC*3) of the seven environmental covariates were used to construct the MLR model based on their values at the 22 sample points (Equation (9)):

$$SOM = 4.61 + 0.02 \times PC1 - 0.01 \times PC2 + 0.01 \times PC3, \left(R^2 = 0.42\right). \quad (9)$$

The summary of the estimation of the coefficients of the MLR model is shown in Table 3. The spatial configuration of the regression residuals at those 22 sample points was also examined, but only a very weak spatial autocorrelation was found (Figure 8).

Uncertainty threshold in iPSM for fair comparison

The MLR uses all available soil samples to predict the content of SOM (%) in the topsoil layer at every unvisited location. However, iPSM only predicts the value of SOM content (%) in the topsoil layer at the locations that can be represented by existing soil samples and assigns '*NoData*' to others. The predictable locations of iPSM are determined by the uncertainty threshold (see the Methods Section). In order to avoid '*NoData*' predictions and to make a fair

Table 1 Statistics of the standardized values of the seven environmental covariates at the 10 sample points, the 22 sample points and the 44 validation sample points, as compared with the entire study area

	10 sample points				22 sample points				44 validation sample points				Study area			
	Min.	Median	Mean	Max.	Min.	Median	Mean	Max.	Min.	Median	Mean	Max.	Min.	Median	Mean	Max.
Elevation	40.7	54.4	52.0	62.6	5.7	56.6	54.0	94.5	17.8	67.8	64.1	96.2	0.0	54.8	52.8	100.0
Planform curvature	−3.3	−1.4	3.3	43.6	−9.2	−0.7	−1.7	2.6	−14.0	−0.2	−0.7	14.7	−50.0	−0.8	−0.8	50.0
Profile curvature	−12.4	3.2	3.7	21.4	−6.1	−0.1	0.8	17.1	−16.4	−2.8	−1.0	23.6	−50.0	−0.5	−0.3	50.0
Relative position index	0.0	44.5	44.7	97.7	2.3	40.4	47.4	100.0	1.9	49.8	52.1	96.1	0.0	37.9	42.3	100.0
Slope gradient	16.0	38.3	38.3	51.3	0.0	33.5	29.0	57.9	7.5	37.7	35.3	61.1	0.0	35.8	33.9	100.0
TWI	28.3	31.1	34.2	65.1	26.7	30.1	37.2	95.1	18.0	29.2	30.5	62.1	0.0	30.3	36.4	100.0
NDVI	4.5	9.5	13.2	28.7	6.1	11.5	13.1	26.2	5.6	11.5	12.8	28.2	−50.0	11.5	13.5	39.8

TWI, topographic wetness index; NDVI, Normalized Difference Vegetation Index.

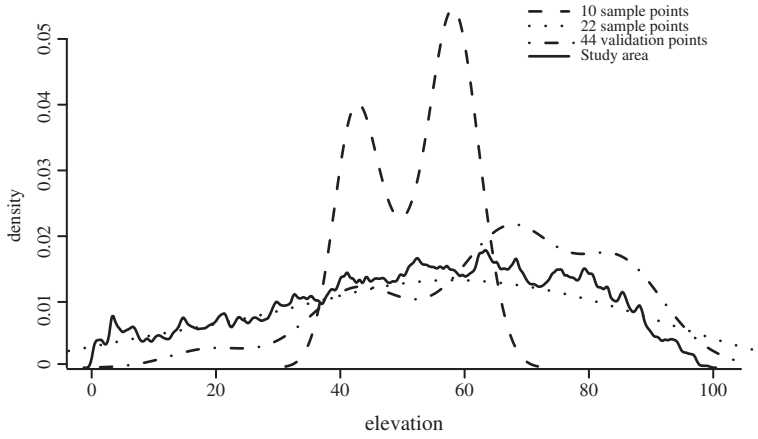

Figure 6 The kernel density estimation of elevation values at the 10, 22 and 44 sample points for model construction and validation, compared with the entire study area.

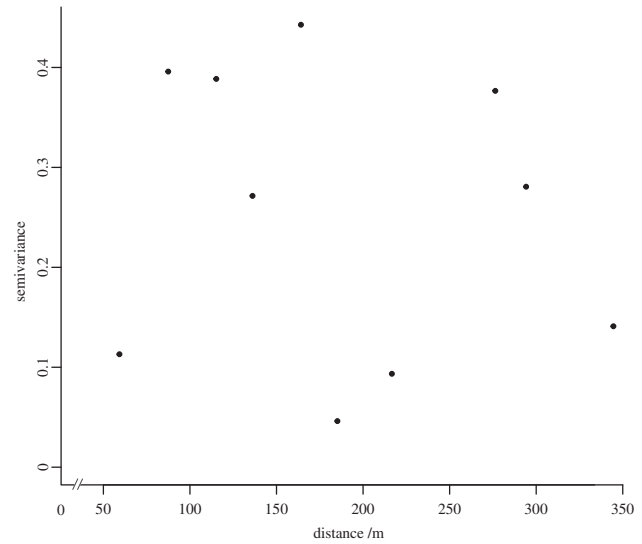

Figure 7 The experimental semivariogram derived from the residuals of the regression of soil organic matter (SOM) content in the topsoil layer (%) using the environmental covariate values at the 10 sample points.

comparison with MLR, the uncertainty threshold in iPSM was set to 1, which forced iPSM to predict a value of SOM content (%) in the topsoil layer at every validation location using all the sample points in each scenario (either 10 or 22 sample points), regardless of how poorly the location was represented by those points.

Results

Under the limited sample scenario

Figure 9(a,b) shows the predicted SOM maps from iPSM under the limited sample scenario with the user-defined uncertainty threshold equal to 1 and 0.6, respectively. The spatial distribution of SOM content (%) in the topsoil layer (Figure 9a,b) matches our understanding of how the terrain and vegetation environment influence SOM content. Thus on upper-to-middle slopes erosive processes tend to be the dominant processes that reduce the soil organic matter content in the topsoil layer; on lower-to-toe slopes which are often gentle depositional processes tend to be the dominant processes that usually lead to larger SOM contents in the topsoil layer. This pattern is obviously disturbed by vegetation conditions: the straight lines with large SOM content (%) correspond to field roads between farmland fields, where trees and grasses are growing.

Figure 9(c) shows the predicted SOM map from MLR under the limited sample scenario. The SOM content (%) in the topsoil predicted by MLR generally had a similar spatial distribution pattern to that from iPSM. The upper-to-middle steep slopes were associated with relatively small values while the lower-to-toe gentle

Table 2 The principal component analysis (PCA) of the standardized values of the seven environmental covariates

			Loadings						
	Eigen-value	Cumulative %	Elevation	Planform curvature	Profile curvature	Relative position	Slope	TWI	NDVI
*PC*1	1414.59	66.63	−0.55	0.06	0.09	−0.75	−0.05	0.33	0.07
*PC*2	376.27	84.35	−0.13	−0.02	−0.02	0.39	−0.70	0.58	0.05
*PC*3	156.96	91.74	0.81	0.17	0.04	−0.47	−0.24	0.21	0.00
*PC*4	61.95	94.66	−0.05	−0.06	0.84	−0.02	−0.36	−0.41	0.02
*PC*5	52.17	97.12	0.08	0.26	0.52	0.24	0.54	0.54	0.13
*PC*6	37.56	98.89	0.06	−0.23	−0.06	−0.02	0.00	−0.06	0.97
*PC*7	23.63	100.00	0.13	−0.92	0.12	−0.07	0.15	0.24	−0.20

TWI, topographic wetness index; NDVI, Normalized Difference Vegetation Index.

Table 3 The estimation of the coefficients of the multiple linear regression (MLR) model under the two sample scenarios

	Estimation
Limited sample scenario	
Intercept	5.12
*PC*1	0.24
*PC*2	0.23
*PC*3	−0.33
Improved sample scenario	
Intercept	4.61
*PC*1	0.02
*PC*2	−0.01
*PC*3	0.01

PC: Principal Component.

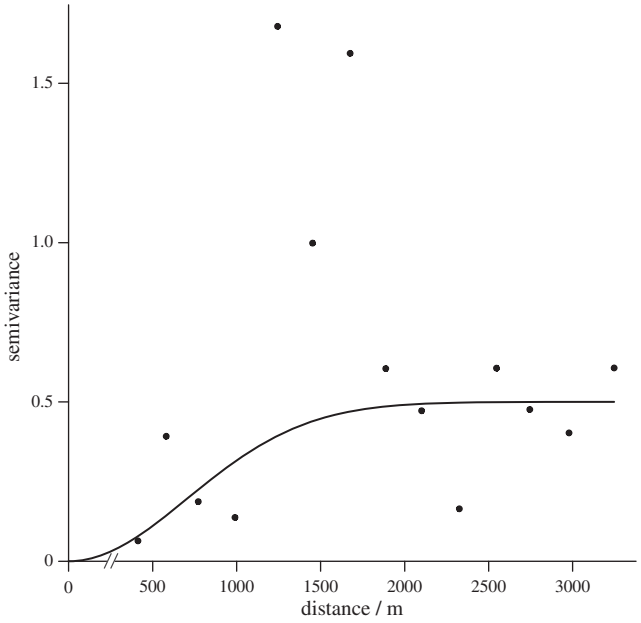

Figure 8 The Gaussian semivariogram (nugget = 0, range = 1000, sill = 0.5) based on the residuals of the regression of soil organic matter (SOM) content in the topsoil layer (%) using the environmental covariate values at the 22 sample points.

slopes were associated with relatively large values. However, some spatial details, such as the sharp changes across the field roads, are not present on the map. In addition, the predicted values of SOM content from MLR ranged from −18.5 to 37.9%, which is beyond the value range of the 10 soil samples. This is because of the limitation of the 10 soil samples, which are not representative of the entire study area. The unrealistic extrapolation of the linear model at the location that cannot be represented by existing samples leads to unrealistic prediction results.

The prediction accuracies of iPSM and MLR evaluated by the 44 validation sample points are compared in Table 4. Those of iPSM with an uncertainty threshold equal to 1 produced better prediction accuracy (1.43 for RMSE and 1.16 for MAE) than those from MLR (8.54 for RMSE and 7.34 for MAE). In addition, when the uncertainty threshold of iPSM was decreased to 0.6, 12 of the 44 validation sample points were predicted and the prediction accuracy of iPSM was increased to 1.06 for RMSE and 0.91 for MAE. The smaller uncertainty threshold improved the prediction accuracy of iPSM but left a greater area as '*NoData*'. The absolute prediction error at each validation location from iPSM was also compared with that from iPSM (Figure 10). At each individual validation location, iPSM produced better prediction accuracy.

Figure 9(d) shows the quantified prediction uncertainty from iPSM under the limited sample scenario. Because the 10 soil samples were mainly located on side slopes (Figure 4) and represent those areas well, prediction uncertainty is generally small in those areas. In contrast, large prediction uncertainty values are assigned to ridges and valleys. In addition, because no sample points were located on the field roads, the prediction uncertainty for the roads is also relatively large.

The values of prediction uncertainty estimated from iPSM are plotted against the absolute prediction residuals at the validation sample locations in Figure 11 (with uncertainty threshold = 1). Overall, there is a positive correlation between the PSM and its prediction residuals (Pearson's $r = 0.54$, $df = 42$). This indicates that the increases in the quantified prediction uncertainty are positively correlated with the increases in prediction residuals produced by iPSM. It further suggests that the quantified prediction uncertainty can be a good indicator of the prediction accuracy

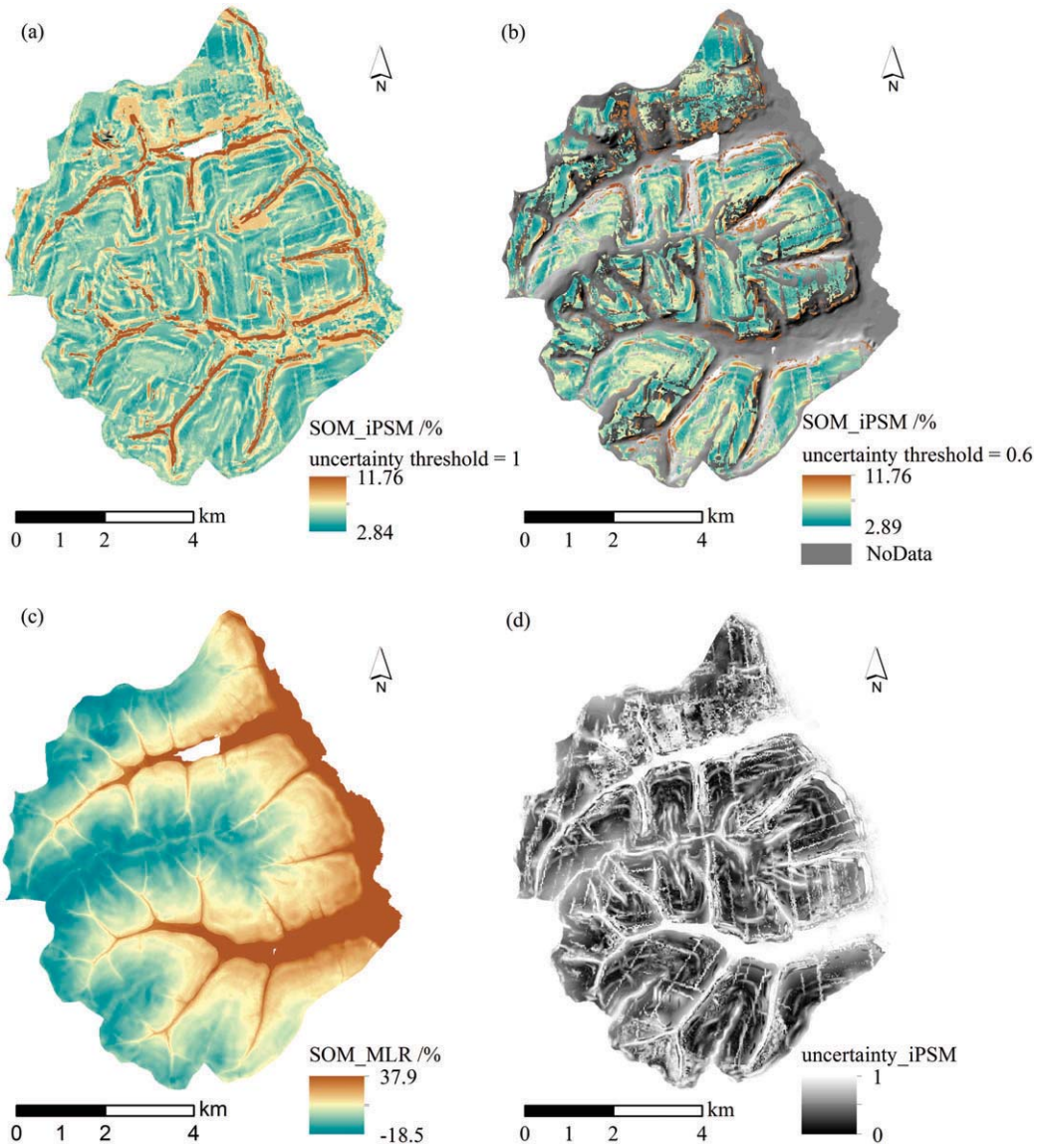

Figure 9 The predicted maps of soil organic matter (SOM) content in the topsoil layer (%) using the 10 sample points: (a) from individual predictive soil mapping (iPSM) with the uncertainty threshold equal to 1; (b) from iPSM with the uncertainty threshold equal to 0.6; (c) from the multiple linear regression (MLR) method; (d) the quantified prediction uncertainty from iPSM.

of iPSM because smaller prediction uncertainty corresponds with smaller prediction residuals at unvisited location.

Under the improved sample scenario

Figure 12(a,b) shows the predicted SOM maps from iPSM under the improved sample scenario (with the user-defined uncertainty threshold equal to 1 and 0.6, respectively). Figure 12(c) shows the predicted SOM maps with MLR.

According to the 44 validation sample points, the prediction accuracy of iPSM with the uncertainty threshold equal to 1 was 1.10 for RMSE and 0.69 for MAE. When the uncertainty threshold was set to 0.6, 38 of the 44 validation sample points were predicted and the prediction accuracy was 0.69 for RMSE and 0.51 for MAE. When compared with the results in the limited sample scenario, the improvement in the soil sample set improved the prediction accuracy. The prediction accuracy of MLR was 1.01 for RMSE and 0.73 for MAE in this scenario (Table 4). The results suggest that iPSM produced comparable prediction accuracy to that of MLR.

Figure 12(d) shows the quantified prediction uncertainty from iPSM. The values of prediction uncertainty estimated from iPSM under this scenario are plotted against the absolute prediction residuals at the 44 validation locations in Figure 13 (with the uncertainty threshold equal to 1). Another positive relationship between the PSM uncertainty and its prediction residuals (Pearson's $r = 0.63$, $df = 42$) was found.

Table 4 The comparison of the prediction accuracies of individual predictive soil mapping (iPSM) and multiple linear regresssion (MLR) based on the 44 validation sample points in the two scenarios

	MAE	RMSE	ME
Limited sample scenario			
iPSM	1.16	1.43	−0.83
MLR	7.34	8.54	4.51
Improved sample scenario			
iPSM	0.69	1.10	0.02
MLR	0.73	1.01	0.02

MAE, mean absolute error; RMSE, root mean squared error; ME, mean error.

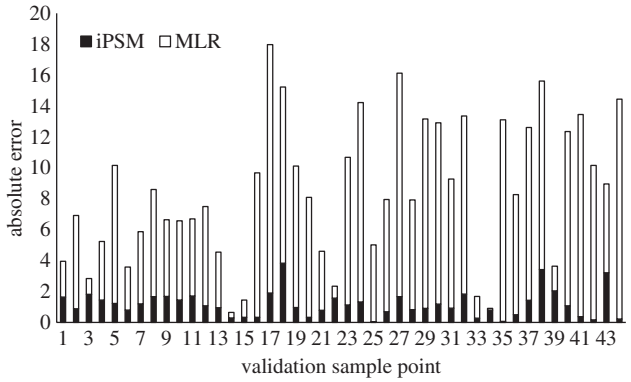

Figure 10 The comparison of prediction residuals from individual predictive soil mapping (iPSM) and multiple linear regression (MLR) at the 44 validation sample points when the uncertainty threshold equals 0.6.

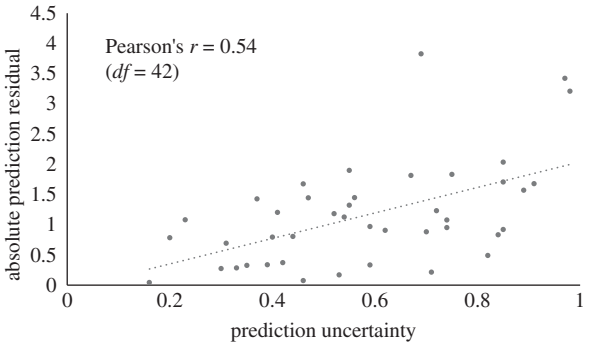

Figure 11 The relationship between prediction uncertainty and absolute value of prediction residual produced by iPSM using the 10 soil sample points (with the uncertainty threshold equal to 1).

Discussion

Choosing an uncertainty threshold

The value of the uncertainty threshold controls the areas that can be predicted with existing soil samples. The areas with a large prediction uncertainty cannot be represented with existing soil samples and the soil conditions over these areas are not predicted in iPSM according to a user-defined uncertainty threshold. This strategy helps to avoid unrealistic extrapolation and makes iPSM use the existing soil samples more appropriately.

As the uncertainty threshold takes smaller values, the area where iPSM can make a prediction shrinks (Figures 9, 12). A small value of uncertainty threshold leads to a strict criterion to determine whether an unvisited location can be represented and thus can be predicted from existing soil sample points. For example, under the limited sample scenario with the uncertainty threshold set to 0.6 (Figure 9b), only the locations where the environmental similarities to the existing soil samples were larger than 0.4 (1 − 0.6) could be predicted. Those locations were mainly on the side slopes, which are the landforms well represented by the 10 soil samples. In contrast, a large value of uncertainty threshold leads to a poor criterion, which means that more areas can be represented and thus predicted by the existing soil samples (Figure 9a).

The value of the uncertainty threshold also influences the prediction accuracy. The overall accuracies of the predicted SOM maps from iPSM by using different values of uncertainty thresholds were evaluated by the 44 validation samples. Figure 14 shows that RMSE and MAE increased as the uncertainty threshold increased under the two scenarios. The number next to each point is the number of validation samples that were predicted when using the corresponding uncertainty threshold. The iPSM method thus has the advantage of controlling the quality of the prediction through its uncertainty threshold; when choosing a value for the uncertainty threshold a trade-off needs to be made between prediction accuracy and the areas that can be predicted using the existing soil samples.

Selection of environmental covariates

When selecting the environmental covariates, the ultimate goal is to describe the environmental conditions associated with the targeted soil property as accurately as possible and all components of Jenny's equation (Jenny, 1941) need to be considered. However, one also needs to consider the effectiveness of environmental covariates in indicating the spatial variation of soil, and the availability of data sources for calculation (Zhu *et al*., 2001). In our study only the environmental covariates describing topographic and vegetation conditions were used. One reason is that the study area was relatively small and climate conditions and parent materials are generally similar over the area; another reason is that it is relatively easy to obtain data characterizing terrain and vegetation conditions. With a more complex study area, more types of environmental covariates should be involved.

Measurement of environmental similarities

The measurement of environmental similarity is a crucial step in iPSM. Currently a Boolean operator is used for nominal and ordinal covariates and a Gaussian-shaped function (Equation (3)) is used for interval and ratio covariates. Certainly other similarity/distance measurements defined in the environmental feature space, such as taxonomic distance (Minasny & McBratney, 2007), Gower similarity coefficient (Gower, 1971), Euclidian distance or Manhattan distance, can also be used. The choice of the method used for

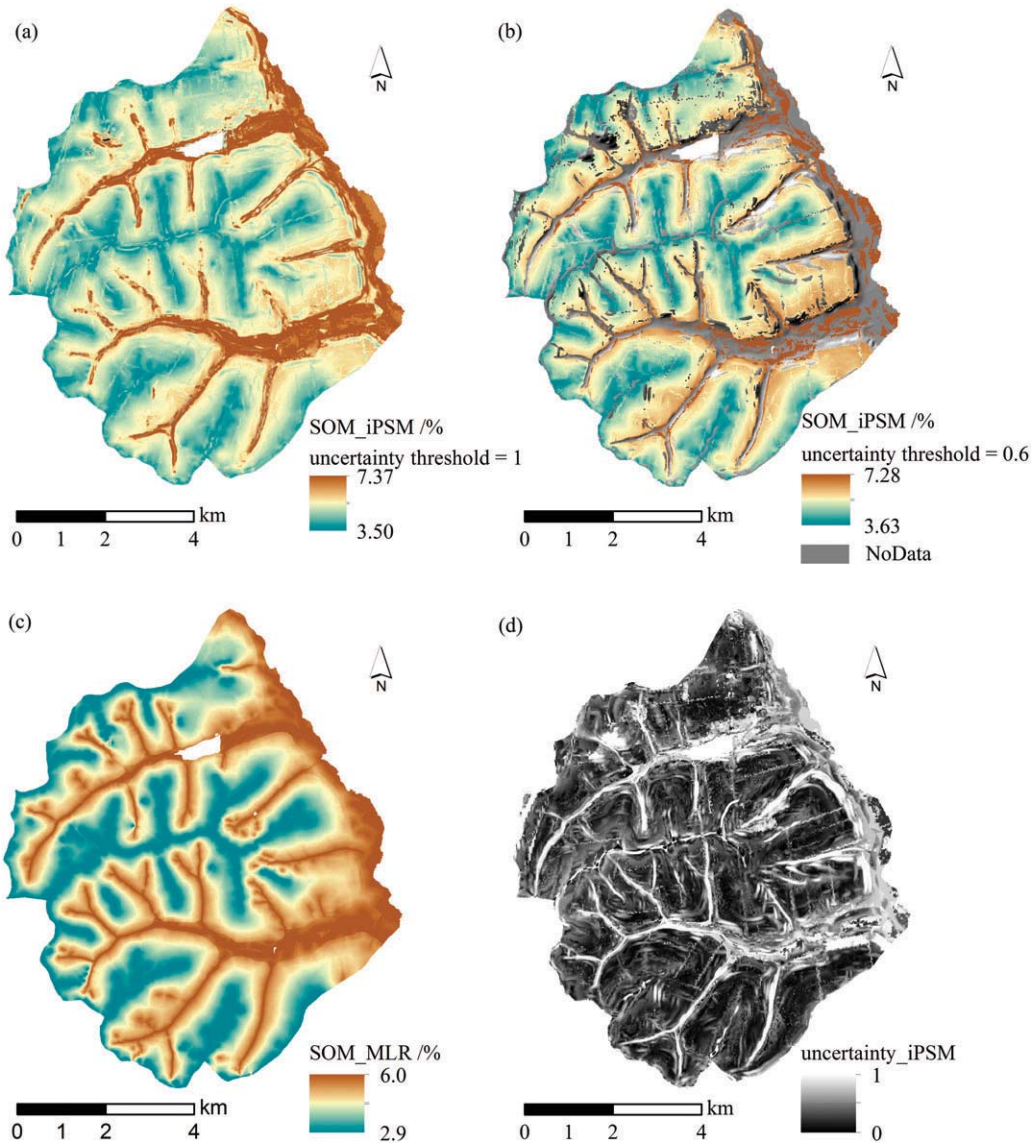

Figure 12 The predicted maps of soil organic matter (SOM) content in the topsoil layer (%) using the 22 sample points: (a) from individual predictive soil mapping (iPSM) with the uncertainty threshold equal to 1; (b) from iPSM with the uncertainty threshold equal to 0.6; (c) from the multiple linear regression (MLR) method; (d) the quantified prediction uncertainty from iPSM.

measuring environmental similarity depends on (i) the measurement form (nominal, ordinal, interval and ratio) of environment covariates and (ii) how environmental similarity changes as the value of the environmental covariate at an unvisited location deviates from the value at the soil sample point.

For the integration of environmental similarities we used a minimum operator following Zhu & Band (1994) and Shi *et al.* (2004). The theoretical basis of using a minimum operator is Liebig's law of the minimum (van der Ploeg *et al.*, 1999), which assumes that the factor (such as a particular nutrient, water or sunlight) in shortest supply will limit the growth and development of an organism or a community. When applied to the integration of an appropriate, although conservative, choice similarities of multiple environment covariates, it is choice when the relative importance among multiple environmental covariates is not clear. By assuming that the least similar soil-forming factor determines the environmental similarity between two locations, it requires no additional information about the relative importance among multiple factors. Nevertheless, more research is needed to find a more reasonable way to integrate the impact of multiple environmental covariates on soil formation.

Application of iPSM over large areas

The limitation of soil sample data is more obvious over large areas, such as those at regional and global scales. This pilot study applied iPSM in a relatively small area to first explore the possibility of

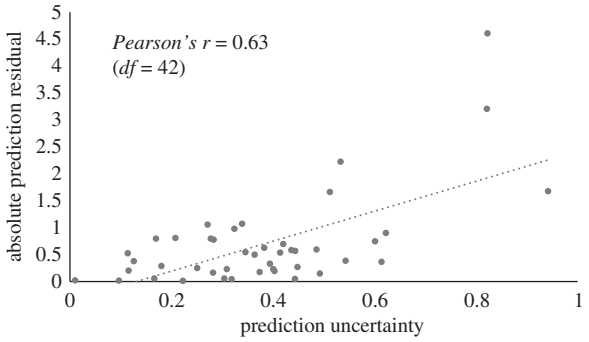

Figure 13 The relationship between prediction uncertainty and absolute value of prediction residual produced by individual predictive soil mapping (iPSM) using the 22 sample points (with the uncertainty threshold equal to 1).

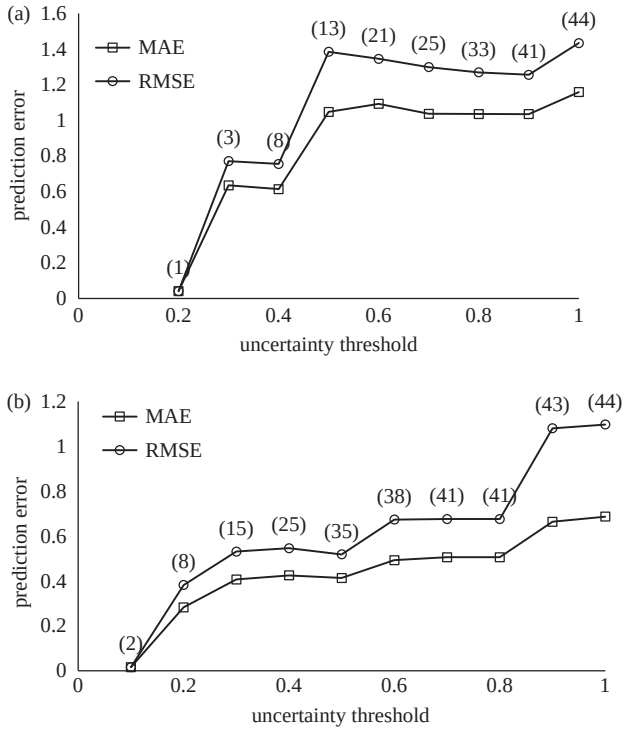

Figure 14 Prediction accuracies of the soil organic matter (SOM) content in the topsoil layer (%) from individual predictive soil mapping (iPSM) with different values of the uncertainty threshold: (a) from the 10 sample points; (b) from the 22 sample points.

using iPSM as an alternative method when available soil sample data are not sufficient to construct a PSM model using existing methods. Given its performance in this case study, iPSM shows promise in large areas when existing soil sample data are limited. The method has several advantages that could guide its suitability. First, it does not require that soil samples collectively represent the study area and thus can be applied to areas where some strata used in regression relationships are under-sampled (as in the case study). This is often the case if only legacy soil samples are available or if field work is too expensive or difficult to collect fully representative samples. Second, iPSM explicitly differentiates between prediction locations that are well represented by available soil samples and those locations that are not by assigning '*NoData*' to those areas. In addition, when considering the positive correlation between prediction uncertainty and prediction residual from iPSM, the prediction uncertainty can be used not only to evaluate prediction accuracy at each location, but also to guide the efficient allocation of additional sampling resources to improve the accuracy of the predicted soil map.

Added value to the existing work on PSM with limited data

How to use limited data for predictive soil mapping has been one of the key research topics in this field (Hartemink *et al.*, 2008). Mallavan *et al.* (2010) suggested that the so-called Homosoil method for quantitative extrapolation of soil information from a reference area and the region of interest could be used. The basic assumption is that the same SCORPAN spatial soil prediction functions can be applied in the regions with similar soil-forming factors. Our paper is based on a similar assumption but focusses more on using each individual soil sample point for prediction. It demonstrates the possibility of using individual soil sample points to only predict unvisited locations that can be well represented. The method is designed to relax the requirements regarding the sufficiency of soil sample data. It is suitable for the scenario where there are not sufficient soil sample data for constructing or calibrating the PSM model using existing PSM methods.

Conclusions

This paper presents an alternative predictive soil mapping method, iPSM, which is suitable for using a limited quantity of soil sample data to predict soil property and quantify prediction uncertainty at unvisited locations. The case study suggests that the method can be used as an effective alternative to existing PSM methods when the available soil sample data have a limited ability to represent the study area. In addition, iPSM also takes into consideration the prediction uncertainty caused by the use of limited sample data. The positive relationship between prediction uncertainty estimated by iPSM and its prediction residuals implies that the uncertainty can not only serve as an effective indicator of prediction accuracy at every location, but also can be used to effectively allocate additional sampling resources. With the aid of the prediction uncertainty, the sampling resources can be used wisely by sampling the under-represented areas with greater priority.

Acknowledgements

This study was supported by the Natural Science Foundation of China (No. 40971236), the Ministry of Science and Technology of China (Project No. 2010DFB24140), the Natural Science Foundation of China (Project No. 41023010), the National High Technology Research and Development Program of China (863 Program, Project No. 2011AA120305), PAPD, the Vilas

Associate Award, the Hammel Faculty Fellow Award, the Manasse Chair Professorship from the University of Wisconsin-Madison, and the 'One-Thousand Talents' Program of China. Chengzhi Qin is grateful for the support from the National Science and Technology Support Program (No. 2013BAC08B03-4). The authors would also like to express their appreciation of the valued comments from Dr D.G. Rossiter from the International Institute for Geo-Information Science and Earth Observation (ITC) of the University of Twente.

References

Beven, K.J. & Kirkby, N.J. 1979. A physically based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P. & Knotters, M. 2006. *Sampling for Natural Resource Monitoring*. Springer, New York.

Florinsky, I.V. 2011. *Digital Terrain Analysis in Soil Science and Geology*. Elsevier/Academic Press, Amsterdam.

Gower, J.C. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.

Hartemink, A.E., McBratney, A.B. & de Lourdes Mendonça-Santos, M. 2008. *Digital Soil Mapping with Limited Data*. Springer Science + Business Media LLC, New York.

Hudson, B.D. 1992. The soil survey as a paradigm-based science. *Soil Science Society of America Journal*, **56**, 836–841.

Jenny, H. 1941. *Factors of Soil Formation: A System of Quantitative Pedology*. Dover Publication, New York.

Johnson, C.J., Hurley, M., Rapaport, E. & Pullinger, M. 2012. Using expert knowledge effectively: lessons from species distribution models for wildlife conservation and management. In: *Expert Knowledge and its Application in Landscape Ecology* (eds A.H. Perera, C.A. Drew & C.J. Johnson), pp. 153–171. Springer, New York.

Kienast-Brown, S. & Boettinger, J.L. 2007. Land cover classification from Landsat imagery for mapping dynamic wet and saline soils. In: *Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science* (eds P. Lagacherie, A.B. McBratney & M. Voltz), pp. 235–244. Elsevier, Amsterdam.

Liu, F., Geng, X.Y., Zhu, A.X., Fraser, W. & Arnie, W. 2012. Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from MODIS. *Geoderma*, **171-172**, 44–52.

Mallavan, B.P., Minasny, B. & McBratney, A.B. 2010. Homosoil, a methodology for quantitative extrapolation of soil information across the globe. In: *Bridging Research, Environmental Application, and Operation* (eds J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink & S. Kienast-Brown), pp. 137–149. Springer, Dordrecht.

McBratney, A.B., Mendonça-Santos, M.L. & Minasny, B. 2003. On digital soil mapping. *Geoderma*, **117**, 3–52.

Mendonça-Santos, M.L., McBratney, A.B. & Minasny, B. 2006. Soil prediction with spatially decomposed environmental factors. In: *Digital SoilMapping: An Introductory Perspective. Developments in Soil Science* (eds P. Lagacherie, A.B. McBratney & M. Voltz), pp. 269–278. Elsevier, Amsterdam.

Minasny, B. & McBratney, A.B. 2006. A conditioned latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, **32**, 1378–1388.

Minasny, B. & McBratney, A.B. 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. *Geoderma*, **142**, 285–293.

Nield, S.J., Boettinger, J.L. & Ramsey, R.D. 2007. Digitally mapping gypsic and natric soil areas using Landsat ETM data. *Soil Science Society of America Journal*, **71**, 245–252.

Qi, F. & Zhu, A.X. 2003. Knowledge discovery from soil maps using inductive learning. *International Journal of Geographical Information Science*, **17**, 771–795.

Qin, C.Z., Zhu, A.X., Shi, X., Li, B.L., Pei, T. & Zhou, C.H. 2009. The quantification of spatial gradation of slope positions. *Geomorphology*, **110**, 152–161.

Qin, C.Z., Zhu, A.X., Pei, T., Li, B.L., Scholten, T., Behrens, T. *et al.* 2012. An approach to computing topographic wetness index based on maximum downslope gradient. *Precision Agriculture*, **12**, 32–43.

Ramsey, C.A. & Hewitt, A.D. 2005. A methodology for assessing sample representativeness. *Environmental Forensics*, **6**, 71–75.

Scull, P., Franklin, J., Chadwick, O.A. & McArthur, D. 2003. Predictive soil mapping: a review. *Progress in Physical Geography*, **27**, 171–197.

Shi, X., Zhu, A.X., Burt, J., Qi, F. & Simonson, D. 2004. A case-based reasoning approach to fuzzy soil mapping. *Soil Science Society of America Journal*, **68**, 885–894.

Stevens, S.S. 1946. On the theory of scales of measurement. *Science*, **103**, 677–680.

van Groenigen, J.W. & Stein, A. 1998. Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, **27**, 1078–1086.

van der Ploeg, R.R., Bohm, W. & Kirkham, M.B. 1999. On the origin of the theory of mineral nutrition of plants and the law of the minimum. *Soil Science Society of America Journal*, **63**, 1055–1062.

Webster, R. & Oliver, M.A. 1990. *Statistical Methods in Soil and Land Resource Survey*. Oxford University Press, Oxford.

Yang, L., Zhu, A.X., Li, B.L., Qin, C.Z., Pei, T., Liu, B.Y. *et al.* 2007. Extraction of knowledge about soil-environment relationship for soil mapping using fuzzy c-means (FCM). *Acta Pedologica Sinica*, **44**, 784–791 (in Chinese).

Zhu, A.X. 1999. A personal constructed-based knowledge acquisition process for natural resource mapping using GIS. *International Journal of Geographic Information Systems*, **13**, 119–141.

Zhu, A.X. 2001. Modeling spatial variation of classification uncertainty under fuzzy logic. In: *Spatial Uncertainty in Ecology* (eds C. Hunsaker, M.F. Goodchild, M. Friedl & T. Case), pp. 330–350. Springer, New York.

Zhu, A.X. & Band, L.E. 1994. A knowledge-based approach to data integration for soil mapping. *Canadian Journal of Remote Sensing*, **20**, 408–418.

Zhu, A.X., Band, L.E., Vertessy, R. & Dutton, B. 1997. Derivation of soil properties using a soil land inference model (SoLIM). *Soil Science Society of America Journal*, **61**, 523–533.

Zhu, A.X., Hudson, B., Burt, J.E. & Lubich, K. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, **65**, 1463–1472.

Zhu, A.X., Yang, L., Li, B., Qin, C., English, E., Burt, J.E. *et al.* 2008. Purposefully sampling for digital soil mapping. In: *Digital Soil Mapping with Limited Data* (eds A.E. Hartemink, A.B. McBratney & M.L. Mendonca Santos), pp. 233–245. Springer-Verlag, New York.

Zhu, A.X., Liu, F., Li, B.L., Pei, T., Qin, C.Z., Liu, G.H. *et al.* 2010. Differentiation of soil conditions over flat areas using land surface feedback dynamic patterns extracted from MODIS. *Soil Science Society of America Journal*, **74**, 861–869.