Supplemental sampling for digital soil mapping based on prediction uncertainty from both the feature domain and the spatial domain

Yan Li ^a, A-Xing Zhu ^{b,c,d,e,f,*}, Zhou Shi ^g, Jing Liu ^f, Fei Du ^f

^a Institute of Land Science and Property Management, School of Public Affairs, Zhejiang University, Hangzhou 310058, China

^b Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing 210023, China

^c State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China

^d Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

^e State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

^f Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA

^g Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Article history:

Received 31 May 2016

Received in revised form 16 August 2016

Accepted 20 August 2016

Available online 7 September 2016

Keywords:

Prediction uncertainty
Supplemental sampling
Feature domain
Spatial domain
Digital soil mapping

ABSTRACT

This paper presents an uncertainty-directed sampling method that can be used to design additional samples for soil mapping. The method is based on uncertainty from both the feature domain (the domain of relationships with environmental covariates) and the spatial domain (the domain of spatial autocorrelation). Existing soil samples are also taken into account. The method comprises three steps: 1) the selection of a ranked list of additional sample locations based on uncertainty from the feature domain using individual predictive soil mapping (iPSM); 2) the selection of a ranked list of additional sample locations based on uncertainty from the spatial domain using ordinary kriging; 3) the determination of a final ranked list created by merging the ranked lists from steps 1) and 2) based on both uncertainties. To evaluate the method, the three lists were used to map soil organic matter (SOM) in a 299.14 km² study area near Fuyang city in the northwest region of Zhejiang Province, China. The mapping accuracy of each list was then calculated and used to assess the effectiveness of the method. Compared with the sampling scheme based on the uncertainty from either the feature domain or the spatial domain alone, the root-mean-squared error (RMSE), with the addition of the final list based on both uncertainties, was found to be the smallest, ranging from 0.829 to 1.126, and the agreement coefficient (AC) was the largest, ranging from 0.634 to 0.737. This confirms that sampling based on two uncertainties is better than sampling based on uncertainty from either the feature domain or the spatial domain alone. The results suggest that the proposed combined additional sampling method is more effective for sampling additional points in soil mapping.

© 2016 Published by Elsevier B.V.

1. Introduction

Soil-property maps provide important information for land resource management, environmental and ecological modeling and precision agriculture. To meet the increasing demand for accurate and reliable maps, research on digital soil mapping (DSM) began to appear in the literature in the early 1990s (McKenzie and Austin, 1993; Moore et al., 1993; Zhu and Band, 1994; Gessler et al., 1995; Zhu et al., 1997; Moran and Bui, 2002; McBratney et al., 2003; Mueller and Pierce, 2003). This research has made it clear that the first step in generating digital soil data is to take full advantage of the large quantity of existing

soil samples collected in the course of traditional soil or land resource surveys (Arrouays et al., 2014).

However, traditional methods of surveying soil are generally empirical and based on the experiences of the surveyor, who may correlate soil with the underlying geology, landscape, vegetation and so on. This may lead to bias in the areas being sampled (Carré et al., 2007). In addition, the soil in a given area has often been sampled at different times and for various purposes, with some areas being relatively over or undersampled. Without following any conventional sampling designs (e.g. stratified random sampling, regular sampling), the collected samples do not provide a good spatial coverage of the study area (Webster and Oliver, 1990; Zhang et al., 2016). In spite of this, these existing samples are important resources for DSM. When additional sampling resources are available, one might want to acquire new samples to improve mapping accuracy, but it is beneficial to include previously acquired samples into the design process

* Corresponding author at: School of Geography, Nanjing Normal University, No. 1, Wenyuan Road, Xianlin University District, Nanjing 210023, China.
E-mail address: azhu@wisc.edu (A.-X. Zhu).

(Arrouays et al., 2014). In this case, designing a reliable sampling scheme to determine where new samples should be taken from presents a challenging problem.

Recently, uncertainty-directed sampling has drawn a great deal of attention (Van Groenigen et al., 1999; Zhu et al., 2015; Zhang et al., 2016). This method is based on uncertainty from both the feature domain (the domain of relationships with environmental covariates) and uncertainty from the spatial domain (domain of spatial autocorrelation).

Uncertainty-directed sampling in the feature domain is based on a method called individual predictive soil mapping (iPSM), which was first proposed by Zhu et al. (2015) to predict soil organic matter (SOM) using limited samples and to quantify the prediction uncertainty. It was then used to improve the additional sampling design of SOM at the county scale (Zhang et al., 2016). The iPSM method assumes that locations with similar environmental conditions have similar soil properties (Hudson, 1992; Mallavan et al., 2010). Therefore, each soil sample point can individually serve as a surrogate, representing locations with similar environmental conditions, and thus can be used to predict the soil conditions for these locations. The prediction uncertainty is inversely related to the similarity to the existing soil samples. If the prediction uncertainty is high, the similarity in environmental conditions between the unvisited location and the existing sample points is low, and the existing soil sample points poorly represent the current location. Additional candidate samples should be located in a region where the prediction uncertainty is high.

Uncertainty-directed sampling in the spatial domain is largely based on geostatistical methods that draw samples by specifying a covariance-based criterion (McBratney and Webster, 1981; Haining, 2003; Simbahan and Dobermann, 2006; Zhu and Stein, 2006; Delmelle and Goovaerts, 2009). The uncertainty of a prediction is determined by the kriging model. Kriging variance can be minimized by configuring the sampling location of the observed data, rearranging the sampling scheme or adding new samples to improve the kriging estimation. For example, Van Groenigen optimized the sampling scheme by minimizing the mean kriging variance (AKV) or maximum kriging variance (MaxKV) using spatial-simulated annealing (SSA) (Van Groenigen and Stein, 1998; Van Groenigen et al., 1999; Van Groenigen, 2000). Zhu and Stein (2006) designed a sampling method for spatial prediction when the covariance parameters have to be estimated from the same data. They incorporated the parameter uncertainty of the semivariogram into a design criterion to represent the uncertainty in prediction and used a simulated annealing algorithm to search for the optimal design. Juang et al. (2008) proposed an adaptive cluster sampling approach based on the regulation threshold and kriging variance to improve reliability in a delineation of a heavy-metal contaminant.

The limitation of recent efforts on uncertainty-directed sampling is that these sampling strategies can only cover one domain, either the feature domain or the spatial domain; few can cover both (Bros and Heuvelink, 2007). There is an urgent need to develop uncertainty-directed sampling that can combine the uncertainty information from both the feature domain and the spatial domain.

This paper proposes a sampling method that can integrate the prediction uncertainty from both the spatial domain and the feature domain. Section 2 introduces the basic idea and procedure of the method. The method is then illustrated in a case study on additional sampling of SOM for digital soil mapping. This is followed by an assessment of the proposed method and a discussion of the results.

2. Methodology

2.1. Basic concept

The basic concept of the sampling design is to use the uncertainty information from both the feature domain and the spatial domain to select additional soil sample locations. The uncertainties from these two domains are incompatible. We obtain this “combination” in two major steps: 1) selecting a sample location based on each of the uncertainties and ranking the sample locations based on the reduction in uncertainty for each location in the respective domains; 2) selecting the final sample by merging the two ranked lists of sample locations based on the sample's relative ranking in its own respective list (Fig. 1).

2.2. Generation of additional samples based on the uncertainty from the feature domain

2.2.1. Production of the uncertainty

The iPSM method developed by Zhu et al. (2015) was introduced to generate the uncertainty distribution map from the feature domain. The key idea is that each field sample reflects an underlying relationship between soil and its relative environmental conditions and that this relationship recurs across the space. It is assumed that locations with similar environmental conditions will have similar soil properties. Therefore, each sample can be considered representative of locations with similar environmental conditions. The level of similarity between an individual sample location and an unsampled location can be approximated by the similarity in environmental covariates. The soil property value at an unsampled location can be predicted by referring to environmentally similar samples, and the uncertainty associated with this prediction at this location can be quantified by assessing the environmental similarities. A more detailed discussion of this method is beyond the scope of this paper. Interested readers can refer to Zhu et al. (2015) for more details. The SoLIM software (<http://solim.geographic.wisc.edu/software>) was used to infer the prediction map of SOM and the uncertainty distribution map.

It is possible that there are locations with little similarity to any of the existing samples. In this case, it would be better to assign a NO-DATA value to these points so that users of the predicted property map are better informed. To do this, a user-defined uncertainty threshold needs to be set, and a NO-DATA value will be assigned to the locations whose uncertainty values exceed this threshold. The choice of uncertainty threshold depends on the user's tolerance regarding the reliability of existing soil samples to represent unvisited locations. A small

Fig. 1. Sampling design flowchart.

Fig. 2. Study area and spatial distribution of training and validation sample points.

value for the uncertainty threshold means that the existing soil samples are used to predict only unvisited locations with very similar environmental conditions (Zhu et al., 2015). Besides, the iPSM method has the advantage of controlling the quality of the prediction through its uncertainty threshold (Zhu et al., 2015). In the present study, the uncertainty threshold is set to 0.3, which is selected based on the trade-off between prediction accuracy and the areas that can be predicted using the existing soil samples.

2.2.2. Generation of additional samples

To select locations for additional samples, we adopted the method proposed by Zhang et al. (2016). Based on the generated uncertainty distribution map, the areas with higher prediction uncertainty (uncertainty value over a prescribed value, >0.3 in the present study) were determined. A fuzzy c-means (FCM) cluster was then performed on the environmental covariates, corresponding to areas with prediction uncertainty >0.3, to generate clusters with similar environmental conditions. Environmental data layers were preprocessed before FCM clustering to remove outliers and to standardize the ranges of each data layer. The standardized environmental data layers were then analyzed with the FCM algorithm to assign a pixel to the appropriate classes based on the minimization of the generalized least-squared error function of each pixel. Several types of cluster validity functions are usually calculated on each fuzzy membership matrix produced by fuzzy c-means since the local minima of generalized least-squared error functions are not consistent with the visually acceptable clustering patterns of the data. For the present study, the fuzziness performance index (FPI) (Odeh et al., 1992; Boydel and McBratney, 1999) and normalized classification entropy (NCE) (Bezdek, 1981) were used to determine the optimal number of clusters. FPI is a measure of the degree of separation (i.e., fuzziness) between fuzzy c-partitions of the fuzzy membership matrix. Values of FPI may range from 0 to 1. Values approaching 0 indicate distinct classes with little membership sharing while values near 1 indicate no distinct class, with a large degree of membership sharing. The NCE models the amount of disorganization of a fuzzy c-partition

of the fuzzy membership matrix (Odeh et al., 1992; Lark and Stafford, 1997). The optimal number of clusters for each computed index is when the index is at the minimum, representing the least amount of membership sharing (FPI) and the greatest amount of organization (NCE) as a result of the clustering process (Fridgen et al., 2004).

Once the optimal number of clusters has been determined, the raster number of each cluster can be calculated by assigning the fuzzy membership threshold. The curve of probability density distribution of the cluster with the largest raster number can then be obtained using kernel density estimation (KDE) (Zhang et al., 2016). KDE is a non-parametric way to estimate the probability density function of a random variable. It spreads the mass of observations around the observed value, and the amount of spreading is governed by a function called the kernel. There are several kernels available. Here, we use the Gaussian kernel (Silverman, 1986). The choice of the bandwidth is the most important issue in implementing a kernel density estimator. Narrow kernels allow nearby observations to have the greatest influence on the density observations, and wide kernels allow more influence on distant observations (Seaman and Powell, 1996). The rule of thumb proposed by Silverman (1986) was used to determine bandwidth for this study.

The peak values in the probability density distribution curve correspond to the specific uncertainty values that show up most frequently in the cluster. Its corresponding raster locations are typical and can represent the cluster well, thus they can be taken as the candidate sampling locations. Among these locations, the first sampling point with

Table 1
Environmental predictors likely to impact SOM distribution.

Variable	Source, description
Elevation	Digital relief map
Slope	From DEMs by first order finite difference
Plan curvature	Curvature of contour drawn through the grid point
Profile curvature	Curvature of the surface in the direction of steepest descent
TWI	See description in the text

convenient accessibility is selected. The first sampling point is then added to the existing sample set to generate the uncertainty to help design the next new sampling location. In this manner, additional

sampling locations can be obtained, providing the ranked list of additional sampling locations based on the uncertainty in the feature domain.

Fig. 3. Spatial distribution maps of selected environmental variables.

2.3. Generation of additional samples based on uncertainty from the spatial domain

2.3.1. Production of the uncertainty

Geostatistics were used to generate the uncertainty distribution map from the spatial domain. Geostatistics, in the form of semivariogram and kriging, have been widely used to model spatial continuity and variability of soil properties in terms of the rate of change of regionalized variables (Webster and Burgess, 1980, 1984; Burgess and Webster, 1980; Webster, 1985; McBratney and Webster, 1986). In this study, ordinary kriging was used to generate a prediction uncertainty distribution map of target soil properties, that is, a map of kriging variance of SOM. The kriging variances or standard errors were produced as byproducts of ordinary kriging estimate and a guide to the reliability of the estimates. Where sampling is irregular, a map of prediction variances may indicate that there are parts of a region where sampling should be increased to improve the estimates (Oliver, 2010).

2.3.2. Generation of supplemental samples

The only factors influencing the kriging variance are the variogram, the number of observations and the locations of the prediction point. This means that it is possible to calculate the kriging variance before actual sampling takes place, provided the variogram is known or can be assumed. This feature is used to optimize spatial sampling schemes for minimal kriging variance. The optimization technique that was used here is spatial simulated annealing (SSA), as developed by Van Groenigen et al. (1999).

SSA is a method of numerical optimization to find the values of a set of parameters of an objective function, ϕ , and then optimize (minimize or maximize) that function. The parameters could be the coordinates of a set of sample points, and the objective function could be the mean prediction variance evaluated over a fine network of points in the sample area. SSA proceeds by random perturbation of an initial set of parameter values. This means moving each sample point randomly in turn. Perturbations that improve the objective function are all accepted, whereas those that make it worse are accepted or rejected at random, with the probability of acceptance depending on a function called the Metropolis criterion, which simulates the statistical mechanics of atoms in a molten metal. Thus, if the proposed change in the system results in a change in the objective function from $\phi(s_i)$ to $\phi(s_{i+1})$, then the probability of acceptance of the change is

$$p_k(s_i \rightarrow s_{i+1}) = \begin{cases} 1 & \phi(s_i) \leq \phi(s_{i+1}) \\ \exp\left\{-\frac{\phi(s_i) - \phi(s_{i+1})}{k}\right\} & \phi(s_i) > \phi(s_{i+1}), \end{cases} \quad (1)$$

where k is a parameter analogous to the temperature of the metal; reducing k is therefore a “cooling” step. k is reduced as optimization progresses. If S_{i+1} is accepted, it serves as a starting point for the next change, S_{i+2} , and the process continues in a similar way (Aarts and Korst, 1989).

The particular advantage of SSA as a method of optimization is that the Metropolis criterion allows the system in effect to jump over a barrier that could trap it at a solution that is only locally optimal. This could not be achieved in a procedure in which any change that makes the objective function worse is rejected. SSA is suitable for completing irregularly scattered, existing data sets, making full use of the existing observation, and it can handle a variety of quantitative optimization criteria. In this paper, we will focus on minimization of the mean kriging variance, and the aim of the optimization criterion is the minimization of the integral of the kriging variance over the study area:

$$\tilde{\phi}_{ok}(S_i) = \int \sigma_{ok}^2(x_0|S_i) \cdot dx_0 \quad (2)$$

In most studies, the kriging predictions are calculated on the nodes of a raster grid. Hence, SSA discretizes the area for evaluation of Eq. (2). The objective function can then be defined as the mean of kriging variances calculated at the nodes of a fine raster grid:

$$\tilde{\phi}_{ok}(S) = \sum_{j=1}^{n_e} \frac{\sigma_{ok}^2(x_{e,j}|S)}{n_e}, \quad (3)$$

where $x_{e,j}$ denotes the j th raster node, and n_e is the number of raster nodes (Sacks and Schiller, 1988; Christakos and Olea, 1992; Van Groenigen et al., 1999).

The steps of the optimization algorithm can be described as follows: (1) a sampling scheme $S_0 \in S^n$ is defined, consisting of a subscheme $S_0^f \in S^{n_f}$ with existing observations, and a randomly drawn subscheme $S_0^v \in S^{n_v}$ that consists of the additional sampling points that should be optimized. (2) The area is discretized and the raster is chosen, yielding n_e raster nodes. (3) Kriging variances for all raster nodes are calculated, and the mean kriging variance $\phi_{ok}(S_0)$ is calculated using Eq. (2). (4) An observation point from the subscheme S_0^f is perturbed at random, yielding scheme S_1 , and the new mean kriging variance $\phi_{ok}(S_1)$ is calculated. S_1 is accepted as a basis for further optimization depending on the Metropolis criterion. (5) The process proceeds at point v , with S_1 replacing S_0 if it was accepted. (6) At a fixed temperature, steps (4) and (5) are repeated until all the observation points are perturbed in turn. (7) Temperature is reduced gradually, and steps (5) and (6) are repeated.

SSA uses a cooling schedule proposed by Kirkpatrick et al. (1983). The initial temperature of the system, k_1 , is chosen so that the

Fig. 4. Distribution map of SOM (left) and prediction uncertainty (right) based on the individual predictive soil mapping (iPSM) method with uncertainty threshold equal to 0.3.

Table 2
Range and standardization for environmental covariates.

Variable	Original range	Range after removing outliers	Range after standardization
Elevation	27–702	30–700	0–100
Slope	0–52.52	0–45	0–100
Planform curvature	–0.65 to 0.72	–0.4 to 0.4	–40 to 40
Profile curvature	–0.0203 to 0.0151	–0.014 to 0.014	–14 to 14
TWI	4.03–16.99	4.2–16	0–100

proportion of proposed changes accepted before the first reduction in temperature is in the range 0.90–0.99 and the new temperature of the system, k_{m+1} , after the m_{th} cooling step, is ak_m , where $0 < a < 1$. The cooling step takes place after a fixed number of perturbations of coordinates of a set of additional observation points. Similarly, the maximum distance that a sample point can be perturbed may also be reduced at each cooling step.

When a given number of successive cooling steps has been completed with no further decrease in the mean kriging variance, the optimization algorithm stops, and the optimized additional sampling points are generated. The sample locations are then ranked based on the reduction of each location in the mean kriging variance, which provides the ranked list of additional sample locations based on the uncertainty in the spatial domain.

2.4. Generation of the final set of additional samples

The main task of additional sampling in the present paper is to generate additional samples that are representative in both feature and spatial domains. We can achieve this by extracting sample locations from the additional sample locations obtained using the two methods above and forming the final set of sample locations.

For the iPSM method, after the first additional sampling point is generated that has the highest representativeness level, it is added to the existing sample points to generate the uncertainty to help design the second sampling point. After the second sampling point is generated, that has the second highest representativeness level, the first and second sampling points are added to the existing sample points to generate the uncertainty to help design the third new sampling point. The process continues until the uncertainty is lower than the predefined uncertainty value. In this manner, the selected sample locations are ranked according to the reduction in uncertainty in the feature domain, which provides one ranked list.

For the SSA method, the optimization of the total sampling scheme consists of a fixed subscheme with existing observations and a variable,

randomly drawn subscheme, which consists of the additional sampling points that should be optimized. A single transformation from S_i to S_{i+1} is achieved by a random perturbation of the coordinates of one of the additional sampling points from the variable subscheme. Each additional sampling point from the variable subscheme is perturbed in turn, and the changes of mean kriging variance are accepted or rejected according to the Metropolis criterion. So, after the additional sampling scheme has been optimized, we can generate the ranked list from the optimized additional sampling points based on the reduction in the mean kriging variance. The optimized additional points that decrease the mean kriging variance mostly have the highest priority and are ranked the highest in the list. In this manner, the selected sample locations are ranked according to their reduction of the mean kriging variance in the spatial domain, which provides another ranked list.

The final set of locations is generated by combining the highest ranked locations from the above two lists such that the sampling points alternate between those based on uncertainty from the feature domain and those based on uncertainty from the spatial domain. For reason of comparison, we keep the number of sample locations in all three lists the same. This means that the final list contains the locations from the first half of the two lists. The generated set of sample locations is believed to be more effective for soil mapping because it better represents both feature and space domains.

2.5. Mapping and validation

The effectiveness of the method proposed in this study was evaluated by mapping soil organic matter (SOM). Soil mapping by iPSM and ordinary kriging were used to generate the distribution maps of SOM and the prediction uncertainty. Independent samples, which were randomly selected, were used to validate the accuracies of the generated soil maps. Root-mean-square error (RMSE) and the agreement coefficient (AC) between measured and predicted SOM (Willmott, 1984) were used as indices of success.

RMSE and AC can be calculated using the following formulas:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n [Z(x_i) - Z^*(x_i)]^2} \quad (4)$$

$$AC = 1 - \frac{n \times RMSE^2}{PE} \quad (5)$$

Fig. 5. Calculated FPI and NCE for the selected cluster.

Fig. 6. Curve of probability density distribution of uncertainty.

$$PE = \sum_{i=1}^n (|Z(X_i) - \bar{Z}| + |Z^*(X_i) - \bar{Z}|)^2, \quad (6)$$

where $Z(x_i)$ and $Z^*(x_i)$ are the measured and predicted value, $\bar{z}$ is the mean of measured value $Z(x_i)$, and n is the number of measured sample points.

3. Case study of digital soil mapping

3.1. Study area, samples and environmental covariates

We applied the sampling strategy and assessed its effectiveness for digital soil mapping over a study area located in Fuyang city in north-west Zhejiang Province, China (Fig. 2), covering a total area of 299.14 km². The area is under the northern subtropical monsoon climate, with mean annual rainfall of 1410 mm. Vegetation in this area includes a mixture of evergreen and deciduous broadleaf trees. The main soil types are Haplic Acrisols. The topography is characterized by a mountainous landscape with an elevation ranging from 30 to 700 m. The main land-use type is forest.

Fig. 8. Experimental variograms (symbols) and fitted exponential models (solid lines) for SOM and the model parameters.

The data were obtained from a soil fertility survey of the agricultural bureau of Fuyang city. A total of 50 soil samples were collected in 2007, and the content of soil organic matter was measured using the wet oxidation method of Walkley and Black (Nelson and Sommers, 1982). Twenty starting samples (existing samples) were selected randomly from the 50 original samples (Fig. 2), and additional samplings were collected to extend the coverage of the samples using the proposed sampling design. To validate the performance of the sampling design, the experiment was repeated five times, using a different set of starting samples each time by selecting randomly existing samples from the 50 original samples.

Elevation, slope gradient, planform curvature, profile curvature and topographic wetness index (TWI) were the selected environmental predictors (Table 1). Other environmental conditions that are usually considered as the covariates for soil, such as macroclimate, parent material

Fig. 7. Spatial distribution of additional sampling scheme generated based on uncertainty from feature domain (black triangle) and from spatial domain (red dot).

Fig. 9. Maps of SOM from ordinary kriging: estimates (left) and uncertainty (right).

and vegetation condition, were not included because they are generally uniform over this small area.

Terrain is an important factor affecting the spatial variation of SOM. A digital elevation model (DEM) provided terrain information and was derived from the 30-m shuttle radar topographic mission (SRTM). Elevation, slope, profile curvature, plan curvature and topographic wetness index (TWI) were directly derived from the constructed DEM. TWI was calculated according to the following equation (Beven and Kirkby, 1979; Quinn et al., 1995): $w = \ln(a/\tan\beta)$, where a is the cumulative upslope area draining through a point (per unit contour length) and β is the slope gradient at the point. The multiple flow direction strategy, MFD-md, was used to calculate the upslope drainage area (a) (Qin et

al., 2007, 2011; Zhu et al., 2010). The spatial distribution maps of the five environmental variables can be seen in Fig. 3.

3.2. Generation of additional sample sets

3.2.1. Additional samples based on uncertainty from the feature domain

The measurement of environmental similarity for each of the five co-variables was first calculated, and the integrated environmental similarities based on individual covariates were then calculated (Zhu et al., 2015). The distribution map of SOM and the prediction uncertainty with an uncertainty threshold equal to 0.3 were then produced (Fig. 4). The uncertainty throughout the study was relatively high. The

Fig. 10. Spatial distribution of combined additional sampling scheme by abstracting half of the points generated based on uncertainty from both the feature domain and the spatial domain (the labeled number indicates the sampling priority, with 1 highest and 20 lowest).

Fig. 11. Prediction uncertainty maps based on 124 existing samples plus 20 combined samples: using iPSM method with the uncertainty threshold equal to 0.3 (left) and using ordinary kriging (right).

highest uncertainty was in regions with higher elevation and boundary regions, where very sparse existing samples were distributed. Uncertainty was the lowest in the middle valley and plain, where there were many existing soil samples.

To standardize the data for fuzzy classification, elevation, slope and TWI were stretched to 0–100. The planform curvature and profile curvature were stretched to –40 to 40 and –14 to 14, respectively, keeping the 0 value (linear along the plane of curvature) unchanged (Table 2). The details of the preprocessing method have been provided in studies by Zhu et al. (2010). Table 2 lists the ranges of standardization.

The fuzzy exponent, a weighting exponent for fuzziness in the FCM algorithm, is 2.0 in this case study, as recommended in a previous study (Zhu et al., 2010). It was iterated through 2–8 clusters, and the results from the two indices of NCE and FPI were graphed in Fig. 5. The minimum FPI and NCE were obtained with three clusters, so three clusters is the optimal cluster number for the region with uncertainty value >0.3.

By using FCM, the fuzzy membership maps of the environmental cluster classes for each iteration are generated based on the method

developed by Zhang et al. (2016). The first candidate sampling point should be designed in the cluster class with the largest area, within which the probability density distribution of uncertainty is calculated using KDE. The obtained curve of the probability distribution function is shown in Fig. 6.

As Fig. 6 shows, the pixels in the range 0.3–0.325 of peak density in the curve are most frequent and are considered the best candidate sampling locations (Zhang et al., 2016). The first sampling point is selected from these candidate locations based on accessibility and is then added to the existing samples to generate the next uncertainty distribution map. The steps above were repeated to design the second new sampling point. In this manner, the first additional sampling point has the highest representative level. With the generation of other additional sampling points, their representativeness level is gradually reduced. The later a sampling point is generated, the lower its representativeness level.

In total, 20 additional sampling points were generated, and their spatial distribution is shown in Fig. 7. The results show that the additional sampled observations were mostly located where the existing

Fig. 12. Change of root-mean-square error (RMSE) when additional samples were added one-by-one to existing samples.

Fig. 13. Change of agreement coefficient (AC) when additional samples were added one-by-one to existing samples.

sampling density was low and were distributed in the mountainous region. The sampling in this region may be representative, and the prediction uncertainty may be significantly reduced.

3.2.2. Additional samples based on uncertainty from the spatial domain

Distributions of SOM using the Kolmogorov–Smirnov statistic were found to be normal. Because the outliers have significant impact on the kriging variance, the largest SOM values were removed from the samples. An experimental variogram was computed for SOM. To explore the data for any anisotropy, experimental variograms were computed in four directions on the raw data for SOM. With four directions, angular discretization was set to 22.5°; there was no sign of any anisotropy, and the variation was treated as isotropic. The results of the structural analysis of SOM are given in Fig. 8. The spatial dependence of SOM in this area clearly illustrated isotropic behavior. SOM exhibited strong spatial dependence, with a 12.1% ratio of nugget variance to sill variance (<25% means strong spatial dependence; Chien et al., 1997).

The model for SOM was used for ordinary kriging; the smoothed contour map and prediction standard error map (uncertainty map) are presented in Fig. 9. There is considerable variation in SOM in this field; high SOM is distributed in the west section and low SOM in the southeastern portion of the study area. The kriging prediction uncertainty map (Fig. 9, right) shows that the smallest uncertainty is in the middle valley and plain and the largest ones are near the field boundary.

The additional sampling scheme was calculated for SOM, and SSA was used to supplement the existing samples for the minimal mean

kriging variance. To avoid degenerate solutions in which pairs of points become vanishingly close, a minimum distance of 1 unit between any two points was enforced. The initial value of the temperature variable k was adjusted by trial and error so that approximately 95% of the initial perturbations of the system were accepted before the first cooling step. The factor a was set so that this acceptance rate would be reduced to somewhat below 0.1% at the final cooling step. In order to compare the performance with that of additional samples obtained from the feature domain, 20 additional samples were randomly drawn for the optimization. Each point was labeled and perturbed at random but subject to the constraint that it remain within the defined region and not approach any other point by an amount less than the specified threshold. The SSA procedure was repeated five times to ensure that convergence to a global minimum had occurred.

Fig. 7 shows the distribution of the optimized scheme of 20 additional sampling points. Several of these points are very close to the boundaries of the area. In contrast to the schemes for existing observations, the sampling points are dispersed evenly over the region. With the 20 additional points, the mean kriging variance could be reduced from 1.0212 to 0.8868 [unit]² using the criterion of minimization of mean kriging variance. Using SSA for minimization of the kriging variance will result in sampling schemes that explicitly take into account the nature of spatial dependence.

3.2.3. Final set of sample locations combining the above two sample sets

For the 20 samples generated based on the uncertainty from the feature domain, which were ranked according to their reduction in uncertainty, we extracted the first 10 additional samples. For the 20 samples generated based on the uncertainty from the spatial domain, we ranked them according to decreasing mean kriging variance, then extracted the first 10 additional samples. Finally, we combined the two sets of 10 additional samples, alternating between locations based on uncertainty in the feature domain and locations based on uncertainty in the spatial domain, to generate the final ranked list, which also includes 20 samples. Fig. 10 illustrates the spatial distribution map of this final ranked list.

4. Assessment of the efficiency of additional soil sampling

Two sets of samples were collected at the centers of the pixels through the two additional sampling strategies: 20 samples were generated based on uncertainty from the feature domain and 20 samples were generated based on uncertainty from the spatial domain. The first 10 points from each of the two lists were selected and merged in order to generate the third set of 20 samples. The three sets of 20 samples were added to the existing samples to form the pooled samples for further digital soil mapping.

Fig. 14. Box plot of the average of RMSE (left) and AC (right) between measured and kriging predicted SOM with the experiment repeated 5 times.

To validate the effectiveness of the proposed additional sampling strategy, 42 samples were collected regularly across the study area (Figs. 7 and 10) for validation. Two soil mapping methods, iPSM and ordinary kriging, were used to produce the prediction uncertainty maps of SOM (Fig. 11).

The prediction uncertainty maps of SOM are clearly different from Fig. 4 (right) and Fig. 9 (right). With the addition of the 20 combined samples, the prediction uncertainty declined significantly. For the iPSM method, the largest uncertainty value decreased from 0.52 to 0.46. For the ordinary kriging method, the mean kriging variance decreased from 1.0212 to 0.8868 [unit]². The results suggest that the additional sampling approach could improve the reliability of SOM estimates.

Validation using the average of RMSE and AC between measured and Kriging predicted SOM showed that when the additional samples were added one-by-one to the existing samples, the RMSE of the ranked list based on uncertainty from the feature domain decreased from 1.364 to 1.204, and the AC increased from 0.386 to 0.534. The RMSE of the ranked list from the spatial domain decreased from 1.318 to 1.061, and the AC increased from 0.464 to 0.599. The RMSE of the final ranked list based on two uncertainties from both the feature and spatial domain decreased from 1.126 to 0.829, and the AC increased from 0.634 to 0.737 (Figs. 12 and 13). Compared with the additional sampling scheme based on the uncertainty from either the feature domain or from the spatial domain alone, the RMSE of the combined additional sampling scheme was the smallest, and the AC was the largest. It can thus be confirmed that when the number of supplemental samples is the same, the proposed combined method of sampling based on the uncertainty from both feature and spatial domains could improve the accuracy of digital mapping of SOM more. In addition, considering that the study was done with the small samples, future research will be carried out in the data-rich area, using rigorous DSM processes to find what improvement to the uncertainty could be achieved and how additional samples would reduce the overall uncertainty, and if the sample size is sufficient to fully test the approach and draw these conclusions when the more rigorous testing would be applied.

To validate the performance of the sampling design, the experiment was repeated five times, each time with a different set of starting samples (existing samples) by selecting randomly existing samples from the 50 original samples. Ordinary kriging was used as the digital soil mapping method, and the average RMSE and AC were calculated to compare the performances of soil sampling based the three different sets of additional samples. The average of RMSE and AC (Fig. 14) revealed that the average value of RMSE was the lowest and AC was the highest for the combined additional sampling scheme, i.e., the additional sampling scheme based on uncertainties from both feature and spatial domains.

The results indicated that when more samples were added, prediction uncertainty was reduced and prediction accuracy was improved. The combined additional sampling scheme performed better than based on uncertainty from either the feature domain or the spatial domain alone. For the three ranked lists, the samples added first had better representativeness, or higher priority, which can reduce the prediction uncertainty more rapidly than the ones added later. The case study shows that the proposed method is an effective way to design additional samples based on uncertainty from both the feature and spatial domains and integrate them with existing soil samples.

5. Discussion

Existing samples should be taken full advantage of in the process of designing additional samples because they contain underlying knowledge of local soil–environment relations. However, commonly used sampling methods, including classical sampling methods (such as simple random sampling, systematic sampling and stratified sampling)

and sampling based on geostatistics are difficult to integrate with existing samples.

In this paper, we present an effective sampling method to design additional samples based on uncertainty from both the feature and spatial domains. The target of the proposed sampling design is to reduce the prediction uncertainty and improve the practical usage of the digital soil maps. The uncertainty was obtained from two methods: the uncertainty distribution map from the feature domain as defined using the iPSM method and the uncertainty distribution map from the spatial domain, obtained using ordinary kriging.

The degree of success of the iPSM method depends on how the selected environmental covariates can depict the spatial variation patterns of the target soil properties. In this study, environmental covariates, mainly topographical factors, are used. Climate condition, parent material and vegetation are not involved because they are generally similar over such a small study area. In addition, the determination of an uncertainty threshold is user defined, which controls the areas that can be predicted with existing soil samples. A small value for the 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. A large value leads to a poor criterion; more areas can be represented and thus predicted by the existing soil samples (Zhu et al., 2015). The uncertainty threshold here was set to 0.3. It can be further lowered to predict only unvisited locations with very similar environmental conditions. Future studies should examine the impact of different uncertainty thresholds on designing additional sample points. This sampling scheme is stepwise; the most highly representative sample is collected first and then is added into the existing samples to generate the uncertainty maps to help design the next new sampling point; that is, every new sample is determined based on the uncertainty map generated from the last iterative process.

For the kriging method, the optimization criterion of minimization of the mean ordinary kriging variance was used. There are some other optimization criteria dealing with optimizing the quality of the kriging interpolator, like minimizing the maximum ordinary kriging variance (Van Groenigen, 2000), minimizing the variance of universal kriging (Brus and Heuvelink, 2007) and minimizing the co-kriging variance when a model of co-regionalization can be defined (McBratney and Webster, 1983). Future work should compare optimized sampling schemes using different optimization criteria.

6. Conclusions

This article presented a new sampling scheme focusing on the design of additional soil samples based on uncertainty from both the feature domain and the spatial domain by taking existing samples into account. Individual predictive soil mapping (iPSM) was introduced to produce the prediction uncertainty map of SOM. Based on this, additional sample points in the feature space were designed. Kriging of geostatistics was used to produce the prediction uncertainty map of SOM. SSA was then used to design the additional sample points in the spatial domain. The final set of additional samples was then obtained by combining the first half of the ranked samples from the two methods above. The performance of three different additional sampling schemes—additional sampling based on uncertainty from the feature domain, from the spatial domain and from a combination of both feature and spatial domains—was compared by adding the three groups of additional samples respectively to the existing samples for digital soil mapping of SOM. Assessment of the efficiency of additional soil sampling showed that the mapping accuracy increased and prediction uncertainty decreased with the addition of more samples. The samples added first have better representativeness and higher priority, which can reduce the prediction uncertainty more rapidly than the ones added later. The performance of the combined additional sampling scheme was better than that based on uncertainty from either only the feature domain or the spatial domain. The results indicate that the

prediction uncertainty from both the feature domain and from the spatial domain can be used to guide the efficient design of additional sampling more effectively to improve the accuracy of the predicted soil map.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (2016YFD0201200), the National Natural Science Foundation of China (grant no. 40701007), the International S&T Cooperation Program of China (2010DFA92720-27), the Program of International S&T Cooperation, Ministry of Science and Technology of China (project no. 2010DFB24140), the National Basic Research Program of China (project no. 2015CB954102), Zhejiang Provincial Natural Science Foundation of China (grant no. Y3090198), Key Project of Zhejiang Provincial Education Department (grant no. Z201121260), the Fundamental Research Funds for the Central Universities, Natural Science Research Program of Jiangsu (14KJA170001), Priority Academic Program Development of Jiangsu Higher Education Institutions, and the Fundamental Research Funds for the Central Universities. The support received by A-Xing Zhu through the Vilas Associate Award, the Hammel Faculty Fellow, and the Manasse Chair Professorship from the University of Wisconsin-Madison and through the “One-Thousand Talents” Program of China is greatly appreciated.

References

- Aarts, E., Korst, J., 1989. Simulated Annealing and Boltzmann Machines—A Stochastic Approach to Combinatorial Optimization and Neural Computing. Wiley, New York.
- Arrouays, D., Grundy, M.G., Hartemink, A.E., et al., 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. *Adv. Agron.* 125, 93–134.
- Beven, K.J., Kirkby, N.J., 1979. A physically based variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* 24, 43–69.
- Bezdek, J.C. (Ed.), 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.
- Boydell, B., McBratney, A.B., 1999. Identifying Potential within-Field Management Zones from Cotton Yield Estimates. In: Stafford, J.V. (Ed.), Proceedings of the Second European Conference on Precision Agriculture. Odense Congress Center, Odense, Denmark, pp. 331–341.
- Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* 138 (1–2), 86–95.
- Burgess, T.M., Webster, R., 1980. Optimal interpolation and isarithmic mapping of soil properties I: the semi-variogram and punctual kriging. *J. Soil Sci.* 31 (2), 315–331.
- 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.
- Chien, Y.J., Lee, D.Y., Guo, H.Y., et al., 1997. Geostatistical analysis of soil properties of mid-west Taiwan soils. *Soil Sci.* 162 (4), 291–297.
- Christakos, G., Olea, R.A., 1992. Sampling design for spatially distributed hydrogeologic and environmental processes. *Adv. Water Resour.* 15, 219–237.
- Delmelle, E.M., Goovaerts, P., 2009. Second-phase sampling designs for non-stationary spatial variables. *Geoderma* 153 (1–2), 205–216.
- Fridgen, J.J., Kitchen, N.R., Sudduth, K.A., et al., 2004. Management zone analyst (MZA): software for subfield management zone delineation. *Agron. J.* 96 (1), 100–108.
- Gessler, P.E., Moore, I.D., McKenzie, N.J., et al., 1995. Soil-landscape modelling and spatial prediction of soil attributes. *Int. J. Geogr. Inf. Syst.* 9, 421–432.
- Haining, R., 2003. Spatial data analysis. Cambridge University Press.
- Hudson, B.D., 1992. The soil survey as a paradigm-based science. *Soil Sci. Soc. Am. J.* 56, 836–841.
- Juang, K.W., Liao, W.J., Liu, T.L., et al., 2008. Additional sampling based on regulation threshold and kriging variance to reduce the probability of false delineation in a contaminated site. *Sci. Total Environ.* 89, 20–28.
- Kirkpatrick, S., Gellat, C.D., Vecchi, M.P., 1983. Optimisation by simulated annealing. *Science* 220, 671–680.
- Lark, R.M., Stafford, J., 1997. Classification as a first step in the interpretation of temporal and spatial variation of crop yield. *Ann. Appl. Biol.* 130, 111–121.
- Mallavan, B.P., Minasny, B., McBratney, A.B., 2010. Homosoil, a Methodology for Quantitative Extrapolation of Soil Information Across the Globe. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Bridging Research, Environmental Application and Operation. Springer, Dordrecht, pp. 137–149.
- McBratney, A.B., Webster, R., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables. *Comput. Geosci. UK* 7, 331–334.
- McBratney, A.B., Webster, R., 1983. Optimal interpolation and isarithmic mapping of soil properties: V. Co-regionalization and multiple sampling strategy. *J. Soil Sci.* 34, 137–162.
- McBratney, A.B., Webster, R., 1986. Choosing functions for semivariogram of soil properties and fitting them to sampling estimates. *J. Soil Sci.* 37 (4), 617–639.
- McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. *Geoderma* 117, 3–52.
- McKenzie, N.J., Austin, M.P., 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. *Geoderma* 57, 329–355.
- Moore, I.D., Gessler, P.E., Nielsen, G.A., et al., 1993. Soil attribute prediction using terrain analysis. *Soil Sci. Soc. Am. J.* 57, 443–452.
- Moran, C.J., Bui, E., 2002. Spatial data mining for enhanced soil map modelling. *Int. J. Geogr. Inf. Sci.* 16, 533–549.
- Mueller, T.G., Pierce, F.J., 2003. Soil carbon maps: enhancing spatial estimates with simple terrain attributes at multiple scales. *Soil Sci. Soc. Am. J.* 67 (1), 258–267.
- Nelson, D.W., Sommers, L.E., 1982. Total Carbon, Organic Carbon and Organic Matter. In: Page, A.L., Miller, R.H. (Eds.), Methods of Soil Analysis. Part 2. Chemical and Microbiological Properties. American Society of Agronomy, Madison.
- Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992. Soil pattern recognition with fuzzy-C-means application to classification and soil-landform interrelationships. *Soil Sci. Soc. Am. J.* 56 (2), 505–516.
- Oliver, M.A. (Ed.), 2010. Geostatistical Applications for Precision Agriculture. Springer, Dordrecht.
- Qin, C.Z., Zhu, A.X., Pei, T., et al., 2007. An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. *Int. J. Geogr. Inf. Sci.* 21 (4), 443–458.
- Qin, C.Z., Zhu, A.X., Pei, T., et al., 2011. An approach to computing topographic wetness index based on maximum downslope gradient. *Precis. Agric.* 12 (1), 32–43.
- Quinn, P.F., Beven, K.J., Lamb, R., 1995. The $\ln(a/\tan\beta)$ index: how to calculate it and how to use it within the TOPMODEL framework. *Hydrol. Process.* 9, 161–182.
- Sacks, J., Schiller, S., 1988. Spatial Designs. In: Gupta, S.S., Berger, J.O. (Eds.), Statistical Decision Theory and Related Topics IV vol. 2. Springer-Verlag, New York, NY, pp. 385–399.
- Seaman, D.E., Powell, R.A., 1996. An evaluation of the accuracy of kernel density estimators for home range analysis. *Ecology* 77 (7), 2075–2085.
- Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, United Kingdom 175 pp.
- Simbahan, G.C., Dobermann, A., 2006. Sampling optimization based on secondary information and its utilization in soil carbon mapping. *Geoderma* 133, 345–362.
- Van Groenigen, J.W., 2000. The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma* 97, 223–236.
- Van Groenigen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. *J. Environ. Qual.* 27, 1078–1086.
- Van Groenigen, J.W., Siderius, W., Stein, A., 1999. Constrained optimization of soil sampling for minimization of the kriging variance. *Geoderma* 87, 239–259.
- Webster, R., 1985. Quantitative spatial analysis of soil in the field. *Adv. Soil Sci.* 3, 1–70.
- Webster, R., Burgess, T.M., 1980. Optimal interpolation and isarithmic mapping. III: changing drift and universal kriging. *J. Soil Sci.* 31, 505–524.
- Webster, R., Burgess, T.M., 1984. Sampling and bulking strategies for estimating soil properties in small regions. *J. Soil Sci.* 35, 127–140.
- Webster, R., Oliver, M.A., 1990. Statistical Methods in Soil and Land Resource Survey. Oxford University Press, Oxford.
- Willmott, C.J., 1984. On the Evaluation of Model Performances in Physical Geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial Statistics and Models. D. Reidel Publ., Dordrecht, the Netherlands, pp. 43–460.
- Zhang, S.J., Zhu, A.-X., Liu, J., et al., 2016. An heuristic uncertainty directed field sampling design for digital soil mapping. *Geoderma* 267, 123–136.
- Zhu, A.X., Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. *Can. J. Remote. Sens.* 20 (4), 408–418.
- Zhu, Z., Stein, M.L., 2006. Spatial sampling design for prediction with estimated parameters. *J. Agric. Biol. Environ. Stat.* 11, 24–44.
- Zhu, A.X., Band, L.E., Vertessy, R., et al., 1997. Derivation of soil properties using a soil land inference model (SoLIM). *Soil Sci. Soc. Am. J.* 61, 523–533.
- Zhu, A.X., Yang, L., Qin, C.Z., et al., 2010. Construction of quantitative relationships between soil and environment using fuzzy c-means clustering. *Geoderma* 155 (3–4), 166–174.
- Zhu, A.X., Liu, J., Du, F., et al., 2015. Predictive soil mapping with limited sample data. *Eur. J. Soil Sci.* 1–13.