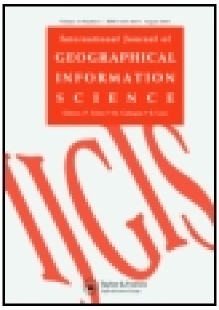

International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping

Lin Yang [a] , A-Xing Zhu [a b] , Feng Qi [c] , Cheng-Zhi Qin [a] , Baolin Li [a] & Tao Pei [a]

[a] State Key Laboratory of Resources and Environment Information System, Institute of Geographical Sciences and Resources Research, Chinese Academy of Sciences, Beijing, PR China

[b] Department of Geography, University of Wisconsin-Madison, Madison, WI, USA

[c] School of Environmental and Life Sciences, Kean University, Union, NJ, USA

Available online: 28 Feb 2012

To cite this article: Lin Yang, A-Xing Zhu, Feng Qi, Cheng-Zhi Qin, Baolin Li & Tao Pei (2012): An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping, International Journal of Geographical Information Science, DOI:10.1080/13658816.2012.658053

To link to this article: http://dx.doi.org/10.1080/13658816.2012.658053

International Journal of Geographical Information Science
i*First*, 2012, 1–23

An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping

Lin Yang[a], A-Xing Zhu[a,b]*, Feng Qi[c], Cheng-Zhi Qin[a], Baolin Li[a] and Tao Pei[a]

[a]*State Key Laboratory of Resources and Environment Information System, Institute of Geographical Sciences and Resources Research, Chinese Academy of Sciences, Beijing, PR China;* [b]*Department of Geography, University of Wisconsin-Madison, Madison, WI, USA;* [c]*School of Environmental and Life Sciences, Kean University, Union, NJ, USA*

(*Received 7 August 2011; final version received 12 January 2012*)

Sampling design plays an important role in spatial modeling. Existing methods often require a large amount of samples to achieve desired mapping accuracy, but imply considerable cost. When there are not enough resources for collecting a large set of samples at once, stepwise sampling approach is often the only option for collecting the needed large sample set, especially in the case of field surveying over large areas. This article proposes an integrative hierarchical stepwise sampling strategy which makes the samples collected at different stages an integrative one. The strategy is based on samples' representativeness of the geographic feature at different scales. The basic idea is to sample at locations that are representative of large-scale spatial patterns first and then add samples that represent more local patterns in a stepwise fashion. Based on the relationships between a geographic feature and its environmental covariates, the proposed sampling method approximates a hierarchy of spatial variations of the geographic feature under concern by delineating natural aggregates (clusters) of its relevant environmental covariates at different scales. The natural occurrence of such aggregates is modeled using a fuzzy *c*-means clustering method. We iterate through different numbers of clusters from only a few to many more to be able to reveal clusters at different spatial scales. At a particular iteration, locations that bear high similarity to the cluster prototypes are identified. If a location is consistently identified at multiple iterations, it is then considered to be more representative of the general or large-scale spatial patterns. Locations that are identified less during the iterations are representative of local patterns. The integrative stepwise sampling design then gives higher sampling priority to the locations that are more representative of the large-scale patterns than local ones. We applied this sampling design in a digital soil mapping case study. Different representative samples were obtained and used for soil inference. We started with samples that are the most representative of the large-scale patterns and then gradually included the samples representative of local patterns. Field evaluation indicated that the additions of more samples with lower representativeness lead to improvements of accuracy with a decreasing marginal gain. When cost-effectiveness is considered, the representative grade could provide essential information on the number and order of samples to be sampled for an effective sampling design.

Keywords: spatial sampling; fuzzy clustering; digital soil mapping; SoLIM

*Corresponding author. Email: azhu@wisc.edu

ISSN 1365-8816 print/ISSN 1362-3087 online

http://dx.doi.org/10.1080/13658816.2012.658053
http://www.tandfonline.com

1. Introduction

Spatial sampling is an essential step in modeling and mapping the spatial distribution of many geographic features. A few geographic features, such as vegetation cover, can be observed and mapped using remote-sensing techniques (Miao *et al.* 2012). For many others such as soil properties or soil classes, however, the only feasible way to map their spatial variations is to collect point sample data in the field and conduct interpolation or predictive mapping (Zhu *et al.* 2001, 2008). Sampling design, therefore, has a significant impact on the accuracy of the mapping results (Brus and de Gruijter 1997, Gregoire and Valentine 2008, Brus and Noij 2008, Heim *et al*. 2009). In this article, we are talking about geographic features which usually vary continuously over space and the spatial distribution of which is highly related with their environmental factors (environmental covariates).

Classical sampling methods (such as simple random sampling, systematic sampling, and stratified sampling) are widely used for geographic survey as the designs are the most straightforward (Cochran 1977; Kish 1985; Grinand *et al*. 2008). These sampling strategies are mainly based on probability theory that allows unbiased estimates of the statistical parameters, such as population or mean (Minasny and McBratney 2006). Spatial interpolation is usually conducted when the collected sampling points are used for mapping geographic features. However, it often requires a large number of samples to comprehensively capture the spatial variation of geographic features. It thus imposes a high cost on studies that involve a large area since field sampling is often labor-intensive and expensive.

Sampling based on geostatistics has also been adopted in many studies, which 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 configuration of sample locations is designed to minimize the average or maximum of the prediction variance over a region, which is determined through a semi-variogram. The construction of a semi-variogram usually requires a prior sampling which also requires a large number of samples (Simbahan and Dobermann 2006) as a random or systematic sampling would do. It has been simulated that at least 100–150 samples are needed to estimate a semi-variogram effectively (Webster and Oliver 1992). Another limitation of this method of sampling is that it is often difficult to verify the stationary assumption, the basis of the geostatistical method, in many complicated field conditions.

Recent applications have started to make use of readily available secondary information to assist sampling designs (Minasny and McBratney 2006, Brus and Heuvelink 2007, Zhu *et al*. 2010, Yang *et al*. 2010, Qin *et al*. 2011b). Secondary information refers to variables which are often called environmental covariates that spatially correlate with the target geographic features, and can be obtained from satellite imagery, aerial photography, digital elevation models, soil sensors, and so on. Although environmental covariates can be helpful when designing sampling schemes, most of the sampling designs are still derived using classical sampling methods or geostatistical techniques (Minasny and McBratney 2006, Minasny *et al.* 2007), and thus possess the limitations mentioned in the above two paragraphs.

When costs associated with field sampling become impractically high in some applications, one may consider dissipating the costs by collecting a number of required samples using resources available for the time being and augmenting these with further samples when more resources become accessible in the future. This is achieved via stepwise or multiple-stage sampling (Beckett 1968, Cochran 1977, Mckenzie and Ryan 1999). An important issue for stepwise sampling is how to design samples for each stage effectively. This article presents an efficient stepwise sampling strategy, which is to collect samples that are most highly representative of the area of interest first, and then collect samples that are less representative. Samples collected later thus provide valid complementary information on the spatial variation of the geographic features of the samples collected first. All samples in a hierarchy of representativeness grade are designed at the beginning, which is why we call this method an integrative sampling design. Section 2 introduces the basic idea and procedure of the method. It is then illustrated with a case study on sampling for digital soil mapping. The proposed method is assessed and the results are discussed in the end.

2. Methodology

2.1. The basic idea

The spatial distribution of geographic features is often highly correlated with the distribution of their environmental covariates, and usually is a result of the spatial process of formation and development of geographic features involving these environmental covariates. For example, soil has long been considered as the result of the interaction of its formative environment, including climate, parent material, terrain, and vegetation conditions (Jenny 1980, Hudson 1992). Therefore, soil distribution has been mapped from the distribution of environmental covariates such as bedrock, slope gradient and aspect, drainage conditions, and so on.

Inherent spatial autocorrelation of the environmental covariates relevant to the distribution of a target geographic feature could lead to the development of natural aggregates or clusters of the target geographic feature spatially. As spatial autocorrelation of environmental covariates happens across different scales, naturally occurring clusters of the geographic feature under concern may exist at multiple scales in a hierarchy: from those representing large-scale spatial patterns (existing at a large spatial extent) to those representing very local ones. For instance, in landform classification, large ridges, side slopes, and valleys may be the main categories on a large scale, existing on the top of a hierarchy across scales. Small draws, convex knobs, on the main ridges, along the side slopes or in the valleys, on the other hand, represent local variations and thus are categories on a lower tier in the hierarchy (MacMillan and Shary 2009).

The basic idea of our stepwise sampling design is to extract samples top-down through such hierarchy of spatial clusters. When sampling at a particular scale, samples are to be chosen based on how representative they are to the prototypes of the clusters/classes. In a previous study (Zhu *et al*. 2010), we developed a methodology to generate clusters from environmental covariates to approximate the naturally occurring aggregates in the target geographic feature. The number of clusters generated has been found in a previous study (Yang *et al*. 2010) to correlate with different spatial scales in a hierarchy generally. While a few clusters could be generated to represent the main patterns on a large scale on the top of the hierarchy, increasing the number of clusters leads to the emergence of new patterns which occur at finer scales. If we take landform classification as an illustration again, three clusters could be generated to represent the main ridge, side slope, and valley features. When we increase the number of clusters to four or five, the clusters generated could include now typical ridge, valley, as well as some local patterns such as a draw along the main slope, a convex shoulder, or a relatively flat ridge top. In other words, the dominant large-scale patterns could always be detected when the number of clusters is small and large, but small-scale local patterns can only be revealed when the number of clusters is larger. This prompted us to approximate the hierarchy of patterns based on the repeated occurrence of such patterns when we change the number of clusters to be generated. We can thus design samples representing large-scale variations by identifying locations which are consistently showing up as part of the large-scale patterns and samples representing local-scale variations by identifying locations which are only showing up as part of local-scale patterns in a stepwise manner based on clustering analysis of multiple numbers of clusters.

2.2. The sampling design

We employ the following three measures to organize the sample candidates into a hierarchy based on which sampling will be performed: representative grade, representative area, and representative level (membership). Representative grade is the frequency with which a location emerges in a cluster analysis below. Representative area refers to the spatial coverage of a given pattern that indicates certain class configuration under one representative grade, and the representative level is the level or degree of specific locations belonging to a given pattern, which is determined by membership values to classes in the pattern. Based on this idea, the sampling design consists of two steps as follows.

2.2.1. Determination of hierarchy

2.2.1.1. Computation of representative grade. The hierarchy of a representative grade for candidate samples is determined through a cluster analysis using a fuzzy *c*-means (FCM) classifier (Bezdek *et al.* 1984). FCM is employed to generate representative grades firstly based on the environmental variables (covariates) relevant to the target geographic features. FCM is a classifier that first optimally partitions a data set (such as the environmental database) into a given set of classes and then computes the membership of an instance associated with each class (Bezdek *et al.* 1984). The FCM classifier is performed in multiple iterations with each iteration identifying a different number of clusters. For example, FCM can be performed to identify 4–10 clusters for a given area. In this case, seven iterations will be performed with Iteration 1 for identifying four clusters and Iteration 2 for five clusters, and so on. Each of these iterations is referred to as 'x cluster iteration,' where x is the number of clusters to be extracted. The rule to choose a range of clustering iterations is to choose a range which could cover both large-scale and local patterns of spatial variation of a geographic feature in the study area. To find an appropriate range, one can test a number of ranges of clustering iterations and look into the scale ranges of the generated environmental clusters. To test how different ranges of clustering iterations will influence the designed sampling scheme, we conducted a sensitivity analysis in the case study in Section 3.5.

The results of FCM under x cluster iteration are as follows: (1) centroids of the x classes, a cluster centroid is a vector of values of environmental covariates, which is computed as the mean of all input data weighted by their membership to the given class and (2) fuzzy membership maps of the study area associated with x classes, in which fuzzy membership values range continuously between 0 and 1, with 0 indicating no membership or completely different from the class prototype and 1 indicating a full membership or total resemblance to the class prototype. Representative locations of a particular class are those bearing high membership values to this class. Operationally, we can define a membership threshold as a cut off value which is usually greater than 0.5 (Yang *et al*. 2010) and then reclassify the fuzzy membership maps into only 0 and 1 values, with 1 indicating representative locations (of any class) and 0 indicating nonrepresentative ones. For each iteration, all the fuzzy membership maps are converted to 0–1 maps in this way for identifying representative locations for each class. There is no certain objective way to define a threshold. We conducted a sensitivity analysis on the membership threshold in the case study in Section 3.5.

Representative grade here refers to the number of times a location emerges as a representative location across all iterations. It is supposed that locations consistently showing up above a membership threshold in clusters at multiple clustering iterations most possibly represent large-scale spatial patterns. If a particular location is repeatedly labeled as being representative at multiple iterations, this point is thus likely a representative point of a certain large-scale pattern and has a high representative grade. On the other hand, a location that is labeled as being representative at only one or two iterations is more likely a representative of a pattern that exists at some local level and therefore the location has a low representative grade. Based on this notion, the reclassified 0–1 valued maps of all iterations are overlaid to produce a frequency map, with values at each location representing the representative grade, the number of times that the location is considered representative of a cluster, a direct indicator of the representativeness of the location in the scale hierarchy from the top-down perspective.

2.2.1.2. Determination of representative area and representative level. It is possible to have locations with the same representative grade representing different environmental clusters. For example, a particular location with a representative grade of 2 might be representative of ‘class 2’ during the three cluster iteration and ‘class 4’ during the four cluster iteration. Another location with the same representative grade 2 might have been selected from ‘class 3’ during the three cluster iteration and then from ‘class 4’ during the five cluster iteration.

We introduced the concept of ‘environmental cluster chain’ to further examine the patterns of class configuration under each representative grade. In this way the different patterns at the same level on the hierarchy can be distinguished. This environmental cluster chain is a chain which records the specific list of classes and the associated cluster iteration. For the example above, the environmental cluster chain for the first location is noted as 3Class2–4Class4 and that for the second location is 3Class3–5Class4.

The representative area and the representative level are determined with reference to an environmental cluster chain. The representative area for a chain is the number of locations (pixels) which is representative for all classes in the chain and the representative level is the average of the fuzzy membership values in all classes in the chain. This is a measure of the average representativeness of the location to all classes it belongs to.

2.2.2. Selection of samples

Considering these three indices, samples are then selected according to the following guidelines:

- Samples are selected at locations with priority from the highest representative grade to lower ones.
- For locations with the same representative grade, those whose environmental cluster chains cover a larger area have the higher priority.
- For each environmental cluster chain, priority is given to samples with higher average membership values.
- The number of samples for each representative grade can be one or more, which could be decided according to the sampling budget.

Figure 1 illustrates the implementation flowchart.

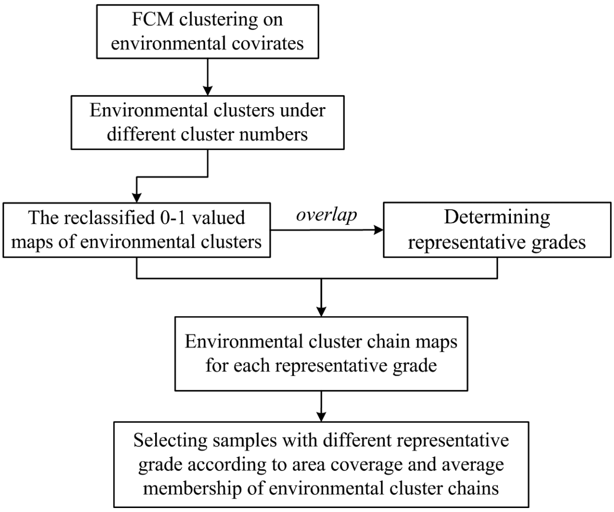
FCM clustering on environmental covirates
Environmental clusters under different cluster numbers
The reclassified 0-1 valued maps of environmental clusters
overlap
Determining representative grades
Environmental cluster chain maps for each representative grade
Selecting samples with different representative grade according to area coverage and average membership of environmental cluster chains

Figure 1. Sampling design flowchart.

3. A case study on digital soil mapping

3.1. Study area and environmental data

We applied the sampling strategy and assessed its effectiveness in a digital soil mapping case study. The study area is 60 km^2 in size, and is located in Heshan farm of Nenjiang County in Heilongjiang Province, China (Figure 2). The elevation in the area ranges from 276 to 363 m with a gentle gradient. The soils in the area are formed on deposits of silt loam loess and have a thick *A*-horizon with high organic matter content. Crops in the watershed

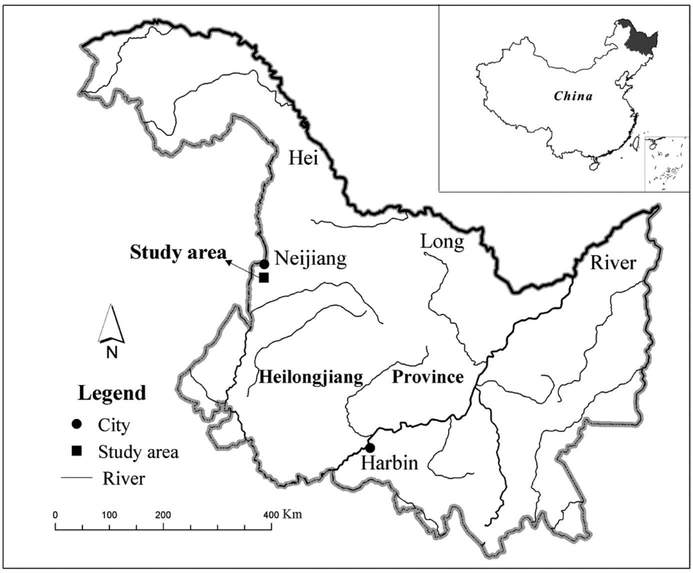

Figure 2. Location of the study area.

at present are generally limited to soybean and wheat (Miao *et al.* 2011). The land use and soil management is generally uniform across the study area.

Four topographic variables (slope gradient, planform curvature, profile curvature, and topographic wetness index) were used in this study as environmental covariates. The reason for including only topographic variables is that other environmental conditions which are usually considered as the covariates for soil, such as macroclimate, vegetation conditions, and parent materials, are overall uniform over this small area (Zhu *et al*. 2010).

A 10-m resolution Digital Elevation Model (DEM) (Figure 3) was created from a 1:10,000 topographic map. Data layers on slope gradient, planform curvature, and profile curvature were derived from the DEM. Topographic wetness index was calculated according to the following equation (Beven and Kirkby 1976; 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. Because the relief of our study area is gentle and the floodplain is wide, we used the multiple flow direction strategy MFD-md to calculate the upslope drainage area (a) (Qin *et al*. 2007, Zhu *et al*. 2010, Qin *et al*. 2011a).

Environmental data layers were preprocessed before FCM clustering to remove outliers and to standardize the ranges of each layer. The outliers were data values with low frequency located at the ends of the data histograms, which were mostly errors introduced during the creation of the DEM and would have a strong impact on the clustering results.

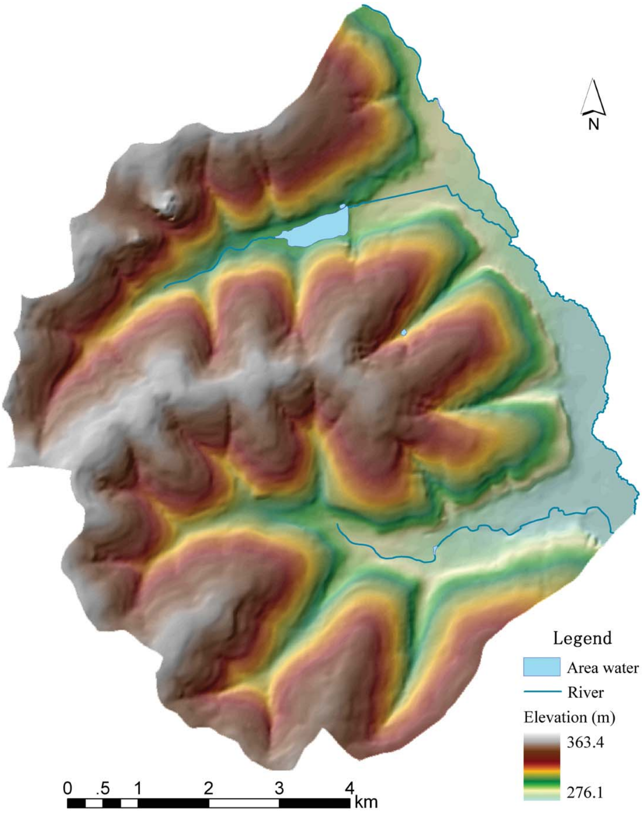

Figure 3. DEM of the study area.

These values were then replaced with values next to these extremes to preserve the integrity of data volume. For standardization, elevation, slope gradient, and topographic wetness index were stretched to 0–100. The profile curvature and planform curvature were stretched to –50 to 50, keeping the 0 value (means linear along the plane of curvature) unchanged. Specifics of the preprocessing method were detailed in our previous studies (Zhu *et al*. 2010).

3.2. Sampling design

FCM clustering was performed on the four environmental data layers. 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). We iterated through 8–13 clusters to cover both large-scale and local patterns of soil types in this study area. Fuzzy membership maps of the environmental cluster classes for each iteration were generated. We also tested the other two ranges of clustering iterations for our study area to test how different ranges of clustering iterations will influence the designed sampling scheme in Section 3.5.

3.2.1. Determination of hierarchy

We used a membership threshold of 0.6 in the case study to identify representative locations (pixels) of the environmental clusters. We also experimented with different threshold values and performed a sensitivity analysis. The results and discussion on these are provided in Section 3.5. Figure 4a shows the fuzzy membership map of environmental cluster class 3 for the 10 cluster iteration is extracted and Figure 4b is the two-valued map showing the representative locations of this class with a threshold of 0.6.

The 0–1 reclassification maps of all cluster classes for a particular iteration were combined first to produce an overall map for that iteration. An 'OR' operation was used to

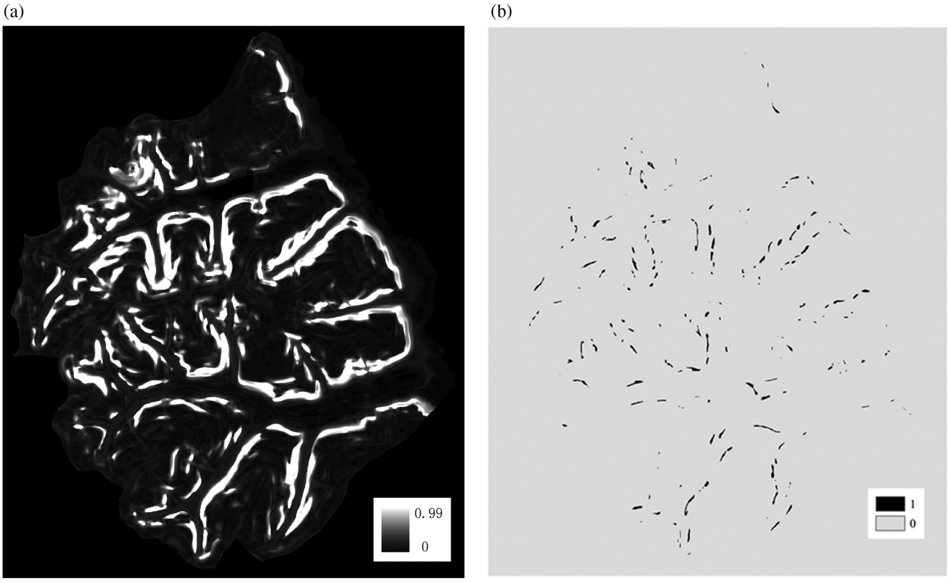

Figure 4. A fuzzy membership map (a) and the reclassified two-valued map (b).

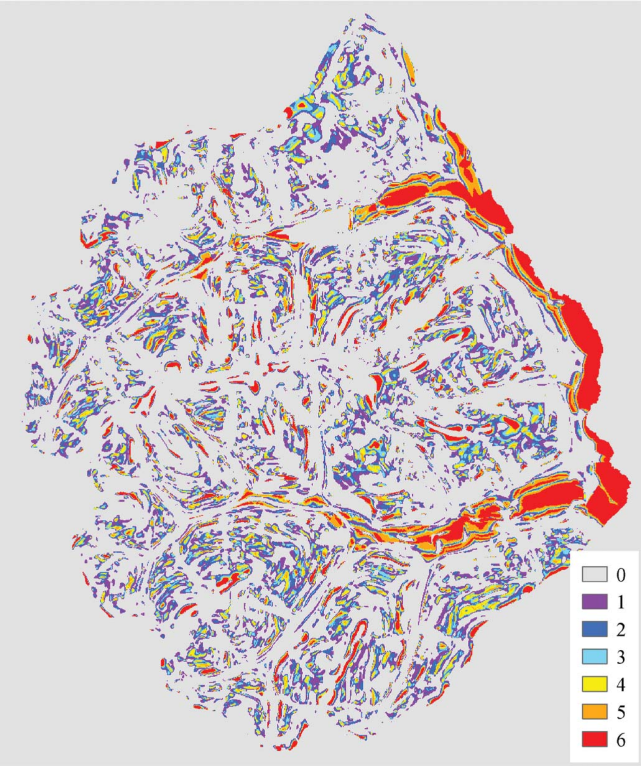

Figure 5. The representative grade map.

combine the individual maps. The results from all six iterations (8–13 clusters) were then overlaid to produce the representative grade map shown in Figure 5. As we iterated through six cluster numbers, a value 6 on the resulting map means that the pixel is considered as a representative location of environmental clusters during all six iterations. The minimum value of a pixel is 0, which means the pixel is not a representative location of any environmental cluster during any iteration.

3.2.2. Sampling

Three hundred and twenty-two different environmental cluster chains were extracted from the representative grade map. Sometimes an environmental cluster chain with a lower representative grade can be the subset of (or is completely embedded in) a longer chain with a higher representative grade. In this case, we eliminate the shorter chain from the list based on the consideration that the longer chain should have covered the pattern represented by the shorter chain but with a higher representative grade. For example, all classes in the chain '9Class7–10Class3–11Class7–12Class5–13Class18 (ID 75)' with representative grade 5 are included in the chain '8Class4–9Class7–10Class3–11Class7–12Class5–13Class1 (ID 10)' with a representative grade 6. We thus eliminate chain ID 75. After this elimination step, 35 environmental cluster chains were left (Table 1). An environmental cluster class ID of N/A in Table 1 means that the pixel is not a representative location of any environmental class under

Table 1. The environmental cluster chains after elimination.

Representative grade	Environmental cluster chain ID	Class ID under iterations 8, 9, 10, 11, 12, and 13 in turn						Number of environmental cluster chains
6	18	1	2	5	1	11	6	9
	21	5	5	2	6	10	3	
	10	4	7	3	7	5	1	
	30	2	9	9	4	6	5	
	1	2	9	9	4	6	2	
	2	7	8	10	8	7	7	
	7	3	1	4	10	8	12	
	13	6	4	8	2	4	4	
	17	6	4	8	3	2	9	
5	8	0	3	6	5	3	13	7
	15	8	0	7	9	1	11	
	20	8	3	6	5	0	13	
	19	5	5	0	3	2	9	
	9	8	3	6	0	1	13	
	32	1	2	5	1	0	5	
	3	7	8	10	0	12	7	
4	12	3	1	4	0	0	8	4
	29	8	3	0	9	1	0	
	22	6	6	0	2	0	4	
	11	6	6	0	0	4	4	
3	5	0	0	0	11	9	10	12
	14	0	4	7	0	0	11	
	28	0	0	0	8	7	8	
	33	3	0	0	0	3	8	
	26	4	7	0	0	0	12	
	16	0	4	7	2	0	0	
	31	1	0	0	0	6	5	
	6	0	6	1	0	0	4	
	4	0	0	7	9	12	0	
	25	0	8	7	0	12	0	
	27	3	0	0	0	7	8	
	35	0	0	6	0	3	8	
2	23	7	0	1	0	0	0	3
	24	0	0	0	2	12	0	
	34	8	8	0	0	0	0	

the corresponding iteration. Figure 6 shows the maps of environmental cluster chains. Environmental cluster chain with pixels less than 200 is not picked in the figure, because it is not clear to display when each map of environmental cluster chain is small. However, if all of the environmental cluster chains for one representative grade are with pixels less than 200, then we keep the environmental cluster chain with the biggest area for this representative grade for integrity of the representative grade. Figure 7 shows an example of the average membership map of an environmental cluster chain.

Based on the sampling design guidelines in Section 2.2, we drew samples in the order from the highest representative grade 6 to the lowest representative grade 2. For those environmental cluster chains whose coverage is very limited, we consider them representing some untypical features and excluded them from sampling. In our experiment, no sample

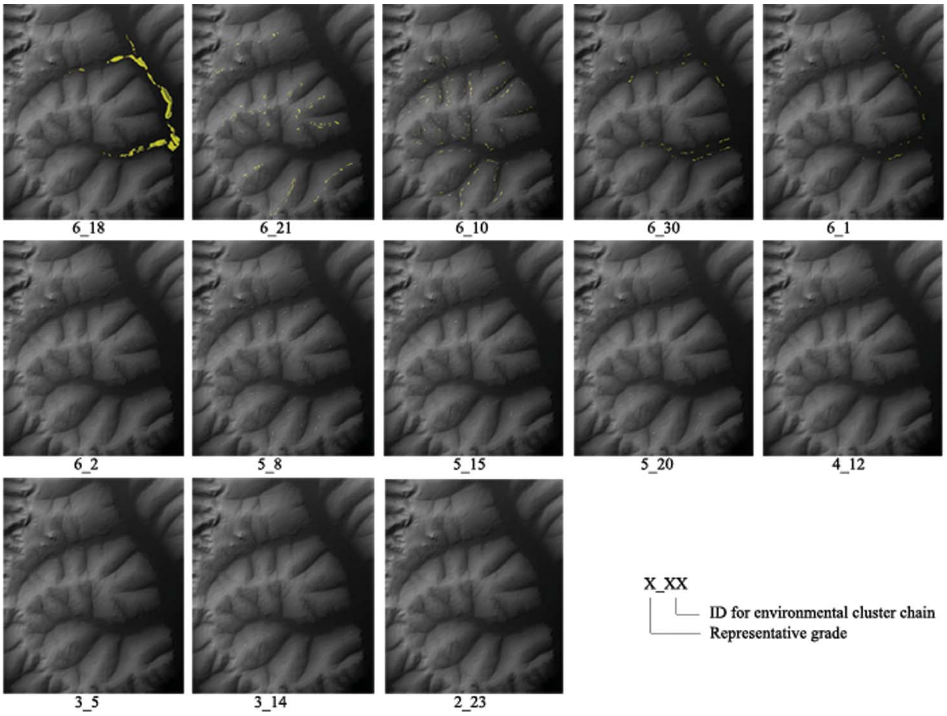

Figure 6. Environmental cluster chains through 8–13 clusters.

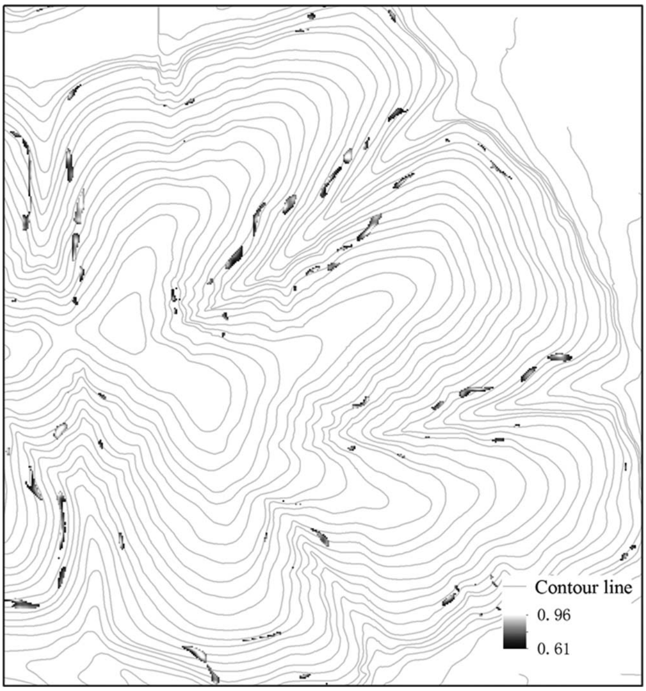

Figure 7. Part of the average membership map for chain ID 10.

Table 2. The order and number of the selected samples.

Representative grade	Environmental cluster chain ID	Number of pixels	Number of samples in order	Total number of samples
6	18	18, 873	3	21
	21	6337	6	
	10	5775	9	
	30	1921	12	
	1	877	15	
	2	231	18	
	7	183	21	
	13	14		
	17	3		
5	8	1078	24	15
	15	613	27	
	20	451	30	
	19	47	33	
	9	45	36	
	32	10		
4	3	4		
	12	49	39	3
	29	19		
	22	13		
	11	1		
3	5	339	42	27
	14	324	45	
	28	193	48	
	33	67	51	
	26	58	54	
	16	46	57	
	31	43	60	
	6	40	63	
	4	31	66	
	25	20		
	27	20		
	35	1		
	23	13		0
	2	24	8	
	34	2		

was drawn from environmental cluster chains with less than 20 pixels (such as IDs 13, 17, 32, and 3). As a result, there is no sample selected from representative grade 2. For each of the remaining environmental cluster chains, we sampled three points with the highest average membership values (representative level). The final sampling order and numbers are listed in Table 2. Stepwise sampling should be following the top-down order in the table. For example, the first row and the fourth column of the table mean that the three points selected for Chain 18 should be collected in field first, then three points for Chain 21, 10, and so on. There is no sampling order between three samples for each environmental cluster chain.

3.3. Sampling and sampled soil types

Soil was sampled at the centers of the pixels selected by the sampling design. Due to the inaccessibility of some points in the original sampling design, field investigations at 48 samples were eventually made (Figure 8). Based on the assumption that samples

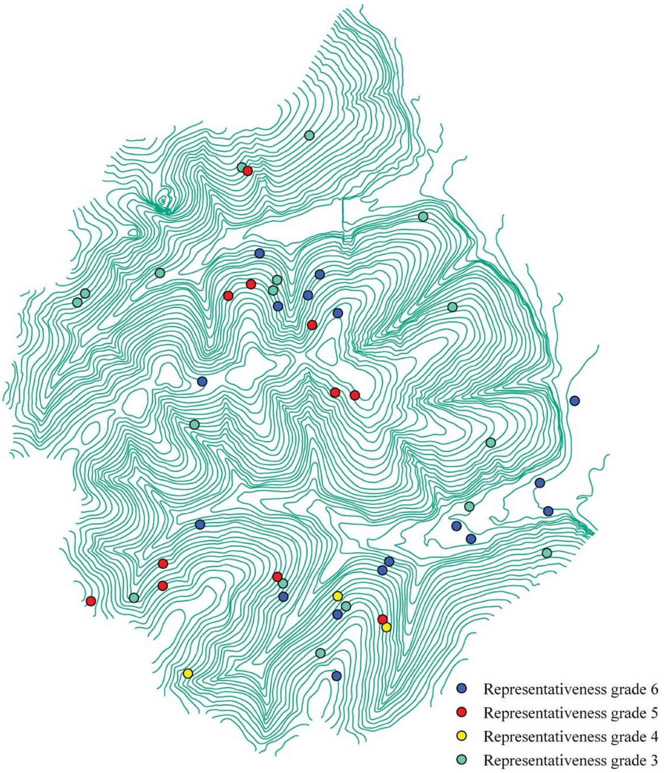
● Representativeness grade 6
● Representativeness grade 5
● Representativeness grade 4
● Representativeness grade 3

Figure 8. Samples observed under different representative grades.

designed on certain environmental cluster chain are all the typical points of this environmental cluster chain, we did not collect more samples to substitute those inaccessible points since each environmental cluster chain has at least one sample but one could represent the environmental cluster chain. This is one of the advantages of this approach because we know the other possible locations which can meet a similar goal. Soil type was identified at the subgroup level in the Chinese soil taxonomy system (Chinese Soil Taxonomy Research Group 2001) by a local soil classification expert at each site. Soil types for samples of representative grade 6 include Mollic Bori-Udic Cambosols, Lithic Udi-Orthic Primosols, Pachic Stagni-Udic Isohumosols, Typic Hapli-Udic Isohumosols, and Fibric Histic-Stagnic Gleyosols-Typic Haplic Stagnic Gleyosols. Samples of representative grade 5 supplemented a new soil subgroup: Typic Bori-Udic Cambosols, while samples of representative grade 4 did not add any new soil types. Samples of representative grade 3 supplemented another instance of Typic Hapli-Udic Isohumosols, which means the Typic Hapli-Udic Isohumosols in this area has two soil instances under different environmental combinations. The soil instance of Typic Hapli-Udic Isohumosols from samples of representative grade 3 was located at footslopes, while the other soil instance of Typic Hapli-Udic Isohumosols obtained from samples of representative grade 6 was located at backslopes, and these two instances were separated spatially (Zhu *et al*. 2010).

In order to investigate the physical meaning of the representative grade, we examined the landscape positions of the selected samples. Specifically, we analyzed how samples of different representative grades might correlate with environmental conditions, in this study

Table 3. The number of points with different representative grades occurring on different terrain positions.

Representative grades included	Summit	Backslope	Steepest slope	Footslope	Floodplain
6	3	6	3	0	5
5	0	11	0	0	0
4	0	3	0	0	0
3	0	15	0	2	0

terrain conditions, at different scales. We counted the number of samples found on each terrain position based on representative grades (Table 3). It shows that samples of representative grade 6 represent summit, backslope, steepest slope, and floodplain, which are the main terrain positions of the study area. Samples of representative grades 5 and 4 represent more detailed patterns on backslope, which is one of the main terrain types with the largest spatial coverage and with small and local features on them such as small draws and divides. Samples of representative grade 3 represent some more detailed patterns on backslope and footslope positions, which are minor terrain features in the study area. Samples of representative grades 6, 5, 4, and 3 altogether represented almost all of the terrain features in our study area. Minor terrain features such as smaller local variations within backslope than those local variations within backslope captured by samples of representative grade 3 were not represented. Since these local variations represent minor terrain features with very small spatial coverages, we do not need extra sampling for them. The above observation shows that samples of different grades do capture terrain features at different levels of scale: high grades corresponding to major patterns and low grades corresponding to minor patterns, also see Figure 6.

3.4. Digital soil mapping and validations

To evaluate the proposed sampling method, digital soil mapping using different groups of samples was conducted. We experimented with four groups. The first group contains only samples of representative grade 6. The second to fourth groups added points of representative grades 5, 4, and then 3, respectively.

In this study, we applied the SoLIM framework, a knowledge-based digital soil mapping approach, to infer soil distribution (Zhu and Band 1994, Zhu 1999, Zhu *et al*. 2010). With this approach, knowledge on the relationships between soils and their environmental conditions is first acquired from the samples in the form of fuzzy membership functions (curves). The discussion on the details of extracting fuzzy membership curves from samples is beyond the scope of this article, but interested readers are referred to Zhu *et al.* (2010) for the specifics.

We used the first group of samples to extract the fuzzy membership curves for the five main soil subtypes (Mollic Bori-Udic Cambosols, Lithic Udi-Orthic Primosols, Pachic Stagni-Udic Isohumosols, Typic Hapli-Udic Isohumosols, and Fibric Histic-Typic Haplic Stagnic Gleyosols). The fuzzy membership curves of the five soil types from grade 6 were refined using the samples from grade 5, and the fuzzy membership curves of Typic Bori-Udic Cambosols were added. The fuzzy membership curves of all the six soil types were refined with samples of representative grades 4 and 3 successively. Thus, we obtained four sets of knowledge on the relationships between soils and their environmental conditions from the four groups of samples. Different fuzzy membership maps of the respective soil

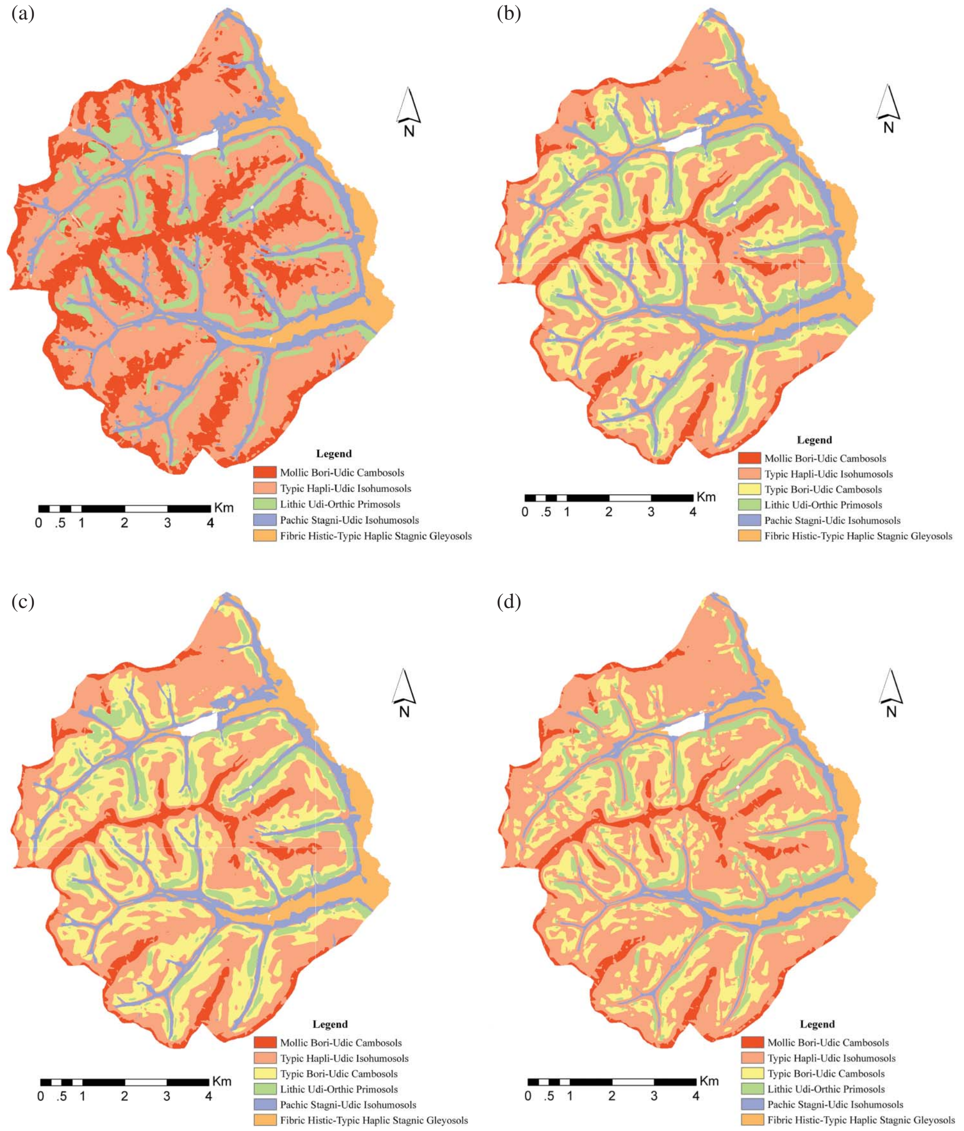
(a)
N
Legend
Mollic Bori-Udic Cambosols
Typic Hapli-Udic Isohumosols
Lithic Udi-Orthic Primosols
Pachic Stagni-Udic Isohumosols
Fibric Histic-Typic Haplic Stagnic Gleyosols
Km
0 .5 1 2 3 4
(b)
N
Legend
Mollic Bori-Udic Cambosols
Typic Hapli-Udic Isohumosols
Typic Bori-Udic Cambosols
Lithic Udi-Orthic Primosols
Pachic Stagni-Udic Isohumosols
Fibric Histic-Typic Haplic Stagnic Gleyosols
Km
0 .5 1 2 3 4
(c)
N
Legend
Mollic Bori-Udic Cambosols
Typic Hapli-Udic Isohumosols
Typic Bori-Udic Cambosols
Lithic Udi-Orthic Primosols
Pachic Stagni-Udic Isohumosols
Fibric Histic-Typic Haplic Stagnic Gleyosols
Km
0 .5 1 2 3 4
(d)
N
Legend
Mollic Bori-Udic Cambosols
Typic Hapli-Udic Isohumosols
Typic Bori-Udic Cambosols
Lithic Udi-Orthic Primosols
Pachic Stagni-Udic Isohumosols
Fibric Histic-Typic Haplic Stagnic Gleyosols
Km
0 .5 1 2 3 4

Figure 9. Hardened soil maps using different groups of samples: (a)–(d) using groups 1–4.

types were inferred using the respective sets of knowledge. These maps were then hardened (Zhu 1997) to create soil type maps by assigning each location the label of the soil type with the highest membership value. Figure 9 shows the four hardened soil maps. Comparing the four maps, it shows that the digital soil maps made using the samples from representative grade 6 (Figure 9a) capture most of the soil types in the study area, while digital soil maps that incorporate the samples from representative grades 5–3 show more spatial details, such as soil variations across the backslopes.

Field validation was conducted to evaluate accuracies of the generated soil maps. Field sites were selected through three sampling strategies: regular sampling, subjective sampling, and transect sampling. Regular sampling was used for collecting validation

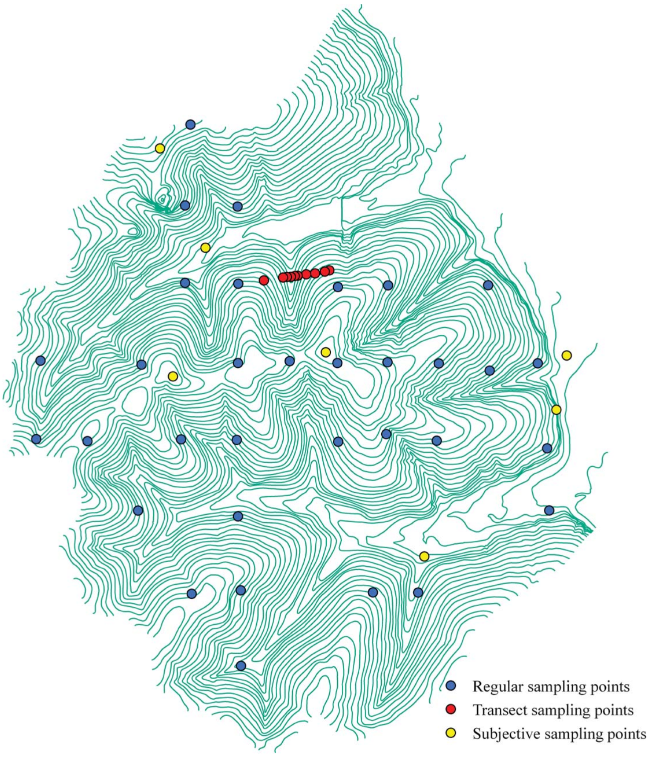
● Regular sampling points
● Transect sampling points
● Subjective sampling points

Figure 10. Location of validation points on contour lines in the study area.

points which were used to validate the overall performance of the inferred soil maps. With an 1100 m × 740 m grid arrangement, 33 validation points were collected (Figure 10, shown in blue). Subjective sampling was conducted to investigate areas with unique characteristics where soils were not covered by regular sampling. These unique areas are locations on summit, steep slope, and floodplain positions. Seven subjective points were collected (Figure 10, shown in yellow). Transect sampling was conducted in such a way that it covered major environmental variations within the shortest distance from ridge top to valley bottom (Figure 10, in red). These were used to evaluate how well the soil maps capture the spatial variation of soil in a catena. The transect we sampled consists of 10 validation points which are on average 100 m apart. In total, 50 validation samples were collected in the study area, and all samples were classified at the soil subgroup level by the local soil classification expert.

The field-observed soil subgroups at these validation sites were compared with the soil subgroups inferred through digital soil mapping at these locations. The overall accuracies of the four groups of samples and the corresponding sample sizes used for the inference are listed in Table 4. It indicates that with only 12 samples in the first group, the overall accuracy of the inferred soil map is 62%. With the addition of samples, the accuracy of soil mapping increases, but the rate of such increase slows down. Tables 5 and 6 show the producer's accuracies and user's accuracies of the soil maps using the four groups of samples separately. With the addition of samples, producer's accuracies for Typic Hapli-Udic Isohumosols and Typic Bori-Udic Cambosols increase with vibrations. With the addition of samples of representative grade 4, the producer's accuracy for Typic

Table 4. The overall accuracies of the soil maps using different groups of samples.

Representative grade	Number of samples	Overall accuracy (%)
6	12	62
6, 5	20	70
6, 5, 4	22	72
6, 5, 4, 3	35	76

Table 5. The producer's accuracies (%) of the soil maps using different groups of samples.

Representative grade	THUI	MBUC	LUOP	PSUI	TBUC	FHTHSG
6	64.29	80	75	100	0	100
6, 5	71.43	80	75	100	28.57	100
6, 5, 4	67.86	80	75	100	57.14	100
6, 5, 4, 3	78.57	80	75	100	42.86	100

Note: THUI, Typic Hapli-Udic Isohumosols; MBUC, Mollic Bori-Udic Cambosols; LUOP, Lithic Udi-Orthic Primosols; PSUI, Pachic Stagni-Udic Isohumosols; TBUC, Typic Bori-Udic Cambosols; FHTHSG, Fibric Histic-Typic Haplic Stagnic Gleyosols.

Table 6. The user's accuracies (%) of the soil maps using different groups of samples.

Representative grade	THUI	MBUC	LUOP	PSUI	TBUC	FHTHSG
6	75	28.57	50	100	0	100
6, 5	83.33	100	42.86	71.43	28.57	100
6, 5, 4	82.61	100	75	83.33	33.33	100
6, 5, 4, 3	81.48	100	75	100	33.33	100

Note: THUI, Typic Hapli-Udic Isohumosols; MBUC, Mollic Bori-Udic Cambosols; LUOP, Lithic Udi-Orthic Primosols; PSUI, Pachic Stagni-Udic Isohumosols; TBUC, Typic Bori-Udic Cambosols; FHTHSG, Fibric Histic-Typic Haplic Stagnic Gleyosols.

Hapli-Udic Isohumosols decreases while that for Typic Bori-Udic Cambosols increases. However, changes in the producer's accuracies of these two soil types are opposite. This is mainly because these two soil types are adjacently located on backslopes and some samples are in transition areas of these two soil types. Comparing with the producer's accuracies, changes in the user's accuracies with the addition of samples are more complex. The user's accuracies for Typic Bori-Udic Cambosols are low.

3.5. Sensitivity analysis

Sensitivity analysis was conducted to test the effect of the different membership thresholds used to extract representative locations from environmental clusters and the ranges of clustering iterations when employing the FCM clustering. To be considered as typical examples of an environmental cluster, Yang *et al.* (2010) suggested that the fuzzy membership values of pixels should be at least greater than 0.5. We thus started with a threshold at 0.6 in our sensitivity testing. On the other hand, if the threshold is set to be too high, such as 0.9, the selection of typical positions of the environmental clusters is often too narrow to allow sufficient overlap during overlay to generate meaningful environmental cluster chains. We thus tested three membership thresholds, 0.6, 0.7, and 0.8 in our study. The accuracies of the

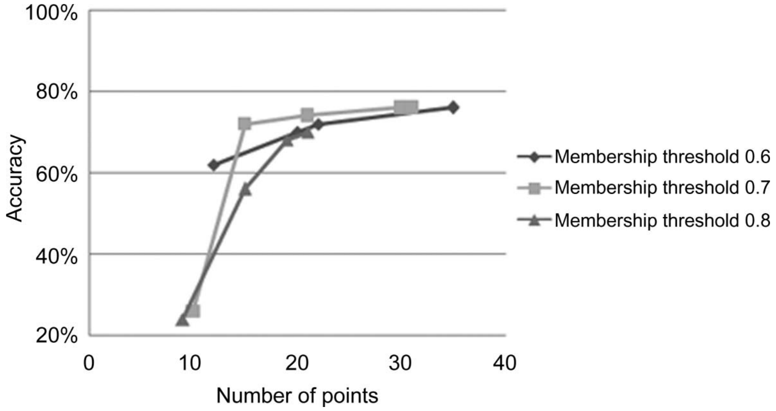

Figure 11. Changes of overall accuracies with sample size under different membership thresholds.

inferred soil maps with thresholds 0.7 and 0.8 are listed in Tables 5 and 6. Compared with Table 4, which lists the accuracies with threshold set to be 0.6, the two tables show the same pattern as that shown in Table 4, which is that mapping accuracies increase with the increase of samples and eventually reach a similar level (Figure 11). It is shown in Tables 7 and 8, however, that for thresholds 0.7 and 0.8, the accuracies are low when only the first group of samples is used in the inference. This indicates that fuzzy membership thresholds do affect the effectiveness of the sampling and the resulting soil maps. One major reason is that the number of pixels considered representative reduces when the fuzzy membership threshold increases, and thus overlaying these pixels from multiple iterations generates different and most possibly environmental cluster chains that are too refined. Table 9 lists the number of environmental cluster chains in correspondence with the landscape positions they represent with the three different thresholds for the highest representative grade 6. It shows that for thresholds 0.7 and 0.8, environmental cluster chains associated with backslope, one of the large-scale landscape positions, are missing from this first group of samples. This is the main reason why accuracies of the soil maps inferred using samples of the highest representative grade are low with thresholds 0.7 and 0.8. It is only when more samples are added from lower representative grade groups that features such as backslope could be detected.

We tested the other two ranges of clustering iterations for our study area except for the range 8–13, which are 7–12 and 9–14. If the generated environmental cluster chains under different ranges of clustering iterations are the same or similar, then the designed samples based on those chains will be the same or similar. Therefore, we compare maps

Table 7. The overall accuracies of the soil maps using different groups of samples when membership threshold is set to 0.7.

Representative grade	Number of samples	Overall accuracy (%)
6	10	26
6, 5	15	72
6, 5, 4	21	74
6, 5, 4, 3	30	76
6, 5, 4, 3, 2	31	76

Table 8. The overall accuracies of the soil maps using different groups of samples when membership threshold is set to 0.8.

Representative grade	Number of samples	Overall accuracy (%)
6	9	24
6, 4	15	56
6, 4, 3	19	68
6, 4, 3, 2	21	70

Table 9. The number of environmental cluster chains with the highest representative grade in correspondence with the landscape positions they represent.

Threshold	Summit	Backslope	Steepest slope	Footslope	Floodplain	Total number
0.6	1	1	1	0	3	6
0.7	1	0	1	0	3	5
0.8	1	0	1	0	2	4

of environmental cluster chains under different ranges of clustering iterations to test how different ranges of clustering iterations will influence the designed sampling scheme.

The generated environmental cluster chains for the ranges 7–12 and 9–14 are shown in Figures 12 and 13. Here, environmental cluster chain with pixels less than 200 is not shown in the two figures because it is difficult to display the environmental cluster chain

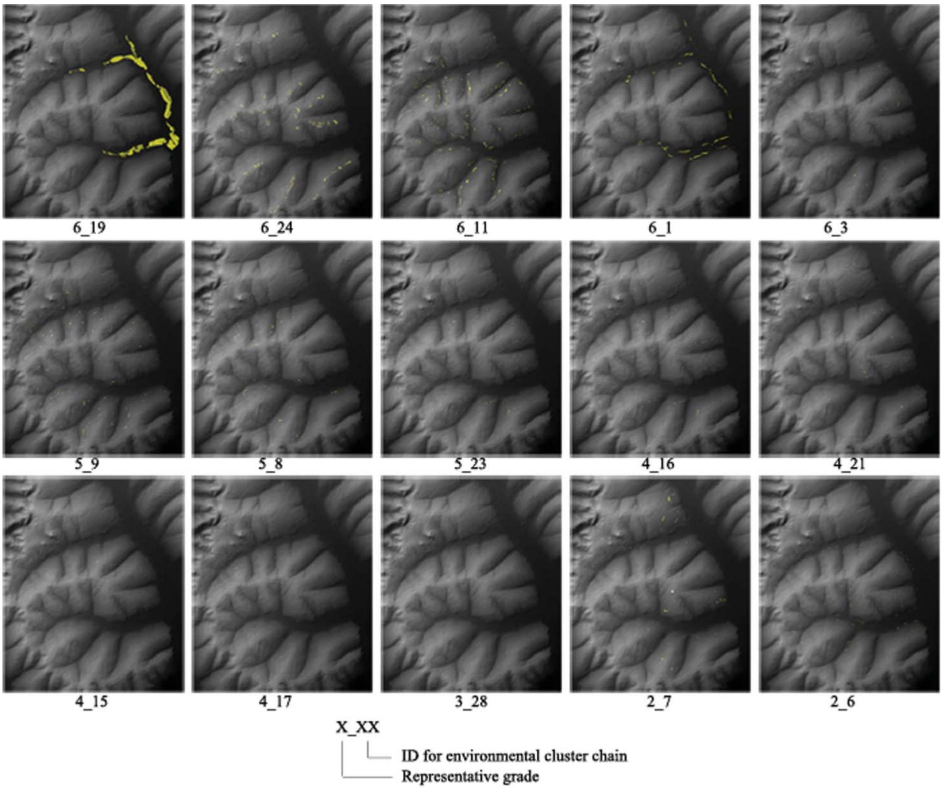

Figure 12. Environmental cluster chains through 7–12 clusters.

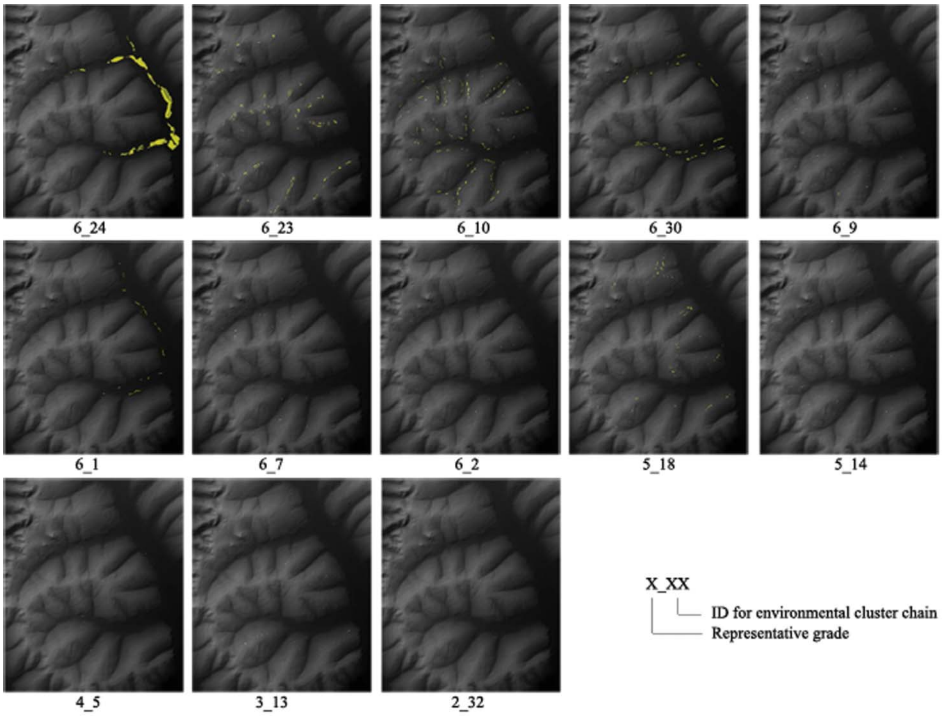

Figure 13. Environmental cluster chains through 9–14 clusters.

when it is small. However, if all of the environmental cluster chains for one representative grade are all with pixels less than 200, then we keep the environmental cluster chain with the biggest area for this representative grade for integrity of the representative grade.

It is shown in Figures 6, 12, and 13 that the environmental cluster chains for the three ranges are very similar, especially those chains with representative grade 6 which represent the large-scale landscape patterns for the study area. Thus, the range of clustering iterations from 8 to 13 can be considered as a reasonable range which could cover both large-scale and local patterns of spatial variation of soil types in the study area. And choosing a range of clustering iterations from clustering iterations 7–14 will not influence the designed sampling scheme much.

4. Observations and implications

The proposed sampling strategy selects samples of different representative grades through approximating a hierarchy of spatial variations of the geographic feature by delineating natural clusters of its relevant environmental covariates at different scales. All the samples are designed all at once. Samples of higher representative grades could represent large-scale spatial patterns of the target geographic feature and those with lower grades represent more detailed spatial patterns. Thus, sampling can be conducted in an integrated, hierarchical, and stepwise approach, which means that with limited resources the limited samples should first be directed toward areas of high representative grades and then toward areas of lower grades when more resources become available successively. In this way, samples from different stages are effectively integrated through the representative grade.

The degree of success of the proposed sampling method highly depends on how the selected environmental covariates can depict the spatial variation patterns of the target geographic feature. Those environmental covariates which influence the formation and development of the target geographic features or highly covary with the target geographic features spatially are a good choice for this sampling method. With geospatial data becoming widely available, richer in volume and types, the proposed sampling strategy would become a vital alternative to the classic sampling strategy.

5. Conclusions

This article presented an integrative hierarchical stepwise sampling design approach based on the three indices: representative grade, representative area, and representative level. Fuzzy clustering was employed on environmental covariates to generate multiple numbers of clusters to determine representative grade of samples. The sampling design was implemented for digital soil mapping in a study area located in northeastern China. Four groups of samples of different representative grades were obtained and used for knowledge-based soil inference under the SoLIM framework. Field validation was done with 50 independent samples. The results showed that the first group of samples (those of the highest representative grade) could produce a soil map with 62% accuracy with a membership threshold of 0.6, and the mapping accuracy increased with a decreasing rate with additions of samples from the lower representative grades. The results indicate that the proposed method is able to differentiate samples that represent the spatial patterns of the target geographic feature at different scales, and that stepwise sampling could be an effective and economical choice for areas sampling can only be done at stages due to the limitation of financial and human resources.

A sensitivity analysis shows that the choice of membership threshold in determining representative samples affects the effectiveness of the sampling result. At a higher threshold, samples that represent large-scale features might not be fully captured in the first group. In this case, lower representative grades may be considered for inclusion in the first group of samples. However, prior knowledge about spatial variation of the geographic features should be used for identifying environmental cluster chains associated with large-scale features.

We employed the SoLIM framework in this case study to infer soil maps, but it is not required to use this soil mapping method. People certainly could use interpolation methods or other soil mapping methods to obtain the soil maps.

In this study, we did not perform the experiment of comparison between our method and the Latin hypercube sampling method or the sampling optimization method by UK variance. We suppose that the biggest difference between our method and the Latin hypercube sampling method is that the designed sampling points through our method are in a hierarchy of representativeness grade, while there is no such difference between sampling points designed by the Latin hypercube sampling method and the sampling optimization method by UK variance. Thus, our method can indicate a sampling order which is useful and necessary for stepwise sampling.

This study indicates that the proposed sampling method could be an effective approach to stepwise sampling. However, investigators could also collect the representative samples determined by the approach at once if sampling cost is not a concern. Our future efforts will focus on applying and evaluating the sampling method in larger areas and for the modeling of other geographic features (such as vegetation), and will also include the comparison between the proposed sampling method and other sampling methods (such as random sampling or systematic sampling).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Project Nos 41001298; 40971236; 41023010); the National High Technology Research and Development Program of China (Project No. 2011AA120305); the Vilas Associate Award; and the Hammel Faculty Fellow Award from the University of Wisconsin-Madison. We thank Mr. Du Guohua for his hard work as our soil classification expert.

References

Beckett, P.H.T., 1968. Method and scale of land resource survey, in relation to precision and cost. *In*: G.A. Stewart, eds. Land evaluation: papers of a CSIRO symposium, organized in cooperation with UNESCO, 26–31 August 1968 Canberra. South Melbourne: Macmillan Company of Australia, 53–63.

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

Bezdek, J.C., Ehrlich, R., and Full, W., 1984. FCM: the fuzzy *c*-means clustering algorithm. *Computers & Geosciences*, 10 (2–3), 191–203.

Brus, D.J. and de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80, 1–44.

Brus, D.J. and Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138, 86–95.

Brus, D.J. and Noij, I.G.A.M., 2008. Designing sampling schemes for effect monitoring of nutrient leaching from agricultural soils. *European Journal of Soil Science*, 59 (2), 292–303.

Chinese Soil Taxonomy Research Group, 2001. *Keys to Chinese soil taxonomy*. 3rd ed. Hefei: University of Science and Technology of China Press.

Cochran, W.G., 1977. *Sampling techniques*. 3rd ed. New York: John Wiley & Sons.

Cressie, N., 1991. *Statistics for spatial data*. New York: John Wiley & Sons.

Delmelle, E.M. and Goovaerts, P., 2009. Second-phase sampling designs for non-stationary spatial variables. *Geoderma*, 153 (1–2), 205–216.

Gregoire, T.G. and Valentine, H.T., 2008. *Sampling strategies for natural resources and the environment*. London: Chapman & Hall/CRC.

Grinand, C., *et al*., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, 143, 180–190.

Haining, R., 2003. *Spatial data analysis*. Cambridge University Press.

Heim, A., *et al*., 2009. Effects of sampling design on the probability to detect soil carbon stock changes at the Swiss CarboEurope site Lageren. *Geoderma*, 149, 347–354.

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

Jenny, H., 1980. *The soil resource: origin and behavior*. New York: Springer.

Kish, L., 1985. *Survey Sampling*. New York: John Wiley & Sons.

MacMillan, R.A. and Shary, P.A., 2009. Chapter 9 landforms and landform elements in geomorphometry. *In*: T. Hengl and H.I. Reuter, eds. *Geomorphometry, volume 33: concepts, software, applications*. Amsterdam: Elsevier, 231–236.

McBratney, A.B. and Webster, R., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables. *Computers & Geosciences*, 7, 331–334.

McKenzie, N.J. and Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. *Geoderma*, 89 (1/2), 67–94.

Miao, C.Y., Yang, L., and Chen, X.H., 2012. The vegetation cover dynamic (1982–2006) in the different erosion regions of Yellow River basin. *Land Degradation & Development*, 23, 62–71.

Miao, C.Y., Yang, L., and Li, S.L., 2011. Recent change of streamflow in the mainstream of the Songhua River basin, Northeast China. *Environmental Earth Science*, 63 (3), 489–499.

Minasny, B. and 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., and Walvoort, D.J.J., 2007. The variance quadtree algorithm: use for spatial sampling design. *Computers & Geosciences*, 33, 383–392.

Qin, C.Z., *et al*., 2007. An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. *International Journal of Geographical Information Science*, 21 (4), 443–458.

Qin, C.Z., *et al*., 2011a. An approach to computing topographic wetness index based on maximum downslope gradient. *Precision Agriculture*, 12 (1), 32–43.

Qin, C.Z., *et al*., 2011b. Mapping soil organic matter in small low-relief catchments using fuzzy slope position information. *Geoderma*. doi:10.1016/j.geoderma.2011.06.006.

Quinn, P., *et al.*, 1995. The ln(a/tanb) index: How to calculate it and how to use it within the TOPMODEL framework. *Hydrological Processes*, 9, 161–182.

Simbahan, G.C. and Dobermann, A., 2006. Sampling optimization based on secondary information and its utilization in soil carbon mapping. *Geoderma*, 133, 345–362.

Webster, R. and Oliver, M.A., 1992. Sample adequately to estimate variograms of soil properties. *European Journal of Soil Science*, 43 (1), 177–192.

Yang, L., *et al*., 2010. A purposive sampling design method based on typical points and its application in soil mapping (in Chinese). *Progress in Geography*, 29 (3), 279–286.

Zhu, A.X., 1997. A similarity model for representing soil spatial information. *Geoderma*, 77, 217–242.

Zhu, A.X., 1999. A personal construct-based knowledge acquisition process for natural resource mapping. *International Journal of Geographical Information Science*, 13 (2), 119–141.

Zhu, A.X. and Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. *Canadian Journal of Remote Sensing*, 20 (4), 408–418.

Zhu, A.X., *et al.*, 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, 65, 1463–1472.

Zhu, A.X., *et al*., 2008. Purposive sampling for digital soil mapping for areas with limited data. *In*: A.E. Hartemink, A.B. McBratney, and M.L. Mendonça-Santos, eds. *Digital soil mapping with limited data*. New York: Springer, 233–245.

Zhu, A.X., *et al*., 2010. Construction of quantitative relationships between soil and environment using fuzzy c-means clustering. *Geoderma*, 155 (3–4), 166–174.

Zhu, Z. and Stein, M.L., 2006. Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological & Environmental Statistics*, 11, 24–44.