Evaluation of Integrative Hierarchical Stepwise Sampling for Digital Soil Mapping

This paper presents an integrative hierarchical stepwise sampling (IHS) method and two case studies to compare it with stratified random sampling (SRS) and conditioned Latin hypercube sampling (cLHS). The first comparison between IHS and SRS was conducted for mapping sand content of two soil layers in a study area in Anhui Province, China. Two sample sets of the same sample size were collected in the field based on IHS and SRS. The second case study is a simulation study, where we compared IHS and cLHS for mapping soil series in the Raffelson watershed in Wisconsin (USA). The study used an accurate and detailed soil series map produced previously as a proxy of the real soil distribution. Virtual samples with nine sample sizes designed by IHS and cLHS were collected on the soil map. For both case studies, an individual predictive soil mapping method was employed and independent validation samples were used to evaluate the mapping accuracies. Results indicate that IHS generally performs better than SRS for capturing distributions of the environmental variables. It obtained higher mapping accuracies than SRS at different sample sizes. On the other hand, cLHS appears to provide a better representation for distributions of the environmental variables than IHS, but the mapping accuracies with IHS are higher than those with cLHS at nearly all sample sizes. Finally, both case studies showed that IHS provides valuable information on representativeness of the samples.

Abbreviations: FCM, fuzzy c-means; IHS, integrative hierarchical stepwise sampling; cLHS, conditioned Latin hypercube sampling; SRS, stratified random sampling.

Soil sampling plays an essential role in soil survey. A sampling design specifies where and how many to sample in the field. Different sampling design strategies will produce different soil distribution maps and greatly affect the mapping accuracies (Brus and de Gruijter, 1997; van Groenigen et al., 2000; de Gruijter et al., 2006; Brus et al., 2006; Gregoire and Valentine, 2007; Heim et al., 2009). Because the collection of soil samples in the field is often time-consuming and resource-intensive, it is important to apply efficient sampling design strategies.

When prior knowledge on an area's local soil variation is limited, simple random sampling has been a commonly used sampling strategy (Oliver and Webster, 1986; Grinand et al., 2008; Walvoort et al., 2010; Pringle et al., 2012; Allbed et al., 1986).

Core Ideas:

- Evaluation of an integrative hierarchical stepwise (IHS) sampling method by comparing it with stratified random sampling (SRS) and conditioned Latin hypercube sampling (cLHS) through two case studies.
- IHS obtained higher mapping accuracies than SRS and cLHS at nearly all sample sizes.
- IHS provides valuable information on the representativeness of samples.
- SRS and cLHS were found to generate unstable results on sample sets and soil maps.
al., 2014). With simple random sampling, each sample location is chosen randomly such that each location has the same probability of being chosen. When an area is divided into geographically compact sub-areas such as different soil parent material units, SRS can be employed. Spatial clustering of the sample locations could be avoided to some extent through stratification, thus stratified random sampling is usually more efficient when compared with simple random sampling (Brus, 1994; Walvoort et al., 2010).

Taking into consideration the correlation between soil and its environmental covariates, some sampling design aims to select sample locations that can represent the feature space of environmental covariates to capture soil variability (Minasny and McBratney, 2006; Mulder et al., 2013; Kidd et al., 2015). A popular sampling method based on this idea is cLHS. This sampling method optimizes the coverage of environmental feature space of the soil by stratifying the full multivariate distribution of environmental covariates. It enables the designed sample distribution to closely replicate the covariate distribution (Minasny and McBratney, 2006). Conditioned Latin hypercube sampling is thus considered as an efficient sampling method and implemented widely (Worsham et al., 2012; Louis et al., 2014; Taghizadeh-Mehrjardi et al., 2014; Pahlavan Rad et al., 2014; Silva et al., 2014, 2015; Kidd et al., 2015).

While soils change from one soil class to another gradually across space, sampling at those locations where the soils are the most typical of the individual soil classes can be another effective way to capture the greatest soil variability. One way to find those “typical” or “representative” sample locations is based on the knowledge of local soil experts. This sampling strategy is often conducted during traditional soil survey and relies on the experience and personal judgement of soil experts (Hengl et al., 2003; Trochim, 2006). When such domain knowledge is not readily available, an IHS strategy has been proposed to design representative samples with the assistance of environmental covariates through a fuzzy clustering approach (Yang et al., 2013). The method is based on the assumption that soil spatial variability exists at multiple scales: from the major large-scale spatial patterns to local ones. It draws samples to represent soil variations from the large to local scales. This results in sample locations with different representative grades. Such representative grade information can direct sampling order: sampling can be conducted in a stepwise fashion when sampling resources are limited. A case study for classifying soil types has showed that this sampling strategy could produce accurate soil maps through a small set of representative sample locations, and the addition of new sample locations in lower representative grade leads to improvements of accuracy with a decreasing marginal gain (Yang et al., 2013). This allows for the samples collected at a sample size for capturing the finest scale of soil variation possible at that size. This provides an efficient way of soil sampling and mapping, especially when field sampling is a limiting factor. It is still unknown, however, how its efficiency compares with other sampling methods, such as the commonly used SRS and cLHS.

The aim of the present study is thus to compare the IHS with SRS and cLHS for digital soil mapping. Two case studies were conducted. The first was conducted to compare IHS with SRS for mapping soil sand content of two soil layers (at 0- to 20-cm and 20- to 40-cm depths) in a study area of 5900 km² in Anhui Province of China. Two sample sets with the same number of sampling points (59) were collected based on IHS and SRS, respectively. Four subsample sets of different sizes were also extracted from the 59 points to test the effect of sample sizes. The second case study was conducted to compare IHS with cLHS for mapping soil classes in a study area of 4 km² in the Raffelson watershed of La Crosse County Wisconsin, USA. This case study is a simulation study, which used an accurate and detailed soil class map (with an accuracy of 83.8%, Qi et al., 2006) produced using a knowledge-based method under the Soil Land Inference Model (SoLIM) framework (Zhu et al., 2001). This soil class map served as a proxy of the real soil class distribution of the watershed, where virtual samples were drawn based on the map locations. A total of nine sample sets with varied sample sizes designed with IHS and cLHS were collected on the map. Comparisons between the two sampling methods were conducted in three aspects: (i) ability to capture the environmental feature space; (ii) accuracies of the predicted soil maps based on the same quantities of samples; (iii) mapping accuracy with increasing sample sizes.

**MATERIALS AND METHODS**

**The Study Areas and Environmental Data**

**The Xuancheng Case Study**

The Xuancheng study area includes Xuanzhou City, Guangde County, and Langxi County, and covers an area of 5900 km² (Fig. 1). The climate is warm and humid in the summer and relatively cool and dry in the winter season, with an average temperature range of 11 to 16°C and annual average precipitation between 1200 to 1800 mm, most of which occurs between April and October. Elevation ranges from 0 to 1039 m with flat plains in the northwest and hills in the northeast and south. Parent materials of the study area are composed of clay-silt-gravel formed during the Quaternary and materials weathered from sandstone, shale, conglomerate, pyroclastic rocks, granite and granodiorite, and limestone. The main soil types in this area are red soil and paddy soil in the Chinese genetic soil classification system (Chinese National Soil Survey Office, 1992), which are Ferric Lixisols and Hydragric Anthrosols in World Reference Base for Soil Resources (WRB), respectively. Land use is mainly cultivated land with rice as the dominant crop, and forest land covered by bamboo, fir, shrub, and other evergreen coniferous or deciduous broad-leaved trees, most of which are secondary or planted. The study area has experienced intensive human activities, especially in flat areas.

Soil sand content was chosen as the soil property for evaluation as soil texture is one of the most important soil attributes, and influences a large number of properties such as the water-holding capacity and hydraulic properties (Heil and
Schmidhalter, 2012) that are important information in land use and management practices for the study area.

The soil forming environment of this study area is very complex mainly because of the complicated soil parent materials and variable landforms. Climate factor also plays a role for soil development because precipitation and temperature are variable throughout the study area. Thus, parent material, terrain characteristics and climate factors were chosen to characterize the environmental conditions for mapping soil sand content in two soil layers (0- to 20-cm depth and 20- to 40-cm depth) based on knowledge of the soil pedogenesis in this area. One categorical variable (parent lithology) and six continuous variables including slope gradient, planform curvature, profile curvature, topographic wetness index, annual average precipitation, and annual average temperature were used to construct the environmental database. Details of these environmental covariates can be found in Yang et al. (2016). The cell size (90 m) of the SRTM DEM (Fig. 1), which was used to derive the above topographic variables, was set as the final cell size for soil sand content mapping.

The Raffelson Case Study

The Raffelson study area is located in the “Driftless area” of southwestern Wisconsin that has remained free of direct impact from the most recent Pleistocene era continental glaciers (Fig. 2). The area has a typical ridge and valley terrain with relatively flat, narrow ridges. There are moderate side slopes with gradients below 20% and steep slopes with high gradient values around 50%. In terms of parent material, there are five distinct types, which are Oneota (dolomite), Glauconite (sandstone), Jordan (sandstone), Wonewoc (sandstone), and alluvial materials. The complexity of parent material over the area has led to the development of 16 different soil series (classes) in the watershed. Most ridges and valleys in the area have been cultivated since late 19th century. Side-slopes are generally forested, though some have been cleared for pasturing.

A detailed and accurate soil series map (Fig. 3) was used as a proxy of the real soil distribution in the study area. The soil series map was generated under the SoLIM framework based on the expert knowledge of a senior soil scientist from the local office of the U.S. Department of Agriculture–Natural Resources Conservation Service (USDA–NRCS; Qi et al., 2006). The soil expert was asked to provide knowledge to construct fuzzy membership functions (curves) for all the soil classes. The fuzzy membership functions describe how similarity between a local soil and the typical case of the given soil class changes as environmental conditions change. The soil series map was then produced based on the fuzzy membership functions using SoLIM inference engine. On the soil series map, the four soil series (Council [coarse-loamy, mixed, superactive, mesic Typic Hapludalfs], Churchtown [fine-silty, mixed, superactive, mesic Mollic Hapludalfs], Norden [fine-loamy, mixed, superactive, mesic Typic Hapludalfs], and Elbaville [fine-loamy, mixed, superactive, mesic Glossic Hapludalfs]) covers 77% of the entire study area. Each of the remaining 12 soil series covers less than 5% of the study area. The overall accuracy of the soil series map was 83.8% based on an independent validation sample set. Compared with a conventional soil map of the study area, the soil series map generated under the SoLIM framework is much more detailed and accurate.
more accurate and with greater spatial details. The conventional soil map contained 12 soil mapping units, of which three are complexes. In contrast, all the different components in the complexes were separately mapped out on the SoLIM soil series map. According to Qi et al. (2006), environmental covariates for mapping the soil series on this map included a categorical variable (soil parent material), and six continuous variables: elevation, slope gradient, profile and planform curvatures, topographic...
wetness index and percentage of colluvium from competing bedrocks. The details of production of these environmental covariates can be found in Qi et al. (2006). The cell size (10 m) of the DEM (Fig. 2), which was used to generate the topographic variables in this area, was set as the final cell size of the predicted soil series map.

**SAMPLING METHODS**

**Integrative Hierarchical Stepwise Sampling**

The basic idea of the integrative hierarchical stepwise sampling method is to identify sample locations that can represent soil spatial variations at different scales with assistance of soil environmental covariates (Yang et al., 2013). In particular, clustering on environmental covariates with a range of cluster numbers is used to approximate soil spatial variations at different scales, that is, a small number of clusters are assumed to represent the major soil variations, and increasing the number of clusters leads to the emergence of minor types. Here soil variations may represent soil classes when IHS is applied in soil class sampling, and represent natural spatial clusters of soil property variations when IHS is applied in soil property sampling. While dominant large-scale soil variations can always be detected regardless of the number of clusters, small-scale local variations are only revealed when the number of clusters becomes large enough. Therefore, samples representing large-scale variations are selected by identifying areas that consistently show up in all clustering iterations, and samples representing local-scale variations are placed at locations that only show up during later stages of iterations. More details of this method can be found in Yang et al. (2013, 2016). An outline of the sampling procedure is as follows.

i) Stratification of the study area. When there are categorical environmental variables such as soil parent material, a stratification of the study area using the categorical environmental variable is implemented first.

ii) Generation of environmental clusters. The fuzzy c-means (FCM) clustering classifier (Bezdek et al., 1984) is performed on the continuous environmental variables to generate environmental clusters for each stratum. In each stratum, FCM clustering is performed multiple times with different cluster numbers ranging from only a few to many more (multiple iterations). The range of cluster numbers should be designed in a way that the clustering results cover large-scale as well as small-scale soil spatial variation types. A few sets of iterations with different cluster number ranges could be tested to find an appropriate range. The output of FCM clustering is fuzzy membership maps with each location filled with a membership value for each cluster.

iii) Determination of representative grades. A location with high membership value (above a user-specified threshold, such as 0.9) in a cluster is considered as a typical location. If a location is consistently identified as being typical at multiple iterations it is considered to be representative of major soil variation types. Similarly locations that are identified less frequently are considered representative of minor types. Representative grade is thus determined by the frequency with which a location is identified as typical across multiple iterations.

iv) Generation of environmental cluster chains. Locations with the same representative grade may represent different environmental clusters under certain iterations (cluster numbers). To examine the class configurations under each representative grade, the concept of ‘environmental cluster chain’ is introduced to record the specific list of clusters and the associated cluster iteration. Each location is then labeled with an environmental cluster chain except for those locations where the representative grade is 0. Thus, the spatial distribution of each chain can be generated to represent different soil variation types under each representative grade. The membership at each location to an environmental cluster chain is calculated by averaging the values of its fuzzy membership to all clusters in the environmental cluster chain. Based on this rule, a membership map is produced for each cluster chain, with each cell value measuring the average representativeness of that location to the chain.

v) Design of sample locations. Sample locations for each stratum are designed according to the following guidelines. All of the environmental cluster chains are first sorted by representative grades. Environmental cluster chains with the same representative grade are sorted by their area coverage. For each environmental cluster chain, one sample is selected at a location with a high average membership value. Finally, sample locations are selected for all the strata.

Sampling locations with the highest representative grades for all strata in the study area are considered to bear the highest overall representative grade. Different strata, however, may lead to different representative grades as their ranges of cluster numbers differ. For example, representative grades for Stratum 1 in a study area could be 3, 2, and 1, while representative grades for Stratum 2 are only 2 and 1. In this case, locations with representative Grade 3 in Stratum 1 and locations with representative Grade 2 in Stratum 2 are all considered as being with the highest representative grade for the study area. Locations with representative Grade 2 in Stratum 1 and locations with representative Grade 1 in Stratum 2 are then considered as being with the second highest representative grade.

Within one representative grade, sample locations can also be ranked in representativeness order. An "area of stratum-area of environmental cluster chain" order is followed in this paper, which means both area sizes of strata and environmental cluster chains of the sample locations are considered to determine the representativeness of each sample location. The first sampling location is always one in the environmental cluster chain with the largest area coverage in the largest stratum in the study area. The second sampling location and forward need to be determined through a comprehensive examination of the area coverages of strata and environmental cluster chains. For example, if the second location should be selected in the stratum with the largest...
area coverage, it should belong to the second biggest environmentally cluster chain in the stratum. If the second location should be selected in the stratum with the second largest area coverage, it should belong to the largest environmental cluster chain in this stratum. The principal to determine which stratum the second location or forward is extracted is to make sure that the number of sample locations selected for each stratum is proportional to the area covered by the stratum.

The IHS strategy was implemented with the SoLIM Solutions software (http://solim.geography.wisc.edu/software/index.htm). Based on the above sampling procedure, a total of 59 environmental cluster chains with three different representative grades were generated for all the parent lithology strata in the Xuancheng study area. These environmental cluster chains include 46 chains with the highest representative grade, 12 with the second highest representative grade and one with the lowest representative grade. Together they represent 59 soil sand content variation types at different scales. Following the sampling design, 59 samples were collected in the field (the squares in Fig. 1). At each sampling location, soil was collected using a soil auger at two different depths, being 0 to 20 cm and 20 to 40 cm. Soil samples were air dried and sieved through a 2-mm mesh sieve, and analyzed using the laser diffraction technique with a Mastersizer 2000 laser particle size analyzer (Malvern Instruments, Malvern, England). Measurements were taken for an interval of sizes ranging from 0.02 to 2,000 μm. Soil sand percentages were then calculated based on these measurements.

To examine how soil mapping accuracies change with the addition of samples and the effectiveness of our derived representativeness (order) information of the different sample locations with the same representative grade, four subsets of samples with different representativeness and increasing sample sizes were created through selecting samples from the original set. The first set contains 32 samples whose locations were drawn from the 46 sample locations with the highest representative grade. The second set contains all from the first set, plus the remaining sample locations with the highest representative grade, which is all the highest representative 46 samples. The third set contains all samples from the second set, plus 12 sample locations with the second highest representative grade. Sample size of the third set is 58. And finally, the fourth set contains all in the third set plus the one last sample which is of the lowest representative grade among all 59 samples.

For the Raffelson case study, soil parent material was used to stratify the study area. Twenty-five environmental cluster chains with three representative grades were generated. And 25 sampling locations were designed, including 17 with highest representative grade, 7 with the second highest representative grade and 1 with the lowest representative grade. Soil series of these sampling locations were extracted from the soil series map. As with the first case study, 9 subsets of samples were extracted, whose sample sizes are 8, 10, 12, 14, 17, 20, 22, 24, and 25, respectively.

Stratified Random Sampling

Stratified random sampling was implemented in the Xuancheng study area for comparison. The study area was stratified using the parent lithology layer which is a determining forming factor for the development of soil sand content. Within each stratum, sample locations were selected randomly without replacement with the number of samples proportional to the area covered by the stratum. At least one sampling point was located in each stratum to ensure all of the soil parent lithology types being represented. A total of 59 samples were collected (the dots in Fig. 1), comparable with the sample size with IHS sampling. The field sampling, preprocessing and measurement of sand content for these soil samples were the same with those for the IHS samples.

As with IHS, four subsets of samples with sample sizes of 32, 46, 58, and 59 were also created. The first set of 32 samples was drawn randomly from the original sample set. Additional samples were then randomly drawn from the remaining samples without replacement and added successively to generate the second, third, and fourth sets. Three replicates of stratified random sample sets with the different sample sizes were generated using the above process.

Conditioned Latin Hypercube Sampling

Latin hypercube sampling is also considered as a stratified random procedure. It stratifies the marginal distributions of the covariates into n equally probably non-overlapping strata and draws one random sample from each stratum. Due to the fact that the sample locations selected by LHS may not exist in the real world, a cLHS is developed to search samples in the real world, which can form a true or approximate Latin hypercube of the feature space (Minasny and McBratney, 2006). The sampling strategy becomes an optimization problem: given environmental data (X), select n sample locations from the real world so that the sample locations x form a Latin hypercube. The objective function of cLHS (Minasny and McBratney; 2006) is as follows:

\[
O = w_1O_1 + w_2O_2 + w_3O_3
\]  

where \(O_j\) is a subfunction for continuous variables, \(O_2\) is a subfunction for categorical variables, and \(O_3\) is to ensure that the correlation of the sampled variables will replicate the original data. \(w\) is the weight given to each component of the objective function. For general applications \(w\) is set to 1 for all components of the objective function.

\[
O_1 = \sum_{j=1}^{n} \sum_{j=1}^{k} [n(q_j \leq x_j < q_j^{\text{ui}}) - 1] 
\]

where \(k\) is the number of environmental variables, \(n(q_j \leq x_j < q_j^{\text{ui}})\) is the number of \(x_j\) that fall in between quantiles \(q_j\) and \(q_j^{\text{ui}}\) of the \(j\)th environmental variable.

\[
O_2 = \sum_{j=1}^{c} \left| \frac{n(x_j)}{n} - k_j \right| 
\]
where \( \eta(x) \) is the number of \( x \) that belongs to Class \( j \) in sampled data, and \( k_j \) is the proportion of Class \( j \) in \( X \).

\[
O_{i,j} = \sum_{i=1}^{k} \sum_{j=1}^{k} |c_{ij} - t_{ij}|
\]

[4]

where \( c \) is the element of \( C \), the correlation matrix of \( X \), and \( t \) is the equivalent element of \( T \), the correlation matrix of \( x \).

In the Raffelson study, we implemented cLHS using the seven environmental covariates used to generate the SoLiM soil series map. The cLHS algorithm was run using Matlab (MathWorks, 2009) with the codes provided by Minasny and McBratney (2006). An annealing simulation schedule (Press et al., 1992) was used to search for sample locations. The same nine sample sizes, we experimented with the HIS, were used here. Three sample sets were generated for each size to examine the stability of cLHS. Because the cLHS sampling strategy is not designed for stepwise sampling, the sample sets of the nine sample sizes were designed independently. One sample set containing 25 sample locations was shown in Fig. 2. Soil series of all the cLHS sampling locations were also extracted from the soil series map as done with the IHS (Fig. 3).

**Soil Inference: An Individual Predictive Soil Mapping Method**

An individual predictive soil mapping (iPSM) method (Zhu et al., 2015) was employed in this study to infer soil sand content maps in the Xuan Cheng study area and soil series maps in the Raffelson watershed. iPSM is a predictive soil mapping method based on the soil-environment relationship at each individual soil sample location under the SoLiM framework. With the assumption that the more similar environment conditions between two locations the more similar soil property values, iPSM predicts soil properties (types) of unsampled locations based on the environmental similarity between unsampled locations and sampled locations (Zhu et al., 2015). One advantage of this method is that there is no requirement on sample size or the representativeness of samples for predicting soil properties (types).

The method includes two main steps. The first step is to calculate environmental similarity. Environmental similarity between an unsampled Location \( g \) and sample Location \( m \) is first evaluated at the individual environmental variable level, and then similarities based on all environmental variables are integrated to represent the overall similarity at the given sample Location \( m \).

For calculating similarity at the environmental variable level, the specific form of function depends on the data type of the environmental variable. If an environmental variable is categorical, for example, parent materials, a Boolean function could be used, which produces 1 or 0 depending on whether or not the value at Location \( g \) is the same as that at Sample \( m \). If the environmental variable is continuous, for example, slope, annual average temperature, etc., Euclidean distance, Gower distance similarity coefficient, Mahalanobis distance, etc. could be used. In this study, a Gaussian distance measure was adopted using the following equation (Eq. [5]) for continuous variables.

\[
E_{v}(\epsilon_{v,g}, \epsilon_{v,m}) = \exp \left[ -\frac{(\epsilon_{v,g} - \epsilon_{v,m})^2}{2(\frac{SD_{v}}{SD_{vm}})^2} \right]
\]

[5]

where \( \epsilon_{v,g} \) and \( \epsilon_{v,m} \) stand for the value of the \( v \)-th environmental covariate at Location \( g \) and at soil Sample \( m \), respectively; \( SD_v \) is the standard deviation of the \( v \)-th environmental covariate in the study area; \( SD_{vm} \) is the square root of the mean deviance of the value of \( v \)-th environmental covariate at all unsampled location \( g \) from that at soil Sample \( m \) (i.e., \( \epsilon_{v,m} \)), and \( np \) is the number of unsampled locations to be predicted.

When integrating variable-level similarities, a limiting factor approach (Zhu and Band, 1994) was adopted based on the assumption that the least favorite environmental condition determines soil formation and thus attributes at a given site. A minimum operator on similarities for all the environmental variables is used to generate the similarity between Location \( g \) and Sample \( m \). The outcome of this step is a similarity vector \( S_g \) at Location \( g \), in which elements were the environmental similarities between location \( g \) and the \( n \) sampled locations.

\[
S_g = (S_{g,1}, S_{g,2}, \ldots, S_{g,n})
\]

[7]

The second step is to predict soil types (properties) based on the environmental similarities. When predicting soil types, the soil series of the sample location which had the maximum similarity in \( S_g \) is assigned to Location \( g \). When predicting soil property variations, a weighted average model (Zhu et al., 1997) was used in this study. In particular, environmental similarities with all sample locations were used as weights to estimate the soil sand content value on location \( g \) (Eq. [7]) (Liu, 2010, Zhu et al., 1997).

\[
P_{g} = \frac{\sum_{m=1}^{n} S_{g,m} \times P_{m}}{\sum_{m=1}^{n} S_{g,m}}
\]

[8]

where \( P_g \) is the predicted soil sand content value on Location \( g \), \( S_{g,m} \) is the environmental similarity of the unsampled Location \( g \) to the soil sample Location \( m \), and \( P_m \) is the soil property value of this sample Location \( m \).

**Validation of the Inferred Soil Maps**

The inferred soil sand content maps of the Xuan Cheng study area were validated using an independent validation set of 56 sampling points, which was collected based on a systematic sampling strategy with 10 km by 10 km grid arrangement (the cross points in Fig. 1). The correlation coefficient \( r \), Lin’s concordance
correlation coefficient $\alpha \alpha$ (Lin, 1989) and root mean square error (RMSE) were calculated. The RMSE tested the prediction errors based on the validation points. Correlation coefficient $r$ tested the correlation trend of the predicted values and observed values of the validation points. Lin’s concordance correlation coefficient measured the agreement between the predicted values and observed values of the validation points.

$$r = \frac{\sum_{i=1}^{n}(P_i - \bar{P})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2 \sum_{i=1}^{n}(Q_i - \bar{Q})^2}}$$  \[9\]

$$\alpha \alpha = \frac{2\sigma_P \sigma_Q}{\sigma_P^2 + \sigma_Q^2 + (\bar{P} - \bar{Q})^2}$$  \[10\]

$$\text{RMSE} = \left[ \frac{1}{n} \sum_{i=1}^{n}(P_i - Q_i)^2 \right]^{0.5}$$  \[11\]

where $n$ is the number of validation points, $P_i$ is the predicted soil sand content value at a validation point $i$, and $Q_i$ is the observed soil sand content value at point $i$. $\bar{P}$ and $\bar{Q}$ are the means of $P_i$ and $Q_i$, and $\sigma_P^2$, $\sigma_Q^2$ are the corresponding variances.

For the Raffelson study area, 100 validation points were selected from the soil series map using a stratified random sampling strategy with parent materials as the strata (Fig. 2). Mapping accuracies were calculated by comparing soil series of the validation locations extracted from the original soil series map with those from the predicted soil series maps using either IHS sampling or cLHS sampling designs.

### RESULTS

#### The XuanCheng Case Study

**Descriptive Statistics of Sand Contents**

Table 1 shows the descriptive statistics of sand percentages for the two soil layers in the XuanCheng study area.

<table>
<thead>
<tr>
<th>Soil layer</th>
<th>No. of samples</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>CV(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil layer 01</td>
<td>IHS samples</td>
<td>59</td>
<td>6.94</td>
<td>85.47</td>
<td>30.91</td>
<td>15.66</td>
</tr>
<tr>
<td></td>
<td>SRS samples</td>
<td>59</td>
<td>3.97</td>
<td>83.33</td>
<td>33.67</td>
<td>16.68</td>
</tr>
<tr>
<td></td>
<td>All samples</td>
<td>174</td>
<td>0.31</td>
<td>85.47</td>
<td>31.98</td>
<td>17.27</td>
</tr>
<tr>
<td></td>
<td>Soil layer 02</td>
<td>IHS samples</td>
<td>59</td>
<td>5.78</td>
<td>83.34</td>
<td>36.65</td>
</tr>
<tr>
<td></td>
<td>SRS samples</td>
<td>59</td>
<td>0.39</td>
<td>77.10</td>
<td>31.30</td>
<td>18.93</td>
</tr>
<tr>
<td></td>
<td>All samples</td>
<td>174</td>
<td>0.39</td>
<td>83.34</td>
<td>34.17</td>
<td>19.47</td>
</tr>
</tbody>
</table>

Comparisons on Representing the Feature Space of Environmental Covariates

The boxplots of the environmental variables built from the original data (a total of 687, 887 pixels) and the sampled locations with the four different sample sizes using the IHS and SRS methods are presented in Fig. 4. Generally, both IHS and SRS are slightly biased in simulating the distribution of the environmental variables. For planform curvature, profile curvature, annual average precipitation and slope, IHS appears better than SRS with nearly all sample sizes. For annual average temperature, IHS performs better with sample sizes of 32 and 46, but not as good with sample sizes of 58 and 59. For topographic wetness index, IHS appears to have under-sampled its large values, while SRS is the opposite. There are, however, clearly great differences among the box plots of the three replicates of SRS sample sets, especially when sample sizes are small. On the other hand, distributions of IHS sample locations are much more stable with different sample sizes, even with very small sample sizes.

**Comparison of Mapping Accuracies and Responses to Sample Sizes**

Figure 5 shows the sand content in the two soil layers inferred using the iPSM method with 59 IHS and SRS samples. The maps share some similarity in high values that are found in the parent lithology units of granite and granodiorite. Apart from this, the maps differ significantly. More differences exist between soil sand content maps of IHS and SRS for the second soil layer than the first soil layer.

Figure 6, 7, and 8 displays the $r$, $\alpha \alpha$, and RMSE values for sand content estimation using the two sampling methods with the four sample sizes. It shows that IHS tends to have higher $r$, $\alpha \alpha$, and lower RMSE than SRS for both soil layers, especially for the second layer. The addition of more samples with lower degrees of representativeness generally leads to an increase of $r$ and $\alpha \alpha$ (and decrease of RMSEs) at a decreasing rate for the first soil layer, but at an unstable rate for the second soil layer. There are large differences of the $r$, $\alpha \alpha$, and RMSE values among the three SRS replicates for both soil layers, especially when sample sizes are small. The values of $r$, $\alpha \alpha$, and RMSE fluctuate greatly for the sample set replicates when additional random samples are added.

#### The Raffelson Case Study

**Comparisons on Representing the Feature Space of Environmental Covariates**

The boxplots of the environmental variables built from the original data (a total of 34, 760 pixels) and the sampled locations with the four different sample sizes (8, 17, 20, and 25) using the IHS and cLHS methods are presented in Fig. 9. It shows that cLHS performs slightly better than IHS when simulating the original environmental feature space. The three replicates of cLHS sample sets, however, show apparent differences among each other. A possible reason may be that the sample sizes employed in this study are all not big enough for this method to capture the environmental feature space well.
Comparison of Mapping Accuracies and Their Responses to Sample Sizes

Soil series maps were created using sample sets with nine sample sizes based on the IHS and cLHS sampling methods. Figure 10 shows the soil series maps created using samples sets with 25 samples based on IHS and cLHS. It is noted that the soil series map generated using IHS sample sets share greater similarity with the original soil series map than those using cLHS sets. A total of 12 soil series were mapped out using IHS sample sets (Fig. 10a). Four soil series, Dorerton, Urne, Orion, and Hixton, were not mapped. According to the soil expert who mapped this area, Elbaville occurs on linear to concave slopes while Dorerton occurs on convex nose positions (Zhu et al., 2001). On the conventional soil series map of this study area, these two components were combined as a complex map unit. The reason why IHS samples did not map out Dorerton may be that the input environmental variables are not able to recognize effectively the convex nose position using FCM clustering. Urne, Orion and Hixton are three soil series that cover only 1.6% of the study area in total. It is thus difficult to capture them using only 25 samples. At the same time, only 9 or 10 soil series were mapped out with the three cLHS sample sets. Except for soil series Elbaville, Churchtown, and Council, the spatial distributions of other mapped soil series show big differences among the three maps (Fig. 10 b, c, and d). The main reason is that with cLHS, samples may be consistently drawn from soil classes that cover large areas but the random nature of cLHS causes small soil classes not consistently sampled especially when the sample size is not big enough to capture the variability of those soil series.

Figure 11 shows the mapping accuracies with nine different sample sizes for the two sampling strategies. We note that the accuracies of maps produced from IHS samples are consistently higher than both the average and single mapping accuracies using cLHS's three replicate sample sets except for the sample size of eight. And mapping accuracies with the addition of IHS samples generally increased with a decreasing rate. At the sample size of eight, two cLHS sample sets generated higher accuracies than that of IHS. The main difference we found was that
clHS designed a sample location on soil series Churchtown (the second biggest soil series in this study area) while IHS did not (Table 2). As discussed before, clHS design tends to select soil series occupying large areas with greater probability. With IHS, the selection of samples depends on representativeness of locations determined by their representative grades and the area coverage of the environmental cluster chains. The Churchtown soil series was represented by several environmental cluster chains using IHS but did not come in the first eight sample locations that both represent a large area and with high degree of representativeness. It was picked up when sample size increased to 10 and above, and the map reached an accuracy of above 85% at merely a sample size of 14 samples. At the same size, clHS’s accuracies were still in the
The highest accuracy achieved with the largest sample size (25) in this study for IHS was 89%, while the accuracies with the same sample sizes for cLHS were 78, 79, and 87% at an average of 81.3%. It is also noted that cLHS produced unstable accuracies. The greatest difference among the three replicate sample sets of cLHS is as high as 26%.

Table 2 tallies the numbers of samples selected for the soil series in the Raffelson study area with IHS and cLHS at different sample sizes. We can see from the table that IHS and cLHS have quite different coverages of the different soil series at the same sample size. Again this is because that cLHS tends to design more sample locations on soil series with large areas. With clustering in each stratum (parent materials) and designing sample locations on clustering results in each stratum, IHS pays more attention on soil series with small areas than cLHS and is able to cover more soil types.

### DISCUSSIONS

The advantage of simple SRS is that it is easy to implement. When the relationship between soil and its environmental covariates for an area is unclear, stratified random sampling with parent materials being the strata is a good choice for preliminary investigation. However, it is shown to be much less efficient and accurate than IHS for mapping soil property in the Xuancheng case study.

The cLHS is also a stratified random sampling as the hypercube space built by all environmental covariates is stratified during stratification.
Soil Science Society of America Journal

ing the process of cLHS. The advantage of cLHS is that the distribution of the designed sample locations simulates the distribution of environmental covariates, which allows the sample set to generate accurate soil maps. In this paper, the average accuracies ranged from 64.7 through 81.3 with relatively sample sizes (8–25 samples). The accuracies, however, are still overtaken by the IHS maps. We did further experiments to keep increasing the sample sizes with cLHS and found that the average mapping accuracy with cLHS’s three sample sets reached 83.7% (individual accuracies with the three sets being 84, 82, and 85%, respectively) at a sample set of 54 samples. Compared with the sample size of 12 for IHS to reach the same accuracy, this suggests that IHS to be a more efficient sampling strategy. Moreover, one other limitation of cLHS is that it may lead to unstable results, meaning that maps resulting from sample sets of the same size at different iterations could be very different especially when the sample size is small. Due to this unstableness, some sample sets could lead to low accuracy while other sets with the same size perform satisfactorily. For example, sample sets cLHS8_2, cLHS14_1, and cLHS20_1 in our experiment (Table 2) resulted in such low accuracies. We investigated these sample sets and found that the reason lies on the soil series Churchtown. Samples of Churchtown were found to be located on the transition areas between Churchtown and other soil types (Fig. 12). We also looked into the fuzzy memberships of these sample locations to the soil series Churchtown from the original membership maps produced by Qi et al. (2006), which lead to the soil series map

Fig. 10. The soil series maps created with 25 samples for the Raffelson study area using a: IHS, b: cLHS_1, c: cLHS_2, d: cLHS_3.

Fig. 11. Mapping accuracies for IHS and cLHS with nine sample sizes.
we used for this case study. These memberships are indications of the typicality of those soil locations classified as Churchtown. The membership values of the four sample locations are 0.1 (for the sample location in cLHS8_2), 0.21 (cLHS14_1), 0.09 and 0.88 (cLHS20_1), respectively. Three of the four sample locations with very low memberships can be considered as “untypical” sample locations for Churchtown. When the designed cLHS sample locations fall into such untypical or transition areas for a soil type with big area coverage, the mapping accuracy is greatly affected.

The unique advantage of IHS is that the designed sample locations are representative and

Table 2. Number of samples selected from the different soil series in the Raffelson study area with IHS and cLHS at different sample sizes.

<table>
<thead>
<tr>
<th>Soil Series ID</th>
<th>No. of the occurring soil series</th>
<th>Mapping accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.0</td>
<td>20.9</td>
</tr>
<tr>
<td>IHS8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>cLHS8_1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>cLHS8_2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>cLHS8_3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H1S10</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>cLHS10_1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>cLHS10_2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>cLHS10_3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>H1S12</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>cLHS12_1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>cLHS12_2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>cLHS12_3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>H1S14</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>cLHS14_1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>cLHS14_2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>cLHS14_3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>H1S17</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>cLHS17_1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>cLHS17_2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>cLHS17_3</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>H1S20</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>cLHS20_1</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>cLHS20_2</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>cLHS20_3</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>H1S22</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>cLHS22_1</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>cLHS22_2</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>cLHS22_3</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>H1S24</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>cLHS24_1</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>cLHS24_2</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>cLHS24_3</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>H1S25</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>cLHS25_1</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>cLHS25_2</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>cLHS25_3</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

Fig. 12. The cLHS sample locations for soil series Churchtown in the Raffelson study area.
can be ordered based on their degrees of representativeness. Such representativeness of sample locations includes two aspects: sample locations representing typical soil variation types at different scales and covering as many soil variation types (such as soil series in the second case study) as possible. The cLHS, on the other hand, generates all independent samples from each other, which does not provide sampling order information.

Both cLHS and IHS make full use of environmental covariates but in two different ways. The results of this paper may indicate that capturing typical types of soil variations is more effective when designing sample locations for soil mapping than focusing merely on covering the environmental feature space.

At last, we should note that the success of both IHS and cLHS depends highly on how the selected environmental covariates can capture the spatial variation patterns of the local soil. Domain knowledge on environmental covariates that influence the local formation and development of soils or co-vary significantly with soils in the study area should be by all means sought for when determining such covariates.

CONCLUSIONS

This paper presented an IHS method and two case studies to compare it with two well-known sampling methods, the SRS and cLHS methods. One study focuses on the soil property mapping (soil sand content, for the Xuancheng case study) and the other on soil classification (soil series mapping, for the Raffelson case study). Results of the Xuancheng case study show that IHS is able to design samples that represent the environmental covariate feature space better than SRS. And IHS obtained higher accuracies of soil sand content estimation for both soil layers at different sample sizes. Results from the second case study demonstrate that cLHS may provide a slightly better representation of environmental covariate feature space than IHS. The mapping accuracies with IHS, however, are often higher than the average (and individual) mapping accuracies with cLHS sets of the same sample sizes. It is also observed that cLHS may be unstable in terms of the sample sets it generates and the resulting mapping accuracies. The finding that the addition of more sample locations with lower representativeness leads to an increase of accuracy at a decreasing rate in the two case studies indicates the effectiveness of the generated representativeness information of sample locations. The representativeness information can be highly valuable for stepwise or supplementary sampling when sampling resources are limited.

ACKNOWLEDGMENTS

This study is supported by National Natural Science Foundation of China (Project No. 41471178; 41431177), by the National Key Technology Innovation Project for Water Pollution Control and Remediation (Project No. 2013ZX07103006), National Basic Research Program of China (Project No.: 2015CB954102), and Natural Science Research Program of Jiangsu (14KJA170001), PAPD. Supports to A-Xing Zhu through the Vilas Associate Award, the Hammel Faculty Fellow Award, the Manasse Chair Professorship from the University of Wisconsin-Madison, and the “One-Thousand Talents” Program of China are greatly appreciated. Thanks to Dr. Yuguo Zhao for improving this manuscript, to Dr. Shujie Zhang and other students in the DGPM group for your help on collecting samples in the field, and to the reviewers for their helpful and considered suggestions on this manuscript.

REFERENCES

recipes in FORTRAN: The art of scientific computing. 2nd ed. Cambridge Univ. Press, Cambridge, UK.


