

An adaptive uncertainty-guided sampling method for geospatial prediction and its application in digital soil mapping

Lei Zhang, A-Xing Zhu, Junzhi Liu, Tianwu Ma, Lin Yang & Chenghu Zhou

To cite this article: Lei Zhang, A-Xing Zhu, Junzhi Liu, Tianwu Ma, Lin Yang & Chenghu Zhou (2023) An adaptive uncertainty-guided sampling method for geospatial prediction and its application in digital soil mapping, International Journal of Geographical Information Science, 37:2, 476-498, DOI: [10.1080/13658816.2022.2125973](https://doi.org/10.1080/13658816.2022.2125973)

To link to this article: <https://doi.org/10.1080/13658816.2022.2125973>



Published online: 26 Sep 2022.



Submit your article to this journal [↗](#)



Article views: 228



View related articles [↗](#)




View Crossmark data [↗](#)



RESEARCH ARTICLE



An adaptive uncertainty-guided sampling method for geospatial prediction and its application in digital soil mapping

Lei Zhang^{a,b} , A-Xing Zhu^{a,c,d,e,f}, Junzhi Liu^{a,c,d}, Tianwu Ma^{a,c,d}, Lin Yang^{b,e} and Chenghu Zhou^{b,e}

^aJiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China; ^bSchool of Geography and Ocean Science, Nanjing University, Nanjing, China; ^cKey Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing, China; ^dSchool of Geography, Nanjing Normal University, Nanjing, China; ^eState Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; ^fDepartment of Geography, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

Sampling design can significantly reduce the uncertainty in geospatial predictions. In this paper, we developed an adaptive uncertainty-guided stepwise sampling (AUGSS) method to select sampling locations to supplement existing legacy sample points whose representation should be improved. The proposed method selects supplemental samples in a stepwise manner as guided by an objective function with two weighted sub-objectives. One reduces the area with high prediction uncertainty, and the other minimizes the overall prediction uncertainty for the entire area. The method takes an adaptive approach to adjust weights for the two sub-objectives and to tune an uncertainty threshold controlling whether a location can be reliably predicted during the sampling procedure. A case study on soil property prediction shows that AUGSS outperforms the stratified random sampling (SRS) and the non-adaptive uncertainty guided sampling method (UGSS) in terms of RMSE and Lin's concordance correlation coefficient with different sample sizes. This study shows that the AUGSS method offers a potential for effectively adding supplemental samples to existing samples which are insufficient for spatial prediction. The adaptive strategy guided by predicted uncertainty provides an efficient support to improve the spatial pattern of samples, which plays a key role in the result accuracy of geospatial predictive mapping.

ARTICLE HISTORY

Received 14 August 2021
Accepted 14 September 2022

KEYWORDS

Spatial sampling; prediction uncertainty; adaptive approach; spatial prediction; digital soil mapping

1. Introduction

Obtaining information on the spatial distribution of natural resources is one of the important issues in geographical information science (Goodchild *et al.* 1992, 1993, Zhu *et al.* 2001, 2018, 2021, Shekhar *et al.* 2011, Jiang and Shekhar 2017, Wadoux *et al.*

2020). Information on spatial variation of some geographical variables, such as water and vegetation cover, can be directly observed by remote sensing techniques. However, spatial variation on some other geographical variables, such as soil classes/properties, hazard susceptibility and habitat suitability, are difficult to be observed directly using remote sensing techniques (Zhu *et al.* 2018). Spatial prediction (spatial predictive mapping) based on field samples is an effective way to obtain the spatial variation of these geographical variables (Webster and Oliver 1990, de Gruijter *et al.* 2006). One of the key requirements for spatial prediction based on field samples is that the field samples used should represent the area under concern well (Zhu *et al.* 2015, Zhu and Turner 2022).

A sample set is a subset of a population. Typically, making a complete enumeration of all elements in the population is impractical because the population is very large. Thus, collecting a manageable size of samples, which can represent the population well, is often desired in spatial prediction is practical and can make inferences from the sample to the population (Peck *et al.* 2015). Much effort has been devoted towards developing sampling methods for spatial prediction, such as simple random sampling based on the classical sampling theory (Kish 1965, Cochran 1977, Brus and de Gruijter 1997), spatial coverage sampling (Royle and Nychka 1998, Brus *et al.* 2006, Ma *et al.* 2020a), and sampling with the help of auxiliary information (environmental covariates) (e.g. Minasny and McBratney 2006, Brus and Heuvelink 2007, Zhu *et al.* 2008, Wadoux *et al.* 2019a, Zhang *et al.* 2022b). Many studies have shown that sampling in the feature space can achieve a higher prediction accuracy with a limited number of samples (Hengl *et al.* 2003, Brus and Heuvelink 2007, Zhu *et al.* 2008, Yang *et al.* 2013, 2016, Wadoux *et al.* 2019a). However, most of the methods mentioned above did not consider how to incorporate legacy samples into the sampling designs when such samples are available. Therefore, studies on supplemental sampling in spatial prediction are desired.

In practice, legacy samples are sometimes available within a study area (Carré *et al.* 2007, Stumpf *et al.* 2016). These samples are often limited in number and with low or biased representation of the study area due to a variety of reasons and the use of these samples along would lead to large uncertainty in spatial prediction. However, legacy sample points are not without use as they inherently contain knowledge and understanding of the local geographical environment. They are valuable resources for spatial predictions, especially when sampling budgets are limited (Rossiter 2008, Zhang *et al.* 2016, 2021). This requires researchers to collect additional samples for their study area to improve the level of representation of these legacy samples (Zhu *et al.* 2015, Li *et al.* 2016, Zhang *et al.* 2016).

One way to guide the supplemental sampling procedure is to use the prediction uncertainty generated from spatial prediction based on the existing legacy sample points. There are two basic approaches to collect additional samples to supplement legacy samples based on the uncertainty of prediction: one is based on the uncertainty related to the spatial distribution of existing samples that have insufficient or uneven coverage in geographical space (Brus and Heuvelink 2007, Stumpf *et al.* 2017, Wadoux *et al.* 2019b), other one is based on the uncertainty as measured by environmental similarity (Zhu *et al.* 2015, Zhang *et al.* 2016). Zhu *et al.* (2015) proposed an

individual predictive soil mapping (iPSM) method for predicting soil maps with limited sample data. Under the basic assumption that the more similar environmental conditions between two locations lead to more similar in soil properties (Hudson 1992), now referred to as the geographic similarity principle (Zhu *et al.* 2018), in that study the similarity between sample points and unvisited points can be calculated by the similarity of their multiple environmental covariates. Then, the predicted value at an unvisited location can be determined by integrating observed values at sample locations with high environmental similarity to that unvisited location. The prediction uncertainty at a location can be quantified and has a negatively relationship with its environmental similarity to sample locations (Zhu *et al.* 2015, Zhang *et al.* 2016). If an unvisited location has a low similarity to the existing samples, the prediction uncertainty for that unvisited location is high and the existing samples do not represent the location well. Since the positive relationship between prediction uncertainty and prediction residuals was significant (Zhu *et al.* 2015, 2018), it is reasonable to prioritize the design of supplemental sample points in the area with high prediction uncertainty. Based on this concept, Zhang *et al.* (2016) proposed a heuristic sampling scheme to select supplemental samples directed by the prediction uncertainty (Zhu 1997, Zhu *et al.* 2015). This method consists of two main stages. The first stage is to select new samples to reduce the 'NoData' area (the area with the uncertainty higher than a certain threshold and the prediction at this area cannot be reliably made due to the poor representative by existing samples) as small as possible, which is also named as the 'gap-filling' stage. The second stage is to further select additional samples to reduce the overall prediction uncertainty as low as possible, named as the uncertainty reduction stage. This sampling method can integrate the legacy samples with additional new samples effectively. In addition, the method can provide the order in which the supplemental samples should be collected based on the contributions to the reduction of uncertainty. Li *et al.* (2016) adopted this method and improved it by considering the prediction uncertainty both from the feature domain and the spatial domain.

Although the earlier proposed supplemental sampling method achieved the purpose of improving the prediction accuracy guided by the prediction uncertainty (Zhang *et al.* 2016), there were still some disadvantages in this sampling method. Firstly, the overall workflow of the sampling method was complicated. Too many parameters needed to be decided subjectively. For example, the clustering algorithm needs to be conducted within each iteration in the uncertainty reduction stage, which is time-consuming to choose the optimal number of clusters by numerous trials. Secondly, the two-stage sampling strategy forces the calculation to be divided into two separate phases. The transition from the first stage to the second stage requires subjective judgment. Thirdly, the previous method lacks an adaptive adjustment strategy for the uncertainty threshold, which controls the 'NoData' areas where the supplemental sample points can be designed. The subjective determination of the threshold leads to the inability of users to complete the sampling design in a fully automated process. These deficiencies make this sampling method difficult to use in practice. Therefore, it is desirable to develop an adaptive supplemental sampling method to increase the effectiveness and applicability of this kind of uncertainty-guided sampling method.

To address the aforementioned problems, we proposed an adaptive uncertainty-guided stepwise sampling (AUGSS) method, which aims to reduce the complexity and the subjective manual interactions in the designing process. The proposed method unifies the two stages in the previous method together and adaptively adjusting the parameters so that the method can achieve the dual goals of reducing the 'NoData' area and reducing the overall prediction uncertainty simultaneously.

The rest of this paper is organized as follows: Section Methodology introduces the theoretical background of the proposed sampling method and describes details of the algorithm; Section A Case study on soil sampling introduces a case study to demonstrate the approach; Section Results examines the method by comparing it with the stratified random sampling (SRS) and the non-adaptive uncertainty guided stepwise sampling (UGSS) methods; Section Discussion discusses the impact of parameters and the applicability of the method; Section Conclusions concludes this paper.

2. Methodology

2.1. Basic concept and overall design

Based on the prediction uncertainty map derived from the environmental similarity between sample locations and unvisited locations (Zhu *et al.* 2015, 2018, Zhang *et al.* 2016), the basic concept of the proposed supplemental sampling method is to select samples at locations with high uncertainty so that it can quickly extend the area that can be reliably predicted and reduce the overall prediction uncertainty. To achieve this goal efficiently, the proposed AUGSS method was designed to simultaneously achieve the dual objective of reducing the 'NoData' area and reducing the overall prediction uncertainty by selecting as few samples as possible. The 'NoData' area is determined by an uncertainty threshold, that is, the locations with the prediction uncertainty higher than a threshold. Since we usually need to avoid excessively high uncertainty in certain regions within a study area, therefore, sampling in the 'NoData' area needs to be given higher priority. Nevertheless, as the final objective is to reduce the overall uncertainty, it is also necessary to consider how much uncertainty can be reduced by the new sample points. Therefore, it is reasonable to combine the two sub-objectives, rather than separate them into two different stages. In the proposed method, these two objectives can be optimized simultaneously in one step. Moreover, as the importance of each of these two sub-objectives will change with the increase of the number of supplemental samples, we included two weighting parameters for controlling the importance to allow the sampling method to adaptively change the weights and the uncertainty threshold at each iteration of sampling. The overall framework of the proposed method is shown in [Figure 1](#) and consists of the following procedures:

1. Calculate the similarity between each unvisited location and the existing sample set. The calculation of the environmental similarity is described in Section Quantification of environmental similarity and prediction uncertainty.
2. Generate the prediction uncertainty map based on the environmental similarity between unvisited locations and the sample set. Refer to Section Quantification of

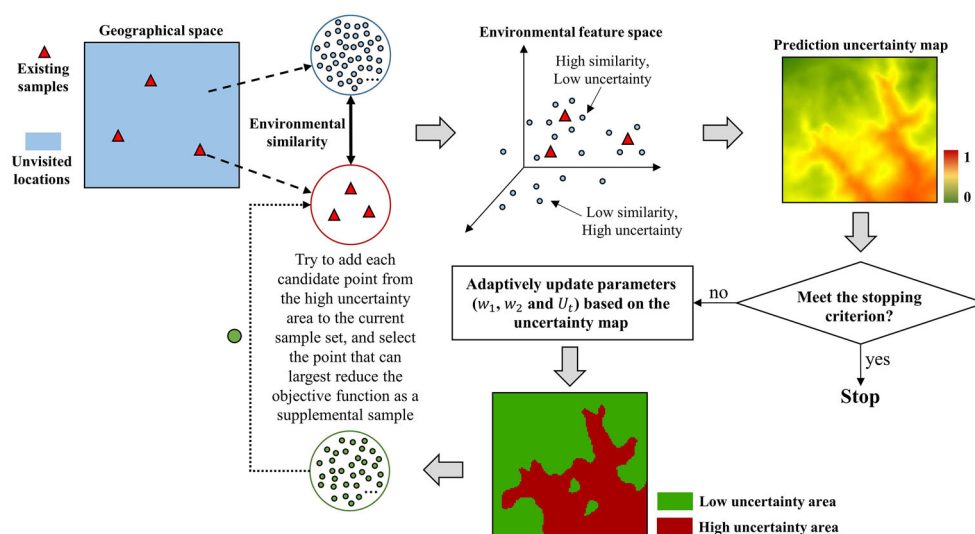


Figure 1. The overall framework of the proposed supplemental sampling method.

environmental similarity and prediction uncertainty for the calculation of prediction uncertainty.

3. Adaptively update the weighting parameters for each sub-objective and the uncertainty threshold. The detail of the adaptive approach for adjusting parameters is described in Section Adaptive approach for adjusting parameters
4. Calculate the objective function derived from a weighted combination of two sub-objective functions. The construction of the objective function is described in Section Construction of objective function.
5. Take the unvisited points in the high uncertainty area, where the prediction uncertainty values are larger than the threshold, as the candidate points. Then go through all candidate points and select the point that can minimize the objective function as a new supplemental sample and include this new sample into the current sample set.
6. Repeat the step (1)-(5) until the change of the objective function value is very small or the number of samples reaches the desired size.

2.2. Objective function based on prediction uncertainty

Based on the basic concept of uncertainty-guided sampling outlined in Section Basic concept and overall design, the selection of supplemental samples is determined by the objective function. The environmental similarity and its derived prediction uncertainty need to be defined for constructing the function. The calculation of the similarities between sample locations and unvisited locations and the construction of the objective function are described in the next two subsections.

2.2.1. Quantification of environmental similarity and prediction uncertainty

Quantification of environmental similarity is the basis for quantifying prediction uncertainty and generating the objective function. The environmental similarity

between the two points is determined by the environmental vector at these two locations. The environmental vector at location i is defined as:

$$\mathbf{e}_i = (e_i^1, e_i^2, \dots, e_i^v, \dots, e_i^m) \quad (1)$$

where m is the number of environmental covariates used. The v th element (e_i^v) in the vector represents the value of the v th environmental covariate at i th location.

The similarity between an unvisited location and a sample set can be determined as follows. The first is at the similarity at the individual environment covariate level and the second is at the location level which integrates the similarities of different environmental covariates (Zhu *et al.* 2015). The calculation is as follow:

$$S(\mathbf{e}_u, \mathbf{e}_j) = P\left(E(e_u^1, e_j^1), E(e_u^2, e_j^2), \dots, E(e_u^v, e_j^v), \dots, E(e_u^m, e_j^m)\right), \quad (2)$$

where e_u^v and e_j^v are the environmental variable values at the unvisited location u and the sampling location j . $E(\cdot)$ is the function for calculating the similarity at the covariate level, and $P(\cdot)$ is the function for the location level. Specifically, $E(\cdot)$ can be defined as:

$$E(e_u^v, e_j^v) = \frac{|e_u^v - e_j^v|}{\max(e^v) - \min(e^v)} \quad (3)$$

The overall environmental similarity between an unvisited location u and a sampling location j can be determined by integrating similarities of all environmental covariates by conducting the function $P(\cdot)$. The relative importance of different types of covariates in influencing the targeted geographical variable needs to be considered for determining the form of $P(\cdot)$. The commonly adopted ways include the weighted average method and a minimum operator based on the Liebig's Law of the Minimum (van der Ploeg *et al.* 1999, Zhu *et al.* 2015). In this study, a minimum operator was used as the function $P(\cdot)$ to integrate the environmental similarities after Zhu *et al.* (1997, 2015) and Shi *et al.* (2004) for digital soil mapping, which is used as the case study in this study. Based on the calculated environmental similarities of a given location i to all samples, an 'environmental similarities vector' at location i can be derived and formulated as shown in Equation (4).

$$\mathbf{S}_i = (S(\mathbf{e}_i, \mathbf{e}_1), S(\mathbf{e}_i, \mathbf{e}_2), \dots, S(\mathbf{e}_i, \mathbf{e}_j), \dots, S(\mathbf{e}_i, \mathbf{e}_n)) \quad (4)$$

The 'environmental similarities vector' at a location can indicate the degree of how well it is represented by the sample set. The prediction uncertainty at each location is inversely related to its environmental similarities. The higher similarity for a location to the sample locations indicates that this location can be represented by samples well, and thus, the prediction uncertainty at this location is lower. Therefore, the prediction uncertainty can be quantified using the following equation:

$$U_i = 1 - \max(S(\mathbf{e}_i, \mathbf{e}_1), S(\mathbf{e}_i, \mathbf{e}_2), \dots, S(\mathbf{e}_i, \mathbf{e}_j), \dots, S(\mathbf{e}_i, \mathbf{e}_n)), \quad (5)$$

where U_i is the uncertainty at location i . It is expected to be large when the existing samples cannot represent the unvisited location well. The prediction uncertainty calculated in this way has been proven to have a positive relationship with the prediction residual (Zhu 1997, Zhu *et al.* 2015), so it can be used as an indicator of the prediction

accuracy. It is necessary to examine whether a location can be reliably predicted by existing samples with an acceptable uncertainty. If a location with an uncertainty higher than a prescribed threshold (user provided), the prediction at this location cannot be reliably made based on user's specific requirement (Zhu *et al.* 2015).

2.2.2. Construction of objective function

The objective function consists of two sub-objective functions: O_1 is the reduction of the 'NoData' area, and O_2 is the reduction of the overall prediction uncertainty. O_1 is quantified by the proportion of the 'NoData' area where the prediction uncertainty is higher than a prescribed threshold. It can be determined as:

$$O_1 = \frac{\sum_{i=1}^N \chi(U_i - U_t)}{N} \quad (6)$$

where U_t is a threshold to control whether a location belongs to the area of high or low uncertainty and U_t will decrease after the selection of each supplemental samples. $\chi(x)$ is a conditional function, when $x > 0$, $\chi(x)$ is equal to 1, otherwise $\chi(x)$ is equal to 0, the numerator counts the total number of locations (cells or pixels) with high uncertainty. N is the total number of cells in the whole area.

The second sub-objective function focuses on the reduction of the overall prediction uncertainty. Thus, O_2 can be defined by the mean value of the overall prediction uncertainty:

$$O_2 = \frac{\sum_{i=1}^N U_i}{N} \quad (7)$$

Finally, the overall objective function is constructed by the weighted combination of O_1 and O_2 . It is formed as:

$$O = w_1 O_1 + (1 - w_1) O_2 \quad (8)$$

where w_1 and $(1 - w_1)$ are the weighting parameter for determining the importance of each sub-objective. The value of w is between 0 and 1. The adaptive approach for adjusting these weights is described in the next section.

2.3 Adaptive approach for adjusting parameters

The key issue for automatically selecting the supplemental samples is to solve the problem of how to design an adaptive approach for adjusting parameters with the addition of each new sample into the existing sample set. To achieve this automatic adaptation, approaches need to be developed for automatically adjusting the weighting parameter of sub-objective functions and the uncertainty threshold areas the iteration of sample selection progresses.

2.3.1. Adaptive approach for adjusting the weights

The two sub-objectives are both important for the sampling design. In most cases, it is necessary to reduce the 'NoData' areas at the beginning, because it is desirable that the whole area can be predicted reliably and the uncertainty at each location can be controlled within an acceptable range. Therefore, O_1 needs to be addressed with a

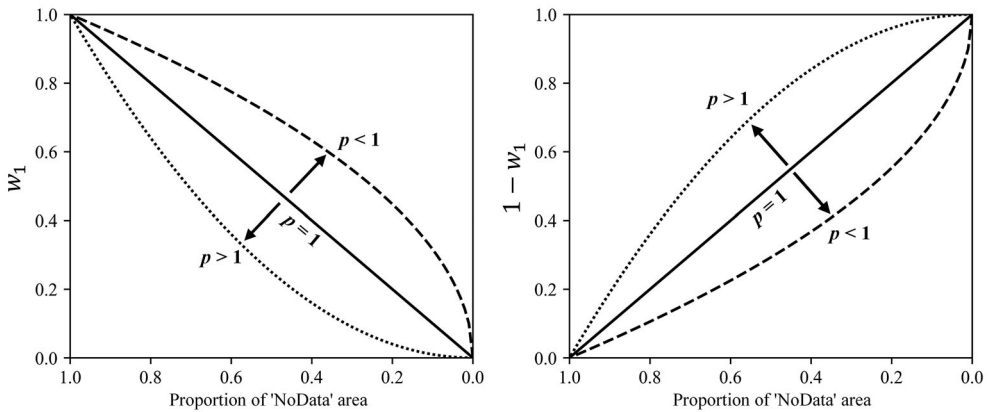


Figure 2. A conceptual illustration of the change of parameter p for controlling two weighting parameters with the change of the proportion of 'NoData' area (left for w_1 and right for $1-w_1$).

higher priority at the beginning of the sampling. Meanwhile, the overall uncertainty also needs to be considered simultaneously because it is important to find the supplemental samples which can reduce the overall prediction uncertainty as much as possible at the same time. Thus, it is reasonable to adjust the w_1 to be smaller with the decrease of O_1 , and enlarge the weight of O_2 (w_1) at the same time. Based on this concept of the importance of two sub-objective functions, we give the adaptive change function of w_1 as the following:

$$w_1 = \left(\frac{\sum_{i=1}^N \chi(U_i - U_t)}{N} \right)^p \quad (9)$$

where p ($p > 0$) is an exponential parameter which controls the relationship between O_1 and w_1 . With the increase of p , the attenuation of w_1 could be faster. Figure 2 shows an illustration of how w_1 change with different values of p .

2.3.2. Adaptive approach for adjusting the uncertainty threshold

Uncertainty threshold is another important parameter in the sampling method because it controls whether a prediction can be made at a location reliably or not. With the inclusion of supplemental samples, the overall prediction uncertainty will gradually decrease, thus, the uncertainty threshold needs to be adaptively adjusted downward with the reduction of the overall prediction uncertainty. For example, if the initial value of the uncertainty threshold is 0.4, and the mean of overall uncertainty has been reduced to about 0.2 with the increase of supplemental samples, the uncertainty threshold will be too large for the current step. This may lead to a problem that the area of relatively high uncertainty would be small and the importance of O_1 would be ignored when the uncertainty threshold is much larger than the mean of overall uncertainty. In this case the areas with relatively high uncertainty but small in size will never receive any additional sample points. Accordingly, the determination of the uncertainty threshold needs to take into account the change of the overall prediction uncertainty in each step. The uncertainty threshold can be adaptively quantified by the following equation:

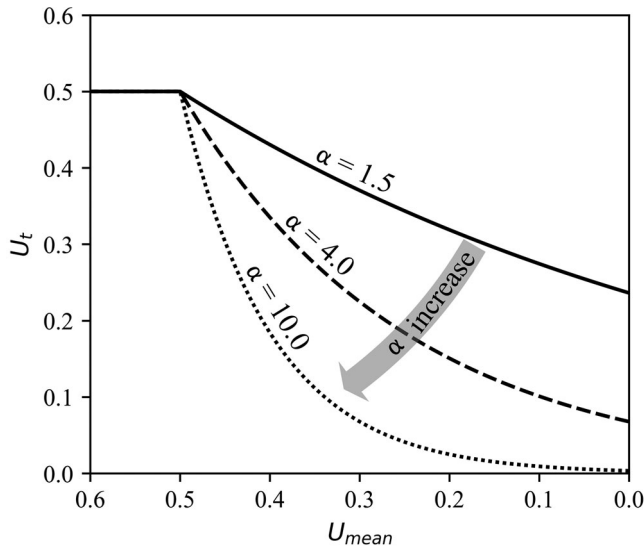


Figure 3. A conceptual illustration of the change of parameter α for controlling the uncertainty threshold (U_t) with the change of the mean of overall prediction uncertainty (U_{mean}). The initial value of the threshold (U_{init}) is 0.5 in this example.

$$U_t = \begin{cases} U_{init}, & U_{mean} \geq U_{init} \\ U_{init} \times e^{-\alpha(U_{init}-U_{mean})}, & U_{mean} < U_{init} \end{cases} \quad (10)$$

where U_t is the uncertainty threshold; U_{mean} is the mean value of the overall prediction uncertainty; U_{init} is a pre-defined initial value of the threshold with which the initial ‘NoData’ areas were determined. It can be set based on the user’s requirement or as the mean of overall uncertainty calculated when only the legacy samples are used (as the default). When $U_{mean} \geq U_{init}$, U_t is determined as U_{init} . When U_{mean} is smaller than U_{init} , U_t is determined by a monotonically decreasing function of U_{mean} . α is a user-specified parameter ($\alpha > 0$) which controls the speed of the reduction of U_t . The reduction speed is going faster if the value of α is larger. An illustration of how U_t change with different values of α is shown in Figure 3. U_t is ensured to be at least larger than zero by this function.

With the adaptive approach for adjusting the uncertainty threshold, the determination of one location whether or not can be reliably predicted could be aligned with the overall prediction uncertainty. This adaptive approach also gives users the ability to control the speed of the reduction of threshold. It makes the proposed method more flexible for the different requirements on the prediction accuracy.

3. A Case study on soil sampling

3.1. Study area and dataset

The proposed method was applied in a case study of digital soil mapping. The study area was located in Xuancheng City, in Anhui Province, China (Figure 4). The study area is about 5,900 km². The range of elevation is roughly 0 to 835 m. The annual average temperature

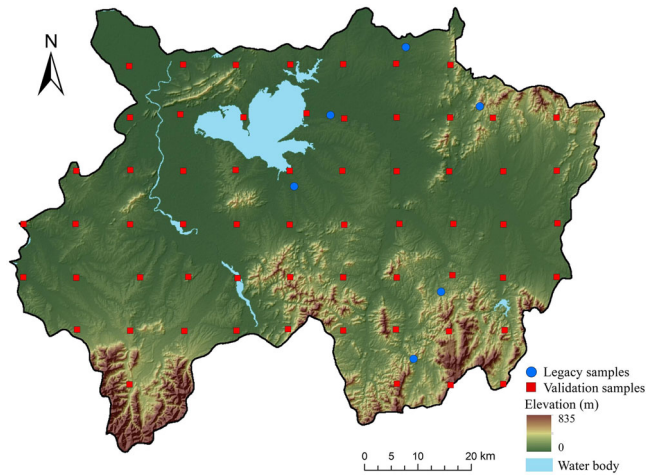


Figure 4. The elevation map and the sample points (legacy samples and validation samples) in the study area.

in this area is about 14°C , and the annual precipitation is about 1,400 mm. The area has a subtropical humid monsoon climate (Yang *et al.* 2021b). Based on previous studies in this area (Yang *et al.* 2016, 2021a; Zeng *et al.* 2016, Zhang *et al.* 2016, 2022a), the following six environmental variables were selected for the sampling experiment in the study area: slope (SLP), profile curvature (PRC), planform curvature (PLC), topographic wetness index (TWI), annual average temperature (TEMP) and annual average precipitation (PRECI). These selected environmental covariates characterize the topographical and climatic conditions of the study area. The resolution of these covariates is 90 m. Raster maps of these covariates are shown in Figure 5, and the descriptive statistics of the environmental covariates are given in Table 1. For PRC and PLC, we scaled values to a range of $0 \sim 0.5$ if values are greater than zero, otherwise, the values were scaled to a range of $-0.5 \sim 0$. All other covariates were normalized to a range of $0 \sim 1$.

There are six legacy soil samples collected in the area from the Second National Soil Survey of China, and 61 independent validation samples collected by a regular grid with a space of 10 km in the study area (Figure 4). In addition, there were additional 181 samples in the area that support our subsequent test of the proposed sampling method. These samples were collected through several field campaigns in 2011, 2015 and 2016, including 59 samples collected using the integrative hierarchical stepwise sampling strategy (Yang *et al.* 2013, Zhang *et al.* 2022b), 30 samples collected by the heuristic uncertainty directed sampling strategy (Zhang *et al.* 2016), 62 samples collected based on a stratified random sampling strategy, and the remaining samples were collected based on the environmental similarity-based recommendation (Ma *et al.* 2020b). The soil organic matter (SOM) content at the surface layer (0–20 cm) was taken as the target property for the soil prediction. The dichromate oxidation method (external heat applied) was used to measure the SOM for each sample (Nelson and Sommers, 1983, Liu *et al.* 1996).

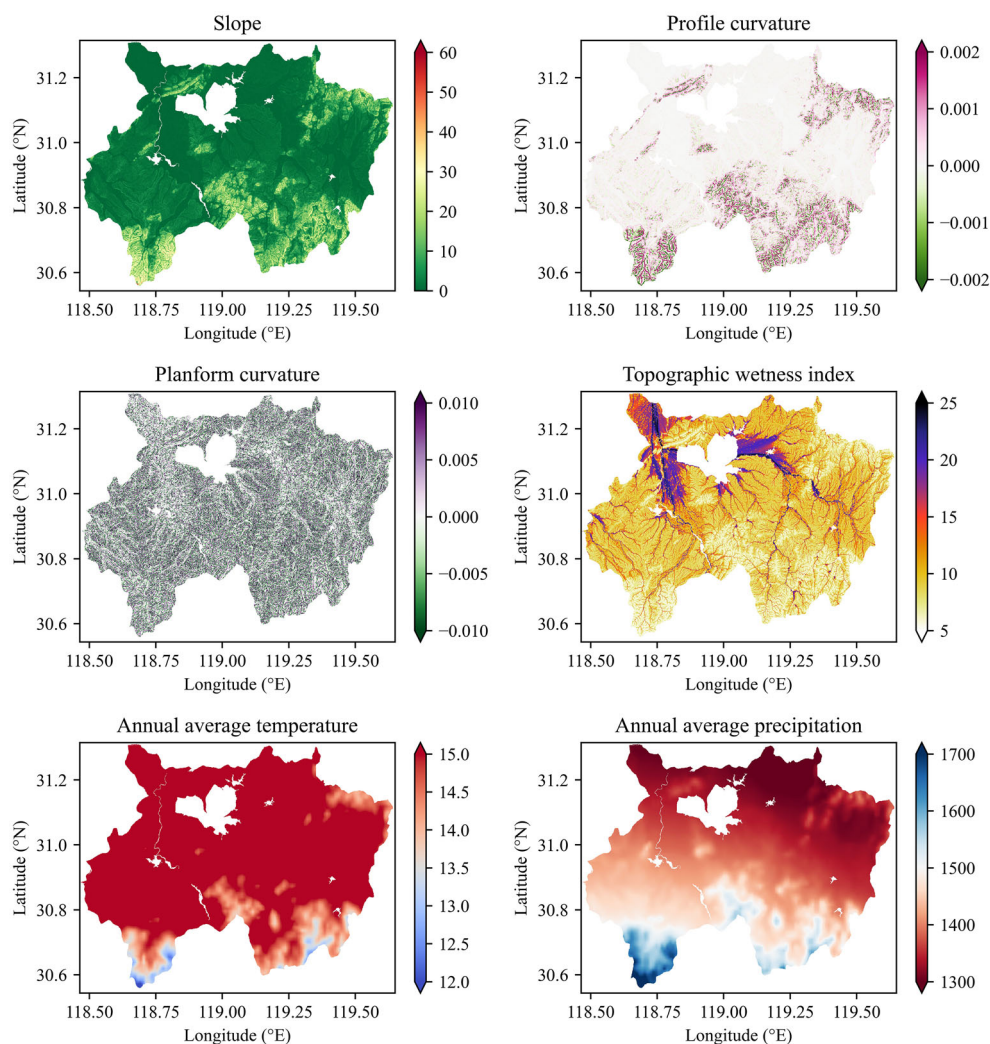


Figure 5. Maps of environmental covariates in the study area.

Table 1. The characteristic (including minimal (min), mean, median, maximal (max) and standard deviation (SD)) values of the environmental covariates.

Covariates	Min	Mean	Median	Max	SD
Slope (°)	0.000	4.376	1.796	64.261	5.992
Profile curvature (°/100 m)	-0.006	0.000	0.000	0.006	0.001
Planform curvature (°/100 m)	-0.304	0.000	0.000	0.409	0.009
Topographic wetness index (unitless)	4.207	10.180	8.996	26.244	3.737
Annual average temperature (°C)	11.575	15.220	15.450	15.779	0.567
Annual average precipitation (mm)	1243.164	1401.603	1389.640	1799.904	81.508

3.2. Experimental design

3.2.1. Parameter setting

As it is obvious that the number of legacy samples is too small to represent this large area (Figure 4), and their spatial distribution also makes it insufficient for digital soil

mapping in the area, the proposed sampling method was performed in this study area. The initial uncertainty threshold U_{init} was set to 0.4, which was used in the previous study (Zhang *et al.* 2016). The parameter p , which controls the weight of O_1 was set to 1.0, and parameter α was set to 4.0, which can make a moderate speed of the reduction of the uncertainty threshold. We selected 30 supplemental samples because the change of the objective function became very small when continuing the sampling procedure.

3.2.2. Evaluating the sampling method based on different predictive models

To evaluate the proposed sampling method, the soil-environmental relationship was generated based on three different machine learning models, including classification and regression tree (CART), random forest (RF), and support vector regression (SVR). The performances of the models were evaluated using the validation samples.

CART is a model to solve classification and regression problems based on a decision tree (Breiman, 1984). It is constructed by splitting subsets of the dataset using all covariates to repeatedly create two child nodes, and it can partition the data into subsets that are as homogeneous as possible in terms of the target variable. RF (Breiman, 2001) is a powerful machine learning model, which is widely used in geospatial predictive mapping (Wiesmeier *et al.* 2011, Heung *et al.* 2014, Hengl *et al.* 2018, He *et al.* 2021, Zhang *et al.* 2021). RF uses the voting or averaging strategy for aggregating the base learners, and it can effectively reduce the risk of overfitting and lead to a good ability for generalization. The rounded down square root of the total number of covariates was used as the parameter value of the number of covariates that randomly selected for each tree building process by default (Breiman, 2001). For another model parameter n_{tree} , which represents the number of trees to be learned in RF, was set as 200, considering the previous studies showed that it is sufficient to obtain stable results (Lopes, 2015, Wadoux *et al.* 2019b). SVR is a popular supervised learning method used for regression (Vapnik, 1995, Drucker *et al.* 1997), and has been successfully adopted for soil prediction (Kovačević *et al.* 2010, Were *et al.* 2015, Heung *et al.* 2016). Its basic concept is derived from the support vector machine (SVM), which uses kernel functions to project the data onto a new hyperspace where complex non-linear patterns can be simply represented. The regularization parameter C , which determines the trade-off between the training errors and model complexity, was set to 5. It was carried out using the grid search method using the training data (Kavzoglu and Colkesen, 2009, Were *et al.* 2015).

The model performances were evaluated using two prediction accuracy measures: the root mean square error (RMSE) and Lin's concordance correlation coefficient (CCC; Lin, 1989). RMSE and CCC are defined respectively as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Z_i - \hat{Z}_i)^2} \quad (11)$$

$$CCC = \frac{2r\sigma_z\sigma_{\hat{z}}}{\sigma_z^2 + \sigma_{\hat{z}}^2 - 2(\bar{Z} - \bar{\hat{Z}})^2} \quad (12)$$

where Z_i and \hat{Z}_i are observed and predicted values respectively; n is the number of validation samples; \bar{Z} and $\bar{\hat{Z}}$ are the averages of n observed and predicted values,

respectively; σ_z and $\sigma_{\hat{z}}$ are the corresponding standard deviations; r is the correlation coefficient value between the predicted and observed values. A smaller value of RMSE or a larger value of CCC indicates a higher prediction accuracy. Considering the stability of the model performance given that the training sample size is not large, the model was repeatedly fitted with different random seeds by the designed sample set 10 times, for reducing the influence of randomness in predictive models. The average RMSE of the ten times was taken as the validation accuracy.

3.2.3. Comparing with other sampling methods

Two other sampling methods, SRS and non-adaptive uncertainty-guided supplemental sampling (UGSS) proposed by Zhang *et al.* (2016), were adopted for comparison. SRS is a commonly used sampling strategy in soil surveys (de Gruijter *et al.* 2006, Brus, 2019). When an area can be divided into sub-areas (strata) such as different soil parent material regions, SRS is a suitable choice, and it is usually more efficient than simple random sampling and can serve as a good benchmark sampling method for comparison. There are eight parent materials in the study area, so we divided the area into eight strata and performed SRS to stepwise collect samples in each stratum. The number of samples in each stratum depends on the area proportion of the stratum in the entire study area. All SRS samples were randomly collected from the existing 62 SRS samples in the area. In order to avoid the randomness caused by SRS, we repeated the SRS method four times and generated four sets of samples for comparison.

The UGSS method is the original proposed uncertainty-guided sampling method based on the similarity between existing sample points and unvisited points in the environmental feature space. The sampling method proposed in this paper was based on its concept but improve it to be an adaptive approach. Thus, it is critical to evaluate the proposed adaptive sampling method against the non-adaptive method. The collected samples by UGSS can be referred to Zhang *et al.* (2016). Constrained by the sampling resources, the target soil properties of samples generated by AUGSS were collected from the existing samples (except the validation samples) with the most similar environmental condition to the designed locations. The similarities between these designed AUGSS sample points and the existing collected samples were all larger than 0.9. The prediction uncertainty and validation accuracy with different sample sizes (six legacy samples plus 10, 20, and 30 supplemental samples) were compared between the three sampling methods.

4. Results

4.1. Evolution of objective functions and adaptive adjustment of parameters

The evolution of the objective function and the adaptive adjusted parameters are shown in Figure 6. In general, the objection function value (O) and two sub-objective function values (O_1 and O_2) generally decreased with the increase of the number of supplemental samples, and w_1 gradually became smaller. At the beginning of sampling, because the proportion of the 'NoData' area was large and the optimizing strategy was addressed as a higher priority to select new samples that can maximally reduce the area with high uncertainty, thus w_1 was larger at the first eight steps,

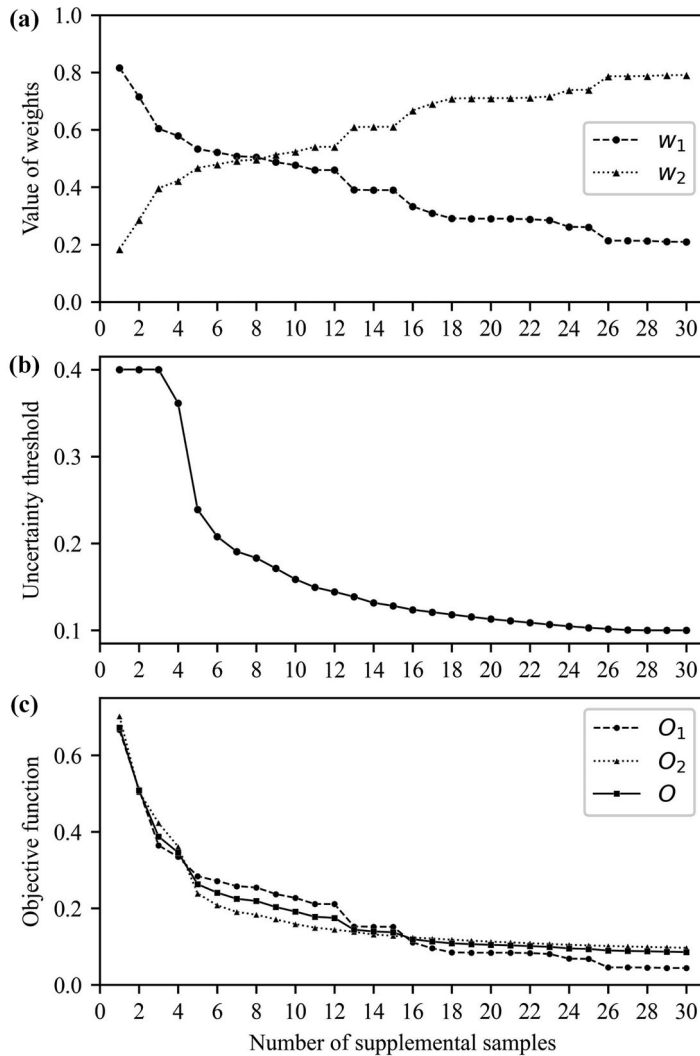


Figure 6. The evolution of the adaptive adjusted parameters (two weighting parameters (a) and uncertainty threshold (b)) and the objective function values (c) with the increase of the supplemental samples.

which was also reflected by the decreasing rate of O_1 was faster than O_2 at the beginning. Meanwhile, with the increase of the supplemental samples, the overall prediction uncertainty was reduced, and the uncertainty threshold was also adaptively adjusted (Figure 6(b)). At the early stage, the uncertainty threshold stayed at 0.4 because the mean uncertainty in the area was larger than 0.4. Then, the uncertainty threshold declined rapidly, which means that the method had effectively found supplemental samples that can greatly reduce the uncertainty. With the sample size continued to increase, the uncertainty threshold became smaller, and this made the demarcation of the area whether or not can be regarded as predictable became more rigorous. Therefore, the decreasing rate of w_1 became slow after the number of the supplemental samples reached 15. After this point, $(1 - w_1)$ increased as a result of the

proportion of the high uncertainty area became stable at a small value, thus the main objective for sampling is to reduce the overall prediction uncertainty. The overall evolution of objective functions and adjusted parameters were shown an expected behavior that could adaptively control the sampling procedure to reduce the area with high uncertainty at first and then reduce the overall prediction uncertainty in the area later.

4.2. Validation and comparison

4.2.1. Comparison of the unpredictable area and prediction uncertainty

The proposed AUGSS method was compared with SRS and UGSS methods. Figure 7(a) shows the variation of the proportion of the area with high uncertainty (pixels with prediction uncertainty larger than 0.2) as the number of the supplemental samples increases. The high uncertainty area was reduced much quickly by the two uncertainty-guided sampling methods. It is an expected behavior since one objective of the uncertainty-guided sampling is to reduce the high uncertainty area as soon as possible, which was not considered in SRS. As the UGSS method dealt with the two sub-objectives separately, it only focused on the reduction of the 'NoData' area at the first stage (i.e. the gap-filling stage), therefore, the decrease of the high uncertainty area by AUGSS was slightly slower than that of UGSS at the first 12 steps (12 additional samples). However, after the middle phase of the sampling procedure, the AUGSS method achieved a lower proportion of the area with high uncertainty than that of UGSS. The reason is that the non-adaptive method did not consider the reduction of the relatively high uncertainty area when it switched into the second stage (i.e. the uncertainty reduction stage), while the AUGSS method still considers both the two sub-objectives over the entire sampling procedure albeit with different weights. The proposed sampling method achieved the lowest proportion of high uncertainty area in the end.

Figure 7(b) shows the evolution of the mean of overall prediction uncertainty in the study area calculated by three different sampling methods. Similarly, it is obvious that the reduction of uncertainty based on the two uncertainty-guided sampling methods was

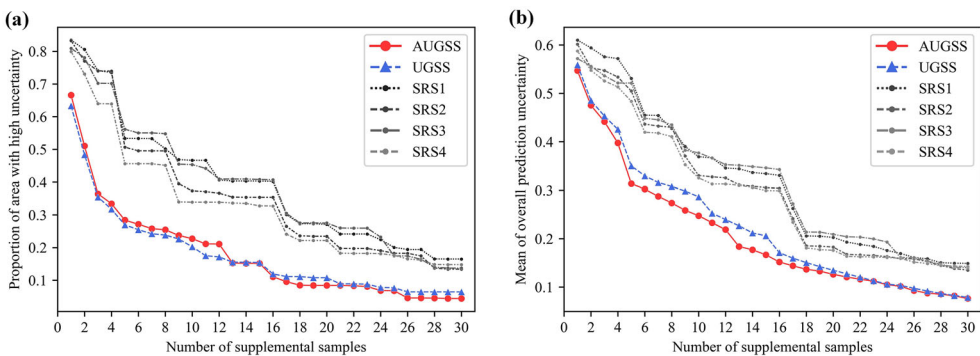


Figure 7. The change of the proportion of area with high uncertainty (a) and the mean of overall prediction uncertainty (b) with the number of supplemental samples by using three sampling methods (adaptive uncertainty-guided stepwise sampling (AUGSS), uncertainty-guided stepwise sampling (UGSS) and stratified simple random sampling (SRS)).

significantly faster than that based on SRS. It is also worth noting that the prediction uncertainty generated by AUGSS was lower than that by UGSS for almost all numbers of supplemental samples. In the beginning, UGSS only focused on the reduction of the unpredicted area, while AUGSS simultaneously consider the unpredicted area and the overall prediction uncertainty, which resulted in a much faster reduction of prediction uncertainty by AUGSS than that by UGSS. After the middle phase of the sampling procedure, the reduction of uncertainty by UGSS became faster, because it moved from the 'NoData' area reduction stage to the uncertainty reduction stage. The prediction uncertainty by AUGSS also continued to decrease with the guide of the adaptively weighted objective function. Finally, the two uncertainty-guided sampling methods obtained very similar values of the mean prediction uncertainty.

From the comparison of the reduction of the unpredictable area and prediction uncertainty across three sampling methods, it is clear that the proposed AUGSS method can effectively achieve the two sub-objectives adaptively, and notably, reduce the uncertainty faster and obtain a lower proportion of the unpredictable than the UGSS method.

4.2.2. Comparison of the validation accuracy

The soil property (SOM) was predicted based on the selected samples generated by three sampling methods. Figure 8 shows the boxplots of predicted accuracies for the three sampling methods by using three models (CART, RF and SVR) with different numbers of samples. For the three models, the average validation accuracies based on

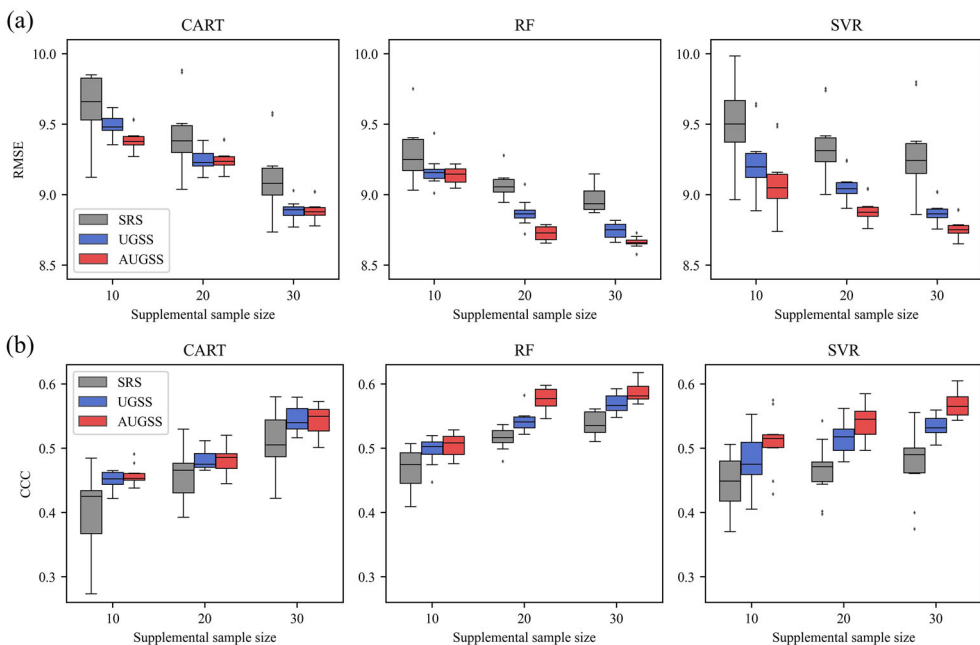


Figure 8. Boxplots of RMSE (a) and CCC (b) for three models (classification and regression tree (CART), random forest (RF) and support vector regression (SVR)) with different numbers of supplemental samples generated by adaptive uncertainty-guided sampling (AUGSS), uncertainty-guided stepwise sampling (UGSS) and stratified simple random sampling (SRS) methods. RMSE: root mean square error; CCC: concordance correlation coefficient.

Table 2. Results of Mann-Whitney *U* test for differences in RMSE and CCC with different sampling methods, predictive models, and sample sizes. Same letters within rows indicate non-significant differences at significance level of 0.05.

Supplemental sample size	RMSE									CCC								
	CART			RF			SVR			CART			RF			SVR		
	SRS	UGSS	AUGSS	SRS	UGSS	AUGSS	SRS	UGSS	AUGSS	SRS	UGSS	AUGSS	SRS	UGSS	AUGSS	SRS	UGSS	AUGSS
10	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
20	a	b	b	a	b	c	a	b	c	a	b	b	a	b	c	a	b	c
30	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c

the two uncertainty-guided sampling methods were both higher than SRS with each number of samples. Generally, the RF model was the most accurate in the study area. When using RF as the predictive model and AGUSS as the sampling method, the average RMSE was 9.1, 8.7 and 8.6, and the average CCC was 0.51, 0.58 and 0.59 with the supplemental sample size of 10, 20 and 30, respectively. The RMSE of AUGSS was 1.6%, 3.8% and 3.5% lower than SRS, and 0.3%, 1.6% and 0.9% lower than UGSS with three sample sizes, respectively. In term of CCC, the improvements of accuracy by using the AUGSS method were 7.7%, 11.8% and 8.9% compared with SRS, and 1.4%, 6.1% and 3.0% compared with UGSS with three sample sizes, respectively. Statistical tests were performed to evaluate whether accuracy differences between different methods are statistically significant under that the assumption of the validation data are representative of the population. It shows that most results of three models with different numbers of samples generated by three sampling methods were significantly different (Table 2). In addition, for all three sampling methods, the variations of the accuracies were all reduced with the increase of the sample size, but AUGSS obtained more stable results than SRS and UGSS in most cases. This result indicated that the AUGSS method can effectively select representative samples and improve the performance of the predictive model.

5. Discussion

5.1. Impact of parameters for the proposed method

There are two important parameters in the method, one is the parameter *p* for controlling two weighting parameters (w_1 and $1-w_1$), another is the parameter α for controlling the uncertainty threshold (U_t) (described in Section Adaptive approach for adjusting parameters). We used RF as the predictive model for testing the prediction accuracy with different values of parameters. Figure 9 shows the model performance (RMSE) with 30 supplemental samples under different combinations of the two parameters, which are expressed in a 2-dimensional grid. The parameter *p* was set from 0.2 to 2.0 with an interval of 0.2. The parameter α was set from 1.0 to 5.5 with an interval of 0.5. It can be noted that the two parameters both had an impact on the prediction result. Most importantly, a balanced combination of a ‘moderate’ value of the parameter could achieve a good result. Extremely large or small values will increase the RMSE of the prediction. From the figure, the optimal combination of the two parameters was 1.2 for *p* and 4.0 for α in this case study.

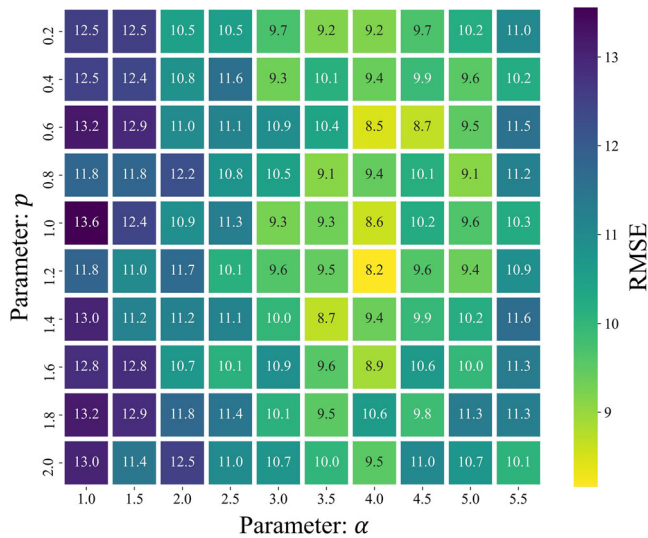


Figure 9. The prediction accuracy (RMSE) of the random forest model based on the samples generated by the proposed sampling method with the different combination values of parameter α and parameter p .

The performance of the prediction impacted by these two parameters in the sampling method probably results from the following reasons. For the parameter p , if p is too small, the decreasing rate of w_1 will be slow. This allows the importance of sub-objective O_1 to remain a high priority during the later phase of the sampling procedure, which is not conducive to reducing uncertainty. If p is too large, the decreasing speed of w_1 will be too fast, which leads to the less importance of O_1 in the early phase. For the parameter α , if it is too small, the reduction speed of the uncertainty threshold U_t will be slow. Thus, U_t will still be large at the later phase of sampling, which makes it difficult to effectively select representative samples due to the mismatch between the threshold value and the overall prediction uncertainty. On the other side, if α is too large, the uncertainty threshold will be reduced rapidly in the early phase, which will greatly limit the space for the selection of supplemental samples. Therefore, there is a moderate range which is appropriate for these two parameters. Though the best combination of parameters obtained in our study area may not simply be applied to other areas, the result here suggests that the reasonable range of α and p for soil prediction could be $3.0 \sim 5.0$ and $0.6 \sim 1.6$, respectively.

5.2. Applicability and implications

The proposed AUGSS method adopted an adaptive approach to select the supplemental samples based on the prediction uncertainty derived from the environmental similarity calculation. The method not only can select new (supplemental) samples incorporating the existing samples, but also can provide the order in selecting supplemental samples. Compared with the previously developed UGSS method, the proposed AUGSS method reduces the subjectivity in using UGSS and makes the whole

computation process run fully automatically and effectively. The uncertainty-guided stepwise sampling strategy could provide the order of additional samples based on their respective contributions to the mapped area coverage and accuracy improvement. The samples ranked in the front are supposed to be representative for the total area because they can largely reduce the area with high uncertainty. The order of supplemental samples provides us a guidance for the priority of field sampling when the resources for sampling are limited.

When there are no legacy samples in a study area, the proposed supplemental sampling method can also be performed. In this case, the entire area can be considered as an unpredicted area and the point that minimizes the high uncertainty area can be selected as the first point. However, the initial uncertainty threshold should be carefully determined by the prior knowledge of the area. In addition, the success of the sampling design also depends on the selected environmental covariates. Choosing environmental covariates that are more comprehensive for predicting the target geographical variable can increase the validity of prediction uncertainty (Zhang *et al.* 2016). The importance of each covariate for predicting a certain soil property is also critical for further improving the performance of the sampling method. As it is hard to know the contributions of different covariates before collecting sufficient samples in an area, taking advantage of expert knowledge or legacy conventional soil maps in the target area could be a potential work in the future studies.

6. Conclusions

This paper presented an adaptive uncertainty-guided supplemental sampling method to generate representative samples for improving geospatial prediction. This method unifies the two important objectives (reduction of unpredictable area and reduction of overall prediction uncertainty) into one overall objective function by using weighting parameters to regulate the importance of each of the sub-objective. Adaptive approaches for adjusting the weighting parameters and the uncertainty threshold were devised in such a way that they are automatically adjusted during the sampling procedure. From the results of a digital soil mapping case study, the proposed sampling method can simultaneously achieve the dual goal of reducing the high uncertainty area and reducing the overall prediction uncertainty by selecting the supplemental samples effectively. Compared with the stratified random sampling and non-adaptive uncertainty-guided sampling method, the proposed AUGSS method reduced the uncertainty faster and obtained a smaller proportion of the high uncertainty area, which indicated the effectiveness of the adaptive approach for adjusting the parameters. Three machine learning models were used for predicting the soil property, the AUGSS method achieved the best validation accuracy with different sample sizes in most cases. The results also suggested that the samples selected by the proposed method could achieve more stable prediction performance by using different predictive models. It is concluded that the AUGSS method offers a potential for effectively sampling for geospatial prediction, and the adaptive approach in this method can greatly reduce the difficulty of using an uncertainty-guided sampling method based on environmental similarity.

Acknowledgments

Lei Zhang thanks to the support from Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX22_0109). Supports to A-Xing Zhu through the NSFC Project (41871300), PAPD, the Vilas Associate Award, the Hammel Faculty Fellow Award, and the Manasse Chair Professorship from the University of Wisconsin-Madison are greatly appreciated. The authors express sincere gratitude to Editor May Yuan, Editor Jennifer Miller and anonymous reviewers, whose valuable comments and suggestions have greatly improved the quality of the paper.

Disclosure statement

No potential competing interest was reported by the author(s).

Funding

The study was supported by National Natural Science Foundation of China [Project No.: 41871300, 41971054, 41901062], the 111 Program of China [Approved Number: D19002], Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX22_0109), and PAPD.

Notes on contributors

Lei Zhang is currently a Ph.D. candidate at the School of Geography and Ocean Science, Nanjing University, and a visiting researcher at Wageningen University. His research interests include vegetation growth and soil carbon dynamics under the impacts of climate change and human activities, spatial predictive mapping, efficient spatial sampling strategy, remote sensing, and machine learning. L.Z.'s homepage: <https://leizhang-geo.github.io>.

A-Xing Zhu is a full professor at the Department of Geography and the Manasse Chair Professor, the University of Wisconsin-Madison. He currently serves as the Editor-in-Chief of *Annals of GIS*. His research interests focus on theoretical and methodological developments in GIS, including artificial intelligence, fuzzy logic, intelligent geocomputing, and their applications in environmental modeling and scenario analysis. His signature work includes the Third Law of Geography and similarity-based spatial prediction.

Junzhi Liu is a professor in the Center for the Pan-Third Pole Environment, Lanzhou University. His research interests mainly include land surface modeling and spatio-temporal data mining. He developed the SEIMS (Spatially Explicit Integrated Modeling System) watershed modeling framework (<https://github.com/lreis2415/SEIMS>), which has been widely used in different types of watersheds.

Tianwu Ma is a Ph.D. candidate at the School of Geography, Nanjing Normal University. His research interests include spatial prediction and sampling design.

Lin Yang is a professor in the School of Geography and Ocean Science, Nanjing University. Her research interests focus on spatial sampling, digital soil mapping, soil carbon cycling.

Chenghu Zhou is an academician of the Chinese Academy of Sciences, and currently a professor in the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. His research interests are broadly situated in research on the application of GIS and remote sensing.

ORCID

Lei Zhang  <http://orcid.org/0000-0002-1090-6338>

Data and codes availability statement

The data and codes that support the findings of this study are available with a digital object identifier (DOI) at: <https://doi.org/10.5281/zenodo.7070697>.

References

- Breiman, L., 2001. Random Forests. *Machine Learning*, 45 (1), 5–32.
- Breiman, L., et al., 1984. *Classification and Regression Trees Belmont*. CA: Wadsworth International Group.
- Brus, D.J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. *Geoderma*, 338, 464–480.
- 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 (1-2), 1–44.
- Brus, D.J., de Gruijter, J.J., and van Groenigen, J.W., 2006. Chapter 14 designing spatial coverage samples using the k-means clustering algorithm. In: P. Lagacherie, A. B. McBratney, M. Voltz, eds. *Developments in Soil Science, Digital Soil Mapping*. Amsterdam: Elsevier, pp. 183–192.
- Brus, D.J., and Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138 (1-2), 86–95.
- Carré, F., McBratney, A.B., and Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma*, 141 (1-2), 1–14.
- Cochran, W.G., 1977. *Sampling techniques*. New York: Wiley.
- de Gruijter, J.J., et al., 2006. *Sampling for Natural Resource Monitoring*. Berlin: Springer.
- Drucker, H., et al., 1997. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9, 155–161.
- Goodchild, M., Haining, R., and Wise, S., 1992. Integrating GIS and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems*, 6 (5), 407–423.
- Goodchild, M.F., Parks, B.O., and Steyaert, L.T., 1993. *Environmental modeling with GIS*. New York: Oxford University Press.
- He, X., et al., 2021. Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. *CATENA*, 205, 105442.
- Hengl, T., et al., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Soil Research*, 41 (8), 1403.
- Hengl, T., et al., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6, e5518.
- Heung, B., Bulmer, C.E., and Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. *Geoderma*, 214-215, 141–154.
- Heung, B., et al., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, 265, 62–77.
- Hudson, B.D., 1992. The soil survey as paradigm-based science. *Soil Science Society of America Journal*, 56 (3), 836–841.
- Jiang, Z., and Shekhar, S., 2017. *Spatial big data science*. Schweiz: Springer International Publishing AG.
- Kavzoglu, T., and Colkesen, I., 2009. A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 11 (5), 352–359.
- Kish, L., 1965. *Survey sampling*. New York: Wiley.

- Kovačević, M., Bajat, B., and Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. *Geoderma*, 154 (3-4), 340–347.
- Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45 (1), 255–268.
- Liu, G.S., et al., 1996. *Soil Physical and chemical analysis and description of soil profile*. Beijing: China Standardization Publishing House, pp. 131–134 (In Chinese).
- Li, Y., et al., 2016. Supplemental sampling for digital soil mapping based on prediction uncertainty from both the feature domain and the spatial domain. *Geoderma*, 284, 73–84.
- Lopes, M.E., 2015. *Measuring the algorithmic convergence of random forests via bootstrap extrapolation. Technical Report*. Davis, CA: Department of Statistics. University of California.
- Ma, T., et al., 2020a. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. *Geoderma*, 370, 114366.
- Ma, T., et al., 2020b. In-situ recommendation of alternative soil samples during field sampling based on environmental similarity. *Earth Science Informatics*, 13 (1), 39–53.
- Minasny, B., and McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32 (9), 1378–1388.
- Nelson, D.W., and Sommers, L.E., 1983. Total carbon, organic carbon, and organic matter. In: *Methods of soil analysis*. Madison, WI: John Wiley & Sons, Ltd., pp. 539–579.
- Peck, R., Olsen, C., and Devore, J.L., 2015. *Introduction to statistics and data analysis*. Boston, MA: Cengage Learning.
- Rossiter, D.G., 2008. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructures. In: *Digital soil mapping with limited data*. Dordrecht: Springer. pp. 69–80.
- Royle, J.A., and Nychka, D., 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences*, 24 (5), 479–488.
- Shekhar, S., et al., 2011. Identifying patterns in spatial information: a survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1 (3), 193–214.
- Shi, X., et al., 2004. A case-based reasoning approach to fuzzy soil mapping. *Soil Science Society of America Journal*, 68 (3), 885–894.
- Stumpf, F., et al., 2016. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. *Journal of Plant Nutrition and Soil Science*, 179 (4), 499–509.
- Stumpf, F., et al., 2017. Uncertainty-guided sampling to improve digital soil maps. *CATENA*, 153, 30–38.
- van der Ploeg, R.R., Böhm, W., and Kirkham, M.B., 1999. On the origin of the theory of mineral nutrition of plants and the law of the minimum. *Soil Science Society of America Journal*, 63 (5), 1055–1062.
- Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory*. New York: Springer Verlag.
- Wadoux, A.M.J.-C., Brus, D.J., and Heuvelink, G.B.M., 2019a. Sampling design optimization for soil mapping with random forest. *Geoderma*, 355, 113913.
- Wadoux, A.M.J.-C., Marchant, B.P., and Lark, R.M., 2019b. Efficient sampling for geostatistical surveys. *European Journal of Soil Science*, 70, 975–989.
- Wadoux, A.M.J.-C., Minasny, B., and Mcbratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210, 103359.
- Webster, R., and Oliver, M.A., 1990. *Statistical Methods in Soil and Land Resource Survey*. Oxford: Oxford University Press.
- Were, K., et al., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators*, 52, 394–403.
- Wiesmeier, M., et al., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil*, 340 (1-2), 7–24.
- Yang, L., et al., 2021a. A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology variables. *International Journal of Applied Earth Observation and Geoinformation*, 102, 102428.

- Yang, L., et al., 2021b. Extracting knowledge from legacy maps to delineate eco-geographical regions. *International Journal of Geographical Information Science*, 35 (2), 250–272.
- Yang, L., et al., 2016. Evaluation of Integrative Hierarchical Stepwise Sampling for Digital Soil Mapping. *Soil Science Society of America Journal*, 80 (3), 637–651.
- Yang, L., et al., 2013. An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping. *International Journal of Geographical Information Science*, 27 (1), 1–23.
- Zeng, C., et al., 2016. Mapping soil organic matter concentration at different scales using a mixed geographically weighted regression method. *Geoderma*, 281, 69–82.
- Zhang, S.J., et al., 2016. An heuristic uncertainty directed field sampling design for digital soil mapping. *Geoderma*, 267, 123–136.
- Zhang, L., et al., 2021. A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data. *Geoderma*, 384, 114809.
- Zhang, L., et al., 2022a. A CNN-LSTM model for soil organic carbon content prediction with long time series of MODIS-based phenological variables. *Remote Sensing*, 14 (18), 4441.
- Zhang, L., et al., 2022b. A multiple soil properties oriented representative sampling strategy for digital soil mapping. *Geoderma*, 406, 115531.
- Zhu, A.X., 1997. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. *Photogrammetric Engineering and Remote Sensing*, 63 (10), 1195–1201.
- Zhu, A.X., et al., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, 65 (5), 1463–1472.
- Zhu, A.X., et al., 2008. Purposive sampling for digital soil mapping for areas with limited data. In: A.E. Hartemink, A. McBratney, M. de L. Mendonça-Santos, eds., *Digital Soil Mapping with Limited Data*. Dordrecht: Springer Netherlands, pp. 233–245.
- Zhu, A.X., et al., 2015. Predictive soil mapping with limited sample data. *European Journal of Soil Science*, 66 (3), 535–547.
- Zhu, A., et al., 2018. Spatial prediction based on Third Law of Geography. *Annals of GIS*, 24 (4), 225–240.
- Zhu, A.X., and Turner, M., 2022. How is the Third Law of Geography different? *Annals of GIS*, 28 (1), 57–67.
- Zhu, A.X., et al., 2021. Next generation of GIS: must be easy. *Annals of GIS*, 27 (1), 71–86.