Additional Sampling Layout Optimization Method for Environmental Quality Grade Classifications of Farmland Soil

Bingbo Gao, Anxiang Lu, Yuchun Pan, Lili Huo, Yunbing Gao, Xiaolan Li, Shuhua Li, and Ziyue Chen

***Abstract*—Farmland soil environmental quality is important for farmland management. To precisely classify the environmental quality grades of farmland soil, additional samples may be required for multistage sampling or supplementary investigations. Compared with the sampling optimization methods used for mapping or estimating global means, environmental quality grade classifications are primarily focused on estimating the relationships between the values of unsampled locations and the thresholds that classify the environment quality grades. Such classifications must use a sampling layout optimization method to distribute additional sampling units into areas with a high risk of misclassification. To resolve such problems, this paper provides an additional sampling layout optimization method that initially develops a classification error index by building a multi-Gaussian model with the predicted values and error variances of unsampled locations and then calculates the probability of a threshold value occurring in the standardized Gaussian distribution. The average error indexes of all locations in the study area are then set as the objectivity function of the additional sampling layout optimization, and the spatial simulated annealing is adopted to obtain the optimized sampling layout by minimizing the objectivity function. The performance of the error index sampling layout optimization method was demonstrated in a case study using chromium concentration data for Hunan Province, China. The results showed that the additional samples generated by the proposed method produce lower and more stable classification error rates than the minimization of the mean of the shortest distances and spatially random sample methods. The proposed method can be used to improve the efficiency of additional sampling for environmental quality grade classifications of farmland soil.**

***Index Terms*—Additional sampling, environmental quality grade classification, layout optimization, multi-Gaussian model.**

I. Introduction

THE environmental quality of farmlands is of increasing public concern in developing countries [1], [2]. To guarantee food security and efficient agricultural production, environmental quality is frequently maintained via scientific farmland management [3], [4]. Soil is the fundamental element of farmlands, and many national standard and scientific reports have defined environmental quality grades according to the concentration of soil pollutants. For example, the environmental quality standard for soils in China (GB 15618-1995) defines three grades divided by thresholds of the main pollutant. Other examples include restricting planting regions for certain type of crops or regions for remediation. Spatial sampling is the most important investigation tool for farmland soil environment [5], [6]. To reduce costs, sampling optimization is performed, and the suitability of sampling methods is determined by the characteristics of the sampling purpose and the investigated population [7], [8].

Gruijter *et al*. [9] described two types of spatial sampling strategies: design-based strategies and model-based strategies. The former strategy is a combination of random sampling and design-based inference, and in this type of spatial sampling strategy, each unit of the population is assigned a probability of selection; this probability is in turn used to estimate the parameters of the population. The latter strategy is composed of purposive sampling and model-based inferences, and in this type of spatial sampling strategy, the sampling sites are chosen for a predefined purpose and a model is employed for estimations [10]. The model-based sampling strategy is more efficient when the values of unsampled locations must be predicted [11]. In model-based sampling strategy, besides interactive sampling design and fixed pattern sampling, such as nested sampling and grid sampling, objectivity function is often designed and optimized with an optimization algorithm to automatically generate a sampling plan. Three types of objectivity functions are frequently used in soil

Manuscript received September 27, 2016; revised March 6, 2017 and May 10, 2017; accepted September 12, 2017. This work was supported in part by the Research projects of agricultural public welfare industry in China under Grant 201403014-04, in part by the Capacity-Building Projects by the Beijing Academy of Agriculture and Forestry Sciences under Grant KJCX20140302 and Grant KJCX20170407, and in part by the National Natural Science Foundation of China under Grant 41601425 and Grant 41531179. *(Corresponding author: Yuchun Pan.)*

B. Gao, Y. Pan, and Y. Gao are with the Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China, and also with the National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China (e-mail: gaobb@lreis.ac.cn; panyc@nercita.org.cn; gaoyb@nercita.org.cn).

A. Lu is with the Beijing Research Center for Agricultural Standards and Testing, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China, and also with Beijing Municipal Key Laboratory of Agriculture Environment Monitoring, Beijing 100097, China (e-mail: luax@nercita.org.cn).

L. Huo is with the Agro-Environmental Protection Institute of Ministry of Agriculture, Tianjin 300191, China (e-mail: huoliliforgood@163.com).

X. Li and S. Li are with the Key Laboratory of Agri-Informatics, Ministry of Agriculture, Beijing 100097, China, and also with Beijing Engineering Research Center of Agricultural Internet of Things, Beijing 100097, China (e-mail: lixl@nercita.org.cn; lish@nercita.org.cn).

Z. Chen is with the College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China (e-mail: zychen@bnu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2017.2753467

sampling: namely, geographical space coverage, feature space coverage, and interpolation error minimization [12], [13]. The geographical space coverage objective is to distribute sampling sites evenly or according to a certain purpose. The minimization of the mean of the shortest distances (MMSD) [14] and the mean squared distance to the sides, vertices, and boundaries [15] are methods that evenly distribute sampling sites, and Warrick and Myers criteria is a method that distributes sampling sites for variogram calculation [16]. Feature space coverage objectivity is employed when ancillary data or historical data are available, and its objective is to distribute sampling sites evenly in the feature space constructed by ancillary data or available historical data. Such methods include equal range design by Hengl *et al*. [17] and the conditioned Latin hypercube method by Minasny and McBratney [18]. When a spatial variogram is available, the average or maximum estimation error can be used as an objectivity function to be minimized, and the average spatial estimation error of universal Kriging [19] or universal co-Kriging [20] and means of surfaces with stratified nonhomogeneity (MSN) error [21] are examples of such methods. However, when additional sampling sites are required to supplement previous sampling data to improve the soil environmental quality grade classification, the above-mentioned methods are not suitable for the following two reasons:

1) The spatial autocorrelation of pollutants introduces degrees of spatial continuity into the environmental quality grades; thus, new sites must be drawn from the grade transition areas, where misclassifications are most likely to happen, and sites far away from grade transition areas are less useful for the environmental quality grade classification [22]. However, these methods set the same weights for the all unsampled sites.
2) Estimated values and estimation errors should be considered at the same time when determining environmental quality grades. Thus, additional samplings should consider the relationship between the confidence interval of the estimation result and the grade threshold.

However, the above-mentioned methods either do not consider both values or consider only the estimation error [8].

The spatial uncertainty theory in geostatistics can be used to resolve the additional sampling layout optimization for farmland soil environmental quality grade classifications [23], [24]. The multi-Gaussian model can reveal the uncertainty of estimated target variables at the unsampled site. Using prior sampling data, the estimated means and variances of all sites in the sampling space can be calculated, and the multi-Gaussian model can then be built to measure the probability of incorrect grade classifications, which can then be used to guide the additional sampling layout. Thus, this paper presents an additional sampling layout optimization method to improve the precision of farmland soil environmental quality grade classifications.

The remainder of this paper is organized into three sections. Section II introduces the additional sampling layout optimization method for farmland soil environmental quality grade classifications. Section III presents a case study demonstrating the performance of the proposed sampling optimization method. Section IV discusses the sampling method and provides the conclusion.

II. Methods

A. Ordinary Kriging and Multi-Gaussian Model

Ordinary Kriging (OK) defines a random variable at each location, and the spatially dependent random variables of the study area constitute a random function, with the spatial location as an independent variable [25], [26]. The value observed at a location is considered one realization of the variable. Under stationary conditions, the mean of the variables is a constant value and the covariance between any two random variables is dependent only on the distance between them; thus, OK can be used to predict the values for unsampled locations. Using OK, the value of an unsampled location is calculated by a linear combination of nearby sampling data as

$$\hat{z}_0 = \sum_{i=1}^{n} \lambda_i z_i \tag{1}$$

where $\hat{z}_0$ is the estimated value of location *0*, z_i is the value at location *i*, and λ_i is the coefficient. To obtain the coefficients, (1) is transformed into the following equation by substituting the values with the corresponding random variables:

$$\hat{Z}_0 = \sum_{i=1}^{n} \lambda_i Z_i \tag{2}$$

where $\hat{Z}_0$ is the estimated random variable at location *0*, Z_i is the random variable at location *i*, and λ_i is the coefficient. OK employs the unbiased condition and the best condition to construct an equation set to solve the coefficients. The unbiased condition sets the mean of the error of the estimation as zero, such as in (3), and the best condition minimizes the variance of the estimation error, which is calculated by (4):

$$E(\text{ERROR}) = E\left(\hat{Z}_0 - Z_0\right) = E\left(\sum_{i=1}^{n} \lambda_i Z_i - Z_0\right) = 0 \tag{3}$$

where $E(\text{ERROR})$ is the expectation of the estimation error

$$\begin{aligned} \text{Var}\,(\text{ERROR}) &= E\left\{\left[\left(\hat{Z}_0 - Z_0\right) - E\left(\hat{Z}_0 - Z_0\right)\right]^2\right\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \text{Cov}\,(Z_i, Z_j) \\ &\quad - 2\sum_{i=1}^{n} \lambda_i \text{Cov}\,(Z_i, Z_0) + \text{Var}\,(Z_0) \end{aligned} \tag{4}$$

where $\text{Var}(\text{ERROR})$ is the variance of the estimation error and $\text{Cov}\,(Z_i, Z_j)$ is the spatial covariance between locations *i* and *j*. $\text{Cov}\,(Z_i, Z_j)$ can be substituted by the spatial variogram, which is easier to calculate. By combining (3) in (4) using Lagrange multipliers, setting the partial first derivatives with respect to each coefficient and setting the Lagrange multiplier to zero, the

equation set can be derived as

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma\left(Z_i, Z_j\right) - l = \gamma\left\{Z_0, Z_i\right\} \quad i = 1, 2, \ldots, n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{5}$$

where $\gamma(Z_i, Z_j)$ is the spatial variogram between locations *i* and *j*, and *l* is the Lagrange multiplier. The coefficients obtained by solving (5) can then be substituted into (1) and (3) to calculate the estimated value of the unsampled location and the variance of the estimation, respectively.

In the Kriging model, the random variable at each location follows a Gaussian distribution, and the variables of all locations together constitute a multi-Gaussian distribution. The mean and variance estimated above can be substituted into a Gaussian distribution to obtain the distribution of the random variable as follows:

$$G\left(x, \mu, \delta\right) = \frac{1}{\delta\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\delta^2}\right)} \tag{6}$$

where x is a random variable, μ is the variable's mean, and δ is the standard variance.

B. Error Index for Environmental Quality Grade Classifications of Farmland Soil

To improve the precision of soil environmental quality grade classifications, sampling points should be established in areas with a high risk of misclassification. For unsampled locations, the results estimated by OK include errors and uncertainty can be modeled by a multi-Gaussian distribution. Thus, the error index is defined to reflect the probability of incorrect classifications of one certain location as follows:

$$\begin{aligned} \text{Index} = G\left(\text{threshold}, \hat{z}_0, \sqrt{\text{Var}\left(\text{ERROR}\right)}\right) \\ /G\left(\hat{z}_0, \hat{z}_0, \sqrt{\text{Var}\left(\text{ERROR}\right)}\right) \end{aligned} \tag{7}$$

where *G* is the Gaussian distribution defined in (6), threshold is the critical value for classifying soil environmental into different quality grades, $\hat{z}_0$ is the estimated value of location 0 using OK, and Var(ERROR) is the variance of the estimation. The numerator of (7) is the probability of a threshold occurring in a Gaussian distribution as defined by the OK estimation result, and the denominator is the maximum probability of the Gaussian distribution. By dividing the maximum probability, Gaussian distributions of different locations are standardized and can be compared with each other.

Fig. 1 presents two standardized Gaussian distributions for two different locations. The blue line with $G(x, 60, 40)/G(60, 60, 40)$ is for location I, and the green line with $G(x, 60, 20)/G(60, 60, 20)$ is for location II. The intersection point for the blue and red lines is the error index value of location I, which is 0.755, and the intersection point for the green and red lines is the error index value of location II, which is 0.325. Although the locations have the same estimated mean value, location I has a larger error index because it has a larger estimation variance, i.e., larger uncertainty.

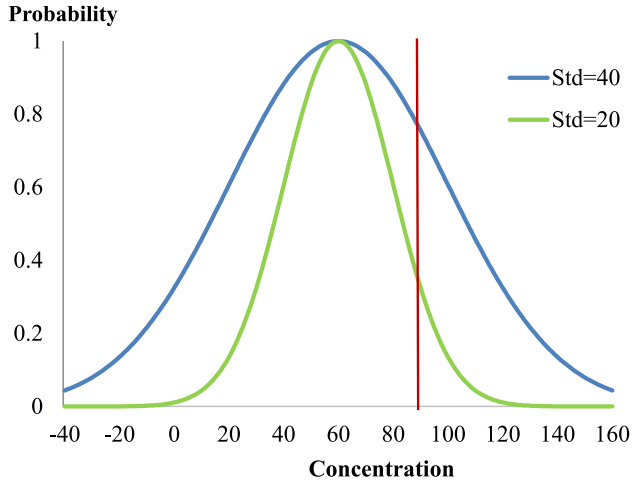


Fig. 1. Error index for two locations with the same mean but different variances.

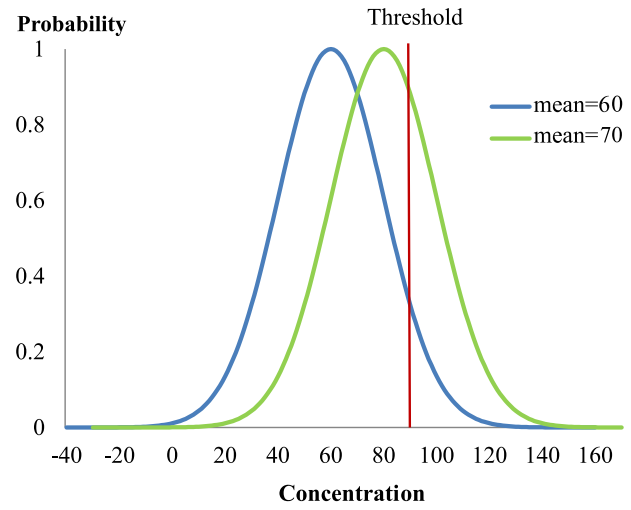


Fig. 2. Error index for two locations with the same variance but different means.

Fig. 2 also presents two standardized Gaussian distributions for two different locations. The blue line with $G(x, 60, 20)/G(60, 60, 20)$ is for location III, and the green line with $G(x, 70, 20)/G(70, 70, 20)$ is for location IV; their error index values are 0.325 and 0.882, respectively, which were obtained from the intersection points with the red line. Although locations III and IV have the same estimation error, location IV has a larger error index because its estimated mean value is closer to the threshold.

The error index considers both the estimation error and the closeness of the estimated mean value to the threshold. Larger estimation errors indicate that the estimated mean is closer to the threshold and the error index is large.

C. Additional Sampling Layout Optimization Method for Environmental Quality Grade Classifications of Farmland Soil

To design an optimized additional sampling plan for farmland soil environmental quality grade classifications, the objectivity function based on the error index of Section II-B is presented

as

$$O = \frac{1}{N * K} \sum_{k=1}^{K} \sum_{i=1}^{N} \text{Index}\,(i; T_k) \tag{8}$$

where O is the objective for minimization, N is the total number of spatial locations in the study area, K is the number of thresholds used to classify the farmland soil environmental quality grades, and $\text{Index}(i; T_k)$ is the error index of the ith location with the kth threshold.

Spatial simulated annealing (SSA) can be adopted to realize the optimization in the following six steps [27]:

1) Initialize the SSA: prepare the sample data from the previous sampling, use the data to fit the variogram and estimate the means and error variances for all unsampled locations, exclude the locations that were previously sampled to form a sampling space, calculate the error index and the initial value of O, and set the initial temperature T and cooling rate α of the SSA.
2) Produce initial additional sample: draw a sample $S1$ from the sample space, set $S = S1$.
3) Update O: use the new sample S to update the variance of each unsampled location and recalculate O.
4) Produce an additional new sample: randomly select one unit from S and replace it with a new unit selected outside of the selected one using a random radius and a random angle to form a new additional sample $S2$, and then calculate the value of the objectivity function O' with $S2$.
5) Determine whether to accept the new additional sample: if $O > O'$, accept $S2$ and set $S = S2$; if not, generate a random number r between 0 and 1 and set $m = \text{Exp}(O - O'/T)$; if $r < m$, accept $S2$ and set $S = S2$; otherwise, discard $S2$.
6) Anneal or terminate: if the stopping criterion is reached, terminate the iteration and output S; otherwise, set $T = T * \alpha$ and then return to step 3.

III. Case Study

A. Study Area and Dataset

The study area was located in the central part of Hunan Province, China, as presented in Fig. 3. In 2011, 807 samples of farmland soil were collected and analyzed to determine the concentrations of heavy metals. Chromium (Cr) was analyzed by flame atomic absorption spectrometry, and it was selected as the target variable in this paper. The environmental quality standard for the soils of China (GB 15618-1995) defines three environment quality grades for farmland soil, and a Cr concentration of 90 mg/kg is the threshold between grade I and grade II. The environmental quality grades for sampling points are shown in Fig. 3, where the red points indicate grade II and yellow points indicate grade I.

B. Experiment

In the experiment, the 807 points were treated as a population because more exhaustive data were not available. To differentiate the efficiency of different additional sampling

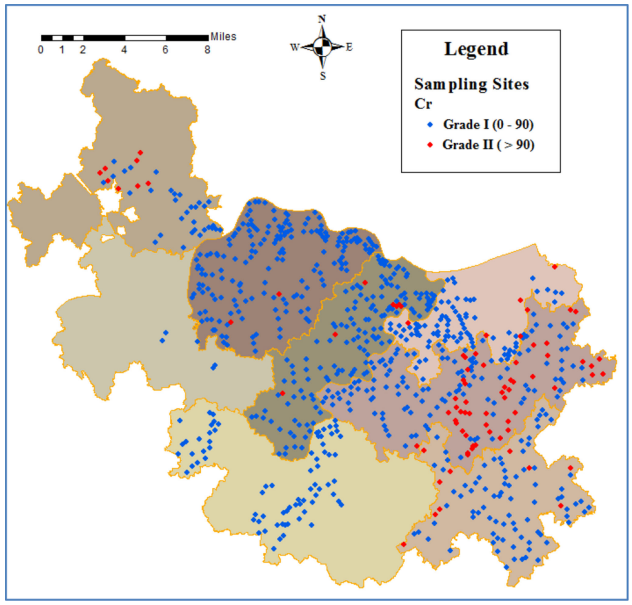


Fig. 3. Study area.

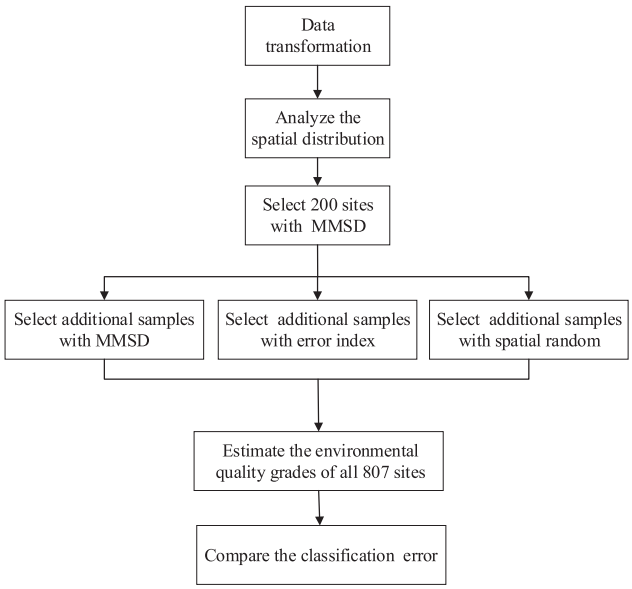


Fig. 4. Flowchart of the experiment.

method clearly, only a small part the data (200 points, about 25% of the data) was used as an initial sample, and the large part was left as sampling candidates. The experiment was conducted in the following five steps as illustrated in Fig. 4:

1) The data were transformed using logarithms because the data were skewed.
2) Analyze the spatial distribution of the data, namely, the degree of spatial autocorrelation and spatial stratified heterogeneity.
3) The initial sample with 200 points was selected from the population using the MMSD sampling method to distribute the sampling points as evenly as possible.
4) A total of 200 points were treated as the compulsory sample, and additional five samples were selected using the

error index method presented in this paper, MMSD, and spatial random sampling method.

5) A total of 200 points were combined with each additional sample to classify all unsampled points to different environment results. Then, the classified results were compared with the true to analyze the classification precision of different methods.

In the second step of the experiment, the spatial variogram was built to measure the degree of spatial autocorrelation. As the nugget of variogram is caused by random variation and the partial sill reflects the variation with distance, the ratio of nugget to sill can reflect the degree of spatial autocorrelation. It is generally thought that if the ratio is smaller than 25%, the spatial autocorrelation is the dominant characteristic. If the ratio is between 25% and 75%, moderate spatial autocorrelation exists. However, if the ration is larger than 75%, the spatial autocorrelation is thought to be weak.

The spatial heterogeneity is also an important character of spatial data and the spatial stratified heterogeneity is very common. Even if the spatial autocorrelation is high, there may be obvious spatial stratified heterogeneity. The Geodetector can be used to measure the degree of spatial stratified heterogeneity [28] as follows:

$$q = 1 - \frac{1}{n\delta^2} \sum_{h=1}^{L} n_h \delta_h^2 \tag{9}$$

where L is the total number of stratum in the study area, δ^2 is the variance of the whole study area, δ_h^2 is the variance of the stratum h, n is the size of the study area, and n_h is the size of the stratum h. The q value is in the range of 0–1. The larger the q value, the more obvious the spatial stratified heterogeneity. The significance of q value can also be tested by F-distribution [29]. The self-organization clustering algorithm considering the spatial continuity and nonspatial similarity proposed by Jiao [30] is used to stratify the data.

If the spatial autocorrelation is strong and the spatial stratified heterogeneity is weak, the prerequisites of OK are satisfied and the error index proposed in the second part can be used.

C. Results

The Cr concentration histograms before and after log transformation are presented in Fig. 5. Before log transformation, the Cr concentration is skewed and presents more small values and few larger values. After log transformation, the distribution is close to a Gaussian distribution and satisfies the OK requirements. The descriptive statistics and results of Kolmogorov–Smirnov (K–S) test of Cr concentration before and after log transformation are listed in Table I. The result also suggests that after log transformation, the distribution Cr concentration is close to a Gaussian distribution.

The variogram of Cr concentration after log transformation is presented in Fig. 6. A spherical model was adopted to fit the scattered point distribution, and it can be observed that the spatial autocorrelation is strong. The parameters of the variogram are listed in Table II . The ratio of nugget to sill is 19.950%, much lower than 25%, and the range of the variogram is 6322.197 m.

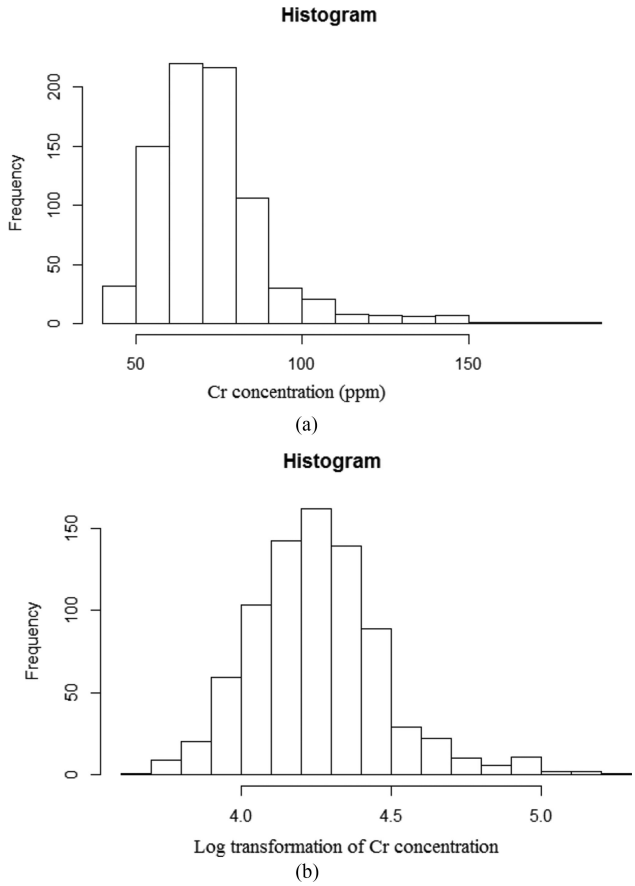


Fig. 5. Histogram of Cr concentrations before and after log transformation. (a) Histogram of the Cr concentrations. (b) Histogram of the Cr concentrations after log transformation.

TABLE I
DESCRIPTIVE STATISTICS OF CR CONCENTRATIONS BEFORE AND AFTER LOG TRANSFORMATION

Data	Std. deviation	Skewness	Kurtosis	Z (K–S)	Significance 2-tailed (K–S)
Cr	18.325	1.898	6.044	3.175	0.000
Ln(Cr)	0.226	0.748	1.527	1.659	0.080

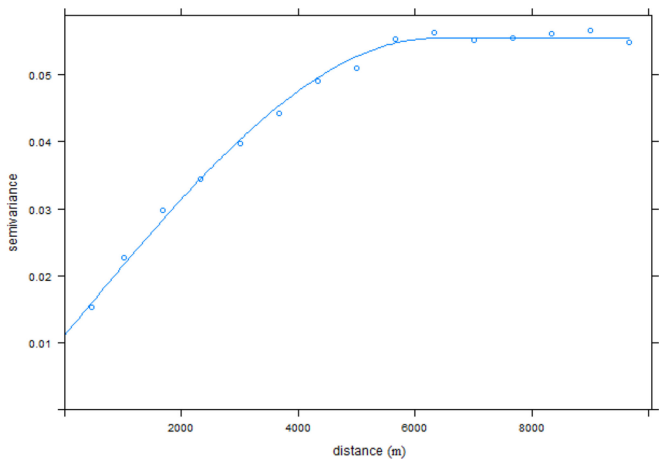


Fig. 6. Variogram of the Cr concentrations after log transformation.

TABLE II
PARAMETERS OF VARIOGRAM OF THE CR CONCENTRATIONS AFTER LOG TRANSFORMATION

Model	Nugget	Partial Sill	Nugget/Sill (%)	Range
Spherical	1.106	4.438	19.950	6322.197

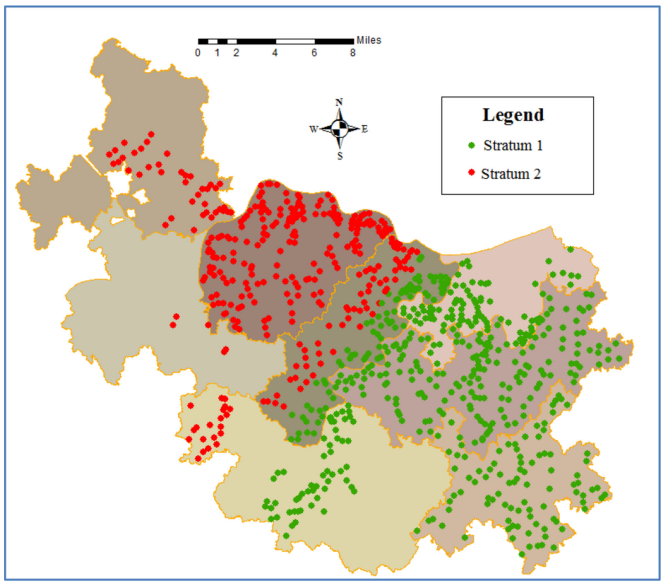


Fig. 7. Stratification results.

TABLE III
ANALYSIS RESULTS OF SPATIAL STRATIFIED HETEROGENEITY

q Statistic	p Value
0.014	0.330

All those results suggest the existence of strong spatial autocorrelation.

The stratification result of the data using the self-organization clustering algorithm is presented in Fig. 7. Two strata were classified. The first stratum locates in the southeast part of the study area, and the second part locates in the northwest part. The analysis results of the spatial stratified heterogeneity are listed in Table III. The q statistic is 0.014 and the significance p value is 0.330. From the results, it can be determined that the spatial stratified heterogeneity is very weak.

It can be determined from Fig. 3 that the grade II points are spatially clustered together and surrounded by grade I points. The transition regions between grade I and grade II have a higher risk of incorrect environment quality classifications and should have higher error indexes. The error indexes of all 807 points estimated by the initial sample are plotted in Fig. 8, where the dark points indicate the initial sample. A comparison of Fig. 8 with Fig. 3 shows that the points with higher error indexes are mostly located in areas with grade transitions and sparse sampling points. Thus, the error index developed in this paper can reveal the risk of incorrect environmental quality classifications.

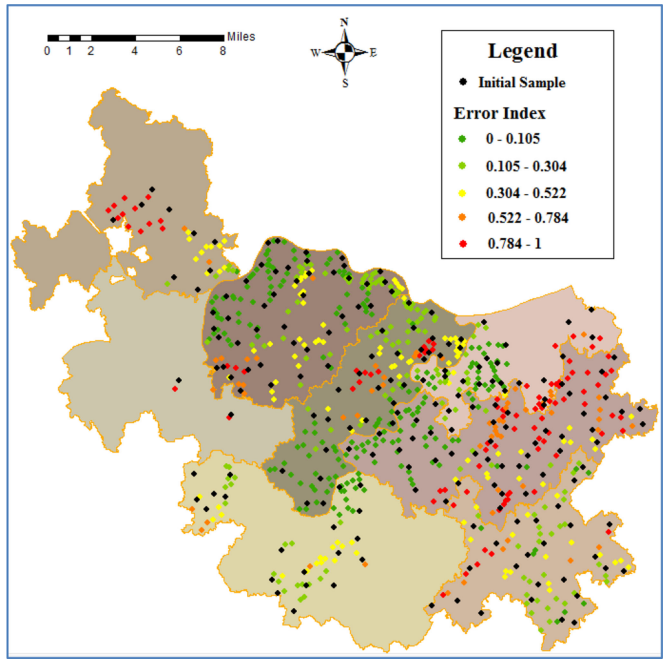


Fig. 8. Error index.

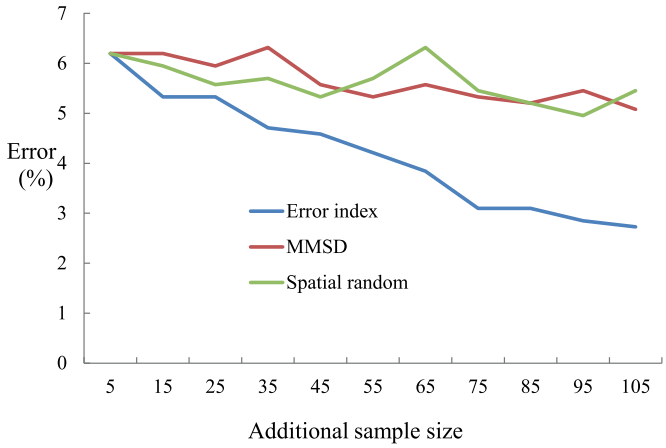


Fig. 9. Classification errors of the different sampling methods.

The errors in the soil environment grade classifications for the 807 points using the compulsory sample and additional samples generated by different methods are presented in Fig. 9. The additional samples from the error index optimization method proposed in this paper yielded lower rates of classification errors than the MMSD and the spatial random sampling method. As the sample size increases, the superiority becomes more obvious. Also, when sampling points are added, the precision of grade classification always gets improved for the error index optimization method, whereas this was not observed from the results of other two methods. When the additional sample size reaches 35, 65, and 95, the classification error of the MMSD method rises and is greater than the error of the samples with the size of 25, 55, and 85 correspondingly. And for the spatial

TABLE IV
SLOPS OF CLASSIFICATION ERROR CURVES OF DIFFERENT SAMPLING METHODS

Segment	Slop of Error Index	Slop of MMSD	Slop of Spatial Random
5–15	–0.173	0.000	–0.050
15–25	0.000	–0.050	–0.074
25–35	–0.124	0.074	0.025
35–45	–0.025	–0.149	–0.074
45–55	–0.074	–0.050	0.074
55–65	–0.074	0.050	0.124
65–75	–0.149	–0.050	–0.173
75–85	0.000	–0.025	–0.050
85–95	–0.050	0.050	–0.050
95–105	–0.025	–0.074	0.099

random sampling method, the classification error rises when additional sample size reaches 35, 55, 65, and 105. Thus, in addition to producing higher classification errors, the classification errors of the MMSD and spatial random sampling method are not stable.

The slops of the segments of classification error curves of different sampling method in Fig. 9 are listed in Table III. The slop of each segment reflects the efficiency in reducing the classification error. From Table IV, it can be found that with the increase of sample size, the reduction of the classification error becomes smaller in general for all three sampling methods. Also, the slop of classification error curves can be used to determine the sample size of additional sample. For the error index optimization method, the reduction of classification error becomes very small when the sample size is larger than 75. Thus, 75 can be suggested as an optimal sample size. For the MMSD and spatial random sampling methods, 25 can be suggested as optimal sample because the classification error increases when sample size reaches 35.

IV. DISCUSSION AND CONCLUSION

When multistage sampling or supplementary investigations require additional samples, sampling layout optimization is especially important because it is an efficient method (with regards to time and costs) that the guarantees inference precision of the new samples. Whether a sampling method is optimal for an investigation depends on the purpose of sampling or the quantities to be inferred. The environmental quality grade classification of farmland soil is a special inference target for sampling optimization that differs from mapping or global mean estimations because it does not emphasize the estimation precision of each location's value or the global mean. Rather, this classification scheme is focused on precisely estimating the relationships between the values of unsampled locations and the thresholds to classify the environmental quality grades. In such cases, it contributes little in improving the grade classification precision to sample in locations where the values are far from the threshold and difficult to misclassify. Thus, sampling units should be set in locations with a high risk of misclassification. The error index proposed in this paper fully considers the risk of misclassification. By transforming the data into a Gaussian distribution and estimating the mean and variance of unsampled locations from prior sampling data, the risk of misclassification can be quantitated by the probability of the threshold occurring in the standardized Gaussian distribution. The error index considers the closeness between the estimated value of an unsampled location and the threshold, and at the same time considers the uncertainty of the estimation. Estimated mean values that are closer to the threshold and have larger estimation errors indicate higher risk of misclassification.

By setting the average error index of all locations as the optimization object, the error index sampling layout optimization method presented in this paper can distribute more sampling sites to locations with higher risks of misclassification while avoiding the clustering of sampling sites in areas of highest risk of misclassification. Therefore, the efficiency of sampling gets improved. As illustrated in the case study, the samples from the error index optimization method generated fewer classification errors than the MMSD and spatial random sampling methods. Additionally, the classification precision was more stable than the other two methods. Thus, the error index sampling layout optimization method is suitable for environmental quality grade classifications of farmland soil.

In the case study, the threshold was defined by the environmental quality standard for the soils of China (GB 15618-1995). In practical usage, thresholds can be values defined by some related standards and research report to classify the study area into different grades. If the thresholds needed are not available, research should be performed to define them. For example, to divide the contaminated area to forbid planting certain crops, the concentration of pollutant in the crops and its health hazards should both be considered in defining the threshold. The error index can also be used in the classifications beyond the soil environment, such as land cover types [31], [32].

To use the error index sampling layout optimization, the following preconditions of OK should be satisfied: the target variable should be stationary and the data should conform to a Gaussian distribution. If the stationary precondition is not met, then the error index should be transformed to adapt to the actual conditions. If the spatial covariance is not stationary and varies with different directions or different subregions, then a combined variogram or multi-variograms should be used to estimate the means and variance, which is similar to mapping with Kriging. If the mean is not stationary, then a trend must be removed before the estimation, and during the calculation of the error index, the local trend should also be subtracted from the threshold. If the data do not follow a Gaussian distribution, then a data transform method, such as a logarithmic, Box–Cox, or Johnson transformation, should be applied. If the distribution of the concentration of soil pollutants is highly skewed and it is hard to transform the data into Gaussian distribution, the nonparametric geostatistical method such as indicator Kriging can be used to constitute the error index [33], [34].

The spatial unit in the case study is points. However, when the spatial units to be classified are polygons, the error index reflecting the risk of misclassification can also be quantitated in a similar way using block Kring (BK) [35], MSN [36], or sandwich estimation [37]. When the study area is large, the spatial heterogeneity cannot be neglected. The q-statistics [29] is a useful tool to measure the degree of spatial stratified hetero-

geneity to help to choose a proper estimation method. If the concentration of soil pollutants is homogeneous in the study area, OK and BK can be adopted to construct the error index. If a certain degree of spatial stratified heterogeneity exists, MSN is suitable to construct the error index. And if the spatial stratified heterogeneity is obvious and the spatial autocorrelation among strata are weak, the sandwich estimation should be used.

The sample size is also a key factor of an additional sampling plan besides the layout of the sampling sites. In this paper, optimal sample size was suggested according to the classification error curves. However, in practical usage, the classification error curves are often not available. In those cases, we suggest choosing the sample size using the error index. It can be carried out in at least following two ways. One is to judge by the number of spatial units whose error index is higher than a benchmark, such as 0.5 or 0.4. To facilitate the judgment, a curve with the above-mentioned number of spatial units on the vertical axis and additional sample size on the horizontal axis can be plotted. The additional sample size should be the one where the number of spatial units with high error index is zero or lower than the predefined percent. Another way is to judge by the thematic map of error index. In this way, the shape and area of clusters of spatial units whose error index are higher than a benchmark can be used to determine the additional sample size. As the additional sample size increases, the shape and area of clusters change. The additional sample size should be the one that makes the clusters thinner or smaller than the spatial precision requirement of environmental quality grade classifications. Also, the error index for sampling layout optimization can be used in multi-objective optimization to classify more than two quality grades or investigate more than one target variable. The authors will endeavor to pursue such work in the future.

REFERENCES

[1] A. Lu, J. Wang, X. Qin, K. Wang, P. Han, and S. Zhang, "Multivariate and geostatistical analyses of the spatial distribution and origin of heavy metals in the agricultural soils in Shunyi, Beijing, China," *Sci. Total Environ.*, vol. 425, pp. 66–74, 2012.

[2] D. Shao, Y. Zhan, W. Zhou, and L. Zhu, "Current status and temporal trend of heavy metals in farmland soil of the Yangtze river delta region: Field survey and meta-analysis," *Environ. Pollut.*, vol. 219, pp. 329–336, 2016.

[3] X. Tang, Q. Li, M. Wu, L. Lin, and M. Scholz, "Review of remediation practices regarding cadmium-enriched farmland soil with particular reference to China," *J. Environ. Manage.*, vol. 181, pp. 646–662, 2016.

[4] Z. Chen, B. Xu, and B. Gao, "Assessing visual green effects of individual urban trees using airborne LiDAR data," *Sci. Total Environ.*, vol. 536, pp. 232–244, 2015.

[5] M. Belli *et al.*, "A soil sampling intercomparison exercise for the ALMERA network," *J. Environ. Radioact.*, vol. 100, no. 11, pp. 982–987, 2009.

[6] X. J. Wang and F. Qi, "The effects of sampling design on spatial structure analysis of contaminated soil," *Sci. Total Environ.*, vol. 224, nos. 1–3, pp. 29–41, 1998.

[7] J.-F. Wang *et al.*, "Design-based spatial sampling: Theory and implementation," *Environ. Model. Softw.*, vol. 40, pp. 280–288, 2013.

[8] Z. Chen, M. Li, and Z. Chen, "The study on adaptive spatial sampling used in allocation of housing price monitoring sites taking Wujin section of Changzhou city as an example," *Geoinformat.: Geospatial Inf. Sci.*, vol. 6753, 2007, Art. no. 67531A.

[9] J. d. Gruijter, D. Brus, M. Bierkens, and M. Knotters, *Sampling for Natural Resource Monitoring*, New York, NY, USA: Springer, 2006.

[10] J. Wang, R. Haining, and Z. Cao, "Sample surveying to estimate the mean of a heterogeneous surface: Reducing the error variance through zoning," *Int. J. Geographical Inf. Sci.*, vol. 24, no. 4, pp. 523–543, 2010.

[11] D. Brus and J. De Gruijter, "Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion)," *Geoderma*, vol. 80, no. 1, pp. 1–44, 1997.

[12] J.-F. Wang, A. Stein, B.-B. Gao, and Y. Ge, "A review of spatial sampling," *Spatial Statist.*, vol. 2, pp. 1–14, 2012.

[13] B. Gao, Y. Pan, Z. Chen, F. Wu, X. Ren, and M. Hu, "A spatial conditioned Latin hypercube sampling method for mapping using ancillary data," *Trans. GIS*, vol. 20, no. 5, pp. 735–754, 2016.

[14] J. W. van Groenigen, W. Siderius, and A. Stein, "Constrained optimisation of soil sampling for minimisation of the kriging variance," *Geoderma*, vol. 87, nos. 3–4, pp. 239–259, 1999.

[15] D. L. Stevens Jr., "Spatial properties of design-based versus model-based approaches to environmental sampling," in *Proc. 7th Int. Symp. Spatial Accuracy Assess. Nat. Resour. Environ. Sci.*, 2006, pp. 119–125.

[16] A. W. Warrick and D. E. Myers, "Optimization of sampling locations for variogram calculations," *Water Resour. Res.*, vol. 23, no. 3, pp. 496–500, Mar. 1987.

[17] T. Hengl, D. G. Rossiter, and A. Stein, "Soil sampling strategies for spatial prediction by correlation with auxiliary maps," *Soil Res.*, vol. 41, no. 8, pp. 1403–1422, 2004.

[18] B. Minasny and A. B. McBratney, "A conditioned Latin hypercube method for sampling in the presence of ancillary information," *Comput. Geosci.*, vol. 32, no. 9, pp. 1378–1388, 2006.

[19] D. J. Brus and G. B. M. Heuvelink, "Optimization of sample patterns for universal kriging of environmental variables," *Geoderma*, vol. 138, nos. 1–2, pp. 86–95, 2007.

[20] Y. Ge, J. H. Wang, G. B. M. Heuvelink, R. Jin, X. Li, and J. F. Wang, "Sampling design optimization of a wireless sensor network for monitoring ecohydrological processes in the Babao river basin, China," *Int. J. Geographical Inf. Sci.*, vol. 29, no. 1, pp. 92–110, 2015.

[21] B.-B. Gao, J.-F. Wang, H.-M. Fan, K. Xu, M.-G. Hu, and Z.-Y. Chen, "A stratified optimization method for a multivariate marine environmental monitoring network in the Yangtze river estuary and its adjacent sea," *Int. J. Geographical Inf. Sci.*, vol. 29, no. 8, pp. 1332–1349, 2015.

[22] M. Van Meirvenne and P. Goovaerts, "Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold," *Geoderma*, vol. 102, nos. 1–2, pp. 75–100, 2001.

[23] P. Goovaerts, *Geostatistics for Natural Resources Evaluation*. New York, NY, USA: Oxford Univ. Press, 1997.

[24] K.-W. Juang, W.-J. Liao, T.-L. Liu, L. Tsui, and D.-Y. Lee, "Additional sampling based on regulation threshold and kriging variance to reduce the probability of false delineation in a contaminated site," *Sci. Total Environ.*, vol. 389, no. 1, pp. 20–28, 2008.

[25] L. Tang and F. Hossain, "Understanding the dynamics of transfer of satellite rainfall error metrics from gauged to ungauged satellite gridboxes using interpolation methods," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 4, no. 4, pp. 844–856, Dec. 2011.

[26] Y. Zhang *et al.*, "Spectral–spatial adaptive area-to-point regression kriging for MODIS image downscaling," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 10, no. 5, pp. 1883–1896, May 2017.

[27] J. Van Groenigen and A. Stein, "Constrained optimization of spatial sampling using continuous simulated annealing," *J. Environ. Qual.*, vol. 27, no. 5, pp. 1078–1086, 1998.

[28] J. Wang *et al.*, "Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China," *Int. J. Geographical Inf. Sci.*, vol. 24, no. 1, pp. 107–127, 2010.

[29] J.-F. Wang, T.-L. Zhang, and B.-J. Fu, "A measure of spatial stratified heterogeneity," *Ecol. Indic.*, vol. 67, pp. 250–256, 2016.

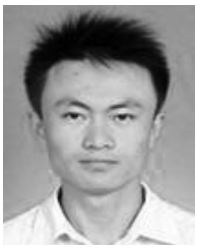
[30] L. Y. Jiao, Li Min, and Z. Bin, "Self-organizing dual clustering considering spatial analysis and hybrid distance measures," *Sci. China Earth Sci.*, vol. 54, no. 8, pp. 1268–1278, 2011.

[31] Z. Chen and B. Gao, "An object-based method for urban land cover classification using airborne LiDAR data," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 7, no. 10, pp. 4243–4254, Oct. 2014.

[32] Z. Chen, B. Xu, and B. Gao, "An image-segmentation-based urban DTM generation method using airborne LiDAR data," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 9, no. 1, pp. 496–506, Jan. 2016.

[33] A. G. Journel, "Nonparametric estimation of spatial distributions," *J. Int. Assoc. Math. Geol.*, vol. 15, no. 3, pp. 445–468, 1983.

[34] B. Gao *et al.*, "Error index for additional sampling to map soil contaminant grades," *Ecol. Indic.*, vol. 77, pp. 129–138, 2017.
[35] E. H. Isaaks and R. M. Srivastava, *An Introduction to Applied Geostatistics*. New York, NY, USA: Oxford Univ. Press, 1989.
[36] J. F. Wang, G. Christakos, and M. G. Hu, "Modeling spatial means of surfaces with stratified nonhomogeneity," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 12, pp. 4167–4174, Dec. 2009.
[37] J.-F. Wang, R. Haining, T.-J. Liu, L.-F. Li, and C.-S. Jiang, "Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface," *Environ. Plan. A*, vol. 45, no. 10, pp. 2515–2534, Oct. 2013.



Bingbo Gao received the B.Sc. degree from Nanjing University, Nanjing, China, in 2006, the M.Sc. degree from the Institute of Geography, Chinese Academy of Sciences, Beijing, China, in 2009, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2015, all in geographical information science.

Currently, he is in the Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. His research interests include spatial statistics and spatial information grid.

Anxiang Lu received the B.Sc. degree in chemistry from Nanjing University, Nanjing, China, in 2001, and the M.Sc. and Ph.D. degrees in environmental sciences from the Chinese Academy of Sciences, Beijing, China, in 2004 and 2012, respectively.

Currently, he is in the Beijing Research Center for Agricultural Standard and Testing, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. His research interest focuses on environmental monitoring.

Yuchun Pan received the B.Sc. degree in geography from Anhui Normal University, Wuhu, China, in 1996, the M.Sc. degree in geographical information science (GIS) from the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences (CAS), Changchun, China, in 1999, and the Ph.D. degree in GIS from the Institute of Geographic Sciences and Nature Resources Research, CAS, Beijing, China, in 2002.

Currently, he is in the Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. His research interests include spatial analysis and land resource management.

Yunbing Gao received the B.E. degree in surveying and mapping engineering from Henan Polytechnic University, Jiaozuo, China, in 2001, the M.E. degree in geographical information science (GIS) from China University of Mining & Technology, Beijing, China, in 2007, and the Ph.D. degree in GIS from China Agricultural University, Beijing, China, in 2016.

Currently, he is with the National Engineering Research Center for Information Technology in Agriculture, Beijing, China. His research interests include spatio-temporal data analysis and quality evaluation of cultivated land.

Lili Huo received the B.Sc. degree in resources and environment urban and rural planning from Northeast Agricultural University, Harbin, China, in 2008, and the M.Sc. and Ph.D. degrees in environmental science from the University of Chinese Academy of Sciences, Beijing, China, in 2013.

Currently, she is in the Agro-Environmental Protection Institute, Ministry of Agriculture, Chinese Academy of Agriculture Science, Tianjin, China. Her research interest focuses on agro-environmental protection.

Xiaolan Li received the B.Sc. degree in surveying and mapping major from Wuhan University, Wuhan, China, in 2011, and the M.Sc. degree in geographical information science (GIS) from the Institute of Geography, University of Chinese Academy of Sciences, Beijing, China, in 2014.

Currently, she is working in the Key Laboratory of Agri-Informatics, Ministry of Agriculture, Beijing, China. Her research interests mainly include GIS technology application and spatial analysis.

Shuhua Li received the B.E. degree in management information system from the Communication University of China, Beijing, China, in 2001, the M.Agr. degree in forestry management from the Institute of Resource information, Chinese Academy of Forestry Sciences, Beijing, China, in 2007, and the Ph.D. degree in agricultural remote sensing from the Institute of Agricultural Resource and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing, China, in 2016.

Currently, she is with the Beijing Engineering Research Center of Agricultural Internet of Things, Beijing, China. Her research interests include application for cloud services and Internet of Things in agriculture.

Ziyue Chen received the B.Sc. and M.Sc. degrees in geographic information science from Nanjing University, Nanjing, China, in 2006 and 2009, respectively, and the Ph.D. degree in geography from the University of Cambridge, Cambridge, U.K, in 2014.

Currently, he is in the College of Global Change and Earth System Science, Beijing Normal University, Beijing, China. His main research interests include the processing and applications of airborne LiDAR data and spatial data analysis.