Transactions in GIS

Research Article Transactions in GIS, 2016, 20(5): 735–754

A Spatial Conditioned Latin Hypercube Sampling Method for Mapping Using Ancillary Data

Bingbo Gao,[*][†][‡] Yuchun Pan,[*] Ziyue Chen,[§] Fang Wu,[**] Xuhong Ren[*] and Maogui Hu[†]

[*]National Engineering Research Center for Information Technology in Agriculture, Beijing
[†]Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, Beijing
[‡]College of Resources and Environmental Sciences, University of Chinese Academy of Sciences, Beijing
[§]College of Global Change and Earth System Science, Beijing Normal University
[**]Information Center, China Waterborne Transport Research Institute, Beijing

Abstract
For obtaining maps of good precision by the spatial inference method, the distribution of sampling sites in geographical and feature space is very important. For a regional variable with trends, the predicting error comes from trend estimation, variogram estimation and spatial interpolation. Based on the cLHS (conditioned Latin hypercube Sampling) method, a sampling method called scLHS (spatial cLHS) considering all these three aspects with the help of ancillary data is proposed in this article. Its advantage lies in simultaneously improving trend estimation, variogram estimation and spatial interpolation. MODIS data and simulated data were used as sampling fields to draw sample sets using scLHS, cLHS, cLHS with x and y coordinates as covariates, simple random and spatial even sampling methods, and the distribution and prediction errors of sample sets from different methods were evaluated. The results showed that scLHS performed well in balancing spreading in geographic and feature space, and can generate points pairs with small distances, and the sample sets drawn by scLHS produced smaller mapping error, especially when there were trends in the target variable.

1 Introduction

In natural resource and environment surveys, maps describing the spatial distribution and variation of the target variable need to be created (Graniero and Robinson 2003). To create maps by sampling, values of the target variable at unvisited locations need to be predicted using the observed data at sampling sites (Brus and De Gruijter 1997). In practice, the target regional variables to be mapped are often non-stationary, for example terrain (Lloyd and Atkinson 2002). In such cases, Universal Kriging is one of the most frequently used mapping methods (Chen and Li 2012). In mapping using Universal Kriging, Brus and Heuvelink (2007) thought that both the trend estimation error and the spatial interpolation errors needed to be minimized. In addition, a precise variogram is of no less importance. Hengl et al. (2004) pointed out that the overall interpolation error depends upon the coverage of the sample in both feature space (also called attribute space, which often consists of target variables or ancillary variables)

Address for correspondence: Yuchun Pan, National Engineering Research Center for Information Technology in Agriculture, Shuguang Huayuan Middle Road 11#, Beijing 100097, China. E-mail: yuchun_pan@sina.com.cn
Acknowledgements: This study was funded by the National Science and Technology Support Program of China (No. 2013BAJ05B01).

doi: 10.1111/tgis.12176

and geographical space. With development of data acquisition technology such as remote sensing, many exhaustive data are available to act as ancillary variables to guide sampling (Martinez et al. 2010; Mulder et al. 2013). Focusing on a sampling method for mapping non-stationary target variables with the help of ancillary data, the scLHS (spatially conditioned Latin Hypercube Sampling) is put forward in this study to obtain samples with good coverage in both spaces, retain points pairs with small distances which are important for variogram estimation, and produce maps with small errors.

The remainder of the article is organized as follows. Section 2 reviews the related work. Section 3 presents details of the scLHS method and the evaluation criteria for distribution of sample sets in feature and geographical space. Section 4 introduces the empirical study using a simulated dataset and a MODIS dataset to illustrate scLHS. Section 5 discusses the advantages and applicable conditions of scLHS. Finally, Section 6 draws the conclusions.

2 Related Work

Sampling is a method for selecting a subset of representative individuals from a population to estimate characteristics of the whole population. It is called spatial sampling in a spatial context, where a two-dimensional area is treated as population. A good spatial sampling method tries to obtain inferences of higher quality with less cost. A sampling strategy consists of two parts, a design p and an estimator t, donated by (p, t). 'Design' refers to the procedure determining the sample selection, while 'estimator' points to the procedure to calculate inferences from the sample (Brus and Heuvelink 2007). Gruijter et al. (2006) thought that the design-based method was a combination of probability sampling and design-based inference, whereas the model-based method was a combination of purposive sampling and model-based inference. In the design-based method, the values in the interested region are regarded as unknown but fixed and the primary source of randomness comes from the chosen sampling sites (De Gruijter and Ter Braak 1990; Särndal et al. 2003). In model-based sampling, on the other hand, the values in the interested region are regarded as one realization of a stochastic model. The randomness in this method is introduced by the stochastic models which define superpopulations (Gruijter et al. 2006; Haining 2003). The design-based method is more suitable for tackling "how much" questions, for example estimating global quantities including the frequency distribution of the target variable and the parameters of this distribution, such as the mean, the standard deviation or quantiles. The model-based method is more suitable for "where" questions, such as predicting values of un-sampled sites, or estimating the parameters of the superpopulation (Wang et al. 2010). Thus, when the sampling purpose is to generate a map, the model-based is frequently employed .

In the model-based method, besides simple purposive sampling such as square grids, triangular grids, transect and nested sampling, optimization models serving certain purposes have also been put forward to generate an optimal sampling plan, such as a minimal Kriging variance model, optimized spatial coverage model and a feature space coverage model (Wang et al. 2012). In the minimal Kriging variance optimization model, one example is postulating an optimum model of average Kriging interpolation error and then generating the sampling pattern using SSA (Spatial Simulated Annealing) (Bertolino et al. 1983; Van Groenigen et al. 1999). It requires an accurate variogram which is often unknown before sample design, so in practical sampling the other two kinds of model are used more often. To place the sampling sites as evenly as possible in space, Stevens (2006) proposed the mean squared distance criteria based on the Thiessen polygon. Van Groenigen and Stein (1998) utilized the SSA to realize the MMSD (Minimization of the Mean of the Shortest Distances) criteria, to minimize the mean distances of un-sampled sites to the nearest sampling site. Chen et al. (2012) proposed to generate even sampling designs within a given irregular polygon via simulating the movements of some ideal homogeneous point charges. For another pattern of spatial coverage, Groenigen (1997) put forward the WMSD (Weighted Mean of the Shortest Distances) criteria to give different weights to different parts. Many spatial sampling methods considering feature space coverage are also suggested. McBratney et al. (1999) introduced the variance quad-tree method in which the study area is divided into four equally sized strata recursively until the variation of each stratum reach certain threshold, and each stratum is then sampled independently. Hengl et al. (2004) proposed the ER (Equal Range stratification) design to allocate points uniformly over the attribute range by sampling proportionally to the distribution of the ancillary variable. The LHS (Latin hypercube sampling) provides an efficient way to sample multiple variables according to their distributions by constituting a Latin hypercube of multi-dimensions (McKay et al. 2000). Based on this, Minasny and McBratney (2006) put forward the more practical cLHS (conditioned Latin hypercube method) to get combinations of the ancillary values that correspond to existing sites. Because coverage in geographic space is not specially considered, the feature space coverage optimization models mentioned above are more suitable for sampling for regression or global parameter estimation, rather than spatial interpolation to create a map reflecting the spatial distribution of the target variable.

For mapping of non-stationary target variables, some efforts have been made in sampling design to improve the precision. Lin et al. (2011) combined the cLHS method and variance quad-tree to sample geographical and feature space at the same time. Similarly Simbahan and Dobermann (2006) proposed to firstly stratify using ancillary data and then combine MMSD+WM (Warrick-Myers criterion) optimization to draw a sample for variogram estimation and interpolation in one time. However, the determinant trends of the target variable cannot be adequately dealt with by simply stratifying. Brus and Heuvelink (2007) proposed a sampling method comprising coverage in geographic and feature space, based on the Universal Kriging model. This method is good at sampling optimization for non-stationary areas, but it requires the Universal Kriging model (variogram and structure of trend) to be known before sampling, a requirement which may not be satisfied in some cases. So there is still a lack of feasible sampling design methods for mapping non-stationary target variables.

3 Methodology

3.1 Conditioned Latin Hypercube Sampling

LHS was initially developed to reduce the number of sample sets required in Monte Carlo simulation by stratifying the variables and drawing representative sample sets (Brus and Heuvelink 2007; Iman and Conover 1980; McKay et al. 2000). It has been widely used to estimate the uncertainty of prediction models and for conditional simulations (Florian 1992; Minasny and McBratney 2006).

LHS first stratifies each variable independently into continuous intervals (strata) according to its CDF (Cumulative Distribution Function) in such a way that the integration of the cumulative distribution of every interval is equal. Then sampling units are drawn randomly from each interval; usually the number of divided intervals equals the sample size and only one unit is selected in each interval. Finally, units obtained for each of the variables are paired with each other either in a random way or based on some rules to constitute a sample.

When using ancillary data to guide sampling in feature space, the LHS method cannot be directly used because it may arrive at combinations of values that do not correspond to any existing

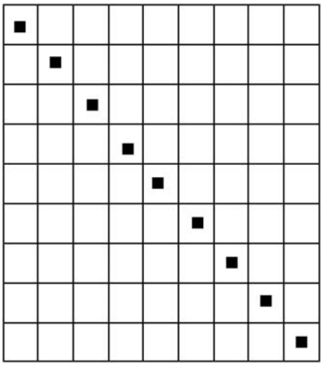


Figure 1 An extreme case of the spatial coverage of cLHS using x and y coordinates as ancillary variables

point in the real world. So the cLHS method was put forward by Minasny and McBratney (2006) to search through the data to find a sample to best meet the requirements of LHS. In cLHS an objective function is designed and optimized using the SSA to achieve optimized sampling patterns.

cLHS samples each ancillary variable independently; therefore, simply treating the x and y coordinates as ancillary variables cannot guarantee even coverage in geographical space which is a combination of x and y coordinates. As illustrated in Figure 1, which presented an extreme case of one best pattern of cLHS using x and y coordinates as ancillary variables, good coverage in x and y coordinates independently cannot guarantee good coverage in geographical space.

3.2 Geographically Stratified Sampling Method

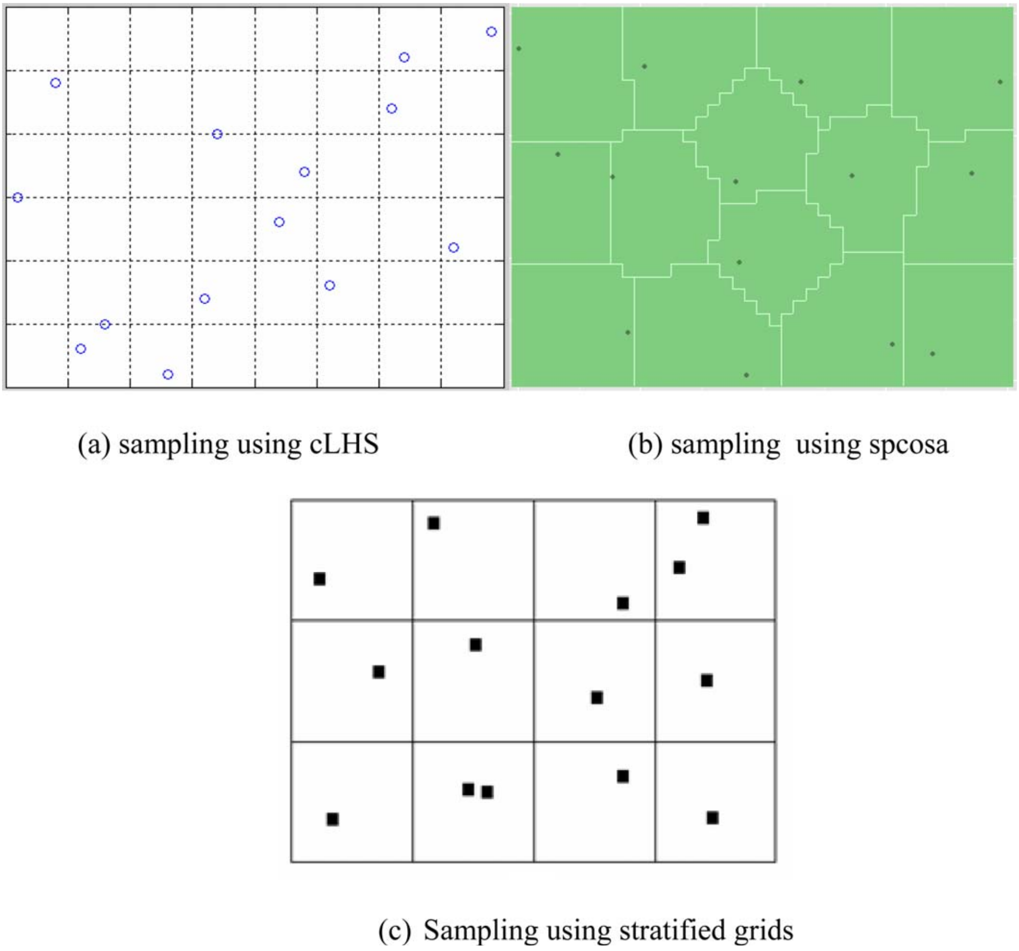
Parallel to cLHS which stratifies the feature space, the geographically stratified sampling method firstly divides the study area into strata and then draws sites from each stratum. One important method is the stratified simple random sampling from compact geographical strata of equal size (Brus et al. 1999), which can be realized by the R package called spcosa developed by Walvoort et al. (2010) based on k-means. The spcosa can divide the study area into an arbitrary number of compact sub-areas of equal size, as shown in Figure 2b. To improve geographical coverage, the number of sub-areas is set equal to the sample size. By partitioning the study area into compact sub-areas with equal size and randomly drawing one site from each, geographical coverage can be improved, especially when the sample size is not very small.

The geographical stratification can also be simplified by using a regular grid, such as square or rectangular, as shown in Figure 2c. For a rectangular study area, this grid stratification sampling can be achieved by following three steps:

Step1: Divide the area in to *Row* by *Col* grids (noted as *Row* Col*), where *Row* and *Col* are the number of rows and columns respectively, and *Row*Col* equals the sample size or is the nearest number to the sample size. The division is realized in following steps:

1. Compute the *Row* and *Col* according to the sample size and the width and height of the whole area:

$$Row = round\left(\sqrt{n * \frac{height}{width}}\right) \quad (1)$$



(a) sampling using cLHS

(b) sampling using spcosa

(c) Sampling using stratified grids

Figure 2 Sampling design results selected by geographically stratified sampling for a rectangular study area

$$Col = round\left(\sqrt{n * \frac{width}{height}}\right) \tag{2}$$

where n is the sample size, and *round* is a function that rounds a decimal number into an integer.

2. Compute the distance step to divide the study area:

$$xstep = \frac{width}{Col} \tag{3}$$

$$ystep = \frac{height}{Row} \tag{4}$$

where $xstep$ is the distance step to divide the area along the lateral axis, and $ystep$ is the distance step along the vertical axis.

3. Divide the study area into *Row* Col* grids. The grid with row number equaling *i* and column number equaling *j* can be defined by two corner points, the left lower: [(minx+(j−1)*xstep); (miny+(i−1)*ystep)], and right upper: [(minx+j*xstep); (miny+i*ystep)].

Step2: Randomly draw one site from each grid.

Step3: If *Row*Col* is larger than sample size, remove (*Row*Col- n*) sites randomly; else draw left sampling sites (*n-Row* Col*) randomly from the whole area.

For study areas that do not take a rectangular shape, grids can also be divided, except that the grids on the boundary would be cut to irregular cells.

Figure 2 shows the sampling results of drawing 14 sites from the study area using cLHS with x and y coordinates as ancillary variables, spcosa and regular grid stratification combined with spatial random sampling. With the restriction of the stratum boundary, the latter two stratified sampling methods can improve the geographical coverage.

3.3 scLHS

The scLHS proposed here is evolved from the cLHS (Minasny and McBratney 2006) by adding spatial evenly distributed objectivity to the objective function. It utilizes the continuous and category auxiliary data to guide a sampling design to approach even coverage in feature space, and at the same time employs the geographically stratified sampling mentioned above to constitute a spatial coverage objectivity to emphasize even coverage in geographical space. The objective function to be optimized is:

$$O = w_{co}O_{co} + w_{ca}O_{ca} + w_sO_s + w_{cor}O_{cor} \tag{5}$$

where O is the overall objectivity to be optimized, O_{co}, O_{ca}, O_s and O_{cor} are the sub-objective function of continuous auxiliary variables, category auxiliary variables, spatial coordinates and correlation respectively. w_{co}, w_{ca}, w_s and w_{cor} are the corresponding weight of each sub-objectivity. O_{co} is computed as follows:

$$\mathrm{O_{co}}=\sum_{\mathrm{i}=1}^{I}\sum_{v=1}^{V}|\eta(\mathrm{q}_v^{\mathrm{i}} \le z_v < \mathrm{q}_v^{\mathrm{i}+1})-1| \tag{6}$$

where $\eta(\mathrm{q}_v^{\mathrm{i}} \le z_v < \mathrm{q}_v^{\mathrm{i}+1})$ is a function to give the number of sampling sites falling between $\mathrm{q}_v^{\mathrm{i}}$ and $\mathrm{q}_v^{\mathrm{i}+1}$, i represents the ith interval and v represents the vth variable, and I and V are the total number of intervals and variables, respectively. The ideal situation is that only one sampling site falls in each interval, so the objective value of O_{co} is zero. O_{ca} is computed by:

$$\mathrm{O}_{ca}=\sum_{v=1}^{V}\sum_{c=1}^{C}\left|\frac{\eta(z_{vc})}{n}-p_{vc}\right| \tag{7}$$

where $\eta(z_{vc})$ represents the number of sampling sites in the cth category of variable v, n is the sample size, P_{vc} is the proportion of category c in variable v, and C is the number of categories. O_s is constituted in a similar way to the auxiliary variables, to place one sampling site in each compact strata. It is computed as:

$$\mathrm{O}_s=\sum_{k=1}^{h}|\eta[(x,y)\in S_k]-1| \tag{8}$$

where h is the number of strata divided and S_k is kth compact stratum, $\eta[(x,y) \in S_k]$ represents the number of sites falling inside S_k. For the simplified square or rectangular grid stratification, O_s can be computed using the following equation:

$$O_s = \sum_{i=1}^{Row} \sum_{j=1}^{Col} |\eta(x_i \le x \le x_{i+1} \quad \text{and} \quad y_j \le y \le y_{j+1}) - \partial| \tag{9}$$

where the η function gives the number of sampling sites falling inside the grid defined by corner points (x_i, y_i) and (x_{i+1}, y_{i+1}), *Row* and *Col* are the row number and column number of the grids, and ∂ is a rate computed by dividing the sample size with the product of row number and column number. It means that the average points in each grid and its value is near one because the number of grids (*Row*Col*) equals the sample size or is the nearest number to the sample size. The O_{cor} is computed

$$O_{cor} = \sum_{v=1}^{V} \sum_{u=1}^{V} |c_{vu} - t_{vu}| \tag{10}$$

where c_{vu} and t_{vu} are the correlation of variable v and u of sample and population, respectively. For the weights of the sub-objectivities, if the distribution in feature space is more important, then w_{co}, w_{ca}, and w_{cor} can be set larger than w_s, otherwise if distribution in geographical space is more important, w_s should be a larger value. If the correlation between the continuous auxiliary variables is larger than the category auxiliary variables, w_{co} should be set larger than w_{ca}, otherwise w_{ca} should be larger. w_{cor} can be set near to w_{co} and w_{ca}, for example, as the average of w_{co} and w_{ca}. If distributions in both feature and geographical space are important, and the correlations between each of the ancillary variables and the target variable are close, equal weights can be adopted.

The optimization algorithm is realized using SSA. As shown in Figure 3, The following five steps should be implemented:

Step 1: Prepare the data: choose related data as ancillary variables, divide the continuous ancillary variable into h (h equals the sample size) intervals according to the CDF (Cumulative Distribution Function), divide the study area into h compact strata with the same size, compute the correlation matrix of the ancillary variables, define the initial temperature T and cooling rate α which are both annealing parameters working together to control the annealing process, set the weights and stopping criterion.

Step 2: Generate the initial sample: select n sites randomly as sample *S1*, set *S* = *S1*; compute the correlation matrix of sample, and use Equations (5) through (10) to compute the overall objectivity *O*;

Step 3: Disturb sample *S* to produce a new sample *S2*; compute the new overall objectivity *Onew*;

Step 4: Judge whether to accept the new sample according to the Metropolis criterion: if the objectivity O is improved, accept the new pattern; if not, perform an annealing schedule to generate a random number *rand* between 0 and 1 and compute *Metro = Exp(O-Onew/T)*, if *rand< Metro*, accept the new sample, Set *S* = *S2*, O = *Onew*, otherwise discard it;

Step 5: Judge the stopping criterion: if the stopping criterion is not satisfied, set T = T*α, and then go to Step 3; otherwise finish optimization and give out the final pattern.

By optimizing the spreading in feature and geographical space, and adding randomness to the location of the sampling sites, scLHS can generate a sample with good coverage in both

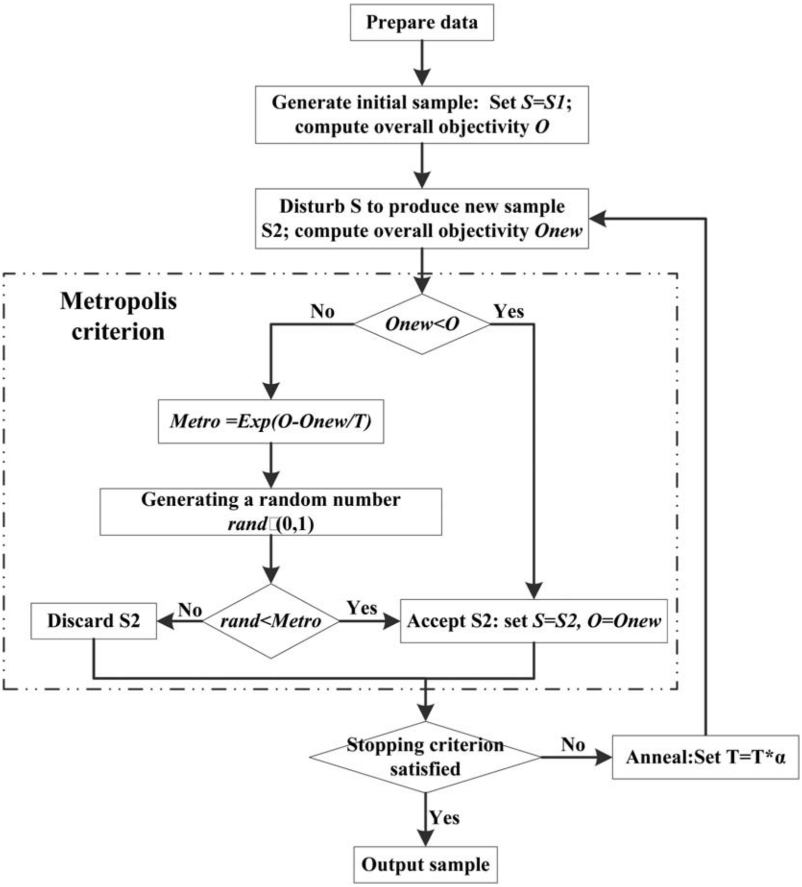

Prepare data
Generate initial sample: Set S=S1; compute overall objectivity O
Disturb S to produce new sample S2; compute overall objectivity Onew
Metropolis criterion
No
Onew<O
Yes
Metro =Exp(O-Onew/T)
Generating a random number rand□(0,1)
Discard S2
No
rand<Metro
Yes
Accept S2: set S=S2, O=Onew
Stopping criterion satisfied
No
Anneal:Set T=T*α
Yes
Output sample

Figure 3 Chart flow of optimization algorithm of scLHS

spaces and retain points pairs with small distances. It aims to give more precise interpolation results with one sampling.

The Q-Q (Quantile-Quantile plot) and P-P (Probability-Probability) plot can compare the distribution of the sample and its population by plotting the quantiles/probabilities of the sample and the corresponding quantiles/probabilities of its population as a scatter diagram. To quantify the representativeness of a sample, the deviation index was defined to be the RMSE obtained from fitting the line y=x to points in the Q-Q plot or P-P plot (Pan et al. 2015):

$$d=\sqrt{\frac{\sum_{i=1}^{n}(q_i-Q_i)^2}{n}} \tag{11}$$

where d is the deviation index, q_i is the quantile/probability of the ith sampling unit, and Q_i is the corresponding quantile/probability of the population. The deviation index quantifies the information of one plot using one number, thus facilitating a comparison between different results and simplifying the results' presentation. The smaller the deviation index, the better the sample represents the distribution of the population in feature space.

The spatial evenness index is defined based on the Theissen polygons around sampling sites to evaluate the spatial coverage by:

$$E=\frac{A}{\sum_{i=1}^{N}\left|S_i-\hat{S}\right|} \tag{12}$$

where A is the total area of the interested region and S_i is the area of the Theissen polygon around sample site i. $\hat{S}$ is the mean value of all S_i. For a certain study area, the numerator A is fixed. The denominator measures the differences among the areas of the Theissen polygons. If the sampling sites are more evenly distributed in geographical space, the differences among the areas of the Theissen polygons are smaller, and E becomes larger. On the contrary, if the sampling sites are less evenly distributed, sites cluster in some part and distribute sparsely in other parts, the Theissen polygons around the sites differs more in area, the denominator becomes larger and E becomes smaller. So, a larger evenness index indicates that the sampling sites are more evenly distributed in geographical space.

The points pairs with small distance are more critical for precise variogram estimation. To evaluate the ability of different methods to retain pairs with small distance, the DIPPSD (Distribution Index of Points Pairs with Small Distances) is defined. Like estimating the variogram, the points pairs are divided into non-overlapping distance groups. The DIPPSD is computed as:

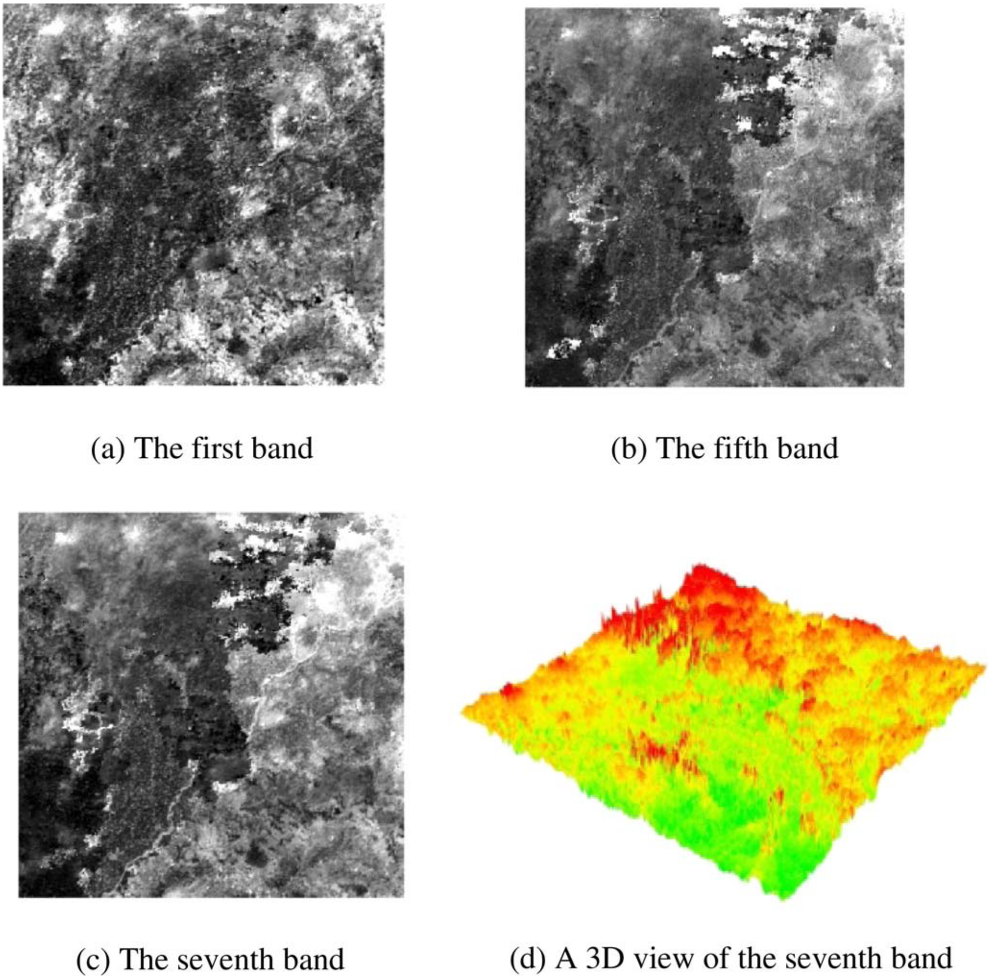
$$\text{DIPPSD}=\frac{p_{g1}*N_g}{\sum_{i=1}^{N_g}p_{gi}} \tag{13}$$

where N_g is the number of distance groups, p_{gi} is the number of points pairs belonging to the group N_i, and p_{g1} is the number of points pairs belonging to be first distance group. The larger the DIPPSD, the more points pairs with small distance are retained. Because only the number of points pairs in the first distance group is counted, DIPPSD cannot reflect the distribution of points pairs.

4 Case Study

4.1 Datasets and Experiments

Two exhaustive datasets were used as the sampling field to demonstrate the efficiency and the applicability of the suggested method. The first one is a 400*400 pixels subset of the level 2 product MOD09 of MODIS over the Henan Province of China on 5th April, 2010 (Figure 4) downloaded from the NASA website (http://ladsweb.nascom.nasa.gov/data/search.html). The seventh band was used as the target variable and the first band and fifth bands are used as ancillary variables. The correlation coefficients between target variable and ancillary variable are 0.75 and 0.89, respectively. The 3D view of the seventh band in Figure 4d shows an obvious trend in which the value goes higher from the nearer corner to the further corner. The other data set is an image with 100*100 pixels in size created using Sequential Gaussian conditional simulation function of Gslib (http://www.gslib.com/). During the simulation, arsenic (As) contents of 78 sites, randomly selected from 400 collected data of heavy metal contents of the topsoil of Shunyi District of Beijing of China, were used as hard data, and a spherical variogram model with Cc equaling 0.8, hMax equaling 15 and hMin equaling 8 was used. Three ancillary


(a) The first band

(b) The fifth band

(c) The seventh band

(d) A 3D view of the seventh band

Figure 4 MODIS data

data were also generated by disturbing the hard data and changing the parameters of the spatial variogram. The correlation coefficients between the target variable and ancillary variables are 0.71, 0.58 and 0.57, respectively. The simulated sampling field and ancillary data are listed in Figure 5. The upper left image is the target variable, and the others are the ancillary data.

MATLAB R2009a was employed to draw sample sets from both datasets using the spatial random, MMSD, cLHS, cLHSXY (the cLHS method that use x and y coordinates as covariates, besides the ancillary variables) and scLHS methods, and to analyse the sampling results. The DACE (Design and Analysis of Computer Experiments), a Matlab toolbox for Kriging models developed by Hans Bruun Nielsen et al. (http://www2.imm.dtu.dk/~hbni/dace/) was used for the mapping. For MODIS data, the variogram of interpolation was estimated from each sample, but for the simulated data, a known variogram was set.

Two experiments were carried out, one is comparing scLHS with other sampling methods to illustrate its efficiency, the other is comparing different weights of the sub-objective function of scLHS to discuss the selection of weights. In the former experiment, the weight of each

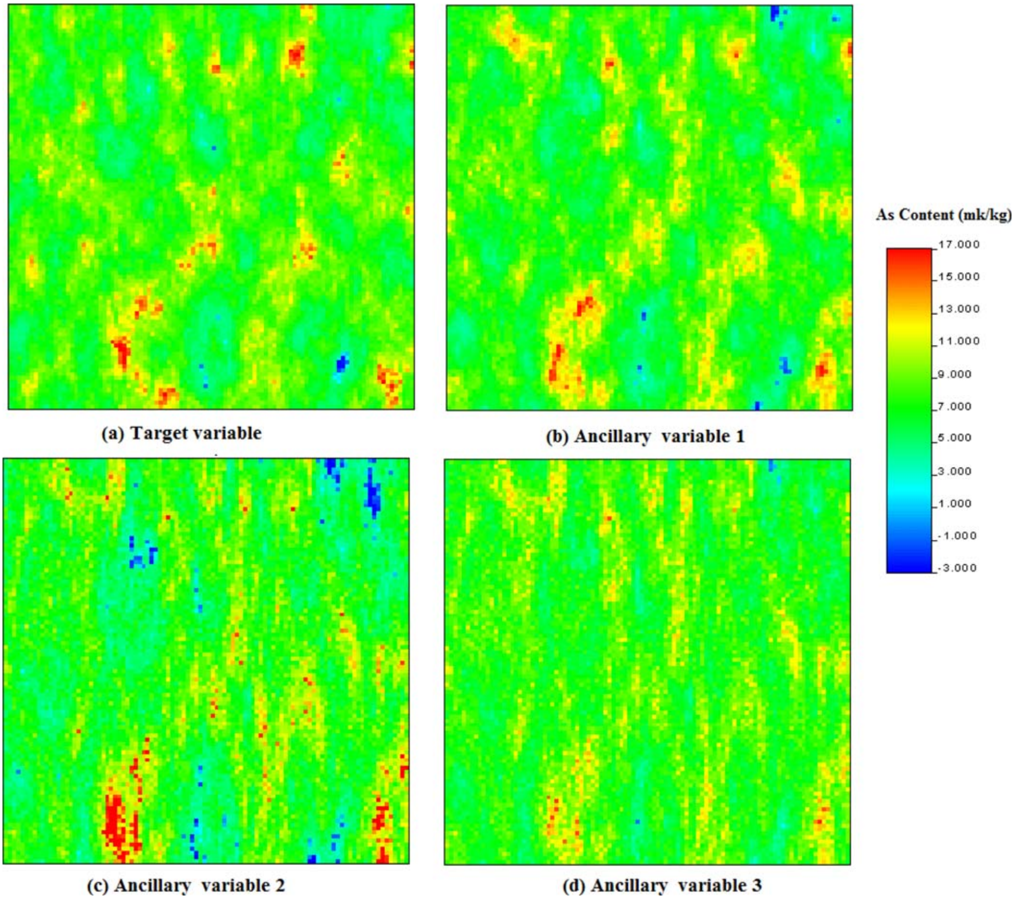


Figure 5 Simulated data

sub-objective were set to 1, and in the latter experiment, different weights was set for the sub-objectivity of the feature and geographical space.

The flow chart of the former experiment is presented in Figure 6. The experiment was implemented in the following four steps. (1) Sampling: drawing five sample sets with size of 200, 250, 300, 350 and 400 from the seventh band of the MODIS data and six sample sets with size of 25, 50, 75, 100, 125 and 150 from the simulated data under different sampling methods; (2) Spreading Evaluation: evaluating and comparing the distribution in feature space and geographical space of sample sets of different methods; (3) Comparing distribution of Points pairs: compute the DIPPSD to compare the ability to retain points pairs with small distances; (4) Inference: interpolating the study area using Universal Kriging with the sample sets and comparing the RMSEs, with a second order trend fitted in the MODIS data, and no trend was removed for the simulated data.

In the latter experiment, the weights of the sub-objectives of the feature space stay the same (from 0.1 to 0.9), but the ratio between them and the weight of sub-objective of the geographical space (from 0.9 to 0.1) keeps varying. It was implemented in the following steps. (1) define a variable w and set w = 0.1; (2) set w_{co}, w_{ca}, w_s and w_{cor} to equal w, set w_s = 1-w, employ scLHS

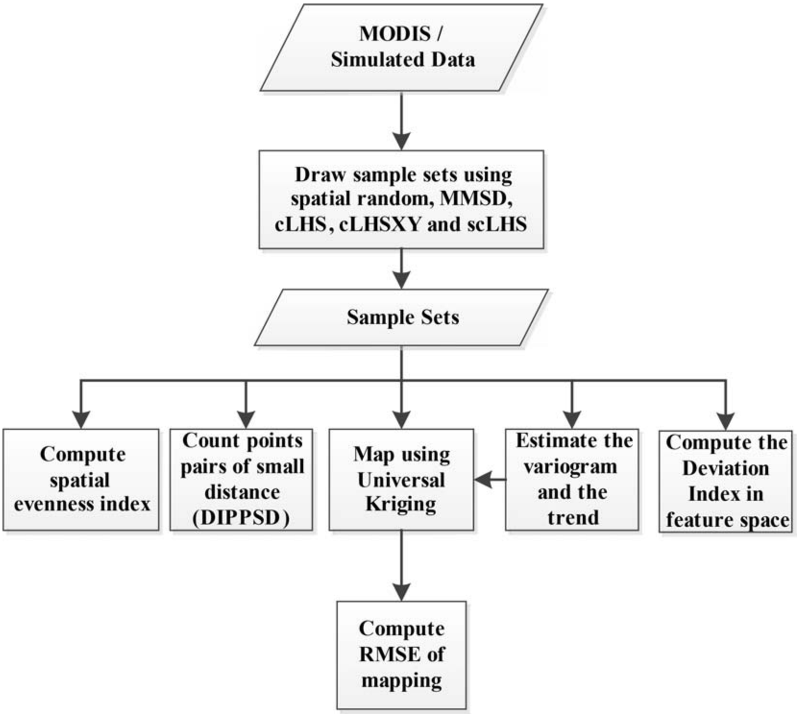

MODIS / Simulated Data
Draw sample sets using spatial random, MMSD, cLHS, cLHSXY and scLHS
Sample Sets
Compute spatial evenness index
Count points pairs of small distance (DIPPSD)
Map using Universal Kriging
Estimate the variogram and the trend
Compute the Deviation Index in feature space
Compute RMSE of mapping

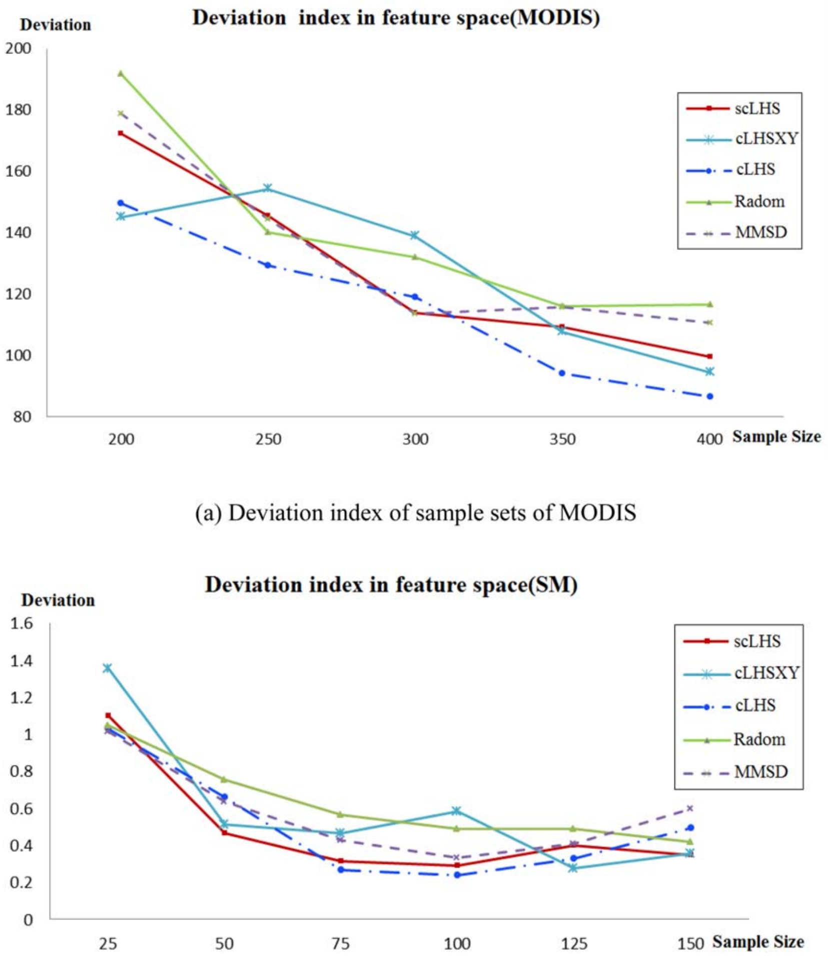
Figure 6 Flow chart of experiment for MODIS and simulated data

to draw five sample sets with size of 200, 250, 300, 350 and 400 from the 7th band of MODIS data and six sample sets with size of 25, 50, 75, 100, 125 and 150 from the simulated data; (3) map the study area using Universal Kriging with the sample sets and compute the mean and standard error of the RMSEs of the different sample sets; (4) if $w \leq 0.9$, set $w = w + 0.1$ and go to step (2), otherwise finish the experiment and output the means and standard error of the RMSEs of the different sample sets.

4.2 Comparing Different Sampling Methods

4.2.1 Distribution in feature and geographical space

The deviation index between the sample sets and the population is presented in Figure 7, where the horizontal axis is the sample size, and the vertical axis is the corresponding deviation index. For both the simulated data and MODIS data, except cLHS which only considers the distribution in feature space, scLHS performs better than cLHSXY, MMSD or the simple random method. For the simulated data, scLHS sometimes performs better than cLHS. The reason may be that the correlation between the ancillary and target variable of the simulated data is low, and according to Tobler's law that "everything is related to everything else, but near things are more related to each other" (Tobler 1970), pursuit of good coverage of geographical space can also improve the coverage of feature space. This also explains why MMSD performs better than the simple random method. The results of cLHSXY are not stable, and are worse than cLHS and scLHS most of the time. Although the only goal of cLHSXY is to evenly cover the distribution of each ancillary variable (including ancillary data and coordinates), due to the



(a) Deviation index of sample sets of MODIS

(b) Deviation index of sample sets of Simulated Data

Figure 7 Deviation index in feature space

low correlation with the target variable, the coordinates in fact play as disturbance factors which most of time decrease the performance of cLHSXY.

To map the study area with the target property, the even distribution of sampling sites in geographical space is very important. The spatial evenness indices of different methods are listed in Tables 1 and 2. Except for the MMSD method, scLHS produces sampling sets with the highest evenness index, while cLHSXY make little improvement in geographical coverage.

By comparing the deviation indexes and spatial evenness indexes of different methods in both datasets, the coverage of sample sets from different methods in geographical and feature space is summarized in Table 3. It demonstrates that scLHS is good at balancing between coverage in geographical and feature space, while cLHS is only optimal in coverage of feature space and MMSD is only optimal in coverage of geographical space. The simple random sampling method performs worst in terms of coverage of both spaces. cLHSXY has limited

Table 1 Spatial evenness index of sample sets from MIDIS data

Method	Sample size = 200	Sample size = 250	Sample size = 300	Sample size = 350	Sample size = 400
MMSD	12.10	10.18	9.04	3.24	53.71
scLHS	3.30	3.25	3.45	3.24	4.05
cLSHXY	2.35	2.54	2.38	2.52	2.57
cLHS	2.50	2.27	2.57	2.37	2.40
Random	2.41	2.38	2.25	2.51	2.37

Table 2 Spatial evenness index of sample sets from simulated data

Method	Sample size = 25	Sample size = 50	Sample size = 75	Sample size = 100	Sample size = 125	Sample size = 150
MMSD	38.48	23.19	55.53	30.40	54.30	23.88
scLHS	4.01	3.25	4.73	4.26	3.46	3.18
cLSHXY	3.34	2.85	2.45	3.006	2.47	2.68
cLHS	2.43	2.57	2.72	2.25	2.00	2.46
Random	1.96	1.93	2.76	2.59	2.16	2.27

Table 3 Distribution in geographical and feature space

Distribution	scLHS	cLHSXY	cLHS	Random	MMSD
Coverage of feature space	****	***	*****	*	**
Coverage of geographical space	****	***	**	*	*****

Note: the number of stars represent the rank, the more stars, the better the coverage

balancing ability, only a little better than MMSD in coverage of feature space, and a little better than cLHS in coverage of geographical space.

The DIPPSD of the sample sets from the MODIS data is shown in Figure 8, where point pairs were classified into 21 distance groups, and the size for each group is 20 pixels except for the last one (points pairs with distance larger than 400 were all classified into the last distance group). scLHS can retain points pairs with small distance which are more critical for a precise variogram estimation, although slightly fewer than cLHS, Random and cLHSXY. MMSD cannot produce such points pairs when the sample size is small.

4.2.2 Prediction error

Sample sets of different size drawn under simple random, MMSD, cLHS, cLHSXY and scLHS methods were used to predict values of un-sampled sites of the study area using Universal Kriging with DACE. For MODIS data because the sample size is large enough for variogram estimation, the parameters of the variograms were estimated from each sample after a trend of

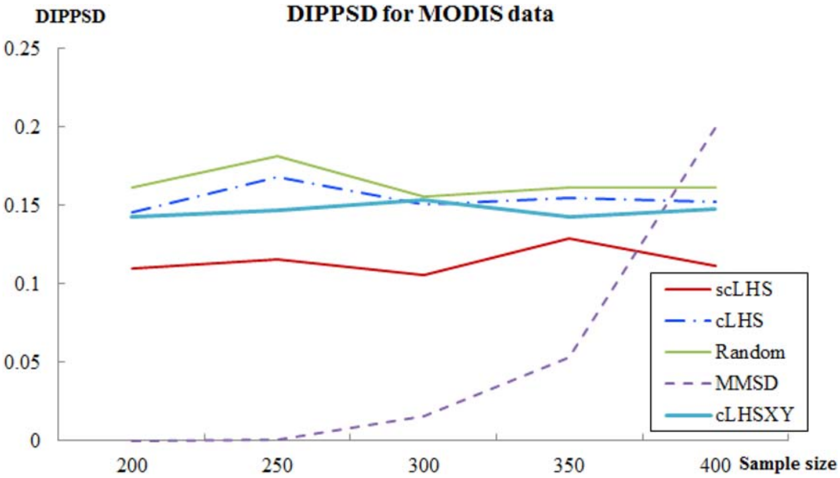


Figure 8 DIPPSD of sample sets from MODIS data

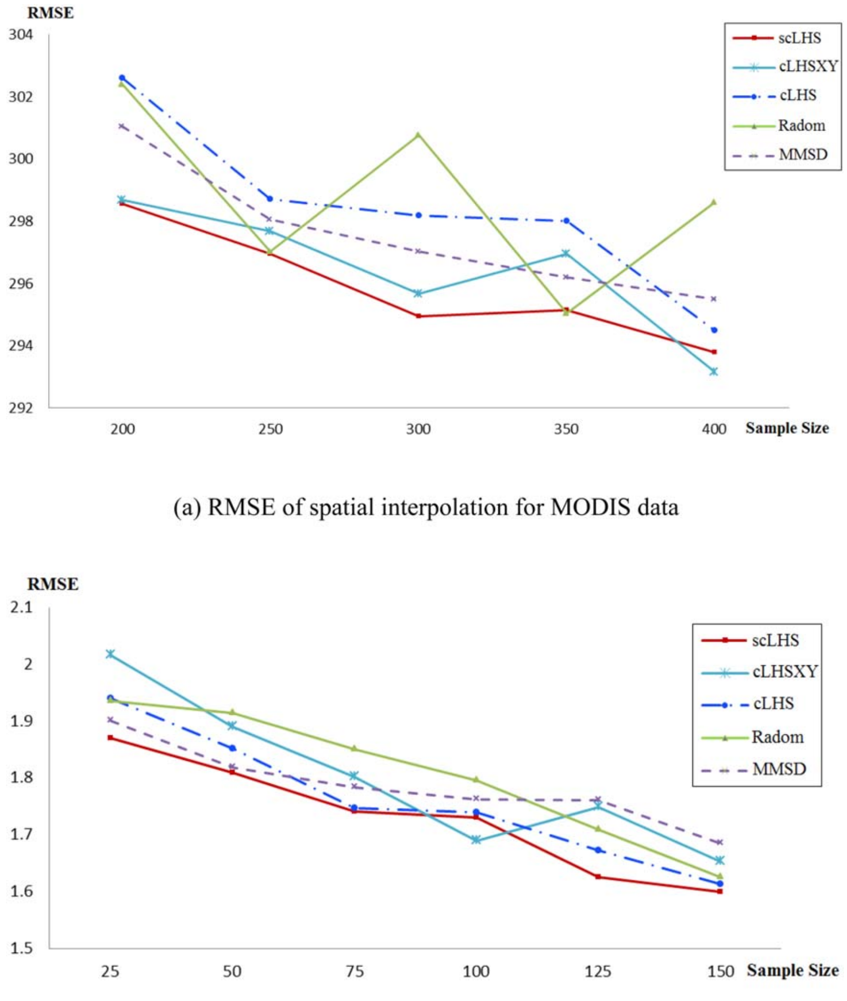
second order is removed. For the simulated data a known variogram is set and no trends were removed because no trends were added when generating the data.

The predicted values were compared to the true value of the original data and the RMSEs were computed. The results in Figure 9 show that scLHS can produce samples with smaller prediction errors in both the simulated data and MODIS data cases. What is more, the advantage over other methods in predicting is more obvious in MODIS data where a universal trend exits. The RMSEs of sample sets of cLHSXY is not stable, because simply treating x and y coordinates as variables in cLHS cannot guarantee good spatial coverage and at the same time impedes the coverage in feature space.

4.3 Comparing the different weights of each sub-objectivity

The mean and standard deviation of RMSEs of prediction using sample sets drawn by scLHS with different weights are presented in Figure 10, where the left and right vertical axis are the mean and standard deviation of the RMSEs, respectively; the lower horizontal axis is the weight of sub-objectives of feature space, and the upper horizontal axis is the weight of sub-objective of geographical space. From Figure 10a it can be found that for MODIS data, equal weights for all sub-objectives produce the best prediction results, and other unequal weights, either larger weights for sub-objectives of feature space, or larger weights for sub-objectives of geographical space, produce worse prediction results. For the simulated data in Figure 10b, when the weights of sub-objectives of feature space are between 0.3 and 0.6, the prediction results are better than other cases. Again, equal weights for all sub-objectives is one of the best weight settings.

The results show that for both MODIS data and the simulated data, it is a good idea to set equal weights for sub-objectivities in scLHS. The results are reasonable for the following reasons. For prediction of MODIS data, a second order trend needs to be fitted besides the spatial interpolation; distributions in both feature and geographical space are important. For the simulated data, although there is no trend, the spatial variation is not as stationary as that required by Kriging and will produce large prediction errors in areas with greater spatial variation if the sampling sites are evenly distributed in geographical space. To avoid such large



(a) RMSE of spatial interpolation for MODIS data

(b) RMSE of spatial interpolation for Simulated data

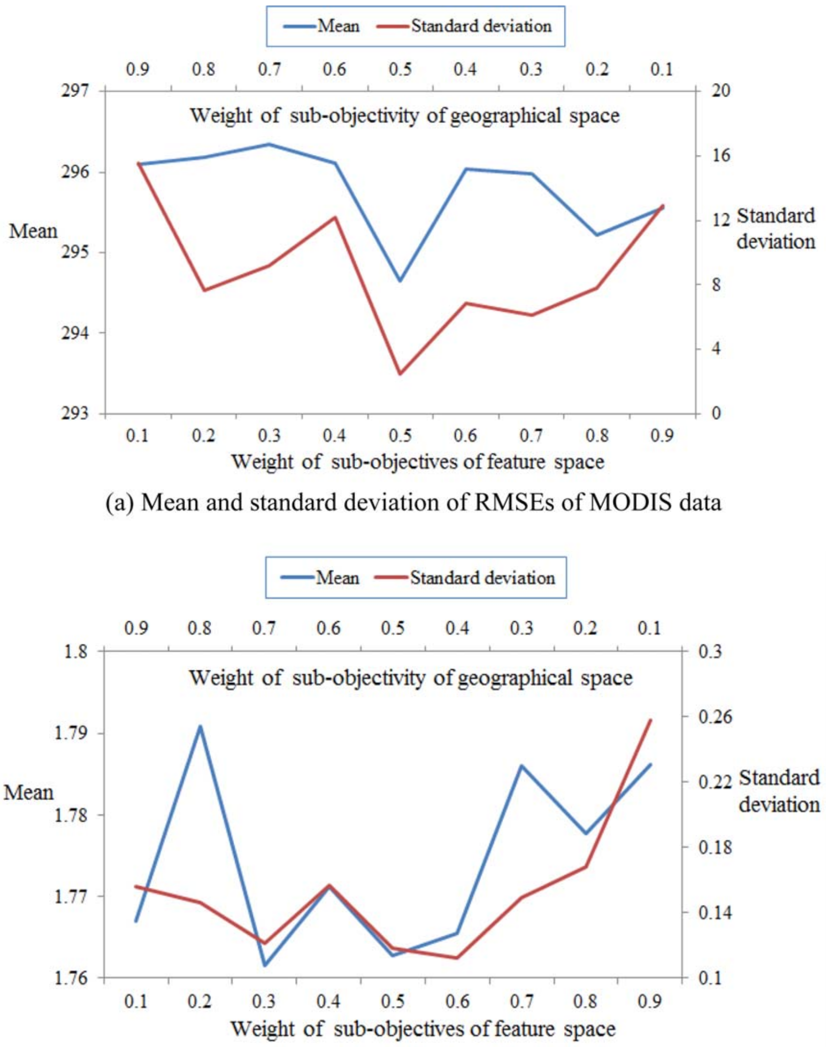
Figure 9 RMSE of prediction

prediction errors, the spreading of the sample in feature space also needs to be emphasized, to place more sites in areas with greater spatial variation.

5 Discussion

5.1 Advantages of scLHS

By utilizing ancillary data, scLHS optimizes the distribution of the sample in both feature and geographical space. Its advantage lies in improving trend estimation, variogram estimation and spatial interpolation at the same time and thus improving the total mapping precision. Compared with optimization methods in feature space such as cLHS, it can avoid spatial clustering which can impede the mapping precision of the whole area. Compared with even coverage



(a) Mean and standard deviation of RMSEs of MODIS data

(b) Mean and standard deviation of RMSEs of Simulated data

Figure 10 Mean and standard deviation of RMSEs of scLHS with different weights

methods such as MMSD, scLHS can depict the feature space better and produce more precise regression models to remove the trends, and at the same time increase the number of points pairs of small distance by allowing the sampling site to locate anywhere in each spatial stratum. Thus the trends-estimation error and the variogram-estimation error are smaller. Also as a purposive sampling method, the consideration of the distribution in feature and geographical space makes it more efficient than probability sampling, such as the spatial random method, most of the time (Brus and De Gruijter 1997; Gruijter et al. 2006). Theoretically, simply adding the x and y coordinates of the sampling sites as covariates to cLHS cannot guarantee good coverage in geographical space, and at the same time impedes the coverage in feature space. The case studies demonstrate that cLHSXY is worse than scLHS in coverage of both geographical and feature space, and also in mapping precision.

In scLHS, the relative importance of different sub-objectives can be adjusted by changing the corresponding weights. By setting the weight of sub-objectivity of geographical space to zero, the scLHS becomes cLHS, and by setting the weights of sub-objectivities of feature space to zero, the scLHS becomes geographically stratified sampling. In many practical cases, the spatial stationary required by Ordinary Kriging cannot be satisfied, thus besides covering the geographical space as evenly as possible, either a trend needs to be fitted or more sampling sites need to be placed in areas with greater spatial variation, to improve the mapping precision. scLHS is flexible in balancing the coverage of geographical and feature space.

5.2 Applicable Conditions

From the deviation index in Figure 7, it can be seen that the feature space coverage of cLHS is much better in MODIS than in the simulated data because the correlations between the ancillary variables and the target variable are higher in the former. Although scLHS can sometimes improve feature space coverage by optimizing geographical space coverage when the ancillary variables have low correlation with the target variable, it still requires ancillary variables of high correlation. In the spatial prediction results in Figure 9, the advantage of scLHS in drawing samples for mapping is more in MODIS data than in the simulated data. The reason is that one of the merits of scLHS is that it reduces the trend estimation error. For the MODIS data, there exist obvious trends, thus the improvement in mapping is obvious. However, for the simulated data, no obvious trend exists and known variograms are used, the benefit is limited.

When the variogram is not known beforehand, to estimate the parameters of the variogram from the sample is also required. To estimate the variogram, Davis and Borgman (1979, 1982) suggest that the number of points pairs in each distance class should be as large as possible, and at the same time the pairs with small distance are more critical for precise estimation. The results show that the scLHS can generate many more points pairs with small distance than MMSD, although fewer than other methods. What needs to be stressed here is that in order to estimate the variogram from sampling results of scLHS, the sample size must be large enough. The basic requirements, as advised by Webster and Oliver (1993), are that 150 locations can suffice in many situations and 225 in most isotropic applications. At the same time to meet the needs of spatial interpolation, the simple size should satisfy the following inequation:

$$n \geq \frac{A}{R_0^2} \tag{14}$$

where A represents the area of the study region and R_0 represents the correlation distance. To use scLHS to estimate the variogram and map the study area, the sample size should be equal or larger than that given by Equation (14) and that suggested by Webster and Oliver (1993).

In the case of MODIS data, because a trend of second order exists and the prediction errors come from trend-fitting and spatial interpolation, the distribution of samples in both feature and geographical space should be emphasized. In the case of simulated data, because spatial variation is not stationary, the coverage in feature space also needs to be considered, although the weights for sub-objectivities of feature space should not be larger than the weight of sub-objectivity of geographical space. The weights of scLHS should be adjusted according to specific condition, and lager weights should be set to the corresponding sub-objectivities if the coverage in feature or geographical space is to be emphasized more. If the study area is stationary and a variogram is known, the weights for sub-objectivities of feature space can be set to zero, and if the regression method is used to produce the map, the weight for sub-objectivity of geographical space should be set to zero.

6 Conclusions

With guidance from ancillary data, scLHS performs well in balancing between spreading in geographic space and feature space, and can retain points pairs with small distance which are crucial for variogram estimation. By considering the spatial coverage, it can be used to draw sample for spatial interpolation using Kriging compared with the ER and cLHS sampling methods which only consider coverage in feature space. Also, the optimal coverage in feature space results in a reduction of the trend estimation error (Brus and Heuvelink 2007) and thus can reduce the error of mapping when trends exist. scLHS completes all these in one step compared with the two-step procedure proposed by Hengl et al. (2004). Unlike the method to reduce Mean Universal Kriging Variance proposed by Brus and Heuvelink (2007), it does not need a Universal Kriging model before sampling design. What is more, in scLHS the relative importance of coverage in feature and geographical space can be adjusted flexibly.

The gain in prediction precision of scLHS can be strengthened when obvious trends exist. In addition, the correlations between ancillary variables and target variable affect the optimal coverage of the feature space of the target variable; the higher the correlation, the better the coverage and the higher the estimation precision. Sufficient sample size is required to estimate the parameters of the spatial variogram using a sample of the scLHS. The cases in this article showed that if distribution in both feature and geographical space should be emphasized, equal weights for sub-objectives are proper.

As far as we can see, three improvements of scLHS need to be further studied in future: (1) adding WM criteria (Warrick and Myers 1987) into the overall objective function to minimize the difference between the actual distribution of distance groups of points pairs and the preselected distribution and to improve the estimation of the variogram; (2) studying the weight of different sub-objectives under different sampling aims and different population characteristics; and (3) giving out the ancillary variable selecting criteria based on the correlation coefficient and employing methods such as PCA (Principal Component Analysis) to reduce the dimension of ancillary variable when they are plenty.

References

Bertolino F, Luciano A, and Racugno W 1983 Some aspects of detection networks optimization with the kriging procedure. *Metron* 41: 91–107

Brus D and De Gruijter J 1997 Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma* 80: 15–44

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

Brus D J, Spätjens L E E M, and de Gruijter J J 1999 A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. *Geoderma* 89: 129–48

Chen B, Pan Y, Wang J, Fu Z, Zhang Y, and Zhou Y 2012 Even sampling designs generation by charges repulsion simulation. *Environmental Monitoring and Assessment* 184: 3545–56

Chen C and Li Y 2012 An adaptive method of non-stationary variogram modeling for DEM error surface simulation. *Transactions in GIS* 16: 885–99

Davis B M and Borgman L E 1979 Some exact sampling distributions for variogram estimators. *Journal of the International Association for Mathematical Geology* 11: 643–53

Davis B M and Borgman L E 1982 A note on the asymptotic distribution of the sample variogram. *Mathematical Geology* 14: 189–93

De Gruijter J and Ter Braak C 1990 Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology* 22: 407–15

Florian A 1992 An efficient sampling scheme: updated Latin hypercube sampling. *Probabilistic Engineering Mechanics* 7: 123–30

Graniero P A and Robinson V B 2003 A real-time adaptive sampling method for field mapping in patchy, heterogeneous environments. *Transactions in GIS* 7: 31–53

Groenigen J W 1997 Spatial simulated annealing for optimizing sampling. In Soares A, Gómez-Hernandez J, Froidevaux R (eds) *geoENVI: Geostatistics for Environmental Applications*. Berlin, Springer: 351–61

Gruijter J D, Brus D, Bierkens M, and Knotters M 2006 *Sampling for Natural Resource Monitoring*. New York, Springer

Haining R P 2003 *Spatial Data Analysis: Theory and Practice*. Cambridge, UK, Cambridge University Press

Hengl T, Rossiter D G, and Stein A 2004 Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Soil Research* 41: 1403–22

Iman R L and Conover W 1980 Small sample sensitivity analysis techniques for computer models with an application to risk assessment. *Communications in Statistics: Theory and Methods* 9: 1749–842

Lin Y-P, Chu H-J, Huang Y-L, Tang C-H, and Rouhani S 2011 Monitoring and identification of spatio-temporal landscape changes in multiple remote sensing images by using a stratified conditional Latin hypercube sampling approach and geostatistical simulation. *Environmental Monitoring and Assessment* 177: 353–73

Lloyd C D and Atkinson P M 2002 Non-stationary approaches for mapping terrain and assessing prediction uncertainty. *Transactions in GIS* 6: 17–30

Martinez B, Cassiraga E, Camacho F, and Garcia-Haro J 2010 Geostatistics for mapping leaf area index over a cropland landscape: Efficiency sampling assessment. *Remote Sensing* 2: 2584–606

McBratney A, Whelan B, Walvoort D, Minasny B, and Stafford J 1999 A purposive sampling scheme for precision agriculture. In *Proceedings of the Second European Conference on Precision Agriculture*, Odense, Denmark: 101–10

McKay M, Beckman R, and Conover W 2000 A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 42: 55–61

Minasny B and McBratney A B 2006 A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences* 32: 1378–88

Mulder V L, de Bruin S, and Schaepman M E 2013 Representing major soil variability at regional scale by constrained Latin hypercube sampling of remote sensing data. *International Journal of Applied Earth Observation and Geoinformation* 21: 301–10

Pan Y, Ren X, Gao B, Liu Y, Gao Y, Hao X, and Chen Z 2015 Global mean estimation using a self-organizing dual-zoning method for preferential sampling. *Environmental Monitoring and Assessment* 187 (3): 121

Särndal C-E, Swensson B, and Wretman J H 2003 *Model Assisted Survey Sampling*. Berlin, Springer

Simbahan G C and Dobermann A 2006 Sampling optimization based on secondary information and its utilization in soil carbon mapping. *Geoderma* 133: 345–62

Stevens Jr D L 2006 Spatial properties of design-based versus model-based approaches to environmental sampling. In *Proceedings of the Seventh International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Lisbon, Portugal: 119–25

Van Groenigen J, Siderius W, and Stein A 1999 Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma* 87: 239–59

Van Groenigen J and Stein A 1998 Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality* 27: 1078–86

Walvoort D J J, Brus D J, and de Gruijter J J 2010 An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers and Geosciences* 36: 1261–67

Wang J-F, Stein A, Gao B-B, and Ge Y 2012 A review of spatial sampling. *Spatial Statistics* 2: 1–14

Wang J, Haining R, and Cao Z 2010 Sample surveying to estimate the mean of a heterogeneous surface: Reducing the error variance through zoning. *International Journal of Geographical Information Science* 24: 523–43

Warrick A and Myers D 1987 Optimization of sampling locations for variogram calculations. *Water Resources Research* 23: 496–500

Webster R and Oliver M 1993 How large a sample is needed to estimate the regional variogram adequately? In Soares A O (ed.), *Geostatistics Tróia'92*. Berlin, Springer: 155–66