Design of Monitoring Network for Cultivated Land Quality in County Area Based on Kriging Estimation Variance

Sai Tang[1], Jianyu Yang[1 *], Chao Zhang[1], Dehai Zhu[1], Wenju Yun[2]

[1] College of Information and Electrical Engineering, China Agricultural University, 100083 Beijing, China; [2] Land Consolidation and Rehabilitation Center, Ministry of Land and Resources, 100035 Beijing, China; *Author for correspondence {757165996@qq.com; ycjyyang@cau.edu.cn}

Abstract. Through the monitoring network for cultivated land quality in county area, distribution and changing trend of the quality should be reflected. Besides, the quality of non-sampled locations should also be estimated with the data of sampling points. Due to the correlation among spatial samples, traditional methods such as simple random sampling, stratified sampling and systematic sampling are inefficient to accomplish the task above. Thus, spatial sampling method based on Kriging estimation variance is presented in this paper. The method considers the spatial structure characteristics of cultivated land and is configurable to restrict the expense or estimation accuracy, making it more flexible to determine the number of samples. What's more, tessellation polygon is used to densify the network and this helps Kriging interpolation to achieve higher efficiency than traditional methods. Furthermore, inference for the population mean of natural quality indices stays in a very precise level at the same time.

Keywords: cultivated land quality, spatial sampling design, monitoring network, Kriging, estimation variance

[1] The research was funded by the development of the information management system for Farmland grading monitoring (201011006-5).

1 Introduction

China is an agricultural country with a large population but not enough cultivated land. Until 2011, the cultivated land per capita in China was 1.38 mu (0.09ha), only 40% of the world average. With the rapid development of economy, industrialization and urbanization, this situation becomes even worse [1]. Therefore, the rate of change of cultivated land is concerns for Chinese governments. At present, methods for monitoring the quantity of cultivated land have become reliable; some researchers analyzed and interpreted remote sensing images to keep track of the quantity condition [2-6]. However, unreasonable land exploitation in China leaves a negative influence on both quantity and quality of cultivated land. When tracking quality condition, the method of remote sensing is not applicable any more so that on-the-spot sampling and investigation become necessary.

As the dynamic monitoring of cultivated land quality is the next task of Ministry of Land and Resources[7], we have sampled seven hundred counties in China and in each county, a network would be set up to monitor the change of cultivated land quality. With the basis of current Land Resource Survey and Cultivated Land Evaluation, the paper centers on how to design the monitoring network in county area.

2 Study Area and Data

The study area Daxing district, located in the south of Beijing (longitude 116°13′-116°43′E, latitude 39°26′-39°51′ N), is a river alluvium plain with a sub-humid warm temperate continental monsoon climate. The terrain slopes gradually from east to west and the span of altitudes is generally between 14 and 52 meters. The average annual temperature and precipitation are 11.6℃ and 556mm.

The data used in our study was the layer of cultivated land quality of Daxing district from last investigation and the corresponding scale was 1: 10000. It included 5182 map spots, with the total area of 416.97km^2. Daxing district belongs to the region of Huanghuanghai and the sub-region of piedmont plain of Yanshan and Taihang Mountains [8]. When designing the network, only natural quality was concerned and natural quality index was used as the attribute variable. The reasons not to take utilization and economy quality into

consideration are as follows: a) Natural quality index could reflect the natural conditions and agricultural facilities of the region. b) After calculating the natural quality index, the corresponding utilization and economy quality indices will be obtained by introducing the coefficients which are assigned by Chinese provinces [8]. c) Utilization and economy quality indices are associated with input cost and grain output, so they are largely affected by human factors and spatial interoperation cannot be applied directly.

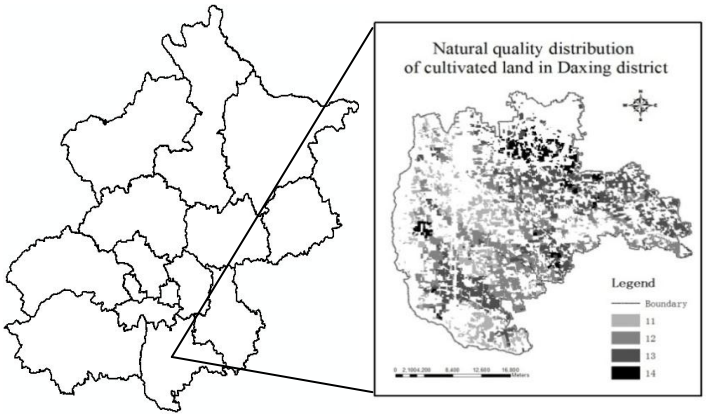

Fig.1. Study area. Daxing district, located in the south of Beijing, has four natural quality grades (from 11 to 14), showing the ladder-like distribution from northeast to southwest. Higher quality land of 14th grade mainly in northeast makes up 7.72%, that of 13th grade 36.19%, that of 12th grade 35.35% and the rest mainly located in middle and south area belongs to 11th grade.

3 Methods

The design of monitoring network is to sample from 5182 blocks of cultivated land and place monitoring points on each sampled block, then infer the quality of non-sampled blocks based on investigation of monitoring points. Currently, most designs are originated from classical sampling models, which mainly include simple random sampling, stratified sampling, systematic sampling and cluster sampling [9].These methods ignore the spatial correlation [10] so that they are not suitable for our study because cultivated land belongs to spatially distributed object and there are correlations among samples. Therefore, sampling method based on Kriging estimation variance is presented in this paper.

Kriging method was first raised by Matheron [11, 12], originally applied to researches of geology mineral resources. Oliver used it in the interpolation of GIS [13] and later on the method was widely spread in many fields. It is built upon the theoretical analysis of semi-variance function, considering the distance between non-sampled and sampled points. Kriging method belongs to optimal unbiased estimation so its prediction is better than that of traditional interpolation methods. By interpolating the data of monitoring points, we can get the relationship among estimation variance, sampling distribution, spatial correlation and number of samples [14].

3.1 Examination of the data distribution

Data examination is necessary before Kriging interpolation because Kriging is based on the assumption of stationarity, which demands all data values come from distributions that have the same variability. Secondly, certain Kriging methods require the data to be approximately normally distributed. Normal QQ Plot was used to examine the data in the study. In the diagram below, Data points fall close to the 45-degree reference line, approximating normal distribution. Therefore, the application of Kriging interpolation is allowed.

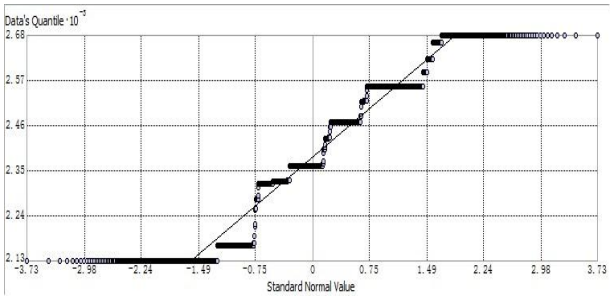

Fig.2. Normal QQ Plot: the quantile values of the standard normal distribution are plotted on the x-axis, and the corresponding quantile values of the dataset are plotted on the y-axis.

3.2 Measures of monitoring network performance

The performance of a monitoring network can be measured using different indices. The following lists three common measures [6]:

Measure 1: *MSE* (mean square error) of the mean estimation

$$\sigma_{\mathrm{n}}^2 = E\left[\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} f(x_i) - \frac{1}{\mathrm{A}}\int_A f(x_i)dx_i\right]^2 \quad (1)$$

Where E denotes the mathematical expectation; n is the total number of samples; $f(x_i)$ is the real value of point x_i in the region A; and term $\frac{1}{\mathrm{A}}\int_A f(x_i)dx_i$ is the true mean of A. In practice, this mean is estimated by the arithmetic mean of n sampling points and the corresponding formula is $\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} f(x_i)$. This measure is quite useful for estimating population mean [15] or total [16].

Measure 2: *MSE* of discrepancy

$$MSE\hat{f}(\mathrm{x}_i) = E\left[\left(\hat{f}(\mathrm{x}_i) - f(\mathrm{x}_i)\right)^2 | A\right] \quad (2)$$

Where $f(\mathrm{x}_i)$ denotes the real value at point x_i and $\hat{f}(\mathrm{x}_i)$ is the estimated value of $f(\mathrm{x}_i)$, which is calculated from interpolation of monitoring points. This measure is suitable to estimate other points by spatial interpolation on existing sampling points [17].

Measure 3: Maximum site error

$$\left\{\max_i \left\| \hat{f}(\mathrm{x}_i) - f(\mathrm{x}_i)\right|, \quad \forall i \in A\right\} \quad (3)$$

The applicable condition of this measure is similar to that of measure 2.

The quality of non-sampled cultivated land should be predicted by interpolation in our study. As administrators only concern the global accuracy so that measure 2 is appropriate, which places sampling points in suitable positions and has the most accurate estimation through Kriging interpolation [18]. Because natural quality was the prior knowledge, formula of measure 2 was redefined like this:

$$MSE = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\hat{f}(\mathrm{x}_i) - f(\mathrm{x}_i)\right)^2 \quad (4)$$

Where n means the total of cultivated land map spots, which was 5182. By the definition, the higher the estimation accuracy of interpolation is, the lower the *MSE* is.

3.3 Design of monitoring network

The steps to set up the monitoring network are as follows: a) Lay out the basic network according to regular grids; the choice of grid size depends on both estimation variance and number of sampling points. b) Adjust the centers of grids which do not fall in the area of cultivated land to form the global monitoring network. c) Densify the global network by adding sampling points in tessellation polygons whose *MSE*s are relatively high. The workflow is displayed in figure 3.

Basic monitoring network. In geographical space, using regular grids to place samples is the common way to quantitatively describe spatial variation and a square grid is the most convenient in practical applications [19]. The reasons to introduce regular grids are listed in the following: a) Sampling points are distributed uniformly so that the entire district could be represented. b) When applying Kriging interpolation, if numbers of point pairs of different lag classes vary greatly, the estimation efficiency of semi-variation function will fall [20]. c) Generally, spatial correlation declines when the distance between monitored objects increases so that the grid sampling is more effective than random sampling in decreasing estimation variation [14].

Referring the bottom left corner of the layer, sampling points are placed according to square grids of different sizes. Using the data of sampling points, a Kriging interpolation is applied to estimate the value of non-sampled cultivated land. Compare the grid sizes and the corresponding *MSE*s; then a suitable grid size can be chosen for the basic network.

Global monitoring network. Becasue the center of a grid may be not located inside a map spot of cultivated land, we create a searching circle around this center. Choose the centroid of a map spot, which is closest to the center and inside the searching circle be a candidate point. As these candidate points may raise the estimation variance, a further observation is needed to decide whether to add them or not. The observation should be: apply Kriging interpolation and calculate the *MSE* after adding each candidate point; add the one that decreases *MSE* by the least to the basic network and take the points which make *MSE* decline as candidates for the next calculation. The loop stops when no point reduces the global *MSE*.

Densification of the network. Adding sampling points in complex areas where natural quality of cultivated land changes intensely can not only increase the accuracy of interpolation but also be quite meaningful because there may be a high chance of quality grades change.

At present, the main approach to densify the network is to use existing points to get weights of non-sampled locations and then add the point with largest weight. A. G. Journel's method was to apply Kriging interpolation at first and get the Kriging variance of every non-sampled point, then set the point with largest Kriging variance as the new one to add [21]. Kriging variance is used to evaluate the uncertainty of Kriging interpolation [22]. But in this study, real values of cultivated land blocks were known so that error of estimation, which shows the deviation of real and estimated value, was more reliable than Kriging variance. Hence, estimation variance was used as the weight in the study.

However, points with large estimation variance may stand as outliers; to solve this problem, tessellation polygon was introduced. Tessellation polygon was proposed by Netherland meteorologist A. H. Thiessen to predict average rainfall based on data of weather stations in discrete distribution. He linked all adjacent weather stations as triangles and drew perpendicular bisector of each line to produce tessellation polygon. At last, the rainfall intensity of polygon region was presented by the only weather station inside the polygon. The tessellation polygon has the following features: a) There is only one discrete point inside each polygon. b) All objects inside a polygon are closest to the corresponding discrete point. c) The objects on the line of a polygon have the same distance to the two points on both sides of the line.

When using monitoring points to build tessellation polygon, every polygon is the control area of one point. If *MSE* of some polygon is relatively high, the corresponding estimation accuracy is low. To some degree, this approach can exclude individual outliers to be added. The details are listed as follows: Selecting the point of maximum variance in a tessellation polygon with a high *MSE*. Add it to the network only if it makes the global *MSE* smaller. The procedure of densification stops until the count of sampling points reaches the specific value or the *MSE* of tessellation polygon is lower than the threshold. Till then, the work to set up the monitoring network is complete.

3.4 Estimate the population mean

After monitoring network is finished, the population mean of natural quality indices of cultivated land will be calculated by following formula:

$$\overline{f}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_i) \tag{5}$$

Where *n* means the total of cultivated land map spots, which is 5182; $\overline{f}(x_i)$is the population mean; $\hat{f}(x_i)$is the estimation value.

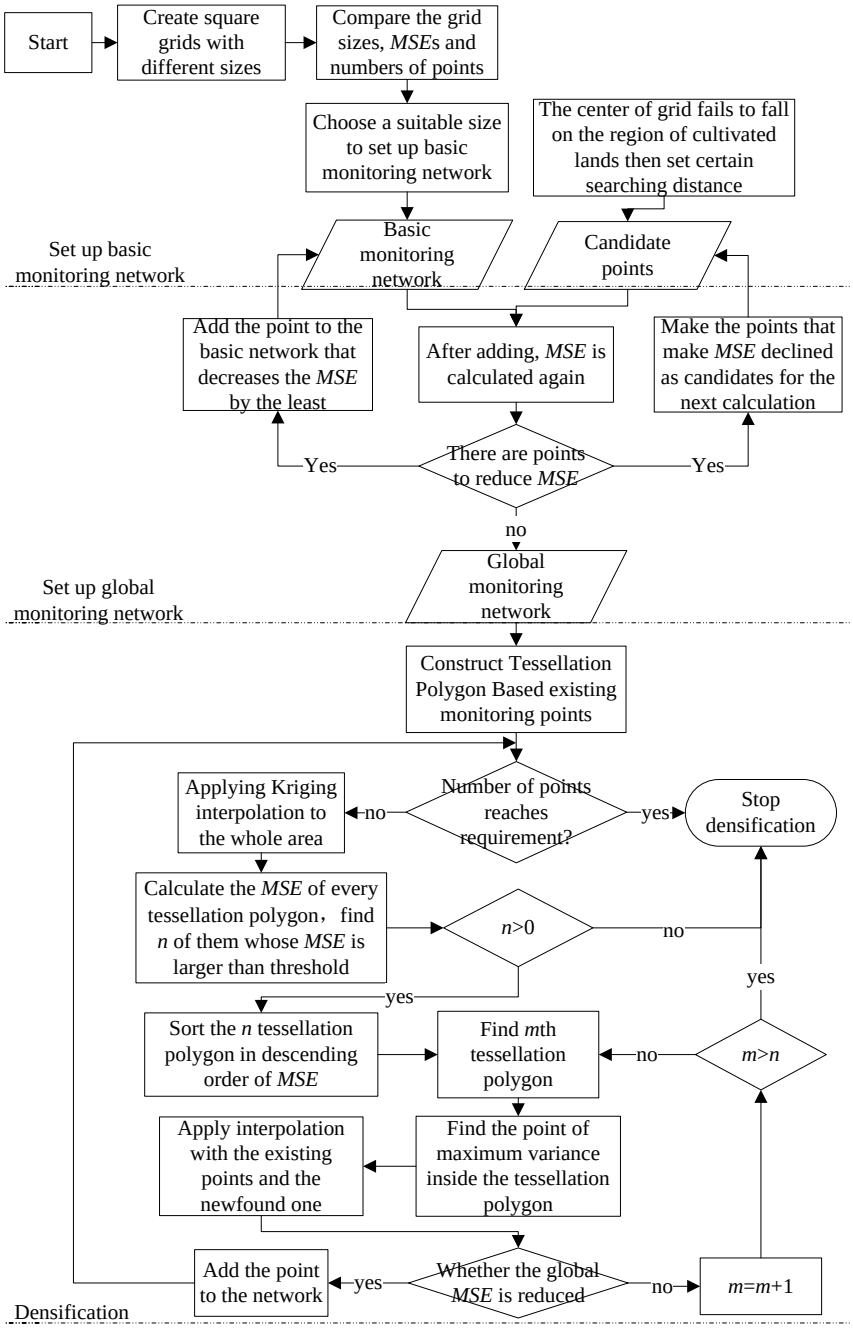

Fig.3. Workflow of setting up the network: *N* could be set to the final count of samples according to the cost (*N*=5182by default). A threshold could be set and new points are added to *n* regions whose *MSE*s are bigger than the threshold (*n* equals the number of global sampling points by default). Variable *m* is initially set to 1.

4 Results and Discussion

4.1 Set up network based on Kriging estimation variance

According to the procedure mentioned above, different sizes of square grid were used and searching circle's radius was set as 100m. In principle, the smaller the size of grid is, the larger the count of points is. But in light of discontinuous distribution of cultivated land, the count was affected by the position relation of grids and map spots. So, it is legal that sometimes smaller grids contain fewer points. Table 1 shows the number of points and *MSE* of each grid size.

Considering both *MSE* and number of points, we found it reasonable to choose the 4000m grid for three reasons: (a) The *MSE* of 4000m grid with 39 points was 13130, smaller than that of 3900m, 3800m, 3600m, 3500m, 3300m and 3100m. This meant that the 4000m grid contained fewer points but had higher accuracy. (b) Although the *MSE*s of 3700m, 3400m, 3200m or sizes below 3000m were smaller, the number of points increased significantly along the way, followed by the augmentation of sampling cost. (c) To leave basic monitoring network for further densification, we made the number of points as small as possible.

Table 1. Number of points and *MSE* of each grid size

Grid size	Number	*MSE*	Grid size	Number	*MSE*	Grid size	Number	*MSE*
4500	34	17077.8	3600	53	13258.8	2700	88	12288.0
4400	34	15527.0	3500	59	13288.6	2600	94	10338.6
4300	36	13763.5	3400	49	12336.4	2500	108	12510.7
4200	37	13995.1	3300	49	13635.9	2400	102	11123.3
4100	31	15932.5	3200	67	12913.5	2300	126	10857.5
4000	**39**	**13130.1**	3100	74	14568.0	2200	141	10602.5
3900	48	13930.6	3000	69	12693.5	2100	154	9505.5
3800	48	14675.4	2900	74	12402.4	2000	161	10046.4
3700	51	12413.1	2800	82	11072.8			

Even though the searching radius was set up as 100m, from the figure 4(a), we can see that sampling points did not cover every grid. To fix this, we set the radius as 500m and got 11 candidate points, as showed in figure 4(b) with the

symbol of triangle. To follow the procedure above, three points was selected with the purpose to increase the estimation accuracy, which came down to a 42 points network. The *MSE* was 12198.37, less than former network with the *MSE* of 13130.05.

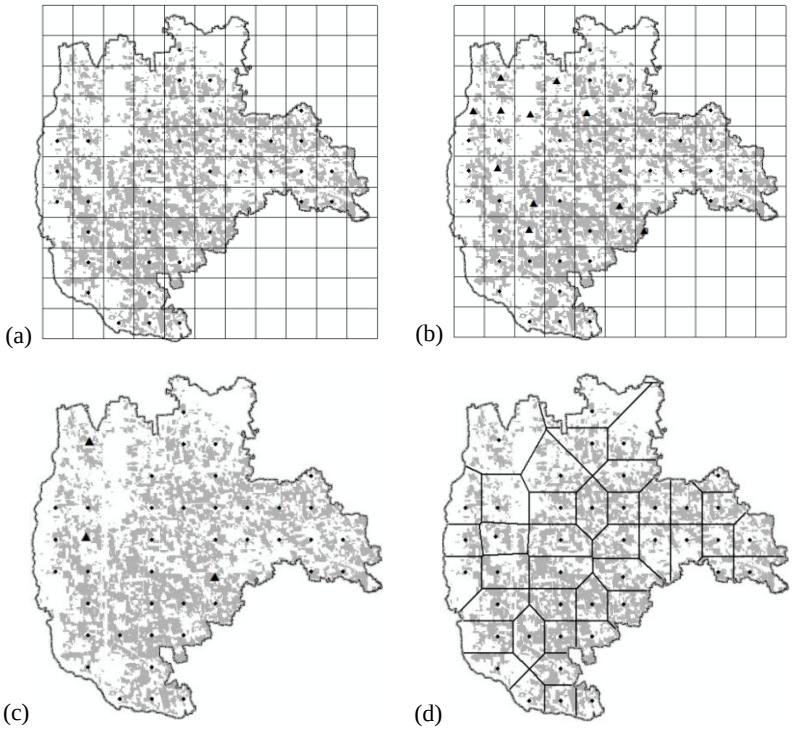

Fig.4. (a) is the basic network of 39 sampling points by 4000m grids. (b) shows the locations of candidate points. (c) is the global network of 42 points. (d) shows the tessellation polygons.

With 42 monitoring points, we constructed tessellation polygons, as showed in figure 4(d). Taking the cost into account, we added six points to the global network and the final *MSE* is 11238.32. The final network is exhibited in figure 5; 48 monitoring points covered all natural quality grades of cultivated land, distributed sparsely in regions of small variance and densely in regions of large variance.

Fig.5. Distribution of monitoring points in Daxing district

According to formula 5, the estimation mean 2384.16 was quite close to the real mean 2385.76 with a very small relative error 0.07%, and the result met the requirement of design.

4.2 Results Comparison

Compared with classical sampling models.

(a) Simple Random Sampling

Table 2 shows twenty groups of results using simple random sampling with 48 points to set up the network. According to the table, the minimal *MSE* was 13171.69, which was bigger than that of our method.

Table 2. Results of simple random sampling

Times	*MSE*	Times	*MSE*	Times	*MSE*	Times	*MSE*
1	13423.81	6	13759.86	11	**13171.69**	16	14000.15
2	14775.30	7	13222.51	12	15108.00	17	14467.08
3	15999.14	8	14419.24	13	15345.55	18	14091.05
4	14154.64	9	14595.11	14	16037.21	19	13274.75
5	13180.20	10	**18165.41**	15	14811.47	20	15200.25

(b) Stratified Sampling

When applying stratified sampling, we divided study area into 4 stratums (from 11 to 14 natural quality grade) and placed 48 sampling points as well. The formula followed [23] was used to assign the number of samples for each stratum:

$$n_h = \frac{W_h S_h / \sqrt{C_h}}{\sum W_h S_h / \sqrt{C_h}} \times n,\ W_h = \frac{N_h}{N} \tag{6}$$

Where h means the hth stratum; N stands for total number of cultivated land; N_h is the number of land in hth stratum; W_h denotes the weight of hth stratum; S_h means standard deviation of natural qualities of h stratum; And C_h is the cost for setting up a sampling point.

Based on the assigned numbers in table 3, the study used random method to place them. Results of twenty experiences are contained in table 4, with the minimal *MSE* being 13990.60 and maximal one being 18546.03, far worse than 11238.32 of our method.

Table 3. Standard Deviation and number of samples in each stratum

Natural quality classification	11	12	13	14
Standard Deviation	19.37	26.57	49.16	19.49
Number of sampling points	7	13	25	3

Table 4. Results of stratified sampling

Times	*MSE*	Times	*MSE*	Times	*MSE*	Times	*MSE*
1	15107.41	6	17697.43	11	14791.12	16	**18546.03**
2	**13990.60**	7	18089.52	12	14441.22	17	17209.28
3	16501.53	8	15171.63	13	14660.60	18	14002.29
4	17101.16	9	16482.26	14	14697.30	19	16447.92
5	14936.49	10	14722.58	15	16422.40	20	16319.17

The results indicated classical sampling models lead to low estimation accuracy of interpolation because of the correlations among samples.

Compared with traditional square grids. From table 1, the count of points of 3800m and 3900m equaled that of our method, but the estimation accuracy was significantly lower than ours. Besides, until sampling points reached one hundred, traditional grids can provide sufficient accuracy. In summary, the densification of our method is valid.

Compared with adding points without tessellation polygon. If we did not use tessellation polygon as a limit and directly added points of maximum variance, the *MSE* was 11680.68.Due to the fact that point of maximum variance could be quite different from surrounding values, directly adding may decrease the estimation accuracy and *MSE* did not monotonically decrease during the procedure of densification. This experience illustrated that accuracy could be improved when we limited the new points to the area of tessellation polygon with relatively high *MSE*.

5 Conclusion

By comparing our approach to traditional sampling methods in cost (count of sampling points), surface fitting (measured by *MSE*) and inference of the population mean, we can conclude that:

(a) Because spatial samples are correlated, traditional sampling methods will reveal their shortages like low efficiency and accuracy. We came up with a method based on Kriging estimation variance, which takes the spatial variation and distribution characteristics into account, and involves interpolation to calculate the value of non-sampled cultivated land.

(b) When densifying the network, points of maximum variance cannot be added directly in case that some outliers appear to decrease accuracy. Introducing tessellation polygon can deal with the problem to some degree.

(c) The paper reports the way to build the monitoring network for cultivated land quality in Beijing Daxing district. By interpolating the whole area, the estimation accuracy is better than traditional methods and the statistical inference of population mean stays in an accurate level as well.

References

1. Shi Shuqin, Chen Youqi, Yao Yanmin, Li Zhibin, He Yingbin: Methodology for Impact Assessment of Regional Cultivated Land Resources Change Based on 3S Technology. Transactions of the CSAE, 24(7), 91--96 (2008).
2. Liu Jianhong, Zhu Wenquan: Accuracy and Efficiency of Different Spatial Sampling Schemes for Cropland Change Monitoring. Transactions of the CSAE, 26(10), 331--336 (2010).
3. Zhang Jinshui, Pan Yaozhong, Hu Tangao, Chen Lianqun, Dong Yansheng: Analysis of Influence Factors about Space Sampling Efficiency of Winter Wheat Planting Area. Transactions of the CSAE, 25(8), 169--173 (2009).
4. Hu Tangao, Zhang Jinshui, Pan Yao zhong, Song Guobao, Dong Yansheng, Jia Bin: Researches on Remote Sensing Area Measurement Based on Different Sampling Methods. Remote sensing for Land & Resources, 3, 37--41 (2008).
5. Jiao Xianfeng, Yang Bangjie, Pei Zhiyuan: Paddyrice Area Estimation Using a Stratified Sampling Method with Remote Sensing in China. Transactions of the CSAE, 22(5), 105--110 (2006).
6. Wang J, Liu J, Zhuan D, et al. Spatial Sampling Design for Monitoring the Area of Cultivated Land. J. International Journal of Remote Sensing, 23(2), 263--284 (2002).
7. Wu Yupeng, Yun Wenju, Li Wuyan: Research on Standard-plot Based Monitoring and Early-warning of Arable Land Quality. China Land Science, 20(4), 40--45 (2006).
8. Ministry of Land and Resources: Regulations for Classification on Agricultural Land (TD/T1004－2003). China Standards Publishing House, Beijing (2003).
9. Kish L.: Survey Sampling. John Wiley and Sons, USA(1985).
10. Wang Jinfeng, et al.: Spatial Analysis. Science Press, Beijing (2006).
11. Matheron G.: The Intrinsic Random Functions and Their Applications. Advances in Applied Probability, 5(3), 439--468 (1973).
12. Matheron G.: Kriging or Polynomial Interpolation Procedures -- a Contribution to Polemics in Mathematical Geology. Canadian Mining and Metallurgical Bulletin, 60, (1967).
13. Oliver M.A., Webster R.: Kriging: a Method of Interpolation for Geographical Information Systems. International Journal of Geographical Information Systems, 4(3), 313--332 (1990).
14. Wang Jinfeng, Robert H., Steve W.: Spatial Sampling Design for Monitoring Drought, Flood and Earthquake Disaster in China. Progress in Natural Science, 9(4), 50--59 (1999).

15.Griffith D.A., Haining R., Arbia G.: Heterogeneity of Attribute Sampling Error in Spatial Data Sets. Geographical Analysis, 26(4), 300--320 (1994).
16.V B.: Elements of Sampling Theory. Edward Arnold, London (1974).
17.Griffith D A, Bennett R J, Haining R P.: Statistical Analysis of Spatial Data in the Presence of Missing Observations: a Methodological Guide and an Application to Urban Census Data. Environment & planning A, 21(11), 1123--1151 (1989).
18.Simbahan G.C., Dobermann A.: Sampling Optimization Based on Secondary Information and Its Utilization in Soil Carbon Mapping. Geoderma, 133, 345--362 (2006).
19.Guo Renzhong: Spatial Analysis. Higher Education Press, Beijing (2001).
20.Jiang Chengsheng, Wang Jinfeng, Cao Zhidong: A Review of Geo-Spatial Sampling Theory. Acta Geographica Sinica, 64(3), 368--380 (2009).
21.Journel A.G.: Nonparametric Geostatistics for Risk and Additional Sampling Assessment. In: Principles of Environmental Sampling. American Chemical Society, 45--72 (1988).
22.van Groenigen J.W., Siderius W., Stein A.: Constrained Optimisation of Soil Sampling for Minimisation of the Kriging Variance. Geoderma, 87, 239--259 (1999).
23.Wang Jinfeng, Jiang Chengsheng, Li Lianfa, Hu Maogui, et al.: Spatial Sampling and Inference. Science Press, Beijing (2009).