SAMPLING SURVEY OF HEAVY METAL IN SOIL USING SSSI

A. H. MA[a], J. F. Wang[b]*, K. L. Zhang[a]*

[a] School of Geography and Remote Sensing Sciences, Beijing Normal University, Beijing 100875, China, maaihua@mail.bnu.edu.cn, keli@bnu.edu.cn
[b] Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, Beijing 100101, China, wangjf@lreis.ac.cn

KEY WORDS: Soil sampling, Spatial dependency, Stratified Sampling, Sampling accuracy, variance, SSSI

ABSTRACT:

Much attention has been given to sampling design, and the sampling method chosen directly affects the sampling accuracy. The development of spatial sampling theory has lead to the recognition of the importance of taking spatial dependency into account when sampling. This text uses the new Sandwich Spatial Sampling and Inference (SSSI) software as a tool to compare the relative error, coefficient of variation (CV), and design effect with five sampling models – simple random sampling, stratified sampling, spatial random sampling, spatial stratified sampling, and sandwich spatial sampling. The five models are simulated 1000 times each with a range of sample sizes from 10 to 80. SSSI includes six models in all, but systematic sampling is not used here because the sample positions are fixed. The dataset consists of 84 points measuring soil heavy metal content in Shanxi Province, China. The whole area is stratified into four layers by soil type、hierarchical cluster and geochronology, and three layers by geological surface.The research shows that the accuracy of spatial simple random sampling and spatial stratified sampling is better than simple random sampling and stratified sampling because the soil content is spatially continuous, and stratified models are more efficient than non-stratified models. Stratification by soil type yields higher accuracy than by geochronology in the case of smaller sample sizes, but lower accuracy in larger sample sizes. Based on spatial stratified sampling, sandwich sampling develops a report layer composed of the user's final report units, allowing the user to obtain the mean and variance of each report unit with high accuracy. In the case of soil sampling, SSSI was a useful tool for evaluating the accuracy of different sampling techniques.

1. INTRODUCTION

Much progress has been made in developing tools to assess estimation accuracy and the relevancy of the chosen model, such as spatial autocorrelation precision (Haining R, 2003; Cressie N,1991; Wang J. *et al.*, 2002; Li L. et al. 2005) or error variance (Jinfeng Wang *et al.*, 2009). Uncertainty in a spatial sampling survey can be reduced with high-quality information, improvement of statistical estimation accuracy, and increasingly precise geo-statistical software. However, the choice of a relevant sampling model is essential to the quality of the analysis and what will be learned in the study.

Many different sampling methods exist, and selecting one relevant to a specific problem is always an important choice for the scientist, who wants to have the most valid statistical inference analysis possible. The user can decide to study the problem with, for instance, a simple random sampling model, which involves selecting a fixed number of units from a population with equal chances for each, or with a stratified sampling model to assess the relation between samples and a stratified information layer (William G. Cochran. 1977). This choice really depends on the problem, on the objective of the study, or on the available dataset. Having some methods to assess the fit of the sampling model with the sample dataset would be useful for the scientist, in order to improve the statistical analysis and refine results.

*corresponding authors

Two approaches for matching sampling effort to accuracy are chosen: a classical approach, which ignores spatial dependence between observations; and a geo-statistical approach, which accounts for spatial dependence (R. D. Hedger et al, 2001). Accounting for spatial dependency decreases the standard error when the environmental variable is spatially correlated (R. B. Haining, 1988; J. F. Wang, 2002).

Sandwich Spatial Sampling and Inference (SSSI) is a professional spatial sampling and statistical inference tool. It can be used for sampling design and statistical inference in environmental, resource, land, ecological, social, and economic sciences and practices (http://www.sssampling.org/). The objective of this study is to evaluate the accuracy of different sampling methods by comparing several indicators, such as relative error, coefficient of variation (CV), and design effect, using SSSI. Because the sampling model and number of samples (or sample size) are the two most important factors determining the sampling efficiency (D.M. Chen, 2009), this paper will compare four models with a range of sample sizes from 10 to 80. The dataset is a survey of heavy metal content in soil in Shanxi province, China.

2. MATERIAL AND METHODS

2.1 Data and site description

The investigation and environmental sampling was carried out in 2002 and 2003. Two counties, Zhongyang and Jiaokou, in

Shanxi Province were selected as the research area. There are seven administrative villages in each county and each has a high incidence of birth defects. Samples were collected using administration villages as units. Sample sites were considered consistent with birth defect epidemiological survey sites, so there exist some gaps within the data set, such as in the western region. Soil samples were collected far away from the soil disturbance plot, which is affected by residences and road, to ensure the representativeness of local environmental characteristics as much as possible. Samples were collected in most villages, with 84 points in all. Well-mixed cultivated land surface soil samples were acquired with a shovel 10 to 20 cm

deep. The samples were dried, cleaned of organic debris and impurities, ground with a mortar, sifted with a 100 mesh aperture, and weighed at about 0.1 g soil 3 copies. They were then mixed with 3 ml of Glass Acid, 1 ml perchloric acid, and 1ml hydrofluoric acid and placed into an oven for baking and nitrification for 5 hours. Finally, they were put on an electric heating board for 1 to 2 hours, then mixed with water to give a uniform volume of 10 ml for testing. An ULTIMA inductively coupled plasma spectrometer was used to measure sixteen elements in the soil: Al，As，Ca，Cu，Fe，K，Mg，Mo，Na，Ni，Pb，Se，Sn，Sr，V，and Zn. In this paper, Mo contents are used as sampling data.
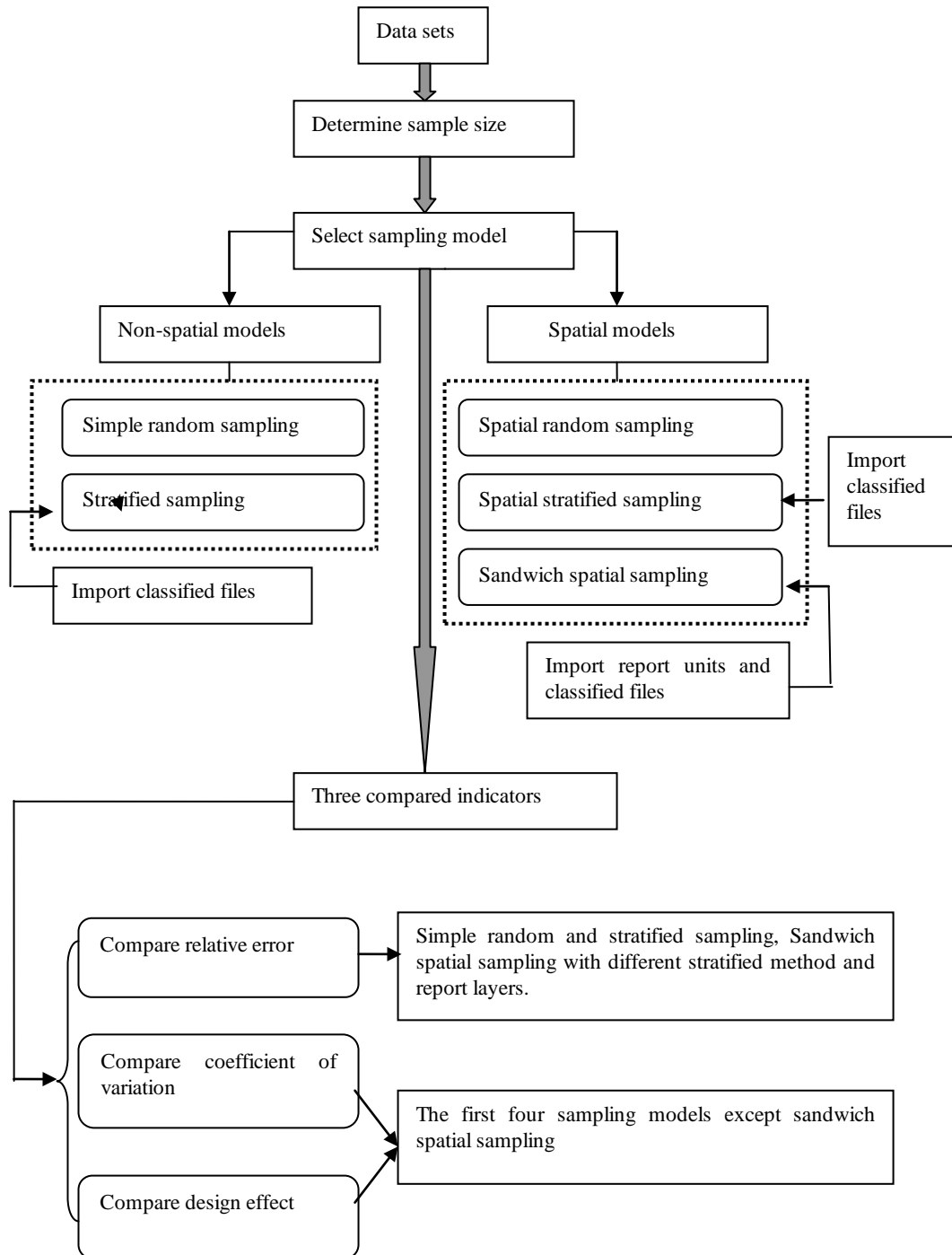

Figure 1 a flow chart of step of this study

Heavy metal contents in the soil depend primarily on the natural levels of soil-forming parent materials and the absence of external contaminants. There are five main soil types in pedogenesis: cinnamon, chestnut cinnamon, loessial soil, regosols soil and Chao soil. In this way, the area is stratified by soil type and geological surface. It can also be stratified geochronologically because elements were introduced at different times. There are eight layers, but in order to keep the sample size reasonable in each layer, it is stratified into four layers. The fourth stratification method is hierarchical clustering accomplished with SPSS.

2.2 Sampling method and algorithm

Five models in SSSI are used: simple random sampling, stratified sampling, spatial random sampling, spatial stratified sampling, and sandwich spatial sampling. The first two models take the classical approach, and the last ones take the geo-statistical approach. The use of different sampling designs and methods may affect the accuracy of the analysis of samples collected in the same area (Paolo de Zorzi *et al.*, 2008). Here, three indicators are selected to compare the efficiency of different sampling models. To get the three indicators, the models are simulated 1000 times each with a range of sample sizes from 10 to 80. Figure1 is a flow chart for the steps which can comprehend the study:

2.2.1   Relative error
Relative error compares the difference between sample mean and its true mean, so the estimated relative error is defined as:

$$\text{relative error} = \frac{\text{abs}(\bar{y}\text{-}\bar{Y})}{\bar{Y}} \qquad (1)$$

where $\bar{y}$ = sample mean

$\bar{Y}$ = observable population mean

The former is estimated by the different models and the latter is acquired by counting the entire observable metal contents.
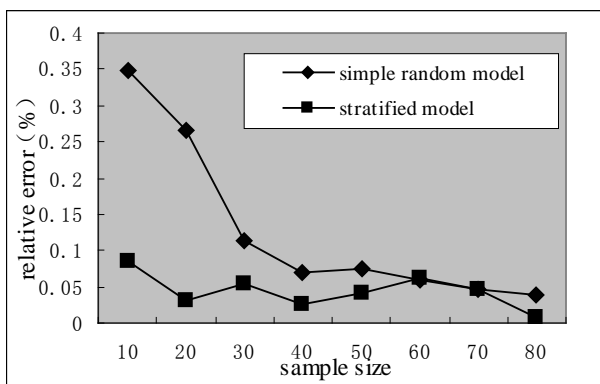

Figure 2 compare relative error between simple random sampling and stratified sampling

Figure 2 compares d_r between simple random sampling and stratified sampling. The sandwich spatial sampling model is the newest method in the SSSI software. It has the same accuracy to the spatial stratified sampling, but it refers to report layers,

which can be in any unit, for example, a county border, provincial boundary, watershed, or artificial grid (http://www.sssampling.org/). Table 1 shows the values of the whole relative error of report layers.

report layer	stratified by soil type	stratified by geochronology
Administrative villages	0.180045	0.085048
grids	0.066463	0.052151

Table 1 relative error of two kinds of report layer

2.2.2   Coefficient of variation
The coefficient of variation is a statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows:

$$\text{coefficient of deviation} = \frac{\text{Standard deviation}}{m} \qquad (2)$$

$$\text{Standard deviation} = \frac{\sum_{i=1}^{n}(x_i)\,\bar{x}^{\,2}}{n\text{-}1} \qquad (3)$$

where   m = sample mean

X_i = value of element content

n = sample size

In this paper it compares the dispersion of each estimator from its true value during the 1000 simulations for each sampling method and sample size (D. M. Chen, 2009).
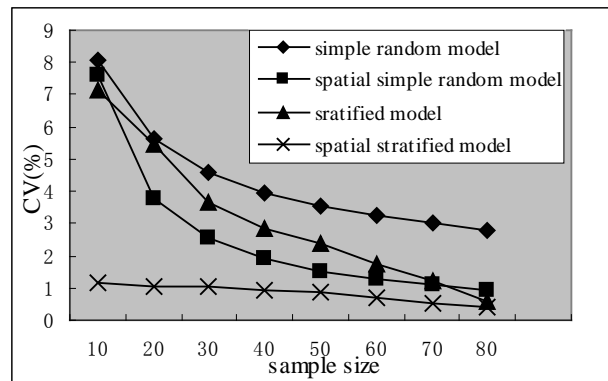

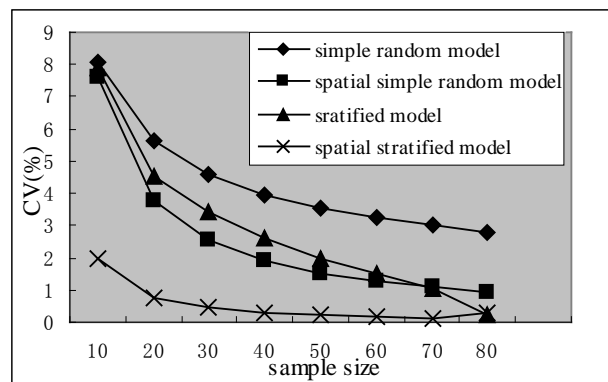Figure 3 Mo (stratified by soil type)

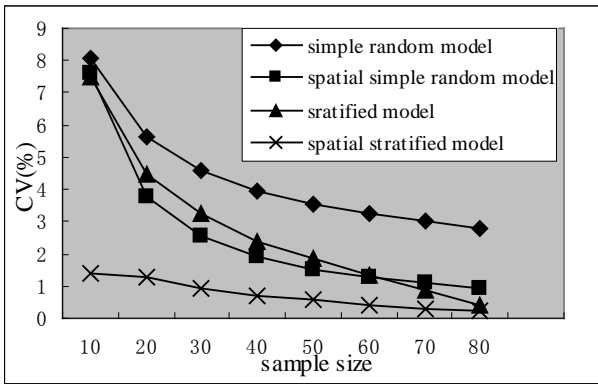Figure 4 Mo (stratified by geological surface)

Figure 5 Mo (stratified by Hierarchical cluster)

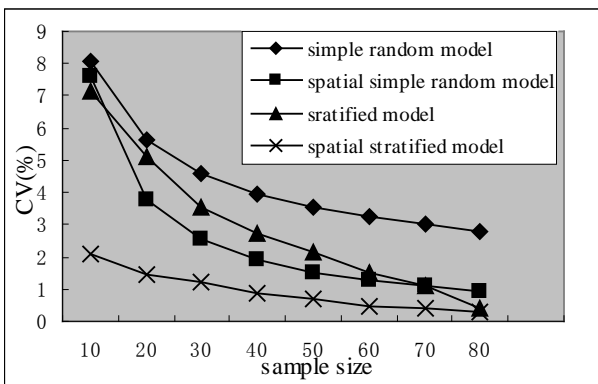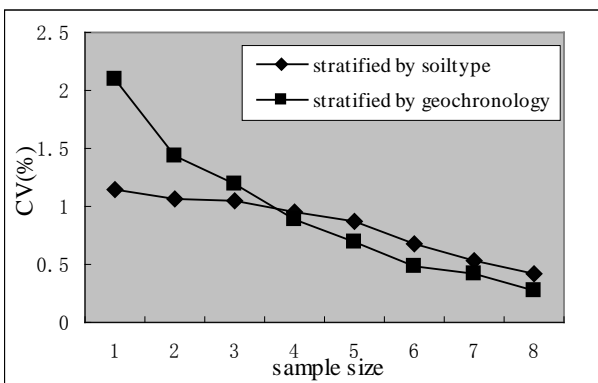
Figure 6 Mo (stratified by geochronology)

Figure7 Mo compare stratified method soil type and geochronology

Figures 3 through 6 show the CV from four sampling models of Mo elements, stratified by soil type, geological surface, hierarchical cluster, and geochronology. And figure 7 shows which stratified method is more efficient.

2.2.3 Design effect:

Design effect is the ratio of variance of the estimate obtained from the (more complex) sample to the variance of the estimate obtained from a simple random sample of the same number of units (William G. Cochran. 1977). This indicator also compares the efficiency of different sampling models.

3. RESULTS AND DISCUSSION

Although sample data is spatially distributed, but if only compare spatial models, maybe it is not enough, so five models are chosen. Spatial random sampling is the same to simple random sampling in evaluating sampling mean, and spatial stratified sampling is also the same to stratified sampling in it, so only compare relative error between simple random sampling and stratified sampling, and because importing report units in sandwich spatial sampling, each report unit also has its relative error, so compare the whole relative error of report layers. And sandwich spatial sampling has the same accuracy to the spatial stratified sampling in coefficient of variation and design effect, so the last two indicators only compare four sampling models.

3.1 Analyses of the relative error

The simple random sampling model is a method in which each sample has an equal chance of being selected. Stratified sampling divides the whole area into groups based on a special factor. This factor affects the distribution of the research object; there is little variation inside the layer but great variation between layers. Figure2 shows that as the sample size increases, the relative error decreases in the simple random sampling model. With smaller sample sizes, the simple random sampling model has lower accuracy than the stratified sampling model, and the interval is large. When the area is stratified by soil type, each sample is representative in every layer, although sample sizes are small. With larger sample sizes, the stratified sampling model fluctuates within a certain range, but is more accurate than the simple sampling random model from 10 to 80. Therefore, the stratified sampling model has higher sampling efficiency than the simple random sampling model and results in a more representative sample.

Table 1 presents two kinds of report units: administrative villages and grids. Under normal circumstances, the layer created with the stratified sampling model is called the knowledge layer and the number of knowledge layers is less than the number of report layers. The mean and variation of each report layer are obtained with SSSI, and using Formula (1), each report layer's relative error can been calculated. We can find the whole relative error by combining them. The values of the table are less than 10%, except for administrative villages stratified by soil type.

3.2 Analyses of the coefficient of variance

Figures 3 though 6 show the coefficient of variance (CV) of Mo for the simple random sampling model, spatial random sampling model, stratified sampling model, and spatial stratified sampling model. CV is the ratio of spatial variation to the sample mean. For small sample sizes, the estimated sample mean is less accurate than for large sample sizes. The CV decreases with an increase in the sample size for all sampling models. But the curve fluctuation ranges are different for each sampling model. The spatial stratified sampling model has the highest precision, indicated by its curve below the others on the graphs, and the simple random sampling model has the lowest precision. The spatial random sampling model has higher precision at the middle sample sizes about from 15 to 70. We can conclude that the heavy metal contents in soil have spatial autocorrelation and the precision of the different sampling models is impacted more

by the spatial dependency than by the areal heterogeneity for most sample sizes.

Taking spatial dependency into account, the nearer the samples are to one another, the more similar their attributes and the greater their spatial correlation is. Spatial mean variation is calculated from the samples with the following formula:

$$\text{var}\left\{\overline{Z} - Z(A)/Z\right\} = \frac{1}{n}\left[\sigma^2 - E\{C(X,Y)\}\right] = \frac{\sigma^2}{n} - \frac{E\{C(X,Y)\}}{n} \qquad (4)$$

where σ^2 = population variation

X 、 Y = random variables subject to uniform distribution in the area A

$C(X,Y)$ = the covariance of random variables X and Y

Classical sampling mean variance is σ^2/n, and spatial sampling mean variance is the above formula with a decrease of $E\{C(X,Y)\}/n$ from classical sampling (Wang J F et al., 2009).

Spatial stratified sampling models take area heterogeneity and spatial dependency into account, not only to keep samples representative, but also to calculate sample autocorrelation. The sampling mean variance is smaller than any other sampling models, but the simple random sampling model does not account for area heterogeneity and spatial dependency, so the efficiency is lower. The stratified sampling and spatial random sampling models account for only area heterogeneity or spatial dependency, so their efficiency is between spatial stratified sampling models and simple random sampling models. For small sample sizes, the stratified sampling model is better than the spatial random sampling model because it shows the dependence of samples is not strong and the samples are more representative. With larger sample sizes, the samples' dependence becomes stronger. Although samples are more representative with the stratified sampling model, the impact degree is not larger than spatial dependence. When the sample size is about 60 to 70 (almost the population size) the spatial dependence reaches its maximum. It tends toward a certain level value, so there are some limitations for spatial dependence at small and large sample sizes.

As shown in Figures 3 through 6, the accuracy of the simple random and spatial random sampling models is the same. There is a difference in the stratified sampling model and spatial stratified sampling model because of different stratification methods. Selecting the best stratification method is important for estimating sampling precision. If the stratification is poor, its accuracy could be lower than simple random sampling. Figure 7 compares the CV of two stratification methods. In the graph, stratification by soil type yields higher accuracy than by geochronology in the case of smaller sample sizes, but lower accuracy in larger sample sizes. At the smaller samples, the sample in one layer is more close in space by stratified by soil type than by stratified by geochronology, so the samples are more homogeneous in the same knowledge layer. When areas are more homogeneous, a single sample can be more representative, leading to high sampling efficiency. But at the larger sample sizes, geochronological stratification is better, possibly because the Mo content has been more strongly affected by geochronology. When the sample size reaches a certain degree, this effect is manifested.

3.3 Analyses of the Design effect

models / sample sizes	Srs	StRs (1#)	SStrs (1#)	StRS (2#)	SStrs (2#)
10	0.945	0.891	0.143	0.899	0.228
20	0.664	0.964	0.187	0.798	0.133
30	0.557	0.801	0.230	0.745	0.100
40	0.485	0.719	0.241	0.661	0.070
50	0.430	0.664	0.244	0.556	0.059
60	0.392	0.536	0.209	0.469	0.048
70	0.361	0.402	0.177	0.351	0.043
80	0.338	0.203	0.152	0.092	0.096

Table 2   design effect of spatial simple sampling, (spatial) stratified sampling by soil type and geological surface

models / sample sizes	StRs (3#)	SStrs (3#)	StRS (4#)	SStrs (4#)
10	0.931	0.172	0.889	0.261
20	0.788	0.228	0.901	0.252
30	0.710	0.203	0.776	0.281
40	0.606	0.178	0.692	0.225
50	0.523	0.156	0.612	0.193
60	0.405	0.127	0.474	0.150
70	0.293	0.106	0.359	0.137
80	0.150	0.087	0.155	0.101

Table 3   design effect of (spatial) stratified sampling by hierarchical cluster and geochronology

Srs: spatial random sampling;
StRs: stratified random sampling;
SStrs: spatial stratified sampling model
1#: stratified by soil type;
2#: stratified by geological surface;
3#stratified by hierarchical cluster;
4#: stratified by geochronology.

Tables 2 and 3 show the ratio of mean variance for different sampling models to the mean variance from the simple random sampling model. The values in the tables are all less than 1, so the simple random sampling model (Srs) is the least efficient and spatial stratified sampling model (SStrs) is the most efficient. Spatial autocorrelation slightly impacted the sampling accuracy.

The spatial stratified sampling model has higher accuracy than any other sampling models. Samples within the same knowledge layer may be very far apart and even separated from the other layer in space. Therefore, spatial stratified sampling keeps the variation small in the same knowledge layer and large between knowledge layers. It requires balancing the samples in the same layer together in space. In these four stratification methods, the second meets the requirement, so from Tables 2 and 3, 2# has lower values at most sample sizes except at 10 and 80, and 3# has higher accuracy at 10 and 80 samples. Samples in the same knowledge layer in space have high spatial dependence, and $E\{C(X,Y)\}/n$ also increases. But at 10 and 80, the spatial dependence shows its limitations. Hierarchical clustering stratifies samples by their Euclidean distance squared, so the properties of each layer is very consistent within the sample

compared to other stratification methods, but they are not together in space, resulting in high sampling efficiency outside the limitations of space.

4. CONCLUSIONS

In this paper, we have looked at both spatial and non-spatial models in the five kinds of sampling models we compared. Of these, sandwich spatial sampling and spatial stratified sampling had the highest efficiency, and the simple random model had the lowest efficiency. The efficiency of spatial random sampling and stratified sampling is dependent on the sample size. The stratified sampling model has higher efficiency than spatial random sampling at large and small sample sizes, but lower efficiency at middle sizes. The efficiency is also up to the stratification method; different methods can affect the sample distribution at the same sample size. Spatial stratified sampling takes spatial dependence into account and keeps samples in the same knowledge layer together in space, so stratification by geological surface has higher efficiency than other stratification methods. By comparing two kinds of stratification methods, by soil type and geochronology, we see that stratification by soil type yields higher accuracy than by geochronology in the case of smaller sample sizes, but lower accuracy in larger sample sizes. Sandwich spatial sampling has the same efficiency as spatial stratified sampling for the sample mean and spatial mean variation, but it can calculate these values for each report layer. These report layers can be divided arbitrarily to meet the user's needs.

There are two kind of sample distribution: sample layout and without layout, in this paper, without sample layout in the circumstances, we compared the sampling accuracy in different sample sizes with different sampling models, analyzing and discussing the reasons. Because the sample distribution is not very satisfactory, there are some gaps in the data set, so the accuracy of the estimates may not be completely satisfactory, but the study fills gaps in the comparison of current spatial sampling, which is important and can inspire readers to understand the present sampling models and methods. Future work should experiment with relatively uniform sample distributions and also deal with sample with layout using this study method.

ACKNOWLEDGEMENT

REFERENCES

Cressive N., 1991. *Statistics for spatial data*. Wiley, New York.

DongMei Chen, Hui Wei., 2009. The effect of spatial autocorrelation and class proportion on the accuracy. *ISPRS Journal of Photogrammetry and Remote Sensing* 64, pp. 140-150.

Haining, R.P., 1988. Estimating spatial means with an applicatio n to remote sensing data. *Communication Statistics – Theory and Methodology*, 17(2), pp. 537-597.

Haining, R.P., 2003. Spatial data analysis: theory and practice. *Cambridge University press*, Cambrdge.

J. WANG, J. LIU, D. ZHUAN, L. LI and Y. GE., 2002. Spatial sampling design for monitoring the area of cultivated land. int. j. *remote sensing*, 23(2), pp. 263–284.

Jinfeng Wang, Robert Haining, Zhidong Cao. 2009. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. DOI：10.1080/13658810902873512.

Jinfeng Wang, chengsheng Jiang, Lianfa Li, Maogui Hu., 2009. *Spatial Sampling and Statistical Inference* (in chinese). Beijing: Science Press. pp. 17-20

Li L, Wang J, Liu J., 2005.Optimal decision-making model of spatial sampling for survey of Chinas land with remotely sensed data. *Sci China Ser D* 48(6), pp. 752-764.

Paolo de Zorzi, Sabrina Barbizzi, Maria Belli, Renzo Mufato, Giuseppe Sartori, Giulia Stocchero., 2008. Soil sampling strategies: Evaluation of different approaches. *Applied Radiation and Isotopes* 66, pp. 1691–1694.

Wang Guanyu, Pan Mao, Liu Xida, Liang Haihua., 1992. On the relationship between the concentrations of elements in soil and the types of soil-forming parental material in Shandong province, China. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 28(4), pp. 475-485.

William G. Cochran., 1977. *Sampling Techniques*. Third edition, Professor of Statistics, Emeritus Harvard Universtiy. pp. 85-88.

Zhang Zhuodong , Zhang Keli , Wu Jilei , Zhang Ting , Zheng Xiaoying.2008. Spatial information analysis of characteristics of soil trace element contents in high incidence areas of birth defects. Journal of Zhejiang University (Agric1 & Life Sci1) 34 (6),pp. 684～691. DOI: 10. 3785/ j. issn. 100829209. 2008. 06. 015.

http://www.sssampling.org/