OPTIMIZED DESIGNS OF FRAMEWORKS AND ELEMENTS IN SPATIAL SAMPLING FOR CROP AREA ESTIMATION

Wang Di, Zhou Qingbo, Liu Jia

Agricultural Resources and Regional Planning Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China -
wangdicaas@hotmail.com; zhouqb@mail.caas.net.cn;

KEYWORDS: Sampling Techniques; Sampling Frame; Sampling Efficiency; Sample-Square Dimension; Sample Size

ABSTRACT:

The experiments were conducted for monitoring crop area by spatial sampling methods in a winter wheat main production region with a size of 42Km×42Km in Hengshui City, Hebei Province in China to optimize designs of the frameworks and elements (sample size, sample-square dimension), based on Remote Sensing (RS) and Geographical Information System (GIS) techniques. 5 sample-square dimension levels (3000m×3000m, 2000m×2000m, 1000m×1000m, 500m×500m and 300m×300m) were selected and 3 sampling techniques (simple random sampling, systematic sampling and stratified sampling) were applied. The experimental results demonstrate that the sampling efficiency by stratified sampling was the maximal (the average relative error was 0.15%, the average sample size varied from 9 to 10); and that of the systematic sampling was inferior (the average relative error varies from 0.74%~2.06%, the average sample size is 229); the sampling efficiency by simple random sampling was the minimum (the average relative error is 2.04%, the average sample size is 229) among 3 sampling techniques. Sampling relative error decreased with sample-square dimension simultaneously. When the sample-square dimension was reduced to a certain extent (the size of sample is 500m×500m), the error was not declining any more. The sampling relative error was the minimum using the sample-square with a size of 500m×500m among 5 sample-square dimension levels.

1. INTRODUCTION

As an important component in the forecasting crop production, the estimation of crop acreage has always been a focus during the assessments of crop production using remote sensing techniques at a large area scale (Xu, 1991; Sun, 1996). At the same time, the accurate estimation of crop area had played a very important role on mastering the crop production status and formulating the reasonable measures.

Sampling methods were often applied in estimating crop area by remote sensing techniques, such as Area Sampling Frame was used in “Large Area Crop Inventory and Experiment” (LACIE) and “Agricultural and Resources Inventory Surveys though Aerospace Remote Sensing” (AGRISTARS) in The United States. Multiple Frames (a kind of sampling technique) was also recommended to survey the crop area in some documents published by FAO. Stratified sampling technique was employed in “The Monitoring Agriculture with Remote Sensing “(MARS) project launched by European Union (EU), 60 sites were sampled and 17 crops were monitored with the sampling method in European counties. Some researches were also conducted using sampling methods estimating crop area in China, such as stratified sampling technique were applied by Chen (2000) etc and Liu (2001) to monitor winter wheat area all over the country. Cotton areas were estimated by Jiao (2002) using a standard relief map with a 1:25000 scale to set up sampling frames. Cluster sampling method and transect sampling frameworks were employed by Wu and Li (2004) to estimate crop area, based on crop stratification all over the country.

Although some progresses have been made in monitoring crop area with sampling methods, there were still some problems required to be resolved, and the problems were summarized as follow:

The sampling methods used at present were simplex. There have not been compared with sampling efficiency among all of sampling methods, and spatial sampling frame has not been optimized.
The calculation of sample size was not reasonable and the formulation of sample-square was not optimized.
Therefore, a classical statistical scheme and RS, GIS techniques were applied in this paper to improve the current sampling methods systems for estimating crop area.

2. METHODOLOGY

2.1 General feature of the experimental site

The experimental site is located in Hengshui City, Hebei Province in China. The region is suited in the warm temperate latitudes, and has a typical temperate continental monsoon climate. The multi-years average rainfall is about 600mm, 70% of the rainfall concentrates in the summer season. There are most of plain and winter wheat is main crop cultivated in the region.

2.2 The design of sampling frame

Firstly, a square with a size of 42Km×42Km was separated as population from SPOT4 image corresponding to the experimental site by the “generate” command in ArcGIS. The sample –square dimension was designed 5 levels (Table 2) referring to the size of sample-square published in the documents on estimating crop area using sampling methods (listed in Table 1), and square was adopted as the shape of sample-square. Then, the image was separated square grids with 5 sizes of 3000m×3000m, 2000m×2000m, 1000m×1000m, 500m×500m and 300m×300m, respectively. After the separation with square grids, there were 5 sampling frame in the image corresponding to the experimental site. Winter wheat area was sum up as observation value of one square grid using ArcGIS, and the sum of observation value was true value of population total. Population size and population total involved in 5 sampling frame were presented in table 3.

Item	NRI project in US	MARS project	MARS project	LUCAS project launched by EU				TERUTI survey project in French	Crop area estimation in Iran	Paddy rice area survey in China
Documents	Dennis（1995）	Tsiligirides（1998）	Gallego（1999）	Delinc etc（2001）				Bettio（2002）	Pradhan（2001）	Zhou（1996）
Unit	m×m	m×m	m×m	m×m	m×m	m×m	m×m	m×m	m×m	m×m
Size of sample	805×805	700×700	707×707	1800×1800	1039×1039	949×949	900×900(*)	1800×1800	1000×1000	500×500
Shape	square	square	square	square				square	square	square

* denotes the size of sample-square was selected at last.

Table 1 A review of sample-square dimensions

Number	1	2	3	4	5
Unit	m×m	m×m	m×m	m×m	m×m
Size	3000×3000	2000×2000	1000×1000	500×500	300×300

Table 2 Results of sample-square dimension designed

Sample frame	3000×3000	2000×2000	1000×1000	500×500	300×300
Unit	m×m	m×m	m×m	m×m	m×m
Population size	14×14 (196)	21×21 (441)	42×42 (1764)	84×84 (7056)	140×140 (19600)
Population total（m^2）	1226924100	1226926800	1226927200	1226925375	1226927232
The total area of experimental site（m^2）			1764000000		

Table 3 Population size and population total included 5 sample frames

2.3 Sampling techniques

2.3.1 Simple random sampling technique:

1) The calculation of sample size *n*
Sample size *n* was calculated through the equation in (1)

$$n_0 = (\frac{t}{r})^2 \frac{S^2}{\overline{Y}^2} \qquad (1)$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad (2)$$

Where
n_0 is initial sample size; n is modified sample size, when n_0/N>0.05, n is modified with the equation in (2); N is population size; t is sampling probability, when the confidence level is 95%, t equals to 1.96; r is relative error, it was 5%; $\overline{Y}$ is population mean; S^2 is population variance. $\overline{Y}$, S^2 was calculated through the equation in (3) and (4).

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad (3)$$

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 \qquad (4)$$

Where
Y_i is the *i*th population units observation. $\overline{Y}$, S^2 was usually estimated with pre-sampling methods (Jin et al, 2002; Du, 2005; Li, 2006). $\overline{Y}$, S^2 was surveyed directly according to all population unit observations in this paper, referring to the method published by Chen (2001).

2) Selection of samples
The samples were selected using Pseudo-Random Number method. Firstly, the pseudo random numbers equating with sample size were generated by a computer. Secondly, the square grids included in the sample frame were numbered, and the numbering means were as follow: the square grid located at the upper left corner of the sample frame was assigned 1, then, other grids were assigned 2, 3, …., which In accordance with the order from left to right. After numbering the grids from the first line, other grids from second, third, …., *n* th line were numbered. At last, all of the grids was assigned a code naming 1, 2,….., *N* in the sample frame. When the number of a grid equated with pseudo random numbers generated in advance, the square grid was drawn as a sample.

3) Setting up estimator and scaling population value

Sample mean $\overline{y}$ was chosen as the unbiased estimator of population mean $\overline{Y}$. Sample mean and variance were calculated through the equations in (5) and (6), the estimator of population total was calculated in equation (7), and the estimation of population total variance was calculated in equation (8).

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (5)$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \quad (6)$$

$$\hat{Y} = N\overline{y} \quad (7)$$

$$v(\hat{Y}) = \frac{N^2(1-f)}{n}s^2 \quad (8)$$

Where

$\overline{y}$, s^2 are sample mean and sample variance, respectively; n is sample size; $\hat{Y}$ is the unbiased estimator of population total; N is population size; $v(\hat{Y})$ is the estimation of the variance of population total estimator; $f = \frac{n}{N}$ is sample fraction.

2.3.2 Systematic sampling technique: There are many kind of sampling patterns during applying systematic sampling method, since the samples are drawn according on "some rule". 2 sampling patterns were chosen in the experiment.

1）Equal-distance systematic random sampling

The programming implemented with the equal-distance systematic random sampling method was as follow: Firstly, sampling interval k was calculated in the equation ($k = \frac{N}{n}$), if N/n was not integer, then k was modified as an integer. Secondly, sampling start point r generated by pseudo random numbers generator in computer was selected among k population units from the first sampling interval. After finishing the selection of r, the start point was numbered as rth code, other samples were chosen in term of the pattern that one sample was selected every k population units, till all of the samples were drawn out. At last, the code of the samples was numbered as $r+k, r+2k, ...r+(n-1)k$. The sample size n was equal to the corresponding value calculated in simple random sampling methods

2）Equal-distance systematic sampling with order

The difference between Equal-distance systematic random sampling with equal-distance systematic sampling with order is that the population unit observations will be listed in ascending sequence using the later method, and other programming are same in two sampling methods. The scaling of population value is equivalent to that of simple random sampling.

2.3.3 Stratified sampling technique: Stratified sampling technique is also known as classification, that is the population was partition some strata, and samples were selected in each stratum, all of the samples were made up of the samples allocated in every stratum. The population values were estimated through sample value from each stratum. Some signals used in the sampling method were demonstrated. 5 stratum were partitioned and subscript h denoted the stratum number (h=1, 2, …, L). The signals involved in hth stratum were presented as follow:

The population size is N_h in hth stratum. $\sum_{h=1}^{l} N_h = N$;

The sample size is n_h in hth stratum. $\sum_{h=1}^{l} n_h = n$;

The ith population unit observation Y_{hi} in hth stratum;

The ith sample unit observation y_{hi} in hth stratum;

Stratum weight, $W_h = \frac{N_h}{N}$;

The sampling fraction in hth stratum, $f_h = \frac{n_h}{N_h}$;

The population mean in hth stratum, $\overline{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi}$;

The sample mean in hth stratum, $\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$;

The population total in hth stratum, $Y_h = N_h\overline{Y}_h = \sum_{i=1}^{N_h} Y_{hi}$;

The sample total in hth stratum, $y_h = n_h\overline{y}_h = \sum_{i=1}^{n_h} y_{hi}$;

The population variance in hth stratum, $S_h^{\,2} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(Y_{hi} - \overline{Y}_h)^2$;

The sample variance in hth stratum, $s_h^{\,2} = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \overline{y}_h)^2$.

1) The selection of strata indicator

The fraction that winter wheat area (in one square grid) accounted for the area of a square was chosen as strata indicator, and the population was partitioned 5 strata, that is 0～20%, 20～40%, 40～60%, 60～80% and 80～100%, respectively.

2) The formulation of strata limits

The strata limits were formulated using Accumulated Square Root of Frequency (put forward by Hodges in 1959). The total square root of strata indicators divided by the number of strata was stratum limit. The programming implemented was demonstrated in a case, where strata were partitioned using the method in the sampling frame with square grids of 1000m×1000m (Table 4).

3) The calculation of sample size n

Sample size n was calculated through the equation in (9)

$$n_0 = \frac{\sum W_h S_h^{\ 2}}{V} \qquad (9)$$

Where
n_0 is initial sample size; V is the variance of population mean estimator,

$$V = (\frac{\gamma \overline{Y}}{t})^2 \qquad (10)$$

Where

$\overline{Y}$ is the population mean, $\overline{Y} = \overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h$.

Population mean $\overline{Y}$ and population variance in every stratum S_h^2 were necessary to calculate n_0, $\overline{Y}$ and S_h^2 were surveyed directly from the population. It is necessary for n_0 to be modified, and the modified equation is the same with that of simple random sampling.

4) Allocations of sample size in each stratum
It was still necessary to allocate the sample size to each stratum, that is n_h（h=1,2,….L）, after the sample size n was calculated. The ratio allocation, that is the allocation in terms of strata weight, was adopted. The equations are as follow

$$\frac{n_h}{n} = \frac{N_h}{N} = W_h \qquad (11)$$

$$n_h = n\frac{N_h}{N} = nW_h \qquad (12)$$

5) Constituting estimator and scaling population value
The unbiased estimator of population mean and population total were calculated through the equation in (13)~(14), and the variance of population total estimator was estimated in the equation (15).

$$\overline{y}_{st} = \sum_{h=1}^{5} W_h \overline{y}_h \qquad (13)$$

$$\hat{Y} = N\overline{y}_{st} \qquad (14)$$

$$v(\hat{Y}) = N^2 v(\overline{y}_{st}) = \sum_{h=1}^{5} N_h^{\ 2} \frac{1-f_h}{n_h} s_h^{\ 2} \qquad (15)$$

Where
$\overline{y}_{st}$ is the sample mean by the stratified sampling method.

Number	Fraction（%）	Frequency f(z)	$\sqrt{f(z)}$	Accumulated $\sqrt{f(z)}$	Referenced strata limits
1	1～5	80	8.94	8.94	34.84
2	5～10	24	4.90	13.84	
3	10～15	21	4.58	18.43	
4	15～20	29	5.39	23.81	
5	20～25	24	4.90	28.71	
6	25～30	29	5.39	34.10	
7	30～35	24	4.90	38.99	69.68
8	35～40	32	5.66	44.65	
9	40～45	48	6.93	51.58	
10	45～50	50	7.07	58.65	
11	50～55	55	7.42	66.07	
12	55～60	78	8.83	74.90	104.52
13	60～65	74	8.60	83.50	
14	65～70	126	11.22	94.73	
15	70～75	113	10.63	105.36	
16	75～80	146	12.08	117.44	139.37
17	80～85	163	12.77	130.21	
18	85～90	180	13.42	143.62	
19	90～95	223	14.93	158.56	174.21
20	95～100	245	15.65	174.21	
Strata limits	Strata limits＝accumulated $\sqrt{f(z)}$ /L=174.21/5=34.84				

Table 4 Strata limits calculated by Accumulated Square Root of Frequency

2.3.4 Evaluations of the sampling results: The relative error *r* was selected as a criterion to evaluate sampling results between the population total estimators with the true population total.

$$r = \frac{\left|\hat{Y} - Y\right|}{Y} \qquad (16)$$

Where
r is the relative error, %; *Y* is the true population total.

3. RESULTS AND ANALYSIS

3.1 Comparisons of the efficiency among sampling methods

The results of sampling relative error were presented in Table 5 using 4 sampling methods (simple random sampling, stratified sampling, systematic sampling with order and disorder) at 5 square grid dimension levels, in order to optimize the spatial sampling frame for estimating winter wheat area. It was found that the relative error by stratified sampling method scaling population value was the minimum (the average relative error was 0.15%), and that of systematical sampling was inferior (the average relative error varied from 0.74%~2.06%), that of simple random sampling was the maximal (the average relative error was 2.04%) among 4 sampling methods. At the same time, it was also found that the variance of sampling error from stratified sampling method was the minimum (0.0003%); and then that of systematical sampling was the less (0.0035%~0.0135%); that of simple random sampling was the maximal (0.027%), compared the results of the variance of sampling error among 4 sampling methods. Otherwise, sample sizes that required in 4 sampling methods were listed in Table 6. It was found that sample size was the minimum using stratified sampling (the mean varied from 9 to 10); and then were those of systematical sampling and simple random sampling (the mean were 229). The sampling efficiency using stratified sampling technique was the maximal among 4 sampling methods, comprehensively compared sampling precision (relative error) and sampling cost (sample size).

Item	Simple random sampling *r*（%）	Systematical sampling with order *r*（%）	Systematical sampling with disorder *r*（%）	Stratified sampling *r*（%）
Sampling frame1*	4.9	4.1	1.7	0.1
Sampling frame2*	1.8	1.2	0.8	0.4
Sampling frame3*	1.6	1.5	0.6	0.002
Sampling frame4*	0.9	1.7	0.5	0.04
Sampling frame5*	1.0	1.8	0.1	0.2
Aver	2.04	2.06	0.74	0.15
Var	0.027	0.0135	0.0035	0.0003
Std	1.64	1.16	0.59	0.16
C_v	81	56	80	107

1* denotes there is a 3000m×3000m of square grid size in the sampling frame;
2* denotes there is a 2000m×2000mof square grid size in the sampling frame;
3* denotes there is a 1000m×1000mof square grid size in the sampling frame;
4* denotes there is a 500m×500mof square grid size in the sampling frame;
5* denotes there is a 300m×300m of square grid size in the sampling frame.

Table 5 Results of sampling error from scaling population value among 4 methods

Item	Simple random sampling *n*	Systematical sampling with order *n*	Systematical sampling with disorder *n*	Stratified sampling *n*
Sampling frame1*	84	84	84	8
Sampling frame2*	122	122	122	9
Sampling frame3*	206	206	206	10
Sampling frame4*	332	332	332	10
Sampling frame5*	401	401	401	10
Aver	229	229	229	9

Table 6 Sample size required in 4 sampling methods

3.2 Comparisons of sample-square dimension

The results of sampling error were presented in Table 7 using 4 sampling methods at 5 square grid dimension levels, in order to optimize the size of sample-square. It was found that sampling error decreased with the size of sample-square simultaneously (the average error varied from 2.7% to 0.8%), when the size of sample-square varied from 3000m×3000m to 500m×500m. However, when the size of sample-square continued to decrease (from 500m×500m to 300m×300m), the sampling error was not declining any more, but it became ascending somewhat (the average error varied from 0.8% to 0.9%). When the size of sample-square was 500m×500m, sampling error was the minimum among 5 sample size levels.

Item	Simple random sampling *r*（%）	Systematical sampling with order *r*（%）	Systematical sampling with disorder *r*（%）	Stratified sampling *r*（%）	Aver（%）	Std（%）	C_v（%）
Sampling frame1*	4.9	4.1	1.7	0.1	2.7	2.20	82
Sampling frame2*	1.8	1.2	0.8	0.4	1.1	0.60	57
Sampling frame3*	1.6	1.5	0.6	0.002	0.9	0.76	82
Sampling frame4*	0.9	1.7	0.5	0.04	0.8	0.70	90
Sampling frame5*	1.2	1.8	0.3	0.2	0.9	0.76	87

Table 7 Results of sampling error at 5 sample-square dimension levels

4. CONCLUSION AND DISCUSSION

4.1 Conclusion

The sampling and scaling scheme experiments for estimating winter wheat area were conducted using 4 classical sampling techniques at 5 sample-square dimension levels in this paper, in order to optimize the designs of sampling framework and elements (sample size and sample-square dimension). The experimental results show as follow:

The sampling efficiency is the maximal using stratified sampling method (the average relative error is 0.15%, the average sample size varies from 9 to 10); and then that of systematical sampling method is the higher (the average relative error varies from 0.74% to 2.06%, the average sample size is 229); that of simple random sampling method is the minimum (the average relative error is 2.04%, the average sample size is 229), comprehensively compared the relative error means, variances and sample sizes in scaling population value by sample value using 4 sampling methods.

The sampling relative error decreases with the size of sample-square simultaneously(the size of sample-square varies from 3000m×3000m to 500m×500m), however, the sampling relative errors do not increase any more, but it becomes ascending somewhat, when the size of sample-square is reduced some degree(the size of sample-square is 500m×500m). The sampling relative error is the minimum using the size of sample-square is 500m×500m among 5 square grid levels.

4.2 Discussion

The main issues investigated in this paper are optimizing designing spatial sampling framework (selecting reasonable sampling method) and sampling elements (sample size and the size of sample-square), but it is still scarce on the reasonable formulation of spatial distribution of samples. In fact, after being drawn out, samples are often adjacent with each other. consequently, there is a spatial correlation among the samples, which causes a violation during using the sampling method scaling population value. Therefore, it is necessary to introduce variograms from Geostatistics as a tool and the range *a* as a criterion to design the reasonable spatial distributions.

REFERENCES

Bettio, M., Delinc, J., Bruyas, P., Croi, W., Eiden, G. 2002. Area frame surveys: aim, principals and operational surveys. *Building Agro-environmental indicators*,.pp. 12-27.

Chen, Z. X., Liu, H. Q., Zhou, Q. B. 2000. Sampling and scaling scheme for monitoring the change of witer wheat acreage in China. *Transactions of the Chinese Society of Agricultural Engineering*, 16（5）, pp. 126-129.

Delinc, J. 2001. A European approach to area frame survey. Processing of the conference on agricultural and environmental statistical applications in Rome（CAESAR）, 2, pp. 1-10.

Dennis, D. C., Lawrence, H. C., Katherine, B. E. 1995. Spatial sampling and the environment. *Technical report number 38*, pp. 10.

Du, Z, F. 2005. *Sampling techniques and practices*. Tsinghua University Press, Beijing, pp. 378-404.

FAO. 1996. Multiple frame agricultural surveys. Volume I: Current surveys based on area and list sampling methods, Rome

FAO. 1998. Multiple frame agricultural surveys. Volume II: Agricultural surveys programs based on area frame or dual frame (area and list) sample designs, Rome

Gallego, F. J. 1999. Crop area estimation in the MARS project. *Conference on ten years of the MARS project*, 4, pp. 1-11.

Jiao, X. F., Yang, B. J., Pei, Z. Y. 2002. Design of sampling method for cotton field area estimation using remote sensing at a national level. *Transactions of the Chinese Society of Agricultural Engineering*, 18（4）, pp. 168-171.

Jin, Y. J., Jiang, Y., Li, X. Y. 2002. *Sampling techniques*. The Chinese People's University Press, Beijing, pp. 40-50.

Li, J. C. 2006. *Applied sampling techniques*. Science Press, Beijing, pp. 40-60.

Liu, H. Q. 2001. Sampling method with remote sensing for monitoring of cultivated land changes on large scale. *Transactions of the Chinese Society of Agricultural Engineering*, 17（2）, pp. 168-171.

Madana , P. 2001. Improving Land Use Survey Method using High Resolution Satellite Imagery. Thesis submitted to the International Institute for Geo Information Science and Earth Observation for the degree of Master.

Pradhan , S. 2001. Crop area estimation using GIS, remote sensing and area frame sampling. *JAG*，3（1）, pp. 86-92.

Sun, J. L. 1996. Pandect on dynamic monitoring and yield estimation for crop in China. Beijing. China Science and Technology Press, Beijing, pp. 105-138.

Tsiligrides, T. A. 1998. Remote sensing as a tool for agricultural statistics: a case study of area frame sampling methodology in Hellas. *Computers and electronics in agriculture*, 20, 45-47.

Wu, B. F., Li, Q. Z. 2004. Crop acreage estimation using two individual sampling frameworks with stratification. *Journal of Remote Sensing*, 8（6）, pp. 551-569.

Xu, X. R. 1991. Bulletin on environment monitoring and crop yield estimation with remote sensing. Beijing University Press, Beijing, pp. 89-97.

Zhou, H. M. 1996. A procedure for sampling investigation of paddy area. *Southwest China Journal of Agricultural Sciences*, 7(3), pp. 100-105.