Theory of double sampling applied to main crops acreage monitoring at national scale based on 3S in China—CT316

Quan WU [a, *], Li SUN [a], Fei WANG [a], Shaorong Jia [a]

[a] *CAAE, Chinese Academy of Agricultural Engineering, 100125, No. 41 Maizidian Street Chaoyang District, Beijing Tel: +86-10-65910066-5007, Fax: +86-10-65929524, E-mail: wuquan95@tom.com;sunli0618@163.com;wangfei@agri.gov.cn*

Abstract: Grain production is of great importance to any country, especially to China. It is very important for local and central governments to get accurate main crops acreage information in time. One noticeable problem is that the estimated accuracy for crops acreage in a certain year is not high, so that Remote Sensing Applications Centre（RSAC） has to use investigated data of crops acreage of two consecutive years to estimate the change rate of crops acreage.

Aiming at the issue, the theory of double sampling based on operational crops acreage investigation was brought forward by RSAC. The paper has given detailed account of the theory and a typical case. In the double sampling method, the first sampling is to estimate the proportion of small features spatially distributing in crops fields in order to purify the samples acting as basis units for calculation in the second sampling. The second sampling is called a kind of stratified sampling which is used to estimate the crops acreage. The test is by adopting the theory of double sampling with 3S to evaluate the planting acreage of cotton and late-rice, acting as representatives of main crops in China, related to operative task and project research. The experiment result described with statistic methods shows that the theory of double sampling applied to main crops acreage monitoring can efficiently improve the estimated accuracy of crops acreage.

Keywords: Theory of double sampling, Small features, Stratified sampling, estimated accuracy, estimated error, 3S

1. Introduction

It is evidently significant that crops acreage has been highly paid attention in China due to the great number of population. For this reason, the estimated accuracy of crops acreage is an important issue discussed or researched by institutes or academies in China. The influence is not clear to the estimated accuracy of crops acreage by improving one or several techniques due to the restrict of RS itself in the operational monitoring system, so RSAC attempts to innovate active technology system to raise the monitoring accuracy by applying the theory of double sampling[1].

1.1. Using the method of stratified sampling to estimate crops acreage

The operational estimation of crops acreage is mainly carried out by RSAC in China. Based on 3S RSAC mainly adopts one method which is called stratified sampling to obtain the acreage information of main crops such as wheat, corn, cotton, soybean, rice, etc. [1] [2][3].

The stratified sampling is a kind of sampling method for estimating investigated collectivity information such as sum of crops acreage [4]. The key step is to select background data to stratify. As soon as the background data for stratified sampling is selected, the sampling units are determined. RSAC mainly selects a sort of background data for stratified sampling which is the latest land-use data in vector format.

When the vector data of land-use is selected as background data, RSAC operate the method as the following steps. Firstly, it is to select the sampling unit such as the frame of relief map used by RSAC. Secondly, the land-use data must be assembled with the frame of relief map in GIS.

Thirdly, it is to calculate the acreage of objective crop coming from background data, which distributes in every frame of relief map. Furthermore, the result data used to stratify should be sorted in ascending or descending order according to the crop area of every unit. The last step is to stratify with statistic software and obtain layers' information.

* **Corresponding author: WU Quan, senior engineer, CAAE, mobile phone:+8613621043045**

The so-called layer is a kind of data set based on the sampling unit. There are obvious differences in sample sizes among all the layers. Meanwhile, the sampling units that belong to a certain layer generally distribute relatively concentrated. For example, the map in figure 5 shows the spatial distribution of six layers of Xinjiang province's cotton plots, in which sampling unit is the quadrangle frame of relief map with the scale 1：25000. When the distribution map of layers has been finished, the next step is to confirm where and how many RS images should be ordered. The position of RS images can be determined using the distribution map and the quantity can be obtained from the parameters table produced by stratified sampling.

The last step is to interpreter the images covering the spatial sampling units with RS and then perform calculation according to the stratified sampling rules using the data produced by RS images interpretation and the parameters table.

1.2. Small features and the theory of double sampling

As set forth, samples used to calculate for estimating investigating collectivity should be paid more attention. Although the calculation process is ruled by the stratified sampling theory, on a large scale, the sample quality will directly affect the investigation result. The sample quality, called sample purity by RSAC, means that the samples probably contain some other features except objective features [5][6]. RSAC has realized that the issue of sample purity mainly originates from small features distributing in crops fields.

The small features mainly include varied field roads, ditches, dykes, graveyards, isolated pools, and other unused small plots. Some small features are shown below by photos in figure 1. Obviously, there is of relevance between small features with the spatial resolution of RS images [7][8][9]. One extreme situation is that there will be not small features if the high resolution RS images such as Quick Bird would be acted as data resource to extract crops acreage because total features would be discriminated and their acreage would be calculated. However, when to select one kind of RS data resource in operational work, the most important factor that need to be considered is spatial resolution, which determines monitoring accuracy, time and cost. RSAC has mainly selected Landsat-TM and SPOT images acted as RS data resource for many years. Therefore, the small features issue all along exists[1] [3].

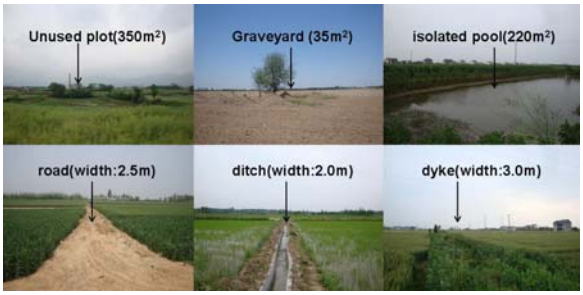

Figure 1. Some small features in fields

Based on the issues of sample purity and small features, RSAC put forward the theory of double sampling. The first sampling is to estimate the acreage proportion of small features spatially distributing in crops fields in order to purify the samples acted as basis units for calculation in the second sampling. The second sampling is stratified sampling mentioned above which is used to estimating the collectivity of crops acreage. The process of double sampling is illustrated by the flow chart shown below in figure 2.

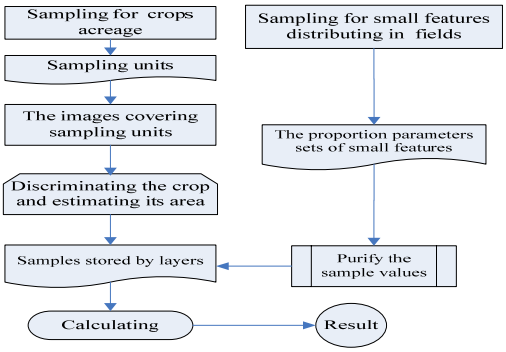

Figure 2. Flow chart of double sampling

2. Methods

The theory of double sampling derived from productive work should be applied to the productive process. It means that the experiment of the theory should be combined with operational monitoring of crops acreage. As the second sampling mentioned above is necessary steps or process in operational work, the main problem needed to be considered is how to design the test of the first sampling aimed at small features in order to purify the samples.

2.1. The scheme of the first sampling for small features

In addition to RS technique, RSAC also adopts ground random sampling to estimate crops acreage every year [4]. Ground Random Sampling (GRS) with GPS is an independent and assistant method. The use of GRS can make up the disadvantage of RS such as images covered by clouds. What's more, ground sampling can

provide independent information of agricultural condition such as crops area. The GRS' sampling unit is designed to polygons that are regularly located on farmland all over the productive regions of main crops by RSAC.

The GRS sample unit structure is mainly made up of natural borderlines such as road, ditch, dyke, ribbing, etc.. The acreage of each polygon unit maybe consisted of many polygons is about 25 hectare. The sampling unit, called a sample, contains varied features of land-use, which are main crops, other crops, and small features. When all the features of the sample collected by GPS are input into GIS, the acreage of small features can be immediately extracted. The sample unit structure is shown below in figure 3. The small features can be easily found in the samples.

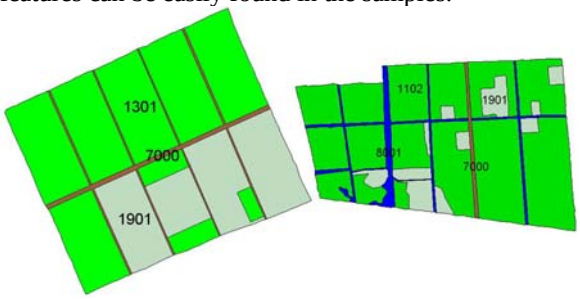

Figure 3. The structure of sampling units

In the figure consisted of two maps, the codes shown in figure 3 indicate different types of land-use. The code of 1102 indicates paddy field for late-rice, the codes of 7000 indicate roods, the code of 8001 indicates a ditch, the code of 1301 indicates cotton field, and the codes of 1901 indicate other plots of crops.

As RSAC has a large number samples coming from GRS distributed on main crops fields in China, the proportions of small features based on these samples are easily calculated by GIS, which can be reasonably used to modify the values of samples coming from the second sampling according to some rules. Thus, the first sampling for small features can be finished by way of using the existing GRS samples collected with GPS by RSAC in the recent years.

2.2. Confirming small features in RS images

Before extracting the acreage of small features, an important issue which is how to confirm small features needs to be resolved. It means that some parameters should be set up. It is clear that the issue of small features is closely related to the spatial resolution of RS images [1]. Only some features can be called small features which can not be discriminated in given RS images acted as one kind of data resource for extracting acreage of objective features such as crops plots. Figure 4 shows the situation of the small features in RS images covered by vector maps of interpretation and ground samples for cotton and late-rice.

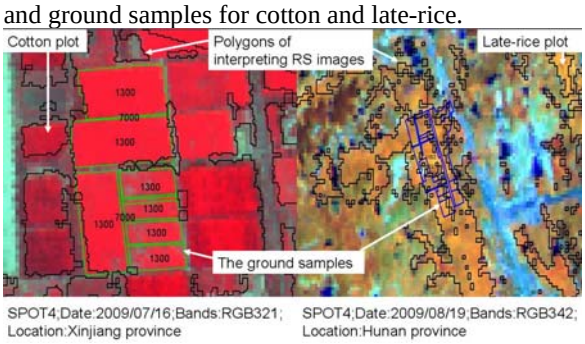

Figure 4. The small features in RS images

RSAC mainly uses RS images of middle spatial resolution such as Land-sat TM, SPOT, CBERS, etc. in the operational work process. According to the productive demands, the threshold values of small features has been decided by researching a large number of RS images used in operational work combining with ground truth got by GPS. The parameter is shown below in table 1.

Table 1. The threshold parameters of small features

Items	The polygonal small feature	The linear small feature
	acreage(square meter)	width(meter)
Values	Below 1600	0.5-20.0

According to the parameters of small features, it should be paid more attention to discriminate small features for calculating their acreage.

2.3 The project of the second sampling for main crops acreage estimation

The second sampling is directed at crops acreage which is called stratified sampling used by RSAC for many years. In this experiment of the double sampling, the second sampling work doesn't need to do repeatedly. The method has been presented above in the introduction, processed data and result produced by RSAC can be directly applied to the experiment.

3. The experiment for the theory of double sampling

The experiment includes two samplings aiming at different objectives. Different methods and data will be used. The experiment result of main crops acreage estimation will be compared with the result directly derived from stratified sampling which can be called traditional method. Cotton selected on behalf of crops on dry land acts as the objective crop which planting acreage in 2009 will be estimated using double

sampling. So does late-rice standing for crops on paddy fields.

3.1. The first sampling for small features on fields of cotton and late-rice

Cotton and late-rice are planted on different farmland which distributes on different geographic and climatic regions in China. The planting habits and methods are also different between the two crops. Therefore, the small features needed to be respectively processed may be different about their acreage proportion on fields..

3.1.1. Sampling for small features on cotton fields

In 2009, RSAC had investigated cotton acreage of 7 provinces of China using 3S. So, all GRS samples of 2009 distributed on the provinces have been checked up and modified if some needs to be amended according to objective situation. On the basis of the threshold values of small features, the work of calculating the acreage proportion of small features has been completed by using computers with manual assisted methods. The first sampling for small features on cotton fields has been accomplished via transferring the GRS samples finished by RSAC in 2009. The statistic result is shown below in table 2.

Table 2. The proportion of the small features on cotton fields

Regions	Samples	Proportions
Middle & East China	629	0.0451
Middle China	188	0.0477
North China	232	0.0295
Provinces		
Xinjiang	209	0.0488
Anhui	57	0.0294
Jiangsu	67	0.0653
Hubei	64	0.0387
Shandong	54	0.0471
Henan	90	0.0339
Hebei	88	0.0140

3.1.2. Sampling for small feature on late-rice fields

Late-rice acreage of 14 provinces of China has been estimated in 2009 using stratified sampling with 3S by RSAC. All GRS samples of 2009 distributed on provinces of 14 have also been checked up and updated to meet the experiment's needs. Adopting the same methods and threshold parameters of the small features, the acreage proportion of the small features has been calculated. Thus, the first sampling for small features on late-rice fields has been completed by means of using the GRS samples achieved by RSAC in 2009. The statistic result is shown below in table 3.

Table 3. The proportion of the small features on late-rice

Regions	samples	Proportions
Provinces of 14	831	0.0418
Middle China	397	0.0451
South China	207	0.0384
Southwest China	107	0.0429
Northeast China	120	0.0357
Provinces		
Zhejiang	73	0.0512
Anhui	60	0.0310
Jiangxi	65	0.0387
Hubei	59	0.0319
Hunan	67	0.0300
Guangdong	66	0.0461
Guangxi	81	0.0296
Hainan	60	0.0419
Liaoning	30	0.0618
Jilin	30	0.0235
Heilongjiang	60	0.0301
Jiangsu	73	0.0617
Chongqing	34	0.0269
Sichuan	73	0.0500

3.2. The second sampling for cotton & late-rice

The second sampling named stratified sampling for cotton & late-rice also had been finished in the process of operational monitoring tasks by RSAC in 2009. Samples of the second sampling are frames of relief map containing vector data of cotton & late-rice derived from RS images, since sampling units were designed to frames of relief map spatially distributed on farmland in China. The process is presented below when cotton is selected to act as an example.

Cotton as a kind of main crops mainly distributes in middle & east of China and Xinjiang province. In 2009, provinces of 7 were investigated for the acreage of cotton. Four statistic collectivities were designed in the light of demands of data analysis and application. As a result, data sets of 4 were produced about the stratified information for statistic. The collectivities of 4 can be easily found in table 2. As an instance, figure 5 shows the state of the second sampling for cotton when Xinjiang province is selected to act as the investigation collectivity. At the same time, the stratified parameters and practical samples in 2009 are shown in table 4.

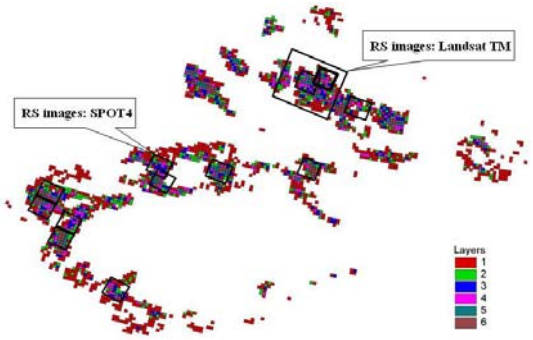

Figure 5. The second sampling information

In this map, the distribution of layers and RS images is shown above at the same time. RS images including Landsat TM and SPOT4 spatially distribute in each layer. Total samples amounts to 404 in 2009. The detailed figure of each layer is shown below in table 4.

Table 4. The stratified and practical information

Layers	samples			Rate of sampling
	Total	Minimum	Practical	
1	1050	23	55	0.0221
2	335	7	66	
3	229	5	64	
4	211	5	77	
5	173	4	75	
6	142	3	67	

3.3 To purify the samples of the second sampling

When the acreage proportions of small features of cotton & late-rice have been achieved, the next step is to update the value of every sample of the second sampling using the parameters presented in table 2 and table 3. It is easy to understand that the effect is better using the parameters coming from smaller spatial regions than larger ones. Accordingly, all samples of every province have been updated via taking off the acreage of small features with the parameters coming from the same province. This process is called samples purification[1].

3.4. Calculating the acreage of cotton and late-rice

The method of calculating the acreage of cotton & late-rice is same as the one adopted by RSAC before. The only difference is that the sample values. Consequently, the estimated results of acreage of cotton & late-rice are also different. The formula is shown below.

$$\hat{Y} = \sum_{h=1}^{L} \sum_{j=1}^{N_h} \left(\frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \right) \qquad (1)$$

Where y_{hi} = the crop acreage of unit i of the h layer

N_h = total number of sampling units of the h layer

$\hat{Y}$ = estimate value of total area of the collectivity

L = total number of the layers

$h = 1, 2, \ldots, L$

n_h = the amount of samples of the h layer

The difference of estimated results of cotton and late-rice, which derived from the difference between the two sets figures produced by the double sampling and traditional method in 2009, is revealed below in the following figures and tables.

4. Results and discussion

It is necessary to compare the result from the double sampling with the one from traditional method produced by RSAC. The figures of two sets should be analyzed with statistic methods. Conclusion of the experiment of the double sampling will be produced with believable reasons.

4.1. t test

The essential point of the double sampling is the fact that the sample values originating from the stratified sampling are updated relative to the one produced without using the double sampling. Due to the sample values existing in the form of pairs of figures, thus, using t test based on pairs of figures it is possible to estimate whether the method of the double sampling is statistically significantly better than the one adopted by RSAC before. The formula of t test based on pairs of figures is shown below.

$$|t| = \left| \frac{\bar{d} - 0}{s / \sqrt{n}} \right| \geq t_{\alpha/2}(n-1) \qquad (2)$$

Where $\bar{d}$ = average of difference between pairs of figures

s = standard deviation of d

n = the amount of samples

α = 0.01

The main figures derived from the process of t test of the experiment of cotton are shown below in table 5.

Table 5. The result of t test to cotton unit:ha.

Region	Samples	$\bar{d}$	s	t
6 Provinces	204	120.7	171.8	10.04
Xinjiang	404	113.0	68.9	32.99
Middle China	78	152.0	188.5	6.84
North China	127	108.6	162.3	7.52

The main figures produced by using the above formula to process data of late-rice are shown below in table 6.

Table 6. The result of t test to late-rice unit:ha.

Region	Samples	$\bar{d}$	s	t
14 Provinces	661	168.8	227.2	19.34
Middle China	240	237.6	212.6	17.28
South China	43	85.8	54.8	11.93
Southwest China	108	192.4	246.6	8.11
Northeast China	270	100.0	204.8	8.02

When the statistically significant level is selected to be 0.01, the critical values of t are shown below.

$$t_{0.005}(n > 45) \approx Z_{0.005} = 2.58$$

$$t_{0.005}(42) = 2.6981$$

From table 5 & table 6, it is clear that two methods have extremely marked difference to all situations of cotton & late-rice.

4.2. Error estimation

Based the t test, it is important conclusion that the estimated acreage results produced by double sampling are more trustworthy than the one produced by traditional method. So, the estimated error can be calculated by using the results acted as truth produced by double sampling and will be assessed from two aspects of time and spatiality in following chapters.

4.2.1 At time sequence

The distribution of small features keeps stable at spatial regions in several years. That means the acreage proportions of small features can be used to recent years relative to the year in which the small features is investigated. So, with the proportion parameters of small features produced by RSAC in 2009 the cotton & late-rice sample data from 2006 to 2008 produced by RSAC in operational tasks has also been updated. The updated data is used to calculate the acreage of cotton & late-rice again. Thus, the estimated errors derived from the small features about cotton and late-rice can be calculated via comparing the several years' data at time sequence. Figure 6 shows three years' estimated errors to cotton acreage and the four years' estimated errors to late-rice acreage is shown in figure 7.

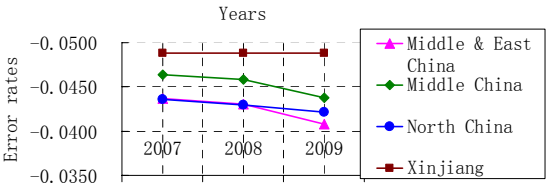

Figure 6. The estimated errors of cotton at time sequence

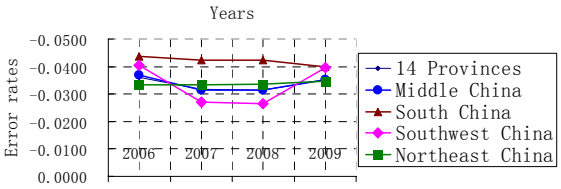

Figure 7. The estimated errors of late-rice at time sequence

From figure 6 and figure 7, on the whole, it is revealed that the estimated errors about the total acreage of cotton & late-rice are relatively stable at time sequence. However, the error rates generally tend to descend although the one of southwest China is not normal. The reason for the descending of errors rates may be that the proportion parameters of the small features bring errors which are smaller in near year than in early years when the small features were investigated.

4.2.2 At spatial sequence

Meanwhile, the estimated errors for the acreage of cotton & late-rice also can be analyzed at spatial points. It is evident that the estimated errors are different along with the difference of spatial geographic regions where crops grow. The errors mainly originate from two points ,which one is the process of estimating the error of the small features and the other is the one of interpreting RS images[10][11][12]. So, it is also significant to make analysis for the difference at spatial regions. Figure 8 shows four regions' estimated errors of cotton acreage and the five regions' estimated errors of late-rice acreage is shown in figure 9.

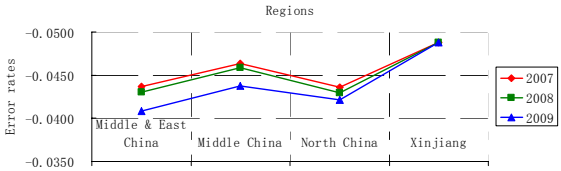

Figure 8. The estimated errors of cotton at spatial sequence

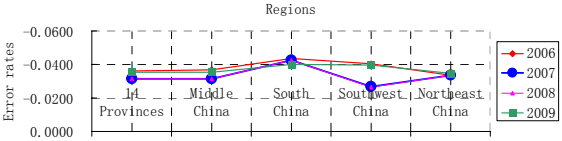

Figure 9. The estimated errors of late-rice at spatial sequence

From figure 8 and figure 9, on the whole, it is revealed that the difference of estimated errors about the total acreage of the cotton & late-rice are not large at spatial sequence. This situation is more evident to late-rice than to cotton. However, the error rates lightly vary among these regions. The estimated error to cotton of North China is the lowest while the one to Xinjiang

province is the highest. For late-rice, the estimated error to South China is relatively evidently higher than the one to other regions.

4.2.3 The interval estimation for the errors

The estimated errors of acreage of cotton & late-rice can be calculated via processing the data derived from the experiment at last. That means the investigating accuracy of cotton & late-rice produced by RSAC can be estimated at numerical level. The interval estimations for the errors are shown below in table 7 and table 8.

Table 7. The estimated errors of cotton

Regions	Average	The interval
Middle & East China	-0.0425	-0.0462 ~ -0.0388
Middle China	-0.0453	-0.0488 ~ -0.0419
North China	-0.0429	-0.0448 ~ -0.0410
Xinjiang	-0.0488	-0.0488 ~ -0.0488
Average	-0.0457	-0.0474 ~ -0.0439

Table 8. The estimated errors of late-rice

Regions	Average	The interval
Provinces of 14	-0.0337	-0.0376 ~ -0.0298
Middle China	-0.0337	-0.0381 ~ -0.0294
South China	-0.0421	-0.0445 ~ -0.0397
Southwest China	-0.0334	-0.0455 ~ -0.0212
Northeast China	-0.0338	-0.0347 ~ -0.0329
Average	-0.0357	-0.0399 ~ -0.0316

5. Conclusion

The theory of double sampling is directed aiming at resolving the problem of small features distributing in fields where crops grow. The proportion parameters of small features probably hold errors produced by the designing of the first sampling including sampling units, sample amounts, sample distributions, and so on. Therefore, the error needed to be estimated would be transferred to the last result of acreage estimation.

Double sampling method is proved statistically better than the traditional means adopted by RSAC in the process of estimating the main crops acreage such as cotton & late-rice at national scale.

The traditional method brought about the estimated error at 4.57% in monitoring cotton acreage on main productive regions in China. It means that the estimated figure derived from traditional method is 4.57% larger than the one produced by using the double sampling. Meanwhile, the figure of late-rice is 3.57%. In other words, the estimated accuracy of acreage for the main crops can be raised in the operational task if the theory of double sampling is applied, and the estimated accuracy of cotton can be raised 4.57% or so while the one of late-rice can be raised about 3.57%.

However, studies on small features derived from operational work have made great progress although new problem appears and needs to be considered.

Reference

[1]Q. Wu, B. J. Yang, etc., "Influence of small features on crop area estimation at a nation scale using remote sensing and a double sampling method", *Chinese Journal of Agricultural Resources and Regional Planning*, 20(3), pp. 130-133.（2004）(in Chinese)

[2]Q. Wu, L. Sun, etc., "The applications of 3S in operational monitoring system of main crops acreage in China", EDITORS: Daoling Li and Simon X. Yang, *Computer and Computing Technologies in Agriculture - Ⅱ*,TSI Press, USA, Volume 24:319-324.（2010）

[3] X. F. Jiao, etc., "Design of sampling method for cotton field area estimation using remote sensing at a national level", *Transactions of the Chinese Society of Agricultural Enginerring*, 18(4), pp. 159-162.（2002）(in Chinese)

[4]Q. Wu, L. Sun, "Sampling methods using RS and GPS in crops acreage monitoring at a national scale in China", In:Chen Jun,Jiang Jie,Alain Baudoin eds., *the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*,Beijing, Vol. XXXVII, Part B7,[WG VII/7]*： 1337-1342.（2008）

[5] Shiro O., "Land use characterization using landcover objects from high resolution satellite image", In:Chen Jun,Jiang Jie,Alain Baudoin eds., *the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*,Beijing, Vol. XXXVII, Part B7,[WG VII/4]*： 671-672.（2008）

[6] P. J. Yu, H. L. Xu, etc., "A net arable land coefficient study of Manas river basin based on remote sensing", *System Sciences and Comprehensive Studies in Agriculture*, 25(3), pp. 365-368.（2009）(in Chinese)

[7]Y. Liu, Q. Z. Zhou, etc., "Linear feature extraction from remote sensing images based on differential geometry", *ComputerEngineering & Science*, 28(8), pp. 40-42.（2006）(in Chinese)

[8]Y. J. He, F. Wang, etc., "Change patterns of linear features in remote sensing images in land use", *Transactions of the Chinese Society of Agricultural Enginerring*, 24(12), pp. 111-115.（2008）(in Chinese)

[9]H. P. Huang, B. F. Wu., "Analysis to the relationship of feature size, objects scales, image resolution", *Remote Sensing Technology and Application*, 21(3), pp. 243-248.（2006）(in Chinese)

[10] Duda T., etc., "Unsupervised classification of satellite imagery: choosing a good algorithm", *Remote Sensing*, 23(11), pp. 2193-2212.（2002）

[11]X. Yang, etc., "Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta,

Georgia metropolitan area". *Remote Sensing*, 23(9), pp. 1175-1798.（2002）
[12]Thomas M. Lillesand, etc., "Remote sensing and imagery interpretation", John Wiley & Sons, Inc,New York, pp. 576-586.（2002）