Application of integration technology to multi-source data in the New Type of Crop Sampling Survey

Ge Wei*
National Bureau of Statistics, Beijing, China – weige@gj.stats.cn

Abstract

The key problem to new type of crop sampling survey(NTCSS) is how to deal with the “remote sensing image”, “text” and other forms of data. Because of the budget constraint we cannot get enough “good” remote sensing image to meet the demand of editing “area” frame, so the “poor” remote sensing image must to be used, and the integration technology become to the method to solve those problems. Further, beside to the “area” frame, the more important goal of the method is to control and convergence estimation error , improve estimation precision. In this paper, I want to introduce the application of to multi-source data on editing the area frame, and through the description of error control, continuous optimization of output results to show the rudiment of NTCSS with the “closed loop ”system.

Keywords: Image; Frame; Accuracy.

1. Introduction

The topics such as “the number of population ”, “Cereal output” are often concerned by national policymakers, of course, has also become an important content of National Statistics. In a normal year, the output quantity of major crops include rice, wheat and corn at national or province level is obtained by the National Bureau of statistics with using the method of crop sampling survey. Nowadays, we have started to use a new type of crop sampling survey (NTCSS) with the “area frame” to replace the method which have to be used with the “list frame”.

No matter how the method changed, the basic process of sampling survey must be followed, such as to edit and build a sampling frame, to select the samples in accordance with the random probability, to implement the sampling survey on field and to obtain the estimation at a promise accuracy.

The obvious difference between NTCSS and used method is Object. The frame for used method is constructed by the people’s list(farmer, holding etc.),and other hand the frame for NTCSS is integrated with “area ”information. Field survey for used method is implemented by interviewers with home visit ,telephone , email and so on, but for NTCSS ,the field samples can be done a observation directly with satellite, airplane, UAV, and field measurement.

The key problem to NTCSS is how to deal with the “remote sensing image”, “text” and other forms of data. In this paper, I want to introduce the application of integration technology to multi-source data on editing the area frame, and through the description of error control, continuous optimization of output results to show the rudiment of NTCSS with the “closed loop ”system.

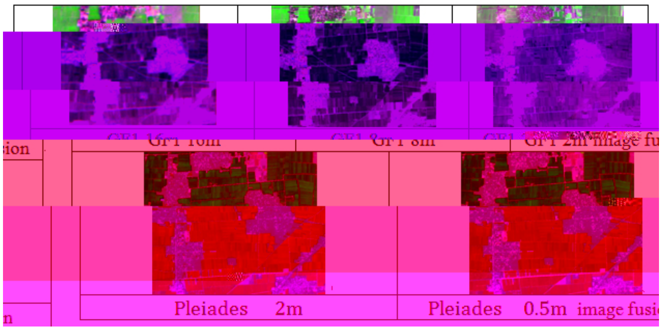
2. Multi-source data

2.1 Remote sensing Image Data

Originating from the satellite image has a large coverage area, there is the appropriate time period, get their way without the technical bottleneck, is a major source of image data. From the resolution of the image, a region can have different resolution image coverage (Figure 1).

The time period from images, with their crop growth period, in a certain period of the growth period may exhibit different optical characteristics and other crops, such as satellite images specific period has very important meaning (Figure 2).

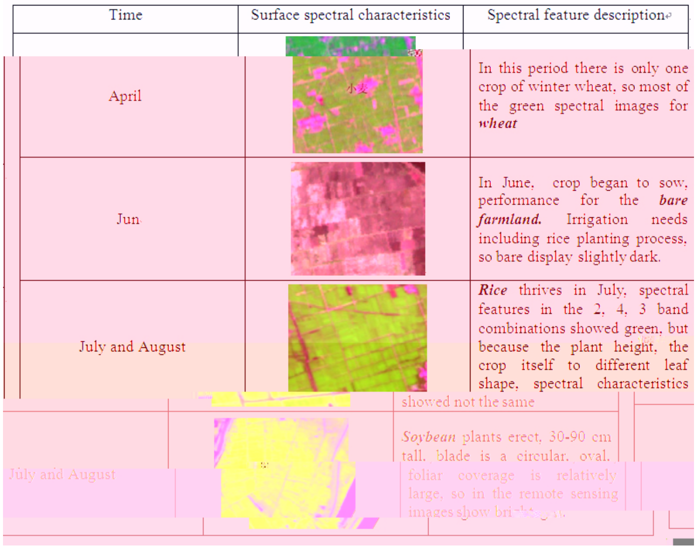
Fig.2 Different crop cover in each resolution satellite image



Of course, we need an optimal resolution and a best time of image, but the reality is very difficult to do, here in addition to technical obstacles, the most important obstacle is the budget constraint. Therefore, strategy is to make full use of *poor* satellite images which may be not timely and enough resolution, but free or cheap , combined with *good* satellite images have to spend money to buy to meet the needs .

At the same time, we have some other image sources, aircraft, UAVs, these are very necessary auxiliary source. But these sources are expensive, is mainly used to sample survey, as modified sampling frame, form an important source of error control system.

Fig.2 The crop in different period of the spectral characteristics



Time	Surface spectral characteristics	Spectral feature description
April	小麦	In this period there is only one crop of winter wheat, so most of the green spectral images for ***wheat***
Jun		In June, crop began to sow, performance for the ***bare farmland.*** Irrigation needs including rice planting process, so bare display slightly dark.
July and August		***Rice*** thrives in July, spectral features in the 2, 4, 3 band combinations showed green, but because the plant height, the crop itself to different leaf shape, spectral characteristics showed not the same
July and August		*Soybean* plants erect, 30-90 cm tall, blade is a circular, oval, foliar coverage is relatively large, so in the remote sensing images show bri[illegible]

2.2 Text data

Text data is a major source of national agricultural census. In china, the period of agriculture census is 10 years, a total of over two times. The first census in 1996, the second national agricultural census in 2006, at present, the third agricultural census is preparing. The information of crop planting structure at village and holding level which can be obtained from the agricultural census, is not only the important auxiliary information sources for "list" frame, but also for the "area" frame.

3. Integration technology (The test of estimation for winter wheat plant area as an example)

From the above description can be seen, the key technology of the preparation of area sampling frame is how to integrate the multi-source satellite images, text statistics data, obtain remote sensing crop area estimation precision. We choose a area of Beijing as the target area, the application of integration technology was tested, the main introduction of the test is as follows.

3.1 The sampling unit

The sampling frame is composed of a sampling unit without duplicated or non-covarge. For NTCSS, the goal is the area of planting crops on the land. Then the sampling frame range must be covared all of the land which crops planted. There are two kinds of sampling unit, one is based on the natural boundary, administrative boundary, another kind is the spatial grid. Here, in target area, we use the spatial grid as sampling unit, size for 100m×100m, so there are 99×98 sampling unit in the target area have been produced.

3.2 Remote sensing image data (TM and QB)

In the target area, we got the TM remote sensing image of the same period (including 6 bands, spatial resolution of 30m) and QuickBird data (resolution 2.4m). The TM and Quickbird data were pre processed, which is also on two image geometric correction, can be integrated into the same coordinate system, the error correction control within 1 pixels. Because TM cannot effectively distinguish image crop, tree, grass and so on, but through the QuickBird data and field survey can be differentiated to obtain the real wheat area, as the test compared with the estimated target accuracy.

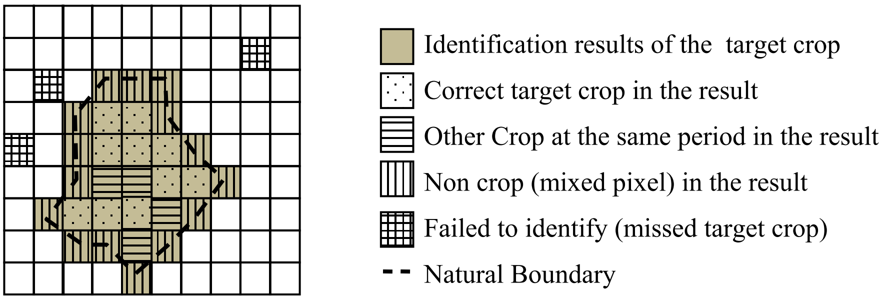
3.3Text Data(Simulation)

Through the agriculture census or statistical reports, it can be got the text data of crop planting structure at village level. In the test area, there is no direct use of census data or statistical reports, but is divided into the same size of the "big" grid in TM and QuickBird images (900m × 700m) each "big" grid as a village level, through the image extraction, get planting structure of each "big" grid as simulatily text data. Each sampling unit of scope within the "big" grid (simulated Village), will be assigned to the same planting structure.

3.4 Identification and Error to the target crop (winter wheat in test area)

The planting area of the target crop (winter wheat in test area) for each sampling unit is result of using maximum likelihood classification recognition, include the following factors: the correct target crops, other crops, non crop at the same period. Non crop and other crop area is wrong area, because caused by the mixed pixel and objects with the same spectrum phenomenon respectively; in addition to the recognition results , also there are missing area, those are mainly because of mixed pixels. Therefore, in this test, the recognition error of target crops for each sampling unit mainly comes from two aspects: mixed pixel, other crop at the same period (Figure 4).

Fig. 3 Sketch map of identification in a sampling unit


Identification results of the target crop
Correct target crop in the result
Other Crop at the same period in the result
Non crop (mixed pixel) in the result
Failed to identify (missed target crop)
Natural Boundary

The relationship between the factors as follows:

$$S_c = W_c + C_c + \sum_{i=1}^{n} A_i \qquad (1)$$

$$W = W_c + W_o \qquad (2)$$

In the formula:

S_c ～ Identification results of the target crop.

W_c ～Correct target crop in the result（S_c）.

C_c ～ Non crop (mixed pixel) in the result（S_c）.

A_i ～No. i Other Crop at the same period in the result（S_c）, n is the number of other crop.

W_o ～Failed to identify (missed target crop).

W ～The Real Area of target crop .

Due to the problem of mixed pixels and the same period crop, identification with remote sensing images of the target crop area S_c can not reflect the real area. The solution is the use of text data may be obtained, that is the target of crop planting structure (r) structure variables(S_S), as follows:

$$S_s = S_c \times r$$

$$= \left(W_c + C_c + \sum_{i=1}^{n} A_i \right) \times \frac{W}{W + \sum_{i=1}^{n} A_i}$$

$$= (W - W_o + C_c + \sum_{i=1}^{n} A_i) \times \frac{W}{W + \sum_{i=1}^{n} A_i}$$

$$= W + \frac{C_c - W_o}{W + \sum_{i=1}^{n} A_i}$$

$$= W + \frac{\Delta}{W + \sum_{i=1}^{n} A_i} \qquad (3)$$

$$\Delta = C_c - W_o \qquad (4)$$

For the mixed pixel, C_c and W_0 represent the wrong area "in" or "out". same condition of remote sensing classification accuracy, a suitable sampling unit size can ensure that C_c and W_0 in formula (4) will be offset, $\triangle$ will be close to 0, thus S_S is more close to the true value of W .

For other crop at the same period, if the true target crop area in the sampling unit is greater, the total of other crop area at the same period $\sum_{i=1}^{n} A_i$ are more smaller, the remote sensing classification for the target crop area S_S is more accurate, and the correction term r is more close to 1, on the contrary, the r more close to 0. This is equivalent to give the different weights to different accuracy remote sensing classification area. Therefore, S_S can get better expression to target crop area in each sampling unite, both to the remote sensing, classification error of mixed pixels and other crop at same period ,and provide supplementary information for the stratification sampling design.

3.5 Sampling Design

3.5.1 Simple random sampling(SRS)

In the edited sampling frame, , the sample of sampling unite (spatial grid) will be random selected according to the consistent sampling probability, sample size *n* calculated by the formula (5):

$$\frac{1}{n} = \frac{1}{N} + \frac{d^2}{Z_{\alpha/2}^2 S^2} \qquad (5)$$

Confidence level(1-α)in 95%, accuracy of 95%, the sample size is 889, the sampling ratio is 14.3%.

3.5.2 Stratified Sampling (SS)

The variables of winter wheat planting area and planting structure made by TM image classification is used as sampling target variable. The area frame was stratified 6 layers, base on "Neyman allocation" total sample and each layer of the sample size was calculated ,the formula is as follow:

$$n = \frac{\sum_{h=1}^{L} W_h S_h^2}{\frac{d^2}{u_{\alpha/2}^2} + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2}$$

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f \qquad (6)$$

Confidence level(1-α)in 95%, accuracy of 95%, the sample size of planting area variable is 245, the sampling ratio is 4%, the sample size of planting structure variable is 190, the sampling ratio is 3.1%. In order to observe the estimation accuracy of sampling varies with the quantity of samples, the sampling ratio by 3% intervals, from 3% to 33%.

3.6 Estimation method

1) Direct estimation method

$$\hat{y}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h \qquad (7)$$

2)The combined ratio estimation method

$$\hat{y}_{RC} = N\frac{\bar{y}_{st}}{x_{st}}\bar{X} = \frac{\bar{y}_{st}}{x_{st}}X \qquad (8)$$

3) United regression estimate method

$$\hat{y}_{lr} = N\bar{y}_{lr} = N(\bar{y}_{st} + \beta(\bar{X} - \bar{x}_{st})) \qquad (9)$$

3.7 Accuracy Measurement

3.7.1 Standard Error (SE)

SE reflects the sample estimate value fluctuation range, in the vicinity of the expectation that stability; higher SE, represent the experimental data more discrete, the more unstable.

1)Measurement for direct estimation

$$\mathrm{SE} = \sqrt{\sum_{h=1}^{L} N_h (N_h - n_h)\ s_h{}^2 / n_h} \qquad (10)$$

2)Measurement for combined ratio estimation

$$\mathrm{SE} = \sqrt{\sum_{h=1}^{L} \frac{N_h{}^2 1 - f_h}{n_h} (S_{yh}{}^2 + R^2 S_{xh}{}^2 - 2R\rho_h S_{yh} S_{xh})} \qquad (11)$$

3)Measurement for united regression estimate

$$\mathrm{SE} = \sqrt{\sum_{h=1}^{L} \frac{N_h{}^2 1 - f_h}{n_h} (s_{yh}{}^2 + b_c^2 S_{xh}^2 - 2b_c s_{xyh})]} \qquad (12)$$

3.7.2 Degree of Accuracy(DA)

DA reflects the sample estimates the extent to which close to the true value, the higher the DA, show the more accurate results.

$$DA = 1 - \frac{(\hat{y} - Y)}{Y} \qquad (13)$$

3.7.3 Evaluation

Fig.4 Precision curve of simple estimation

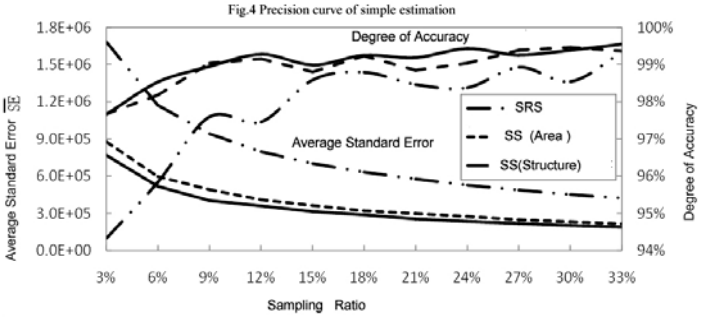


Fig.4, Fig.5, and Fig.6 shows that with simple estimation, ratio estimation, regression estimation, three types of sampling methods in the different sampling ratio estimation accuracy.

1) Stability to the ratio estimation and regression estimation has better than the simple estimator.

2) The sampling ratio 3%, 6% is the accuracy of the best rise period, and also is the standard error decreased period.

3) Stratified Sampling method with the variables of planting area and planting structure, is obviously better than simple random sampling.The best performance is stratified sampling with planting structure variables, along the accuracy rising period it is at the top, along the standard error declining period it is on the bottom.

Fig.5 Precision curve of ratio estimation

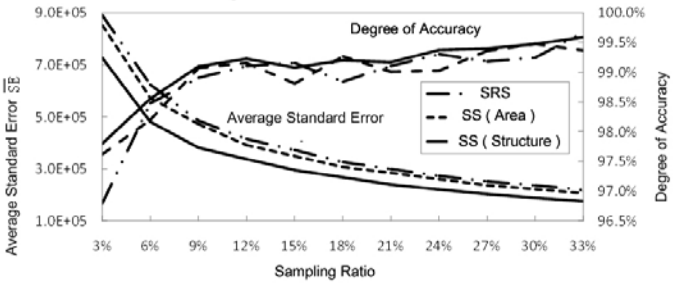
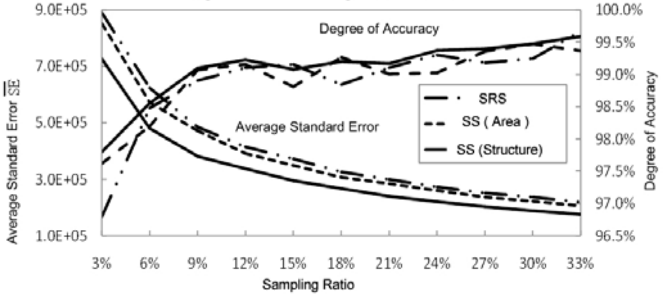


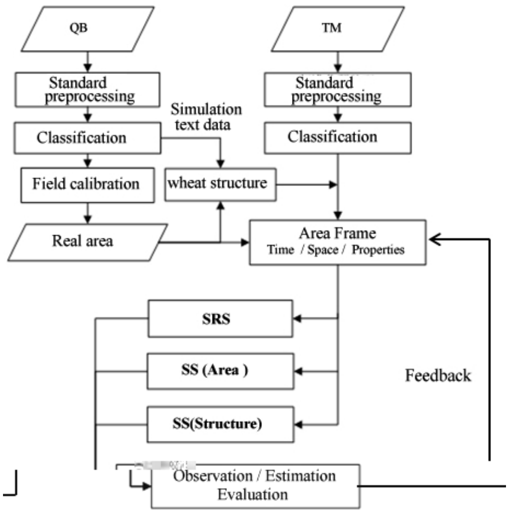
Fig.6 Precision curve of regression estimation



4. "Closed loop" system

Regard to the work flow, the NTCSS is a system based on the integration of the current and historical remote sensing images, text data (agricuture census, statistic report, and so on) , from the sampling frame to sample selection, field survey, estimated, and then feedback to the sampling frame system for further error correction. Especially in the field survey the use of UAV and PDA, can obtain more accurate image data and the measured data, and provide reliable basis for correcting the error of sampling frame. This process is uninterrupted, with the crop growing period from sowing to harvest, the "closed loop" system is developed gradually along with the estimation error gradually brought under control and convergence, estimation precision gradually improve.

Fig.7 The Flowchart of the Test



5. Conclusions

The above described NTCSS conducting for relatively large area and single crop. Based on the "closed loop" system, we added other large area crops (rice, corn and cotton), and got similar results. In order to reduce the burden of work on field survey, MPPS and PPS method are used for sampling design, the identification results crop and land become to the important auxiliary variable. Currently the pilot and promotion is conducted in 6 major grain producing province, 1 cotton producing province, in those region, the planting area and yield data of the main crops were obtained by NTCSS.

A new round of Chinese agricultural census will launch in 2017, this has provided an opportunity to application of remote sensing technology in crops. In the census can add the spatial information technology help, more comprehensive, and useful auxiliary information well be provided.

References

Carfagna E, Gallego F J. （2005）Using remote sensing for agricultural statistics. International statistical review. 73(3).

Gallego J, Bamps C. （2008）Using CORINE land cover and the point survey LUCAS for area estimation. International journal of applied earth observation and geo_information. 10(4SI).

Sharma S A, Panigrahy S, Parihar J S. (2011) Sampling Design for Global Scale Mapping and Monitoring of Agriculture. Journal of the Indian society of remote sensing. 39(3SI).

Gallego F J. (2004)Remote sensing and land cover area estimation. International journal of remote sensing. 25(15).

Wang shuang, Zhuxiufang, Panyaozhong, et al. (2009) Corn area estimation by combining SPOT 5 image with sampling theory. Journal of Remote Sensing. 13(4).

Broich M, Stehman S V, Hansen M C,et al. (2009)A comparison of sampling designs for estimating deforestation from Landsat imagery: A case study of the Brazilian Legal Amazon. Remote Sensing of Environment. 113(11).

Stehman S V, Sohl T L, Loveland T R. (2003)Statistical sampling to characterize recent United States land-cover change.Remote Sensing of Environment. 86(4).