Crop Monitoring as an
E-agricultural tool in
Developing Countries

ACCURACY ASSESSMENT REPORT ON THE SPATIAL EXTRAPOLATION

Reference: E-AGRI_WP5_Acreage Assessment_1.0
Author(s): Zhongxin Chen, Di Wang, Jia Liu, Limin Wang
Version: 2.0
Date: 12/09/2013

DOCUMENT CONTROL

Signatures

Author(s) : Zhongxin Chen
Di Wang
Jia Liu
Liming Wang

Reviewer(s) : Qinghan Dong

Approver(s) :

Issuing authority :

Change record

Release	Date	Pages	Description	Editor(s)/Reviewer(s)
1.0	28/05/2013	37	Initial draft	
2.0	12/09/2013		Review	Qinghan Dong

LIST OF ACRONYMS

AFS	Area Frame Sampling
PRN	Pseudo-random number
PSU	Primary sampling units
SSU	Secondary sampling units
CV	Coefficient of variation
SQRT	Square root
SRS	Simple random sampling
EIS	Stratified sampling with equal interval between strata
CSR	Cumulative square root method is used in stratification sampling
RWG	The fraction that winter wheat area in a sampling unit accounts for the area of the sampling unit

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

Executive summary

According to the requirements of the WP51, CAAS is responsible for crop area estimation in Huaibei Plain. Mengcheng County in Anhui Province was selected as the study area. Six spatial sampling methods (simple random sampling, spatial random sampling, classical systematic sampling, spatial systematic sampling, stratified sampling and two-stage sampling) are formulated, and winter wheat acreage in the study area was estimated using these 6 spatial sampling methods. In order to improve the accuracy of population values, Three estimator options were designed based on the samples data from ground survey and remote sensing images. The report presents the results of the best spatial sampling scheme and the estimator option for estimating the winter wheat area of Mengcheng County, as well as the accuracy assessment of spatial extrapolation models of population values.

1 OPTIMIZATION OF SPATIAL SAMPLING SCHEMES FOR CROP ACREAGE ESTIMATION

1.1 Introduction

The approach is made up of 4 steps: the first step is the collection of basic geo-information data (basic geographic information data and spatial distribution data of winter wheat sown area in the study area); the second is the formulation of spatial sampling scheme including basic sampling unit definition and spatial sampling method selection. The of basic sampling unit definition refers to the selection of shape and size of basic units. Five spatial sampling schemes (simple random sampling, spatial random sampling, classical systematic sampling, spatial systematic sampling and stratified sampling) are used to draw samples, to extrapolate population value as well as to estimate errors; the third step consists of evaluating the efficiencies of 5 sampling methods according to relative error, coefficient of variation (CV) and sampling cost. The last step is to optimize spatial sampling scheme from 5 sampling methods, based on the results of sampling efficiency evaluation.

1.2 Methodology

1.2.1 Study area

Mengcheng County is located between N32°55'29 "-32°29'64" and E116°15'43 "-116°49'25". The total area of Mengcheng County is about 2149Km^2, including a width of 40Km from east to west, and a length of 60Km from south to north. A warm temperate semi-humid monsoon climate governs Mengcheng County, with an annual average temperature around 14.7°C, an average sunshine of 2320h, and an average annual precipitation about 822mm. The cropland is exploited for planting wheat, maize and soybeans in Mengcheng County.

1.2.2 Data

The experimental data include the basic geographic information (GIS) datasets and spatial distribution map of crop sown area. The basic GIS dataset includes administrative boundary of Mengcheng County with a scale of 1:250000 (shapefile). The spatial distribution map of crop sown area is based on winter wheat sown area of 2009 in Mengcheng County (derived from ALOS image with a spatial resolution of 10m). In addition, cropland data with spatial resolution of 250mare also collected. Figure 1.1 shows the administrative boundary of the study area and spatial distribution of winter wheat sown area.

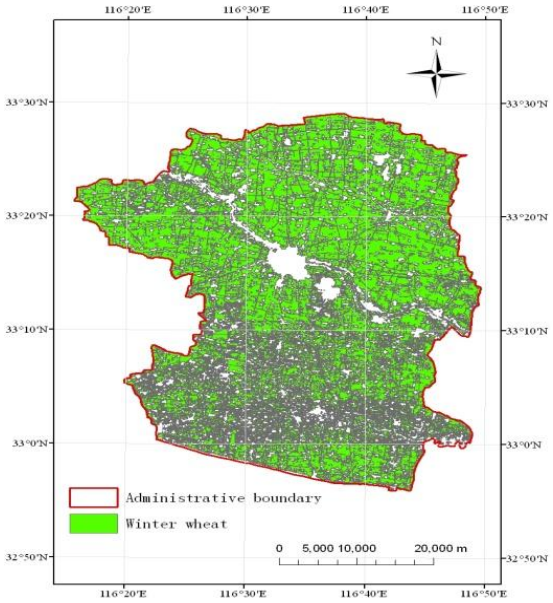

Figure 1.1 The administrative boundary of the study area and spatial distribution of winter wheat sown area

1.2.3 Sampling scheme

(1). **Basic sample unit**

A square grid is drawn to generate basic sampling unit and the size of single grid is 5 arc minute×5 arc minute, in order to improve the convenience for sample surveys . Then, the study area is gridded to construct a sampling frame. The winter wheat sown area in each square grid is calculated as the population unit value using the ArcGIS software.

(2). **Spatial sampling methods**

Simple random sampling: Sample size is calculated according to the formulas (1.1) ~ (1.4).

$$n_0 = (\frac{t}{r})^2 \frac{S^2}{\overline{Y}^2} \quad (1.1)$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (1.2)$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad (1.3)$$

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \quad (1.4)$$

Where n_0 is the initial sample size; n is modified sample size, when $n_0/N > 0.05$, n_0 is modified according to (1.2); t is the sampling probability degrees, when the confidence level is 95%, t is

equal to 1.96; *r* is the relative error, 5% is designed as *r* in the study; $\overline{Y}$ is the population mean; S^2 is the population variance; *N* is the population size; Y_i is the *i*-th population unit observations.

After sample size is determined, the samples are drawn using the pseudo-random number method. First, all population units in the sampling frame are encoded with a number; then pseudo-random number (PRN) equivalent to the sample size is generated by SPSS software; Finally when the code of a population unit agrees with the random number generated by SPSS, then the population unit is drawn as a sample.

Simple estimator is used to extrapolate population value and estimate error. Population total is calculated according to (1.5) and (1.6). The variance of population total estimator is estimated according to (1.7) and (1.8).

$$\hat{Y} = N\bar{y} \tag{1.5}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.6}$$

$$v(\hat{Y}) = \frac{N^2(1-f)}{n}s^2 \tag{1.7}$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1.8}$$

Where $\bar{y}$, S^2 is sample mean and sample variance, respectively; *n* is the initial sample size; *N* is the population size; $\hat{Y}$ is the estimated value of population total; $v(\hat{Y})$ is the unbiased variance of population total estimator; *f* is sampling fraction.

Spatial random sampling: The impact of spatial correlations between population units on sampling size and population extrapolation are taken into account in spatial random sampling. Sample size is calculated according to (1.9) ~ (1.12).

$$n_{\text{spatial}} = n_{\text{simple}}(1-r) \tag{1.9}$$

$$r = \frac{C(Z_i, Z_{i+h})}{\sigma^2 p} \tag{1.10}$$

$$C(Z_i, Z_{i+h}) = \frac{1}{N(h)}\sum_{i=1}^{N(h)}[Z(X_i, Y_i) - \overline{Z(X_i, Y_i)}][Z(X_{i+h}, Y_{i+h}) - \overline{Z(X_{i+h}, Y_{i+h})}] \tag{1.11}$$

$$\sigma_p 2 = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \overline{Z_i})^2 \tag{1.12}$$

Where $n_{spatial}$ is sample size of spatial random sampling; n_{simple} is sample size of simple random sampling; *r* is spatial correlation coefficient; $C(Z_i, Z_{i+h})$ is covariance between the population units in the sampling frame; *N (h)* is the numbers of population units with a distance of *h*; Z (X_i, Y_i) is winter wheat sown area in *i*-th population unit; $\sigma^2{}_p$ is dispersion variance of all population units in sampling frame.

Samples selection is same with that of simple random sampling method. Population total is extrapolated according to (1.13) ~ (1.14), the variance of population total estimator is estimated according to (1.15).

$$\hat{Z} = N\overline{Z}_{spatial} \tag{1.13}$$

$$\overline{Z}_{spatial} = \frac{1}{n}\sum_{i=1}^{n} Z(X_i, Y_i) \tag{1.14}$$

$$v(\hat{Z}) = N^2 \frac{1-f}{n} s^2(1-r) = N^2 \frac{1-f}{n}\left[s^2 - COV(z(x,y))\right] \tag{1.15}$$

Where $\overline{Z}_{spatial}$ is population mean; Z (X_i,Y_i) is winter wheat sown area in *i*-th population unit; s^2 is dispersion variance of sampled units; $\mathrm{cov}(z(x,y))$ is covariance between the sampled units.

Classical systematic sampling: sample size is estimated according to that of spatial random sampling, due to the complexity of calculating sample size in classical systematic sampling.

ID numbers of all population units are sorted from small to large, the first sample is drawn randomly from sampling frame, and then the rest of samples are drawn with a same sampling interval. The population unit that ID number is 2 is selected as the first sample, and sampling interval is equal that population size is divided by sample size.

Population extrapolation of classical systematic sampling is the same as that of simple random sampling, because the samples drawn by classical systematic sampling are random. The variance of population total estimator is estimated according to (1.16).

$$v(\hat{Y}) = N^2 \frac{1-f}{n} \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 \tag{1.16}$$

Where $v(\hat{Y})$ is the unbiased variance of population total estimator; *f* is sampling fraction; Y_i is the *i*-th sample observation; Y_{i+1} is the *i*+1-th sample observation.

Spatial systematic sampling: samples are evenly distributed in sampling frame in the space, and the first sample is drawn randomly. The population unit that ID number is 2 is selected as the first sample, and sampling interval is one sampling basic unit in the research. Sample size is equal to

the amount of samples that can be evenly distributed in whole sampling frame. Population extrapolation and error estimation are the same as those of simple random sampling.

Stratified sampling: the percentage that winter wheat sown area accounting for a sampling basic unit area in the basic unit is selected as stratification mark, in order to improve the stratified efficiency. In addition, stratum numbers are formulated into 5 levels to investigate the impact of it on the efficiency of stratified sampling method. 5 stratum numbers is 2, 3, 4, 5 and 6 respectively. Total sample size is calculated according to (1.17) ~ (1.20). Total sample size is allocated into each stratum according to stratum weighting.

$$n_0 = \frac{\sum W_h S_h^{\ 2}}{V} \tag{1.17}$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \tag{1.18}$$

$$V = (\frac{\gamma \overline{Y}}{t})^2 \tag{1.19}$$

$$\overline{Y} = \overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h \tag{1.20}$$

Where n_0 is the initial sample size; v is the variance of population mean estimator; $\overline{y}_{st}$ is the samples mean in stratified sampling; L is strata number; $\overline{y}_h$ is samples mean in i-th stratum; W_h is weighting in h-th stratum; S_h^2 is the variance of population units in h-th stratum; N_h is population size in h-th stratum. Simple random sampling method is used to draw the samples in each stratum. Population extrapolation and error estimation are calculated according to (1.21)~(1.23).

$$\hat{Y} = N\overline{y}_{st} \tag{1.21}$$

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h \tag{1.22}$$

$$v(\hat{Y}) = N^2 v(\overline{y}_{st}) = \sum_{h=1}^{L} N_h^{\ 2} \frac{1 - f_h}{n_h} s_h^{\ 2} \tag{1.23}$$

Where f_h is the sampling fraction in h-th stratum; $v(\overline{y}_{st})$ is unbiased estimate of samples mean variance; $v(\hat{Y})$ is the unbiased variance of population total estimator; n_h is sample size in h-th stratum; S_h^2 is samples variance in h-th stratum.

Relative error (r) and coefficient of variation of population total estimator (CV) are selected as indices to evaluate the results of population extrapolation and error estimation from 5 spatial

sampling methods. Relative error and CV are calculated according to (1.24) and (1.25), respectively.

$$r = \frac{\left|\hat{Y} - Y\right|}{Y} \times 100\% \tag{1.24}$$

$$CV(\hat{Y}) = \frac{\sqrt{v(\hat{Y})}}{\hat{Y}} \times 100\% \tag{1.25}$$

Where *Y* is the truth value of population total, *Y* is obtained with ArcGIS software summarizing all winter wheat sown area in sampling frame; $CV(\hat{Y})$ is coefficient of variation of population total estimator.

1.3 Results and Analysis

1.3.1 Comparison of sampling efficiencies among 5 spatial sampling methods

Table 1.1 summarizes the results of population extrapolation and error estimation from 5 sampling methods, in order to compare their efficiency. Regarding the relative error and coefficient of variation it is found that stratified sampling obtained the lowest values, before simply random sampling, while the spatial systematic sampling approach has the highest value With respect to sample size, although that of spatial systematic sampling has the smallest one, , the relative error and CV from spatial systematic sampling are very large (16.64% for relative error, 13.66% for CV is). On the other hand, sample size of stratified sampling is 19, just a little higher than that of spatial systematic sampling. Therefore by comprehensively evaluating sampling error and sampling size, the efficiency of stratified sampling has the highest score in 5 sampling methods. Figure1.2 to Figure1.7 show spatial distributions of Sampled units drawn by simple random sampling, spatial random sampling, classical systematic sampling, spatial systematic sampling and stratified sampling, respectively.

Table1.1 Results of population extrapolation and error estimation from 5 sampling methods

Sampling methods	Population size -	Sample size -	Sampling fraction *(%)*	Relative error *(%)*	CV *(%)*
Simple random sampling	45	41	91.1	2.35	2.65
Spatial random sampling	45	25	55.6	6.53	4.54

Sampling methods	Population size	Sample size	Sampling fraction	Relative error	CV
Classical systematic sampling	45	22	48.9	6.83	6.83
Spatial systematic sampling	45	12	26.7	16.64	13.66
Stratified sampling	45	19	42.2	1.57	2.42

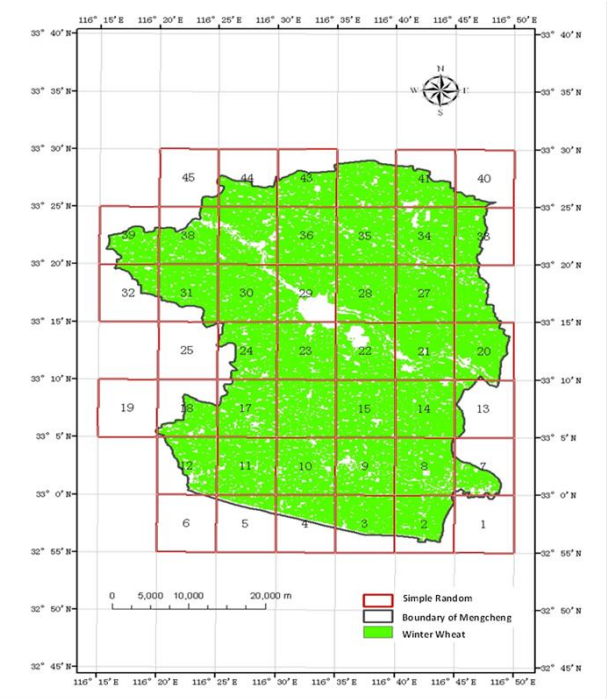

Figure 1.2 Samples spatial distributions of simple random sampling

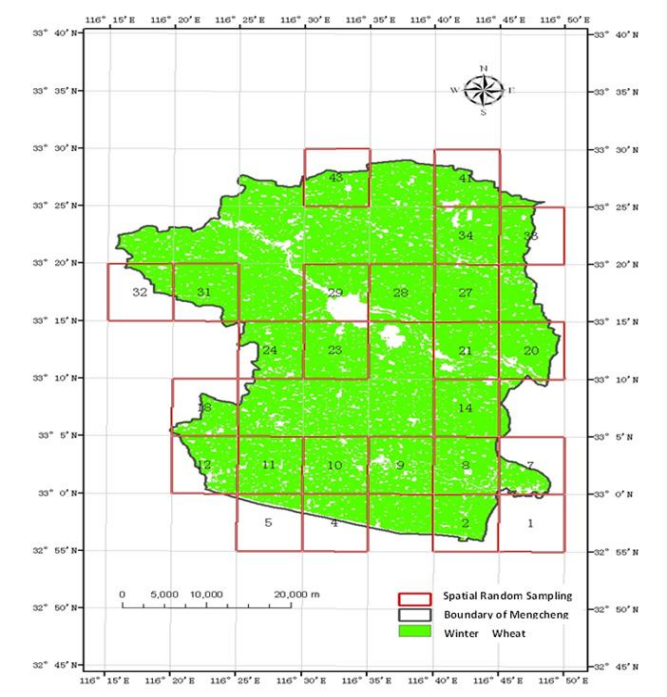

Figure 1.3 Sample spatial distributions of spatial random sampling

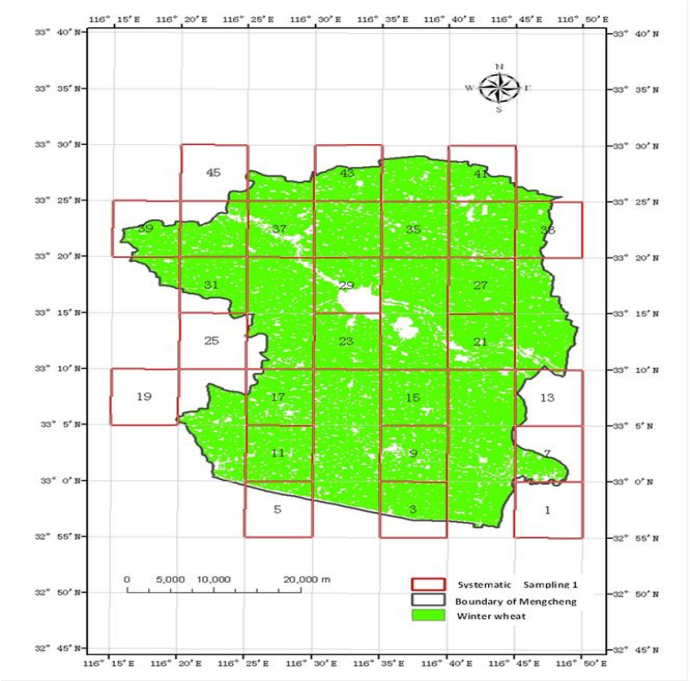

Figure 1.4 Sample spatial distributions of systematic sampling(sorted by the ID of population units)

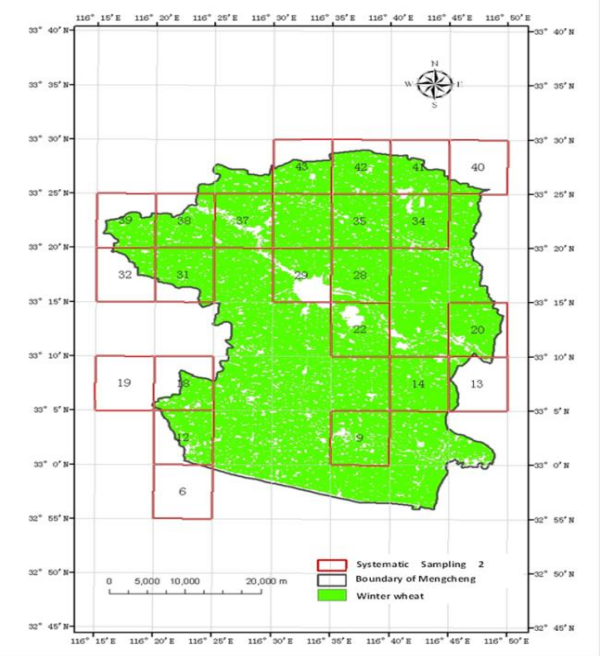

Figure 1.5 Samples spatial distributions of systematic sampling(sorted by winter wheat area in population units)

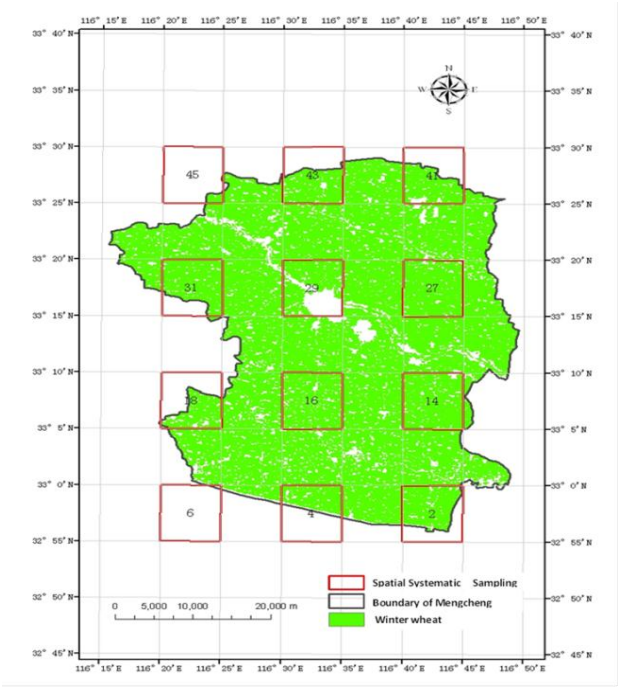

Figure 1.6 Samples spatial distributions of spatial systematic sampling

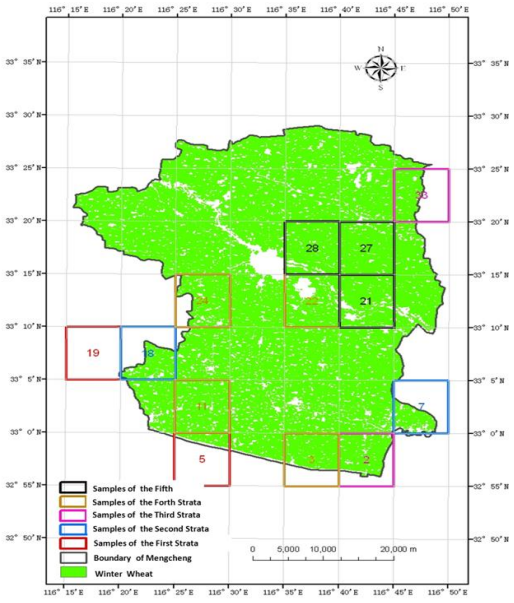

Figure 1.7 Samples spatial distributions of stratified sampling

1.3.2 Impact of the numbers of stratum on stratified sampling efficiency

Table 1.2 summarizes the results of population extrapolation and error estimation from stratified sampling by making 4 strata, in order to investigate the impact of stratum number on the efficiency of the stratified sampling method. It is found that sample size, relative error and CV decrease with increasing stratum numbers (Table 1.2), when stratum numbers are smaller than 5. On the contrast, when stratum numbers are greater than 5, although sample size still maintains a decreasing trend, relative error and CV increase. When the stratum number is 5, stratified sampling efficiency reaches the maximum.

Table1.2 Results of population extrapolation and error estimation from stratified sampling under 4 stratum numbers levels

Stratum numbers	Stratified sampling			
	Stratum interval (%)	Sample size	Relative error(%)	CV(%)
2	41.04	33	3.03	3.50
3	27.36	32	2.68	2.65
4	20.52	19	1.57	2.42
5	16.42	13	1.63	2.50
6	13.68	11	2.89	4.27

Note: Stratum interval is represented by the percentage that winter wheat sown area accounting for a sampling basic unit area in the basic unit

1.4 Conclusion

The experiment on optimizing spatial sampling schemes for estimating winter wheat sown acreage was conducted to improve the current crop acreage sampling survey systems, combining Remote Sensing and Geographic Information Systems with the traditional sampling methods as well as Geostatistics theory. Mengcheng County was selected as study area. 5 spatial sampling methods (simple random sampling, spatial random sampling, classical systematic sampling, spatial systematic sampling and stratified sampling) were used in the experiment. The experimental results demonstrate :

(1). The efficiency of stratified sampling method reaches the highest score among 5 spatial sampling methods(relative error is 1.57%; CV is 2.42%; sample size is 19), when relative error, CV and sample size are compared.

(2). when stratum number is smaller than 5, sample size, relative error and CV decrease with an increasing stratum number; otherwise, although sample size decreases with an increasing stratum number, relative error and CV increase with an increasing stratum number, When the stratum number is equal to 5, stratified sampling efficiency reaches its maximum.

2 A TWO-STAGE SAMPLING METHOD FOR WINTER WHEAT AREA ESTIMATION

2.1 Materials and methods

2.1.1 Introduction

Two-stage sampling method is formulated to estimate winter wheat area, combining remote sensing (RS), Geographic Information Systems (GIS) with the traditional sampling methods. It consist of, collecting basic data for spatial sampling design; defining basic sampling units (including the shape and size of sampling units); designing and optimizing two-stage sampling approach;; and finally, extrapolating population value and estimating sampling error. Figure 2.1 shows the technical approach of the two-stage sampling method.

2.1.2 The study area and the auxiliary data

The study area and the auxiliary data are available from the investigation described in the chapter one.

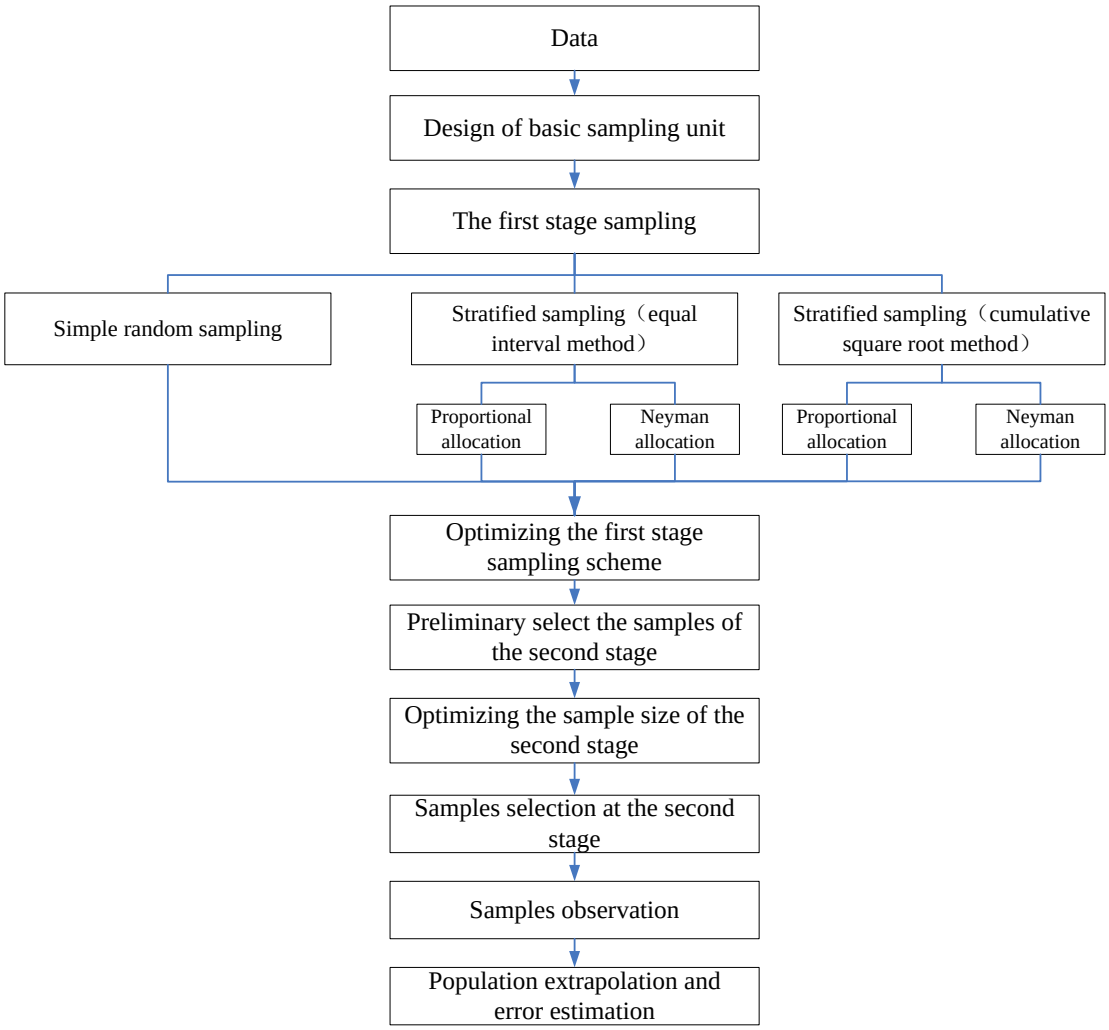

Figure 2.1The technical approach of two-stage sampling method

(1). Basic sampling unit

The square grid is selected as the shape of primary sampling units (PSU), and the size of PSU is defined as 8500m×8500m, referring to the size of 5 arc minute × 5 arc minute used in the previous study, Meanwhile, taking into consideration that the number of second sampling unit (SSU) included in one PSU should be integer. The shape of SSU still is square grid, and the size of SSU is 500m×500m to facilitate field investigation of the samples. Figure 2.2 shows that the spatial distribution of PSU and SSU in the study area.

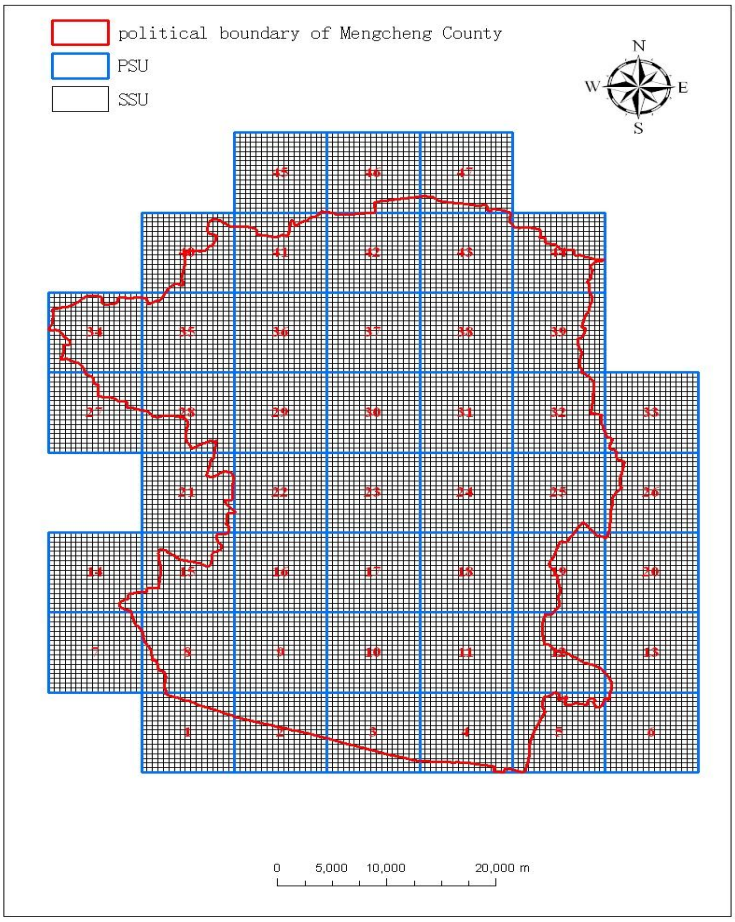

Figure 2.2 Spatial distributions of PSU and SSU in the study area

(2). Spatial sampling method

1) The first stage sampling strategy

The sampling frame of the first sampling stage is constructed based on spatial distribution data of winter wheat area in 2009. Simple random sampling method (SRS), stratified sampling method (with equal intervals between stratums, EIS), stratified sampling method (cumulative square root method is used in stratification, CSR) are selected to optimize the sampling scheme at the first stage. The stratum number is 4, comprehensively taking into account the variance of sampling units and following the rule that the sample size is not less than 2 in each stratum. The ratio of winter wheat area in a sampling unit (that is a square grid in the experiment) accounting for the area of the sampling unit (RWG) is selected as stratification symbol. Since the range of RWG varies from 0.00% to 82.01%, therefore, the intervals of each stratum are 20.50% in the stratified sampling method with equal intervals between stratums. The stratum numbers of stratified sampling method (CSR) are the same with that of stratified sampling method (EIS), and the intervals of each stratum is formulated using the method of cumulative square root method. The stratification processes are as follows: firstly, 5% of RWG is served as the statistical interval, the frequency distribution of all basic sampling units are summarized in the sampling frame; secondly,

the frequency values are conducted square root (SQRT) in each statistical interval; thirdly, all of SQRT of frequency values are summed up; finally, the sum of all SQRT is divided by 4, and the result is stratification interval. 2 patterns are used in the allocation of sample size into each stratum, one is the proportional allocation, that is the sample size is allocated into each stratum according to the stratum weight; the other is Neyman allocation, and the sample size of each stratum is calculated according to equation (2.1).

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \quad (2.1)$$

Where, n_h is the sample size in the *h* stratum; *n* is sample size of stratified sampling method； W_h is stratum weight in the *h* stratum; N_h is population size in the *h* stratum; S_h is standard deviation of population units in the *h* stratum; *L* is the numbers of strata.

3 kinds of sampling methods mentioned ahead are used to draw samples, extrapolate population value and estimate error. Relative error, coefficient of variation and sample size are selected as evaluated indices to optimizing the spatial sampling scheme of the first stage from 3 kinds of sampling methods. Furthermore, the samples of the first stage are drawn using the optimized sampling scheme.

2) The second stage sampling strategy

The sampled units of the second stage are drawn from the selected first stage sampling units. SRS method is used at the second sampling stage to ensure the stability of population extrapolation and simplify the process of sampling errors estimation. Pre-sampling method is used to optimizing the sample size of the second stage. The optimized processes are as follows: firstly, SRS method is used to pre-draw some samples of the second stage, the number of pre-drawing samples is determinate as 4, referring to the literature[3]; secondly, the variance between PSU (S_1^2) and variance of PSU (S_2^2) are estimated based on the pre-drawn samples value; thirdly, the ratio of survey cost between one PSU and SSU is evaluated; finally, the optimal sample size is calculated according to the equation (2.2) at the second stage. Figure2.3 shows the spatial distribution of the selected PSU and pre-drawn SSU in the study area.

It is assumed that *n* samples are drawn from *N* PSU at the first stage, and *m* samples are drawn from each selected PSU at the second stage. Each PSU includes *M* SSUs. The signs used to calculate the optimal *n* are as follows:

Y_{ij} is the observation value of the *j*-th SSU in the *i*-th PSU of population units；
y_{ij} is the observation value of the *j*-th SSU in the *i*-th PSU of sample units；

$f_1 = \frac{n}{N}$ is sampling fraction at the first sampling stage；

$f_2 = \frac{m}{M}$ is sampling fraction at the second sampling stage；

$Y_i = \sum_{j=1}^{M} Y_{ij}$ is the sum of observation values of all SSU in i-th PSU in population units；
$y_i = \sum_{j=1}^{M} y_{ij}$ is the sum of observation values of all SSU in i-th PSU in sample units；

$\bar{Y}_i = \frac{1}{M} Y_i$ is the mean of observation values of all SSU in i-th PSU in population units;

$\bar{y}_i = \frac{1}{M} y_i$ is the mean of observation values of all SSU in i-th PSU in sample units;

$\bar{\bar{Y}} = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i$ is the mean of observation values of all SSU in population units;

$\bar{\bar{y}} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i$ is the mean of observation values of all SSU in sample units;

$S_1^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{\bar{Y}})^2$ is the variance between PSUs in population units;

$s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2$ is the variance between PSUs in sample units;

$S_2^2 = \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(Y_{ij} - \bar{Y}_i\right)^2$ is the variance of PSUs in population units;

$s_2^2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(y_{ij} - \bar{y}_i\right)^2$ is the variance of PSUs in sample units.

The optimal m is calculated according to Cauchy-Schawarz inequality.

$$m_{opt} = \frac{S_2}{S_u} \sqrt{\frac{c_1}{c_2}} \tag{2.2}$$

$$S_u^2 = S_1^2 - \frac{S_2^2}{M} \tag{2.3}$$

$$\hat{S}_1^2 = s_1^2 - \frac{1-f_2}{m_0} s_2^2 \tag{2.4}$$

$$\hat{S}_2^2 = s_2^2 \tag{2.5}$$

Where, c_1 and c_2 is the survey cost of the first sampling stage and the second sampling stage, respectively, $\frac{c_1}{c_2}$ is 6 in the study, referring to the average levels of survey cost from one PSU (surveyed with remote sensing) and SSU(ground survey).

Make $m' = [m_{opt}]$, i e., m' is the integer part of m_{opt}, then

If $m_{opt}^2 \geq m'(m'+1)$, then $m'=m'+1$

If $m_{opt}^2 \leq m'(m'+1)$, then $m_{opt}=m'$

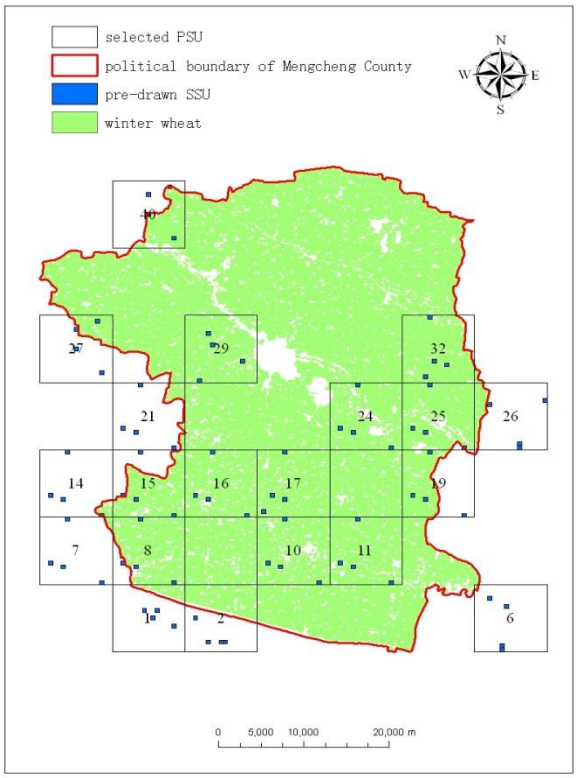

Figure 2.3 Spatial distributions of the selected PSU and pre-drawn SSU in the study area

2.1.3 Samples observations

Samples observations are obtained with ArcGIS software calculating winter wheat sown area in the sampled units, when spatial distribution data of winter wheat (retrieved by ALOS image) is overlapped with sampled units in the sampling frame.

2.1.4 Population extrapolation

The estimate of population total is calculated according to equation (2.6), and the variance of population total estimator is defined as equation (2.7).

$$\hat{Y}_{st} = (\sum_{h=1}^{L} N_h M_h)\bar{\bar{y}}_{st} \tag{2.6}$$

$$v(\hat{Y}_{st}) = (\sum_{h=1}^{L} N_h M_h)^2 v(\bar{\bar{y}}_{st}) \tag{2.7}$$

$$\bar{\bar{y}}_{st} = \frac{\sum_{h=1}^{L} N_h M_h \bar{\bar{y}}_h}{\sum_{h=1}^{L} N_h M_h} = \sum_{h=1}^{L} W_h \bar{\bar{y}}_h \tag{2.8}$$

$$W_h = \frac{N_h M_h}{\sum_{h=1}^{L} N_h M_h} \tag{2.9}$$

$$\bar{\bar{y}}_h = \frac{\sum_{i=1}^{n_h}\sum_{j=1}^{m_h} y_{hij}}{n_h m_h} \tag{2.10}$$

$$v(\bar{\bar{y}}_{st}) = \sum_{h=1}^{L} W_h^{\,2} \left(\frac{1-f_{1h}}{n_h} s_{1h}^{\,2} + \frac{f_{1h}(1-f_{2h}) s_{2h}^{\,2}}{n_h m_h}\right) \tag{2.11}$$

Where,

L are the numbers of strata;

h is the stratum number;

W_h is the weight of the *h*-th stratum;

N_h is the numbers of PSU in the *h*-th stratum;

M_h is the numbers of SSU included in one PSU in the *h*-th stratum;

y_{hij} is observation value of the *j*-th selected SSU in *i*-th PSU in the *h*-th stratum;

f_{1h} is sampling fraction of the *h*-th stratum at the first stage, $f_{1h} = \frac{n_h}{N_h}$;

f_{2h} is sampling fraction of the *h*-th stratum at the second stage, $f_{2h} = \frac{m_h}{M_h}$;

$s_{1h}^{\,2} = \frac{1}{n_h - 1}\sum_{i=1}^{n_h} (\bar{y}_{hi} - \bar{\bar{y}}_h)^2$ is the variance between PSUs in sample units in the *h*-th stratum;

$s_{2h}^{\,2} = \frac{1}{n_h(m_h - 1)}\sum_{i=1}^{n_h}\sum_{j=1}^{m_h} (y_{hij} - \bar{y}_{hi})^2$ is the variance of PSUs in sample units in the *h*-th stratum.

Relative error (*r*) and coefficient of variation of population total estimator (CV) are selected as indices to evaluate the results of population extrapolation and error estimation from 5 spatial sampling methods. Relative error and CV are calculated according to (2.12) and (2.13), respectively.

$$r = \frac{|\hat{Y} - Y|}{Y} \times 100\% \quad (2.12)$$

$$CV(\hat{Y}) = \frac{\sqrt{v(\hat{Y})}}{\hat{Y}} \times 100\% \quad (2.13)$$

Where Y is the truth value of population total, Y is obtained with ArcGIS software summarizing all winter wheat sown area in sampling frame; $CV(\hat{Y})$ is coefficient of variation of population total estimator.

2.2 Results and analysis

2.2.1 Efficiency comparison for the first sampling stage

Table.2.1 shows that relative errors, coefficient of variation and sampling fraction using 3 different spatial sampling methods to extrapolate the population of winter wheat area . It is found that the relative errors and CV are all small (less than 5%), using 3 sampling methods to extrapolate population at the first stage., However the sample size of SRS (sample size is 44, and sampling fraction is 93.62%) is greater, comparing to sample size of 18 and 20 for EIS and CSR stratified sampling methods () respectively. Therefore, we consider that the efficiencies of EIS stratified sampling method and CSR stratified sampling method are higher. Furthermore, it is also found that the relative error and CV using stratified CSR sampling method () for which Neyman allocation is used (relative error is less than 1%) are smaller than those from EIS stratified sampling method () for which a proportional allocation is used,. By comprehensively comparing the relative errors and stability of population extrapolation, the efficiency of CSR stratified sampling method using Neyman allocation reaches the maximum.

Table 2.1 Results of population extrapolation and error estimation using 3 kinds of spatial sampling methods

	Stratified sampling(EIS)			Stratified sampling(CSR)			SRS
	PV	PA	NA	PV	PA	NA	Sample size
Stratum number	N_h	n_h	n_h	N_h	n_h	n_h	n
1	14	5	5	15	6	8	—
2	4	2	2	9	4	5	—
3	12	5	5	14	6	5	—
4	17	6	6	9	4	2	—
sum	47	18	18	47	20	20	44
f(%)	—	38.30		—	42.55		93.62
r(%)	—	2.34	—	—	1.51	0.52	0.89
CV(%)	—	3.15	3.86	—	2.39	2.30	2.16

Note: PV is population value; PA is proportional allocation; NA is Neyman allocation; N_h is the numbers of population units in the *h*-th stratum; n_h is the numbers of samples units in the *h*-th stratum; *n* is sample size; *f* is sampling fraction; *r* is relative error; CV is coefficient of variation.

2.2.2 Population extrapolation and error estimation of two-stage sampling method

Table2.2 shows the results of population extrapolation and error estimation using two-stage sampling method to estimate winter wheat area. It is found that the optimal sample size in the second stage is 3 in each selected PSU; the relative error and CV of population extrapolation are 2.10% and 3.18% (the true value of total population is obtained based on spatial distribution map of winter wheat in 2009) respectively, both less than 5%. Furthermore, the sampling fraction is only 0.44%. This shows that it is suitable for sampling survey business to estimate winter wheat area using two-stage sampling scheme formulated in the experiment. Figure 8.4 shows the spatial distribution of samples drawn by two-stage sampling method in the study area. It is found that some selected SSUs are located outside of the political boundary of Mengcheng County, due to that some PSUs (PSUs with a lower RWG) are allotted at near the boundary of the study area, therefore, this part of samples units do not need to be surveyed, which reduce the survey cost in fact. The numbers of samples that located outside the political boundary are 26, namely, the numbers of samples that need to be investigated are 36. Figure 8.5 shows the spatial distribution of samples that need to be investigated.

Table 2.2 Results of population extrapolation and error estimation using two-stage sampling method

Stratum number	W_h	M_h	m_h	*n*	*f*(%)	*r*(%)	*CV*(%)
1	0.3191	289	3	60	0.44	2.10	3.18
2	0.1915	289	3				
3	0.2979	289	3				
4	0.1915	289	3				

Note: W_h is the weight of the *h*-th stratum; M_h is the numbers of SSU included in each PSU in the *h*-th stratum; M_h is the numbers of SSU included in each selected PSU in the *h*-th stratum.

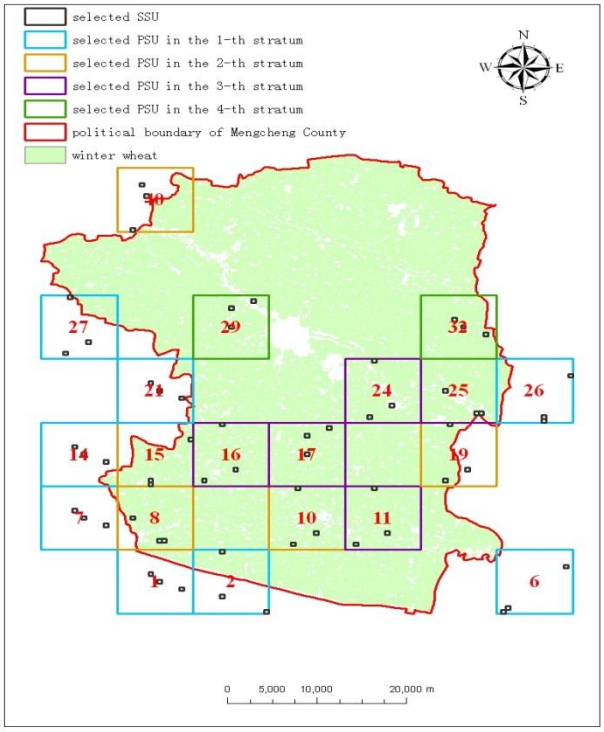

Figure 2.4 Spatial distributions of samples drawn by two-stage sampling method

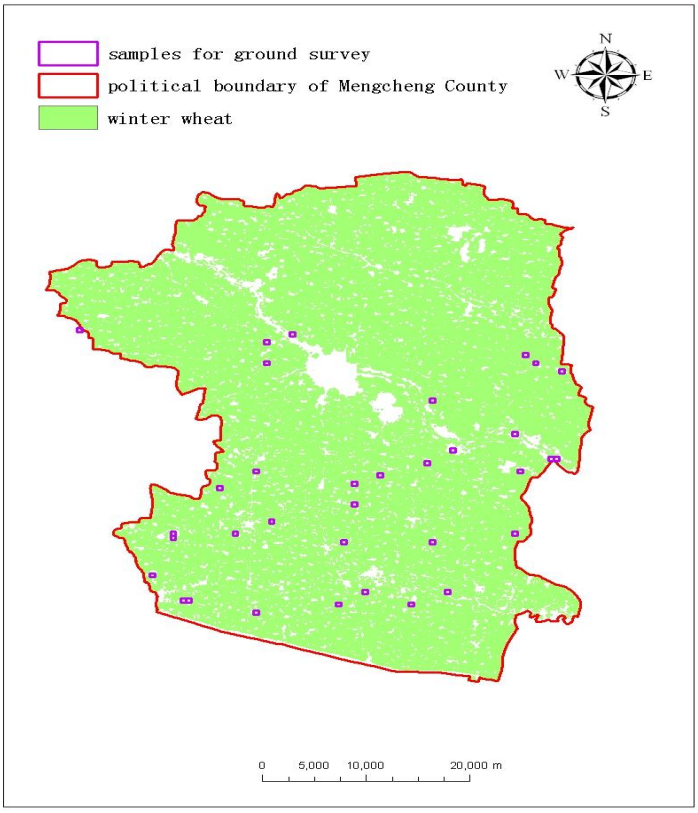

Figure 2.5 Spatial distributions of samples that need to be investigated

2.3 Conclusions

The experiments on two-stage sampling design for estimating winter wheat area is conducted to improve the current crop acreage sampling survey systems., 3 spatial sampling methods including SRS, EIS stratified sampling method and CSR stratified sampling method are used The experimental results demonstrate as follows:

(1) At the first sampling stage, the efficiency of CSR stratified sampling method with Neyman allocation reaches the highest among 3 sampling methods (relative error of population extrapolation 0.52%, CV 2.30% andsample size 20); and the EIS stratified sampling method using the proportional allocation is on second position.

(2) The optimal sample size of the second stage is 3 in each selected PSU; the relative error and CV of population extrapolation are 2.10% and 3.18% respectively, both less than 5%. Furthermore, the sampling fraction is only 0.44%. This shows that two-stage sampling method is appropriate for sampling survey in this winter wheat area estimation study.

3 COMPARISONS ON EXTRAPOLATION ACCURACY OF 3 ESTIMATORS TO ASSESS WINTER WHEAT AREA

3.1 Materials and methods

3.1.1 Introduction

The estimators are often used to extrapolate population values in the sampling survey for crop area estimation. 3 estimators (the simple estimator, the ratio estimator and the regression estimator) are used to estimate crop area The estimators are constructed using two independent variables: the area estimate derived from ground surveys and the area estimated derived from remote sensing image classification.

3.1.2 Experiment data

The auxiliary data were available from previous study. The ground survey data of winter wheat sown area are collected within 12 samples frames, and each sample frame is composed by 5-17 parcels with physical boundary. Table 3.1 shows winter wheat area obtained from the ground survey and from the classification of remote sensing images covering these 12 samples. Figure 3.1 shows the spatial distribution of winter wheat derived from ground survey and image classification.

Table 3.1 Samples data from ground survey and remote sensing

No.	Winter wheat area from remote sensing(m^2)	Winter wheat area from ground survey(m^2)
1	362800.00	548470.38
2	265000.00	293638.72
3	252600.00	365182.48
4	359600.00	400783.99
5	385600.00	380576.00
6	261900.00	299451.80
7	407000.00	391385.75
8	448300.00	471567.61
9	270400.00	265859.51
10	262400.00	224652.94
11	239500.00	243794.42
12	240000.00	286387.12

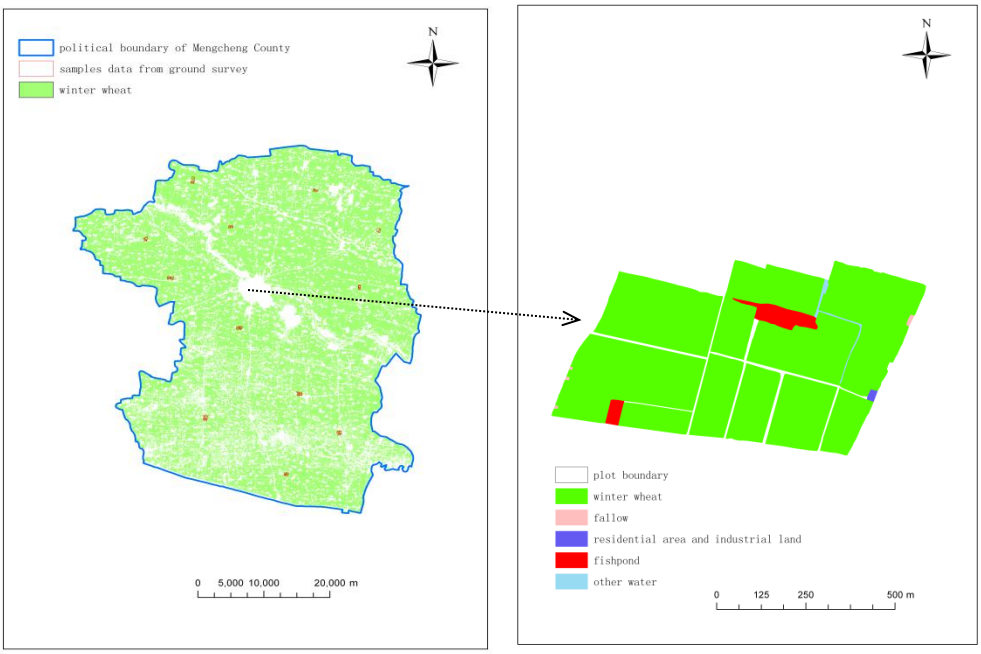

Figure 3.1 Spatial distribution of samples from ground survey in Mengcheng County

3.1.3 Construction of estimators

1) Simple estimator

It is assumed that 12 samples are drawn by simple random sampling method from sampling frame, then, the mean values of 12 samples are considered as the simple estimator of population mean. The sample mean is calculated according to equation (3.1)

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.1}$$

Then the simple estimator of population total is defined as equation (3.2)

$$\hat{Y} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i \tag{3.2}$$

The unbiased estimate of the variance of simple estimator of population total is defined as (3.3)

$$v(\hat{Y}) = N^2 v(\bar{y}) = \frac{N^2(1-f)}{n} s^2 \tag{3.3}$$

where $\bar{y}$ is sample mean; n is sample size; y_i is the winter wheat sown area in the *i-th* sample; $\hat{Y}$ is the simple estimator of population total; N is population size; $v(\hat{Y})$ is the variance of simple estimator of population total; f is sampling fraction; s^2 is the variance of samples.

2) Ratio estimator

The population of winter wheat area from ground survey is considered as a dependent value Y, while the population of winter wheat area from remote sensing as an independent value X. As the independent value X is determinate beforehand the ratio estimator can be constructed according to the equation (3.4)and(3.5)

$$\bar{y}_R = \hat{R}\bar{X} = \frac{y}{x}\bar{X} \tag{3.4}$$

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}}X = \frac{\sum y_i}{\sum x_i}X \tag{3.5}$$

The estimate of the variance of ratio estimator of population total is defined as (3.6)

$$v(\hat{Y}_R) \approx \frac{N^2(1-f)}{n}(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}) \tag{3.6}$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.7}$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.8}$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \tag{3.9}$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum y_i}{\sum x_i} \quad (3.10)$$

Where $\hat{Y}_R$ is the ratio estimator of population total of winter wheat area from ground survey in the study area; X is population total of winter wheat area from remote sensing data; y_i is the i-th sample observation from ground survey; x_i is the i-th sample observation from remote sensing data, x_i is measured through overlapping the spatial distribution data of winter wheat in 2009 and the boundary of samples from ground survey; $\bar{y}$ is the mean of samples from ground survey; $\bar{\mathbf{y}}$ is samples mean from remote sensing data; s_y^2 is samples variance from ground survey; s_x^2 is samples variance from remote sensing data; s_{xy} is the covariance of samples observations from ground survey and remote sensing data; $\hat{R}$ is the estimator of population ratio.

3) Regression estimator

In order to improve extrapolation accuracy of population values, the regression estimator is constructed based on the samples observations from ground survey and remote sensing data. When the simple random sampling method is used, the regression estimator of population mean and population total are defined as equation (3.11) and equation (3.12), respectively.

$$\bar{y}_{lr} = \bar{y} - b(\bar{x} - \bar{X}) \quad (3.11)$$

$$\hat{Y}_{lr} = N\bar{y}_{lr} \quad (3.12)$$

$$b = \frac{s_{yx}}{s_x^2} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (3.13)$$

Where $\bar{y}_{lr}$ is the regression estimator of population mean of winter wheat area from ground survey in the study area; b is regression coefficient based on samples data; $\bar{X}$ is population mean of winter wheat area from remote sensing data; $\hat{Y}_{lr}$ is the regression estimator of population total of winter wheat area from ground survey; s_{xy} and s_x^2 is defined as the same with those of ratio estimator.

The estimate of the variance of regression estimator of population total is defined as (3.13)

$$v(\hat{Y}_{lr}) = N^2 v(\bar{y}_{lr}) \quad (3.13)$$

$$v(\bar{y}_{lr}) = \frac{1-f}{n(n-2)}\sum_{i=1}^{n}[(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \quad (3.14)$$

Where $v(\hat{Y}_{lr})$ is the variance of regression estimator of population total of winter wheat area from ground survey in the study area; $v(\bar{y}_{lr})$ is the variance of regression estimator of population mean of winter wheat area from ground survey; f is the sampling fraction.

4) Relative error and coefficient of variation

Relative error (*r*) and coefficient of variation of population total estimator (CV) are selected as indices to evaluate the results of population extrapolation and error estimation based on 3 types of estimators. Relative error and CV are calculated according to (3.15) and (3.16), respectively.

$$r = \frac{\left|\hat{Y} - Y\right|}{Y} \times 100\% \quad (3.15)$$

$$CV(\hat{Y}) = \frac{\sqrt{v(\hat{Y})}}{\hat{Y}} \times 100\% \quad (3.16)$$

Where *Y* is the truth value of population total, *Y* is obtained with ArcGIS software summarizing all winter wheat sown area in sampling frame; $CV(\hat{Y})$ is coefficient of variation of population total estimator.

3.2 Results and analysis

3.2.1 Comparison of extrapolation using 3 estimators

The efficiency for extrapolating population of winter wheat area using 3 different estimators are shown in Table 3.2. It is found that the relative errors of population extrapolation using ratio estimator and regression estimator (relative error of 11.01% and 13.11%, respectively) are much smaller than that obtained by using simple estimator (relative error of 34.91%). Furthermore, the coefficient of variation using ratio estimator and regression estimator are 6.20% and 8.04%, respectively comparing that of simple estimator (9.61%), showing high accuracies of first two estimator applications.. However, the relative errors of population extrapolation remain higher (more than 10%) in this experiment, due to the fact that the sample size used to extrapolate population of winter wheat area is very little (sampling fraction is only 0.21%) in the study area.

Table 3.2 Results of population extrapolation of winter wheat area using 3 kinds of estimators in the study area

Estimator	Sample size(%)	Estimates of population total(m^2)	Standard error	Relative error (%)	Coefficient of variation(%)	*b*
Simple estimator	0.21	2010436196	161944955	34.91	9.61	
Ratio estimator	0.21	1654207641	102556176	11.01	6.20	
Regression estimator	0.21	1685454239	161694407	13.11	8.04	1.01

Note: Total area of the study area (that is the area included in political boundary of Mengcheng County) is divided by

the area of one sampling basic unit, and the result just is population size. The area of sampling basic unit is the average of 12 samples area; the truth value of population total is obtained based on spatial distribution data of winter wheat in 2009 in the study area.

3.2.2 Correlation analysis of samples data based on ground survey and remote sensing image classification

Figure3.2 shows the scatter plot of winter wheat area of 12 samples derived between ground survey and image classification. A linear relationship with a Significance factor F=15.207 which is higher than F0.01(1, 10)=10.04, indicating a significant. In addition, from the regression equation y=1.1062x (R2=0.598), a direct proportional relationship is found between winter wheat areas estimates from ground survey and those of image classification, demonstrating the higher accuracy of the ratio estimator for population extrapolation .

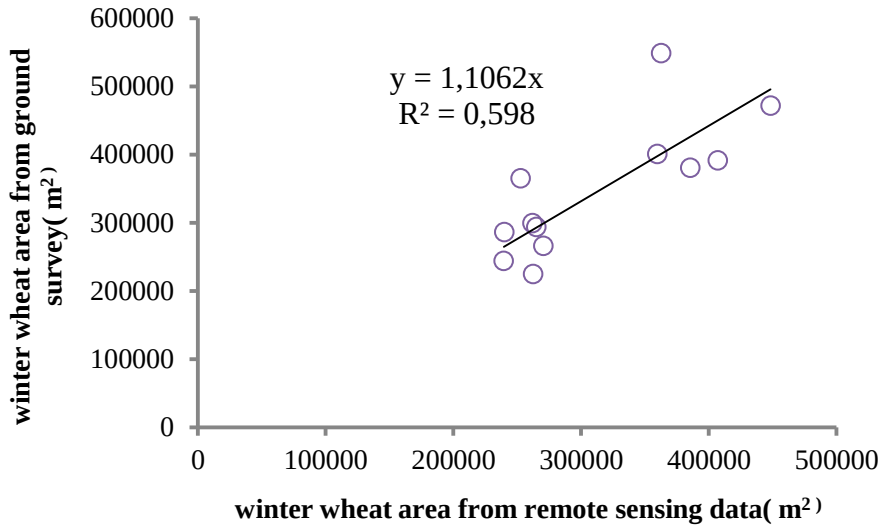

Figure 3.2 The scatter plot of winter wheat area from ground survey and remote sensing data

3.3 Conclusions

The experiments on the extrapolation efficiency of the estimators are conducted in Mengcheng County:

(1) The extrapolation efficiency using the ratio estimator is the highest among 3 estimators; while the efficiency using simple estimator is the poorest, revealed by the relative errors and coefficients of variation of in the exercise of population extrapolation.

(2) A direct proportional relationship between winter wheat area estimates from ground survey and those from image classification was significant, which indirectly proved the higher accuracy obtained by using ratio estimator.