Taylor & Francis
Taylor & Francis Group

International Journal of Remote Sensing

ISSN: 0143-1161 (Print) 1366-5901 (Online) Journal homepage: https://www.tandfonline.com/loi/tres20

An optimized two-stage spatial sampling scheme for winter wheat acreage estimation using remotely sensed imagery

Di WANG, Zhao-Liang LI, Qing-Bo ZHOU, Peng YANG & Zhong-Xin CHEN

To cite this article: Di WANG, Zhao-Liang LI, Qing-Bo ZHOU, Peng YANG & Zhong-Xin CHEN (2019) An optimized two-stage spatial sampling scheme for winter wheat acreage estimation using remotely sensed imagery, International Journal of Remote Sensing, 40:5-6, 2014-2033, DOI: 10.1080/01431161.2018.1516321

To link to this article: https://doi.org/10.1080/01431161.2018.1516321

Published online: 10 Sep 2018.

INTERNATIONAL JOURNAL OF REMOTE SENSING
2019, VOL. 40, NOS. 5–6, 2014–2033
https://doi.org/10.1080/01431161.2018.1516321

An optimized two-stage spatial sampling scheme for winter wheat acreage estimation using remotely sensed imagery

Di Wang, Zhao-Liang Li, Qing-Bo Zhou, Peng Yang and Zhong-Xin Chen

Key Laboratory of Agricultural Remote Sensing, Ministry of Agriculture and Rural Affairs, Beijing, China; Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing, China

ABSTRACT

Timely and reliable information on crop acreage is essential for formulating grain production policies and ensuring national food security. The combination of available satellite-based remotely sensed images and traditional sampling methods offers the possibility of improved crop acreage estimation at a regional scale. Due to the administrative convenience, reduced survey cost and workload, two-stage sampling has widely been used for crop acreage survey at the large-scale regions. However, compared with single-stage sampling, the two-stage sampling can introduce larger estimation errors, since it has multiple sampling stages. This study's aim is to optimize the two-stage sampling scheme using satellite-based remotely sensed imagery to improve the accuracy of crop acreage estimation. Taking Mengcheng County, Anhui Province, China, as the study area, this study explored the influence of stratum boundary and sample selection method on the sampling efficiency at the first sampling stage, analysed the impact of sample size on population extrapolation accuracy and then optimized the sample size of the second sampling stage using crop thematic map retrieved by ALOS (Advanced Land Observing Satellite) AVNIR (Advanced Visible light and Near Infrared Radiometer)-2 images in 2009. The results showed that the relative error (RE), coefficient of variation (CV), standard error (SE) of population extrapolation, and sampling fraction (*f*) using the cumulative square root of frequency (CSRF) method is the minimum among three methods for the stratum boundary determination at the first sampling stage, followed by the equal interval (EI) and equal sample size (ESS) method. Moreover, the RE, CV, and SE of population extrapolation using the ST sampling method is the minimum, compared with simple random (SI) and systematic (SY) sampling method. Therefore, the sampling scheme of the first stage can be optimized by CSRF method for stratum boundary determination and stratified sampling (ST) sampling method for samples selection. At the second sampling stage, RE and CV values of population extrapolation decrease as the sample size increases. Comprehensively considering the accuracy, stability of population extrapolation and sampling cost, the most cost-effective sample size for estimating the winter wheat acreage of the study area is 4. From the perspective of the reasonable selection of sample selection methods, sample size and determination of stratum boundaries, this study provides an important basis for formulating a cost-effective two-stage spatial sampling scheme for crop acreage estimation.

ARTICLE HISTORY
Received 6 December 2017
Accepted 21 August 2018

CONTACT Zhao-Liang Li ✉ lizl@unistra.fr Key Laboratory of Agricultural Remote Sensing, Ministry of Agriculture and Rural Affairs, Beijing 100081, China

1. Introduction

Timely and reliable information on crop acreage is essential for ensuring national food security (Chhikara, Houston, and Lundgren 1986; Quarmby 1992; Reynolds, Yitayew, and Slack 2000; Tao, Masayuki, and Zhan 2005; Song et al. 2017). In China, the traditional multilevel list sampling method (in which sample counties are drawn from the provinces, and then sample villages are drawn from the sampled counties, and finally sample households are drawn from the sampled villages) has been employed by the national statistical department to acquire crop acreage data since 1984 (National Bureau of Statistics of the People's Republic of China 2002). However, the problems facing the list sampling survey include the facts that the update rate of the sampling frame is too slow and that because spatial information on crop plantings has not been used in the sampling survey, the survey results are easily influenced by human operators. Consequently, the accuracy and timeliness of crop acreage data remain poor.

As a modern spatial information acquisition technique, satellite-based remotely sensed imagery, owing to its real-time and wide coverage, has the unique advantage of providing continuous space–time information on crop cultivation and growth at various regional scales. Therefore, such imagery has been widely exploited to monitor crops and associated environmental variables (Mahey et al. 1993; Cihlar 2000; Cohen and Shoshany 2002; Ramankutty et al. 2008; Thenkabail et al. 2009; Portmann, Siebert, and Dool 2010; You et al. 2014). Although crop acreage information is often acquired through the identification of crop types using satellite observation data, because of the restricted availability of satellite-based remotely sensed data and the diversity of cropping systems, crop recognition using remotely sensed imagery remains a technical challenge. In order to achieve accurate and timely crop monitoring, satellite-based remotely sensed data have often been combined with classical sampling methods – that is, spatial sampling – to estimate crop acreage over a large region (Gallego, Delince, and Rueda 1993; Gonzalez et al. 1997; Tsiligrides 1998; Delince 2001; Gallego 2012; Das and Singh 2013; Stehman 2014).

The use of satellite data in the formulation of a sampling scheme for crop acreage estimation was introduced in the early 1970s. The Large Area Crop Inventory Experiment (LACIE), conducted jointly by the United States Department of Agriculture (USDA) and the National Aeronautics and Space Administration (NASA) in 1974, is a typical example. Satellite images were used to derive a thematic map of wheat distribution, and this was the basis of a stratified sampling scheme (Macdonald and Hall 1980). More recent projects include the Agriculture and Resource Inventory Surveys through Aerospace Remote Sensing (AGRISARS) (Benedetti et al. 2010) and the Monitoring Agriculture with Remote Sensing (MARS) project sponsored by the European Union (EU). The latter used stratified sampling to monitor and estimate the acreage of 17 crops. In that project, satellite images were employed to formulate a stratification scheme and to measure crop acreages within the sampled units (Gallego 1999; Carfagna and Gallego 2005).

Although stratified sampling is a common choice for sampling surveys, for crop acreage estimation at a large-scale, two-stage sampling may be more appropriate in many practical situations. Advantages include its convenient sampling frame formulation, flexible sample selection process, and reduced survey labour (Stehman et al. 2009). In the Land Use/Cover Area Frame Statistical Survey (LUCAS) programme, also sponsored by the

EU, the efficiencies of simple random sampling, stratified sampling, and two-stage sampling were quantitatively evaluated. Two-stage sampling was ultimately chosen to estimate the acreage of major crops in 15 EU member countries (Jacques and Gallego 2006; Gallego 2004; Gallego and Bamps 2008). Stratified two-stage sampling and satellite-based remotely sensed data continue to be used in combination by the National Agricultural Statistics Service (NASS) of the United States (US) for monitoring and estimating staple crops acreages across the country (Fisette et al. 2013; Boryan, Yang, and Mueller 2011). Based on the spatial sampling design used in the US and the EU, Pradhan (2001) developed an operational system in which two-stage sampling, remote sensing (RS) and a geographic information system (GIS) were combined to estimate the crop acreage in Hamadan Province, Iran. To decrease the sampling cost and field survey workload, a stratified, two-stage cluster sampling design was proposed by Song et al. (2017) for collection of field data for national soybean area estimation in the US.

Although two-stage sampling has been widely used to estimate crop acreages over large regions, however, studies on sampling scheme optimization (population stratification, sample selection method and sample size) at each stage remain very rare. This impedes the further improvement of the spatial sampling efficiency for crop acreage estimation. Since multiple sampling stages may introduce larger estimation errors than single-stage sampling (e.g., simple random sampling, stratified sampling), with the aid of satellite-based remotely sensed imagery, the main objectives of this study were to (i) analyse the influence of the population stratification and sample selecting methods on the sampling efficiency at the first sampling stage; (ii) investigate the impact of sample size on the population extrapolating accuracy at the second sampling stage; and (iii) propose an optimized two-stage spatial sampling scheme to improve the sampling survey efficiency of the winter wheat acreage.

2. Materials and methods

2.1. *Study area*

Mengcheng County is located in the northwest of Anhui Province, China (32°55′29″ – 32° 29′64″ N, 116°15′43″ – 116°49′25″ E), and has a total land area of 2,091 km^2. This includes 1.53 × 10^5 ha of cultivated land. Mengcheng County has a subhumid continental monsoon climate and four distinct seasons. The annual average temperature is 14.7°C, the average frost-free period is 216 days, and the average annual precipitation is 872.4 mm. Priority is given to the cultivation of food crops, and cash crops are supplemental. Food crops include wheat, maize, rice, soybeans, and cotton. Winter wheat is the most important food crop in Mengcheng County and occupies around 70% of the total cultivated land in the study area.

2.2. *Data*

The experimental data comprise two parts: (1) Basic geographic data, the administrative boundary data for Mengcheng County (originally with a scale of 1:250,000 and in a vector format); and (2) Spatial crop distribution data, the spatial distribution of winter wheat in 2009. These latter data are derived from an ALOS (Advanced Land Observing

Satellite) AVNIR (Advanced Visible light and Near Infrared Radiometer)-2 image (number: 162,652,930; date: 12 February 2009; spatial resolution: 10 m). Figure 1 shows the spatial distribution of winter wheat in the study area in 2009.

2.3. *Design of spatial two-stage sampling scheme*

2.3.1. *Design of sampling units*

We employed a two-stage stratified sampling design for estimating winter wheat acreage within the study area. For the second stage, in which a field survey was conducted to determine the winter wheat area in the sampled units, and after considering a reasonable ground survey workload, we chose a 500 m × 500 m grid as the secondary sampling unit (SSU). In the first stage, a 5 km × 5 km block was selected as the primary sampling unit (PSU). These sizes met three criteria. Firstly, an applicable sample size for the first stage sampling can be acquired using the size. Secondly, the sampling schemes of each stage can be easily optimized using the satellite-based remotely sensed data. Thirdly, the number of SSUs within each PSU can be an integer. The entire study area was divided into a regular grid of 5 km × 5 km blocks (that is, PSUs), with a total of 113 PSUs, and then each PSU was divided by a square grid such that each PSU uniformly contains 100 SSUs. The spatial distribution of PSUs and SSUs covering the study area is shown in Figure 2. It is important to note that although some PSUs, such the PSUs numbered as 24, 42, 51, 60, 101, and so on, are partially covered by the administrative

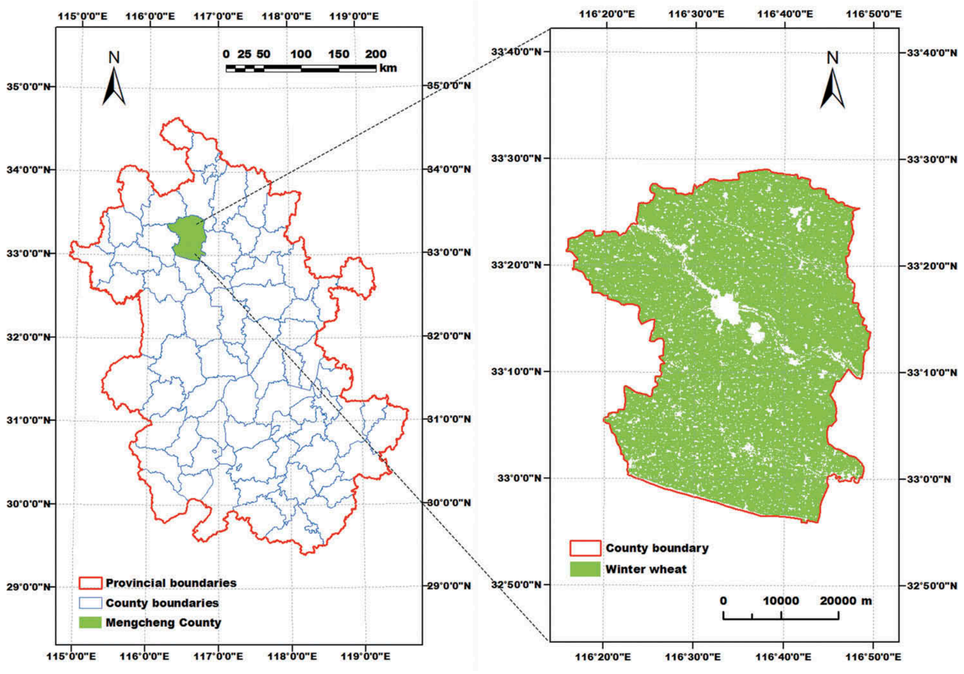


Figure 1. Spatial distribution of winter wheat in Mengcheng County in 2009.

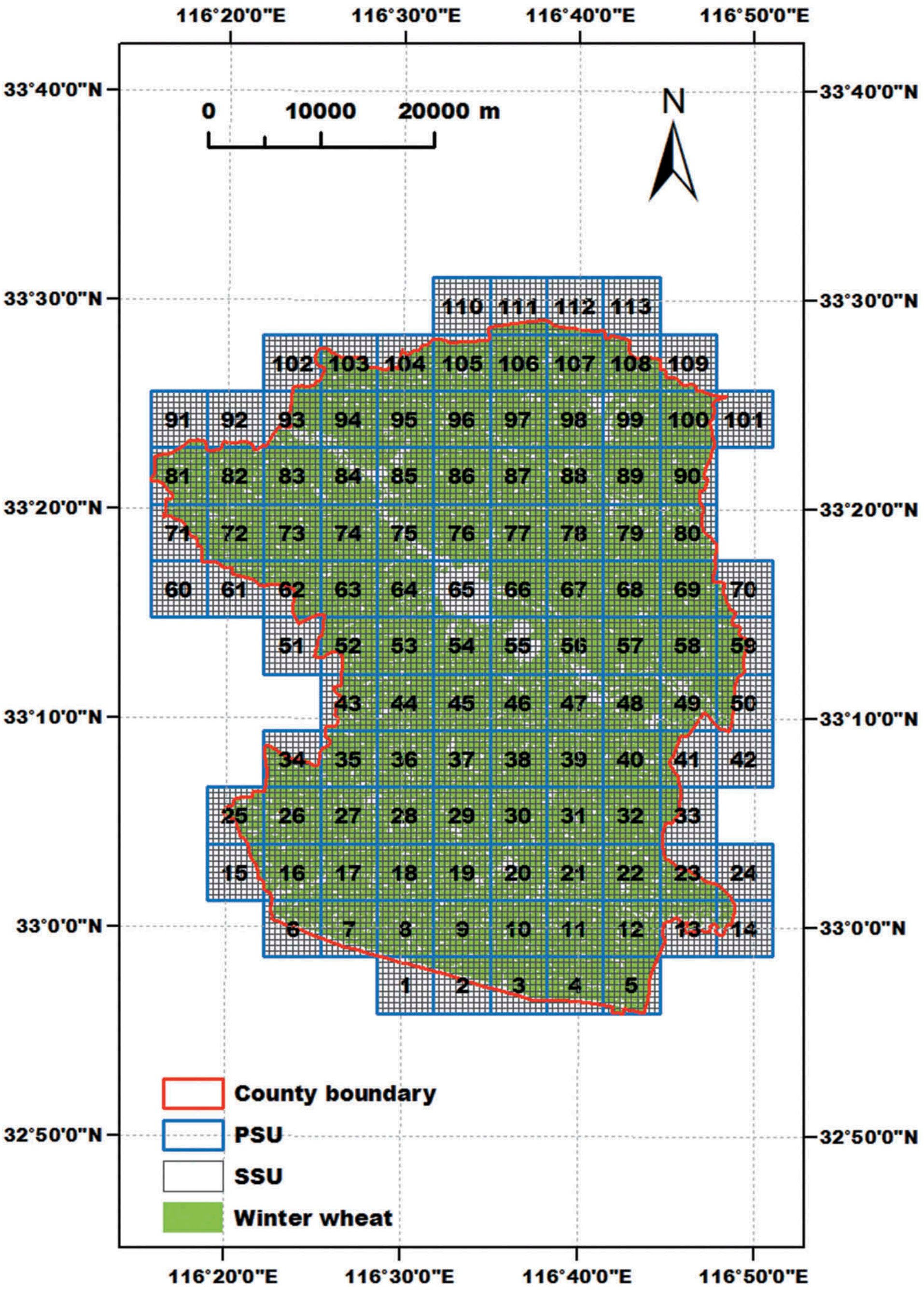


Figure 2. The spatial distribution of PSUs and SSUs covering the study area. Blue hollow boxes denote the PSUs, and small black hollow boxes denote the SSUs.

boundary of the study area, there is still some winter wheat in the covered regions belonging to these PSUs. In order to accurately estimate the winter wheat acreage in the whole study area, these PSUs should be included in the sampling frame.

2.3.2. *The sampling scheme at the first stage*

In designing a stratified sampling scheme to select the sample PSUs, we first needed an effective stratification criterion. The winter wheat acreage percentage (WAP) for every PSU was calculated based on the distribution data of winter wheat in the study area in 2009. This was chosen as the stratification criterion because there is a clear correlation between WAP and winter wheat area (that is, the target variable). We then used three methods to formulate the boundary of each stratum. They were the equal interval method (EI), the equal sample size method (ESS), and the cumulative square root of frequency method (CSRF). Because the dispersion variance of winter wheat acreage across all the PSUs was not large, we sorted them from the smallest to the largest WAP and then divided the population into four strata. For the EI stratification method, the differences of WAP across each stratum are equal. For the ESS method, the numbers of population units in each stratum are as equal as possible. The CSRF method, described by Du (2005) and Bhagia, Rajak, and Patel (2011), is used to calculate the boundary of each stratum by calculating the WAP within one PSU, and the occurrence frequency of WAP in each (5%) statistical interval. The reference stratum boundary (RSB) is equal to the cumulative square roots of the occurrence frequency of WAP divided by the number of strata, and using the RSB and the number of strata, the actual stratum boundary can be determined (see Table 1). Table 2 shows the stratum definition, for winter wheat acreage estimation in the study area, using the three stratification meth-

Table 1. Population stratification using the CSRF method.

No.	WAP (%)	*f*(*z*)	$\sqrt{f(z)}$	$\sum\sqrt{f(z)}$	RSB
1	1–5	10	3.16	3.16	
2	5–10	4	2.00	5.16	
3	10–15	3	1.73	6.89	
4	15–20	4	2.00	8.89	
5	20–25	3	1.73	10.63	10.26
6	25–30	2	1.41	12.04	
7	30–35	2	1.41	13.45	
8	35–40	3	1.73	15.19	
9	40–45	3	1.73	16.92	
10	45–50	5	2.24	19.15	
11	50–55	1	1.00	20.15	20.53
12	55–60	12	3.46	23.62	
13	60–65	12	3.46	27.08	
14	65–70	13	3.61	30.69	30.79
15	70–75	14	3.74	34.43	
16	75–80	12	3.46	37.89	
17	80–85	10	3.16	41.06	41.06
Stratum boundary		$\mathrm{RSB} = \sum\sqrt{f(z)}/L = 41.06/4 = 10.26$			

f(*z*) is the occurrence frequency of WAP in the sampling frame; RSB is the abbreviation of reference stratum boundary; *L* is the numbers of strata.

Table 2. Stratum definition using three stratification methods for winter wheat acreage estimation.

	EI		ESS		CSRF	
Stratum	Range of WAP (%)	Number of PSUs	Range of WAP (%)	Number of PSUs	Range of WAP (%)	Number of PSUs
1	0 – 20.69	22	0 – 35.70	29	0 – 22.21	24
2	21.57 – 42.16	10	39.53 – 62.16	28	26.30 – 52.68	16
3	44.05 – 63.01	29	62.32 – 73.31	28	56.51 – 69.60	37
4	64.25 – 84.36	52	73.32 – 84.36	28	70.67 – 84.36	36

ods. The spatial distributions of the PSUs belonging to different strata, using the three stratification methods, are shown in Figure 3.

In addition to the stratified sampling (ST) method, we also used the simple random (SI) and systematic sampling (SY) methods to select the sample PSUs and then to extrapolate the sample results to the full population of PSUs. For SI sampling, the pseudo-random number method is used to select the samples. The process of sample selection is as follows: First, all PSUs are encoded in 1 – N order using ArcGIS 10.2 software. Second, pseudo-numbers with a maximum upper boundary identical to the sample size are generated by Statistical Product and Service Solutions (SPSS 16.0) software. Finally, PSUs are chosen as samples if their codes match the random number generated by SPSS. SI sampling is also used to draw the samples belonging to each stratum for the ST method. In SY sampling, all PSUs are sorted in ascending order of ID numbers. A sampling interval k is then determined, equal to the integer value of the population size N divided by the sample size n. If the population is divided into n sections, then every section includes k PSUs. One PSU is selected randomly from the k PSUs in the first section as the starting point, and another PSU is selected every k units until n PSUs have been chosen. Figure 4 shows the spatial distribution of the sample PSUs selected using the three sampling methods. Using each of the three sampling methods, a simple estimator, which means that the sample arithmetic mean is served as the estimator of the population mean, is used to extrapolate the sample results to the whole PSU population and estimate the error. Sample size, population values (i.e. acreage), and the sampling error can be calculated as described by Cochran (1977). Specifically, for the SI and SY sampling, the sample size is calculated according to Equation (1) – (4). For the ST sampling, the sample size is calculated by Equation (2) and Equation (5) – (7).

$$n_0 = \left(\frac{t}{r}\right)^2 \frac{S^2}{\bar{Y}^2} \tag{1}$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \tag{2}$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{3}$$

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

$$n_0 = \frac{\sum W_h {S_h}^2}{V} \tag{5}$$

$$V = \left(\frac{r\bar{Y}}{t}\right)^2 \tag{6}$$

$$W_h = \frac{N_h}{N} \tag{7}$$

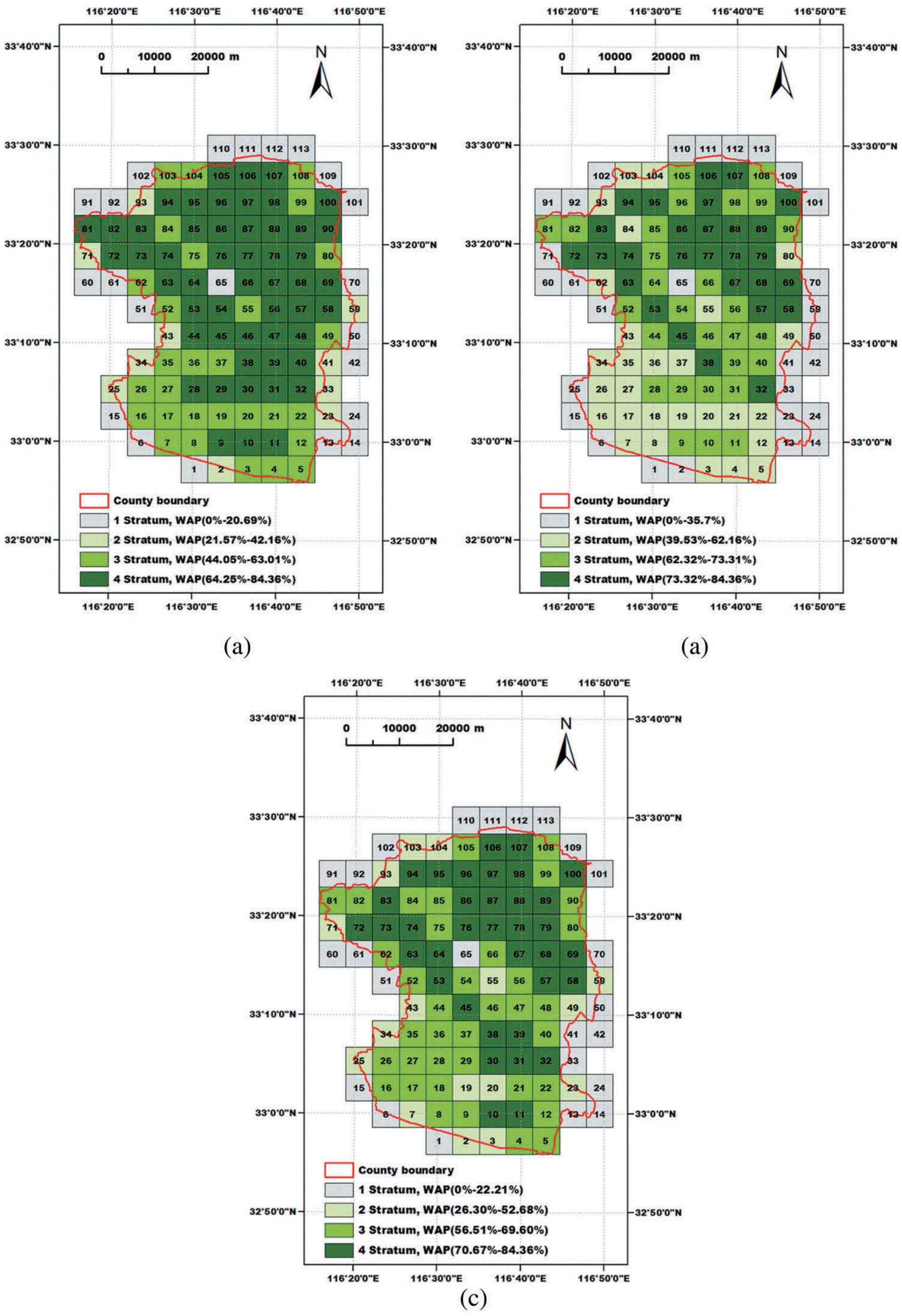
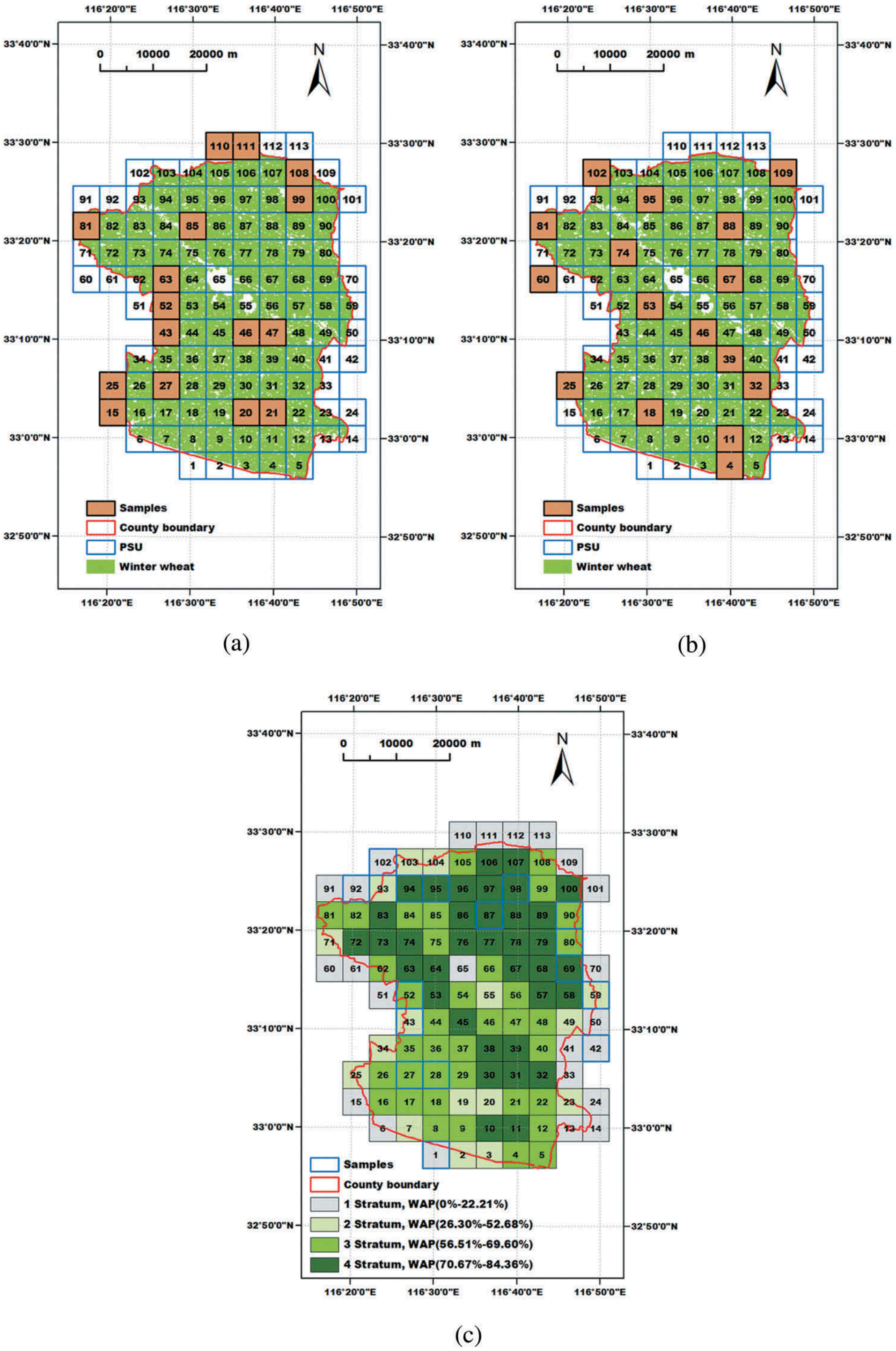

116°20'0"E
116°30'0"E
116°40'0"E
116°50'0"E
33°40'0"N
33°30'0"N
33°20'0"N
33°10'0"N
33°0'0"N
32°50'0"N
0 10000 20000 m
N
County boundary
1 Stratum, WAP(0%-20.69%)
2 Stratum, WAP(21.57%-42.16%)
3 Stratum, WAP(44.05%-63.01%)
4 Stratum, WAP(64.25%-84.36%)
(a)
County boundary
1 Stratum, WAP(0%-35.7%)
2 Stratum, WAP(39.53%-62.16%)
3 Stratum, WAP(62.32%-73.31%)
4 Stratum, WAP(73.32%-84.36%)
(a)
County boundary
1 Stratum, WAP(0%-22.21%)
2 Stratum, WAP(26.30%-52.68%)
3 Stratum, WAP(56.51%-69.60%)
4 Stratum, WAP(70.67%-84.36%)
(c)

Figure 3. Spatial distributions of PSUs belonging to different strata using three stratification methods. (a) EI method; (b) ESS method; and (c) CSRF method.



(a)

(b)

(c)

Figure 4. The spatial distribution of the sampled PSUs selected by three sampling methods. (a) SI sampling method; (b) SY sampling method; (c) ST sampling method based on CSRF stratification. For the SI and SY method, brown boxes denote selected PSUs. For the ST method, blue hollow boxes denote selected SSUs.

where n_0 is the initial sample size; t is the degree of sampling probability (when the confidence level is 95%, t is equal to 1.96); r is the relative error (10% is used in this study); S^2 is population variance; $\bar{Y}$is the population mean; n is modified sample size (when $n_0/N > 0.05$, n_0 is modified according to Equation (2)); N is the population size; andY_i is the winter wheat acreage proportion in the ith population unit; N_h is population size in the hth stratum; ${S_h}^2$ is the variance of population units in the hth stratum; V is the upper limit of the estimated variance of sample mean; W_h is the weight of population units in the hth stratum.

To evaluate quantitatively the efficiencies of the three methods, we chose relative error (RE), the coefficient of variation (CV), and standard error (SE) as indicators and ensured that the sample size was the same for the three sampling methods. The RE and SE are estimated using Equation (8) and Equation (9), respectively, and CV is calculated using Equation (10) (Cochran 1977; Du 2005):

$$\mathrm{RE} = \frac{\left|\hat{Y} - Y\right|}{Y} \times 100\% \tag{8}$$

$$\mathrm{SE} = \sqrt{v\left(\hat{Y}\right)} \tag{9}$$

$$\mathrm{CV}\left(\hat{Y}\right) = \frac{\sqrt{v\left(\hat{Y}\right)}}{\hat{Y}} \times 100\% \tag{10}$$

where Y is the true value of the population total, which can be obtained by calculating the winter wheat acreage in all PSUs based on the spatial distribution data for winter wheat in 2009; $\hat{Y}$ is the estimate of population total, which is different for every sampling method; $\mathrm{CV}\left(\hat{Y}\right)$ is the CV of the population total estimator; and $v\left(\hat{Y}\right)$ is the unbiased estimate of the variance of the population total estimator.

2.3.3. *The sampling scheme at the second stage*

To simplify the process of population extrapolation and error estimation for the two-stage sampling scheme, the SI method was used to select the sample SSUs. With SI sampling, the only element that can be optimized in the second stage is the sample size m. The optimal m can be calculated using Equation (11) (Du 2005):

$$m_{\mathrm{opt}} = \frac{S_2}{\sqrt{S_1^2 - \frac{S_2^2}{M}}}\sqrt{\frac{c_1}{c_2}} \tag{11}$$

$$S_2 = \sqrt{\frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{i,j} - \bar{Y}_i\right)^2} \tag{12}$$

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 \tag{13}$$

$$\bar{Y}_i = \frac{1}{M} Y_i \tag{14}$$

$$\bar{\bar{Y}} = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i \tag{15}$$

$$Y_i = \sum_{j=1}^{M} Y_{i,j} \tag{16}$$

where m_{opt} is the optimal sample size at the second stage; ${S_1}^2$ is the variance of winter wheat acreage between the PSUs; ${S_2}^2$ is the variance of winter wheat acreage within all the PSUs; N is the number of all PSUs; M is the number of SSUs within one PSU; $Y_{i,j}$ is the winter wheat acreage of the jth SSU in the ith PSU, it can be measured by overlapping the winter wheat spatial distribution data of the study area and the square grid that makes up the SSU; Y_i is the winter wheat acreage of i-th PSU; $\bar{Y}_i$ is the mean of the winter wheat area of all SSUs in the i-th PSU; $\bar{\bar{Y}}$ is the mean of the winter wheat area of all SSUs; c_1 is the cost of winter wheat retrieving in a sampled PSU using one ALOS AVNIR remotely sensed imagery. Since the price of an image is about 2960 China Yuan, c_1 is then thought as this price. c_2 is the ground survey cost of winter wheat in one sampled SSU, and it is approximately 185 China Yuan, considering the labour cost of measuring the winter wheat acreage within a SSU. Since c_1 and c_2 are 2960 and 185 China Yuan, $c_1 {c_2}^{-1}$ is approximately 16 in this study.

If the calculated m_{opt} is not a positive integer, then it must be rounded to the nearest integer. The rules for rounding off the m_{opt} are as follows: Assuming that m' is the integral part of m_{opt}, (a) if $m^2{}_{\text{opt}} \geq m'(m'+1)$, then $m_{\text{opt}} = m' + 1$; (b) if $m^2{}_{\text{opt}} \leq m'(m'+1)$, then $m_{\text{opt}} = m'$; (c) if $m_{\text{opt}} \geq M$ or $S_1^2 - \frac{S_2^2}{M} < 0$, then $m_{\text{opt}} = M$.

After m_{opt} was determined, based on the procedures above, we additionally formulated several sample sizes for the second- stage sampling scheme to verify that the calculated m_{opt} is indeed optimal. To ensure the stability of population extrapolation using the selected SSUs, five sets of samples were drawn for every designed sampling size.

As we used a stratified two-stage sampling design, we also used a stratified two-stage estimator to extrapolate the population and to estimate the errors. The estimated population total and its variance were calculated using Equation (17) and Equation (18) (Du 2005):

$$\hat{Y}_{\text{st}} = \left(\sum_{h=1}^{L} N_h M_h \right) \bar{\bar{y}}_{\text{st}} \tag{17}$$

$$v\left(\hat{Y}_{\text{st}}\right) = \left(\sum_{h=1}^{L} N_h M_h \right)^2 v\left(\bar{\bar{y}}_{\text{st}}\right) \tag{18}$$

$$\bar{\bar{y}}_{\text{st}} = \frac{\sum_{h=1}^{L} N_h M_h \bar{\bar{y}}_h}{\sum_{h=1}^{L} N_h M_h} = \sum_{h=1}^{L} W_h \bar{\bar{y}}_h \tag{19}$$

$$W_h = \frac{N_h M_h}{\sum_{h=1}^{L} N_h M_h} \tag{20}$$

$$\bar{\bar{y}}_h = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{m_h} y_{h,i,j}}{n_h m_h} \tag{21}$$

$$v\left(\bar{\bar{y}}_{\text{st}}\right) = \sum_{h=1}^{L} W_h^2 \left(\frac{1 - f_{1h}}{n_h} s_{1h}^2 + \frac{f_{1h}(1 - f_{2h}) s_{2h}^2}{n_h m_h} \right) \tag{22}$$

$$f_{1h} = \frac{n_h}{N_h} \tag{23}$$

$$f_{2h} = \frac{m_h}{M_h} \tag{24}$$

$$s_{1h}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\bar{y}_{h,i} - \bar{\bar{y}}_h \right)^2 \tag{25}$$

$$s_{2h}^2 = \frac{1}{n_h (m_h - 1)} \sum_{i=1}^{n_h} \sum_{j=1}^{m_h} \left(y_{h,i,j} - \bar{y}_{h,i} \right)^2 \tag{26}$$

$$\bar{y}_{h,i} = \frac{y_{h,i}}{m_h} \tag{27}$$

$$\bar{\bar{y}}_h = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{m_h} y_{h,i,j}}{n_h m_h} \tag{28}$$

$$y_{h,i} = \sum_{j=1}^{m_h} y_{h,i,j} \tag{29}$$

where $\hat{Y}_{\text{st}}$ is the estimate of population total for the stratified two-stage sampling; $v\left(\hat{Y}_{\text{st}}\right)$ is the unbiased estimate of the variance of the population total estimator; L is the number of the stratum at which the population is stratified at the first stage; h is the stratum number, h = 1,2, …, L; W_h is the weight of the h-th stratum; N_h is the number of PSUs in the h-th stratum; M_h is the number of SSUs included in each PSU in the h-th stratum; $y_{h,i,j}$ is the winter wheat area of the j-th selected SSU in the i-th sampled PSU belonging to the h-th stratum; f_{1h} is the sampling fraction of the h-th stratum at the first stage; f_{2h} is the sampling fraction of the h-th stratum at the second stage; ${s_{1h}}^2$ is the variance of the winter wheat area between all sampled PSUs in the h-th stratum; ${s_{2h}}^2$ is the variance of the winter wheat area between all sampled SSUs in the h-th stratum; $y_{h,i}$ is the sum of winter wheat area of the sampled PSUs in the h-th stratum; and $\bar{\bar{y}}_h$ is the mean of winter wheat area of all sampled SSUs in the h-th stratum.

3. Results

3.1. *The influence of stratum boundary on the efficiency of the stratified sampling scheme at the first sampling stage*

To assess the influence of stratum boundary on the efficiency of the stratified sampling scheme at the first sampling stage, the results of population extrapolation and error estimation using three methods for determining stratum boundaries were compared. Table 3 summarizes the calculated sample size and estimated errors for the three stratification methods, on the basis of the same designed relative error RE and the degree of sampling probability *t*. The RE and CV are all small (less than 6%), indicating that the required sample size for achieving this accuracy is less than 23. In other words, the *f* is less than 20% when any of the three types of stratification are used in the first stage to extrapolate population and to estimate errors. Furthermore, when the RE, CV, SE, and *f* using the three methods to extrapolate population are sorted from the smallest to the largest, it can be seen that the four values (RE, CV, SE, and *f*) are minimized using the CSRF method, followed by the EI and ESS methods. This indicates that the accuracy of population extrapolation is the highest when the CSRF stratified sampling method is used to estimate the winter wheat acreage of the study area. Moreover, as the *f* using the CSRF method is the smallest, the sampling cost is also the lowest. In summary, evaluating the accuracy and cost of a sampling survey for the winter wheat area estimation, CSRF is the most efficient of the three stratification methods.

3.2. *Comparison of different sample selection methods at the first stage*

Within a two-stage sampling approach, optimization of the sampling scheme of the first stage is more important than that of the second stage (Du 2005). To improve further the sampling efficiency of the first stage, besides the stratification method, the sample selection method should also be optimized. Table 4 summaries the results of population

Table 3. Summary of calculated sample size and estimated errors results for three methods for determining stratum boundaries.

Method	Stratum	*n*[a]	*N*[b]	*f*[c] (%)	RE (%)	CV (%)	SE
EI	1	3	22	13.64	-	-	-
	2	2	10	20.00	-	-	-
	3	5	29	17.24	-	-	-
	4	8	52	15.38	-	-	-
	Sum	18	113	15.93	4.18	2.42	34495307.08
ESS	1	6	29	20.69	-	-	-
	2	5	28	17.86	-	-	-
	3	5	28	17.86	-	-	-
	4	6	28	21.43	-	-	-
	Sum	22	113	19.47	5.79	2.57	36059065.17
CSRF	1	4	24	16.67	-	-	-
	2	2	16	12.50	-	-	-
	3	5	37	13.51	-	-	-
	4	5	36	13.89	-	-	-
	Sum	16	113	14.16	3.46	1.99	28543310.99

[a] Sample size.
[b] Population size.
[c] Sampling fraction.

Table 4. Results of estimated errors of population total for three sample selection methods.

Sample selection method	*n*	*N*	*f* (%)	RE (%)	CV (%)	SE
SI	16	113	14.16	7.75	11.65	183029601.20
SY	16	113	14.16	6.90	10.14	155387549.64
ST	16	113	14.16	3.46	1.99	28543310.99

extrapolation and error estimation using three sample selection methods to estimate the winter wheat acreage of the study area. When the sample size is identical for the three sample selection methods, the RE, CV, and SE values using the ST sample selection method are the smallest, followed by the SY and SI methods. This indicates that the accuracy of population extrapolation is best when the ST method is used to select the samples during the first sampling stage. In Paragraph 3.1, we established that the optimal ST method is CSRF stratified sampling. Therefore, we can identify the CSRF stratified sampling method as the optimal method for estimating the winter wheat area at the first sampling stage of the two-stage sampling scheme.

3.3. *Influence of sample size on the relative error and CV of population inference at the second sampling stage*

Based on the calculation procedure mentioned in Paragraph 2.3.3, the optimal sample size (that is m_{opt}) of the second stage is found to be 4 in each selected PSU. To verify whether the calculated m_{opt} is indeed optimal for the second-stage sampling survey, we also tested eight other sample sizes: 2, 3, 5, 6, 7, 8, 9, and 10. Figure 5 shows the RE and CV of population extrapolation using the nine different sample sizes in the second sampling stage and the CSRF stratified sampling method is used at the first sampling stage. It can be seen that the RE and CV values decrease as the sample size increases. The changes in RE and CV with sample size can be summarized into two phases. In the first phase, as the sample size increases from 2 to 4, RE and CV decrease rapidly from more than 20% to less than 5%. In the second phase, although the sample size increases from 4 to 10, the reduction of RE and CV is very small. In addition, the standard deviations of RE and CV change in a similar way with the sample size. This indicates that when considering accuracy, reliability of population extrapolation, and sampling survey cost, the most cost-effective sample size for estimating the winter wheat acreage of the study area is 4, which confirms that the calculated m_{opt} is indeed the optimal sample size for the second sampling stage. Taking 2, 4, 6, 8, and 10 as examples of sample size, Figure 6 shows the spatial distribution of s PSUs elected by the CSRF stratified sampling method and SSUs for these five levels of sample size. For the optimal sample size, it can be seen in Figure 6(b) that, because the sampled PSUs containing these SSUs overlap the boundary of the study area, some sampled SSUs are located outside the administrative boundary of Mengcheng County. Therefore, these SSUs do not need to be surveyed, which reduces the field survey cost and workload. A total of 16 sampled SSUs are found in Figure 6(b) to be located outside the boundary of the study area. Consequently, the number of SSUs that need to be investigated on the ground is only 48.

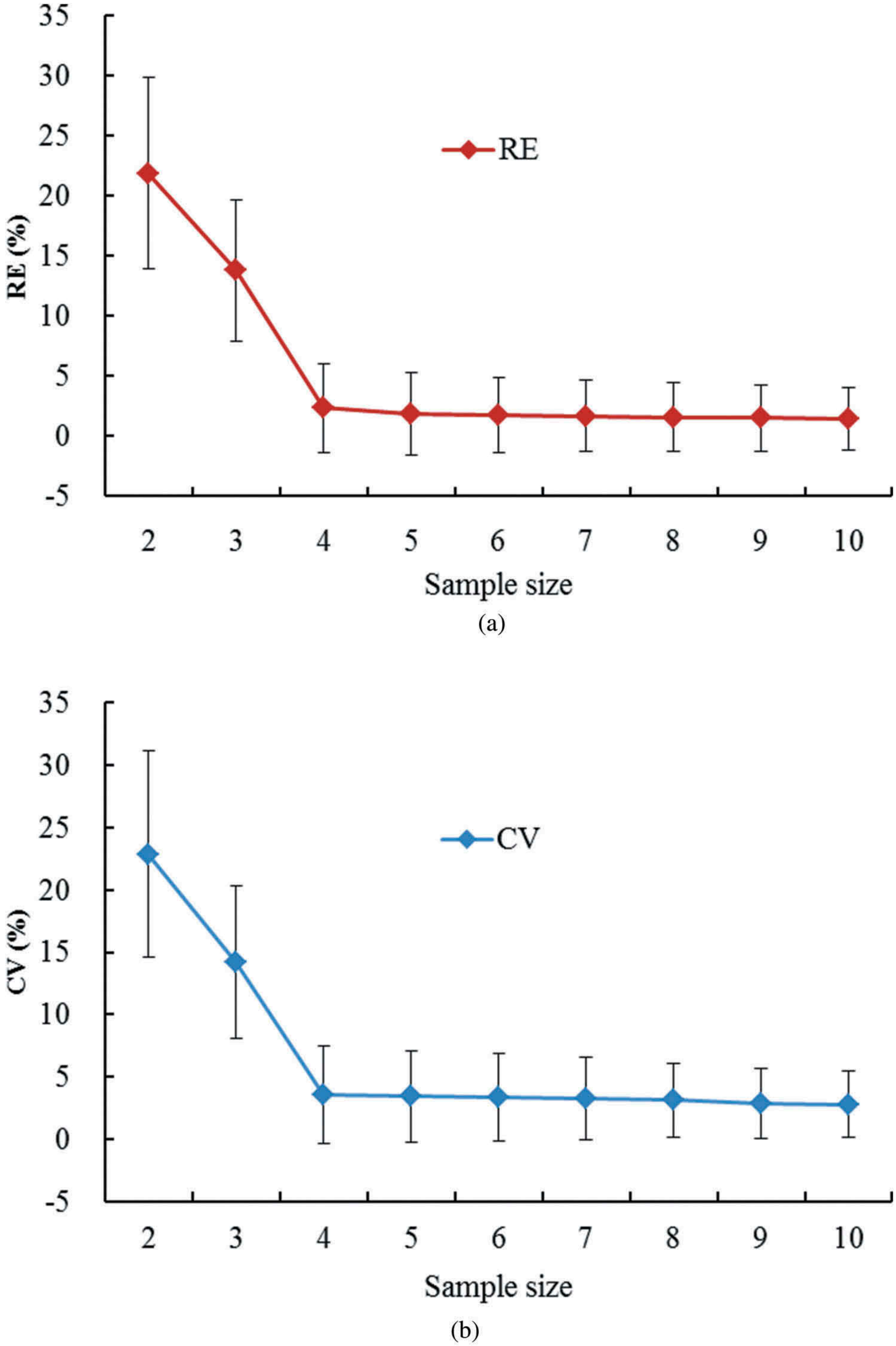


Figure 5. RE and CV of population extrapolation for nine sample sizes. Every diamond point in the figure denotes the mean value based on five sets of samples.

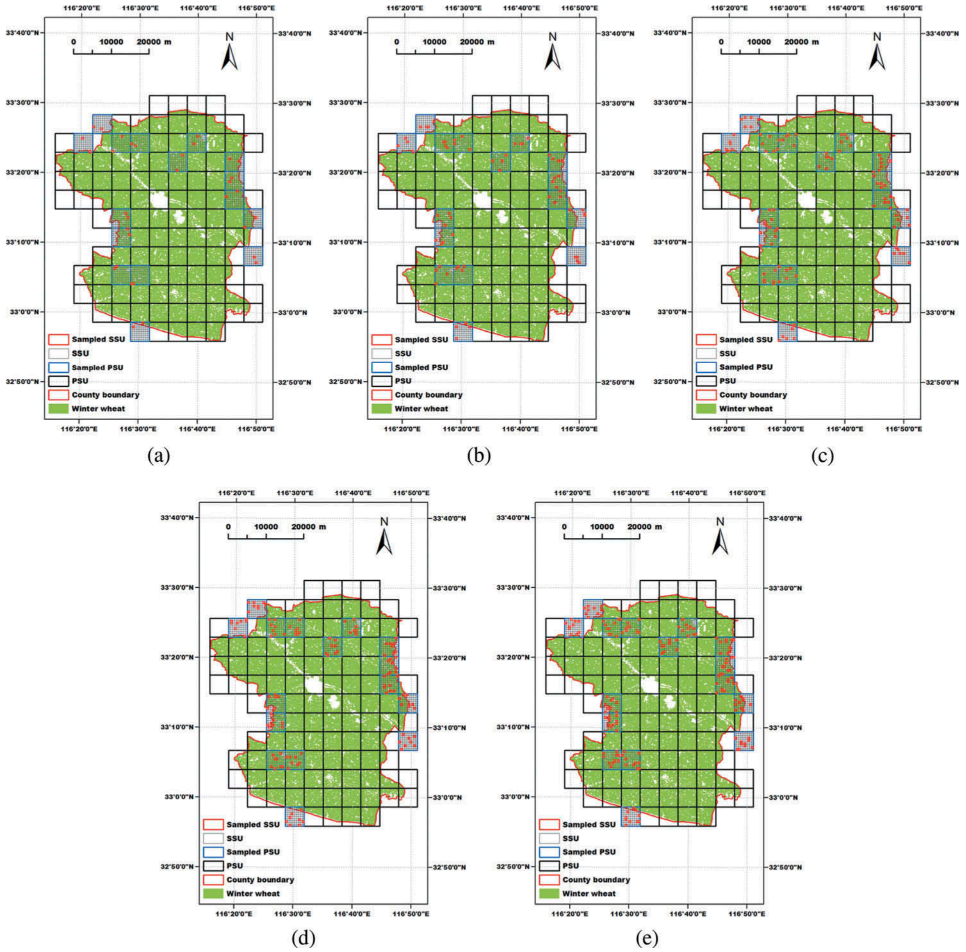


Figure 6. Spatial distributions of the selected PSUs and SSUs for five levels of second-stage sample size. (a) Sample size is 2; (b) sample size is 4; (c) sample size is 6; (d) sample size is 8; (e) sample size is 10.

4. Discussion

Crop acreage information is essential for formulating national food policies and economic planning. The combination of satellite-based remotely sensed imagery and traditional sampling methods provides an effective way for crop acreage estimation at the regional scale. Compared with single-stage sampling (e.g., simple random sampling, stratified sampling, and so on), two-stage sampling is more suitable for crop acreage survey at the large-scale regions, due to its convenient sampling frame formulation, flexible sample selection process and reduced survey workload. However, using the two-stage sampling can introduce larger estimation errors than single-stage sampling, since it has multiple sampling stages. To improve the accuracy of crop acreage estimation, this study analysed the influence of stratum boundary and sample selection method on the sampling efficiency at the first sampling stage and optimized the sample size of the second sampling stage using crop thematic map retrieved by ALOS AVNIR imagery. The results indicated that the *RE*, *CV*, *SE* and *f* of population extrapolation using the CSRF method to determine the stratum boundaries is the minimum at the first sampling stage, followed by the EI and ESS method, when the stratification criterion, stratum numbers are the same in the three methods. Moreover, compared with SI and SY sampling method, the RE, CV, and SE of population extrapolation using the ST sampling method that the CSRF method is employed to calculate the stratum boundaries is the minimum, when the sample size is the same among the three sample selection methods. At the second sampling stage, RE and CV values of population extrapolation decrease as the sample size increases. Comprehensively considering the accuracy, the stability of population extrapolation and sampling cost, the most cost-effective sample size for estimating the winter wheat acreage of the study area is 4. Overall, from the perspective of the reasonable selection of sample selection methods, sample size and determination of stratum boundaries, this study provides an important basis for formulating a cost-effective two-stage sampling scheme for crop acreage estimation, based on satellite-based remotely sensed imagery.

Due to the administrative convenience, reduced survey cost and workload, two-stage sampling may be more appropriate survey scheme for the crop acreage estimation in the large-scale regions where a satisfactory sampling frame of the ultimate units is not available. Although the two-stage sampling combined with satellite-based remotely sensed imagery have widely been employed to estimate the crop acreage in many previous studies, however, these studies mainly focus on how to formulate a simple and feasible sampling frame or reduce the workload of the samples field survey using the two-stage sampling. For example, in order to quickly estimate the annual changes of main crops acreage in 15 EU member countries, two-stage sampling was combined with SPOT (Systeme Probatoire d' Observation de la Terre) remotely sensed imagery to construct an available sampling frame and extrapolate the population values of crops acreage in LUCAS programme (Gallego and Bamps 2008). To date, stratified two-stage sampling has still been used by NASS for monitoring and estimating staple crops acreages across the country, due to its administrative convenience and cost-efficiency (Boryan et al. 2014). Recently, Song et al. (2017) put forward to a stratified two-stage sampling scheme for estimating national soybean acreage across the US (United States), in order to reduce the cost and workload of the sample ground survey. Compared with these previous studies, we not only took full advantage of the two-stage sampling to estimate crop acreage at the regional scale, but also optimized the sampling scheme from the perspective of stratum boundary, sample selection methods and sample size at each sampling stage to further improve the accuracy of crop acreage estimation, which is the innovation aspect of this study.

Crop planting structures and spatial distributions generally vary between regions; moreover, they may become more complex as the regional scale increases. To improve the applicability of this study, more research is required to verify the efficiency of this optimized two-stage sampling method for crop acreage estimation and explore the crop spatial distribution characteristics in other and larger crop-producing regions. Using the satellite-based remotely sensed imagery, this study analysed the influence of stratum boundary, sample selection methods and sample size on the efficiency of the sampling scheme and proposed the optimized sampling scheme at each stage, which thus provides an important basis for improving the efficiency of the sampling survey for crop acreage estimation. However, besides the stratum boundary, sample selection methods, there are some components, such as sampling unit size of PSU and SSU, stratification criterion, sample layout et al., to be taken into account in the design of two-stage sampling scheme. Therefore, a study that analyses the influence of these components on the efficiency of two-stage sampling for crop area estimation will be a future research focus.

5. Conclusions

This study shows that the RE, CV, SE and *f* of population extrapolation using the CSRF method is the minimum among three methods for the stratum boundary determination at the first sampling stage, followed by the EI and ESS method. Moreover, the RE, CV and SE of population extrapolation using the ST sampling method is the minimum, compared with SI and SY sampling method. Therefore, the sampling scheme of the first stage can be optimized by CSRF method for stratum boundary determination and ST sampling method for samples selection. At the second sampling stage, RE and CV values of population extrapolation decrease as the sample size increases. Comprehensively considering the accuracy, the stability of population extrapolation and sampling cost, the most cost-effective sample size for estimating the winter wheat acreage of the study area is 4.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [41471365,41531179].

References

Benedetti, R., M. Bee, G. Espa, and F. Piersimoni. 2010. *Agricultural Survey Methods*. Chippenham: John Wiley and Sons.

Bhagia, N., D. R. Rajak, and N. K. Patel. 2011. “Improvement in Precision of Crop Acreage Estimation by Remote Sensing Using Frequency Distribution Based Stratification.” *Journal of Indian Society of Remote Sensing* 39 (2): 153–160. doi:10.1007/s12524-011-0098–y.

Boryan, C., Z. W. Yang, L. P. Di, and K. Hunt. 2014. “A New Automatic Stratification Method for U.S. Agricultural Area Sampling Frame Construction Based on the Cropland Data Layer.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (11): 4317–4327. doi:10.1109/JSTARS.2014.2322584.

Boryan, C., Z. W. Yang, and R. Mueller. 2011. “Monitoring US Agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program.” *Geocarto International* 26: 341–358. doi:10.1080/10106049.2011.562309.

Carfagna, E., and F. J. Gallego. 2005. “Using Remote Sensing for Agricultural Statistics.” *International Statistical Review* 73: 389–404. doi:10.1111/j.1751-5823.2005.tb00155.x.

Chhikara, R. S., A. G. Houston, and J. C. Lundgren. 1986. “Crop Acreage Estimation Using a LANDSAT-based Estimator as an Auxiliary Variable.” *IEEE Transactions on Geoscience and Remote Sensing* 24: 157–168. doi:10.1109/TGRS.1986.289545.

Cihlar, J. 2000. “Land Cover Mapping of Large Area from Satellite: Status and Research Priorities.” *International Journal of Remote Sensing* 21 (6): 1093–1114. doi:10.1080/014311600210092.

Cochran, W. G. 1977. *Sampling Techniques*. Third ed. New York: John Wiley & Sons.

Cohen, Y., and M. Shoshany. 2002. "A National Knowledge-Based Crop Recognition in Mediterranean Environment." *International Journal of Applied Earth Observation and Geoinformation* 4: 75–87. doi:10.1016/S0303-2434(02)00003-X.

Das, S. K., and R. Singh. 2013. "A Multiple-Frame Approach to Crop Yield Estimation from Satellite Remotely Sensed Data." *International Journal of Remote Sensing* 34: 3803–3819. doi:10.1080/01431161.2012.762697.

Delince, J. 2001. ""A European Approach to Area Frame Survey" Processing of the Conference on Agricultural and Environmental Statistical Applications in Rome (CAESAR)." *Rome, European Commission GI & GIS* 2: 463–472.

Du, Z. F. 2005. *Sampling Techniques and Practices*. Beijing: Tsinghua University Press.

Fisette, T., P. Rollin, Z. Aly, L. Campbell, B. Daneshfar, P. Filyer, A. Smith, A. Davidson, J. Shang, and I. Jarvis. 2013. "AAFC Annual Crop Inventory: Status and Challenges." In *Proceedings of the Second International Conference on Agro-Geoinformatics*, Fairfax, WV, USA, 12–16.

Gallego, F. J. 1999. "Crop Area Estimation in the MARS Project." Conference on ten years of the MARS Project, Brussels, 1–11.

Gallego, F. J. 2004. "Remote Sensing and Land Cover Area Estimation." *International Journal of Remote Sensing* 25: 3019–3047. doi:10.1080/01431160310001619607.

Gallego, F. J. 2012. "The Efficiency of Sampling Very High Resolution Images for Area Estimation in the European Union." *International Journal of Remote Sensing* 33: 1868–1880. doi:10.1080/01431161.2011.602993.

Gallego, F. J., and C. Bamps. 2008. "Using CORINE Land Cover and the Point Survey LUCAS for Area Estimation." *International Journal of Applied Earth Observation and Geoinformation* 10: 467–475. doi:10.1016/j.jag.2007.11.001.

Gallego, F. J., J. Delince, and C. Rueda. 1993. "Crop Area Estimates through Remote Sensing: Stability of the Regression Correction." *International Journal of Remote Sensing* 14: 3433–3445. doi:10.1080/01431169308904456.

Gonzalez, F., J. M. Cuevas, R. Arbiol, and J. M. Baulies. 1997. "Remote Sensing and Agricultural Statistics: Crop Area Estimation in North-Eastern Spain through Diachronic Landsat TM and Ground Sample Data." *International Journal of Remote Sensing* 18: 467–470. doi:10.1080/014311697219213.

Jacques, P., and J. Gallego. 2006. "The LUCAS 2006 project-A New Methodology. Joint Research Centre□European Commisio." http://mars.Jrc.it/Bulletins-Publications/The-LUCAS-2006-project-A-newmethodology

Macdonald, R. B., and F. G. Hall. 1980. "Global Crop Forecasting." *Science* 208: 670–679. doi:10.1126/science.208.4445.670.

Mahey, R. K., R. Singh, S. S. Sidhu, R. S. Narang, V. K. Dadhwaal, J. S. Parihar, and A. K. Sharma. 1993. "Pre-Harvest State Level Wheat Acreage Estimation Using IRS-IA LISS-I Data in Punjab(India)." *International Journal of Remote Sensing* 14: 1099–1106. doi:10.1080/01431169308904398.

National Bureau of Statistics of the People's Republic of China. 2002. *The Operation Manual on Rural Statistics and Survey*. Beijing: China Statistics Press.

Portmann, F. T., S. Siebert, and P. Dool. 2010. "MIRCA2000-global Monthly Irrigated and Rainfed Crop Areas around the Year 2000: A New High-Resolution Data Set for Agricultural and Hydrological Modeling." *Global Biogeochemical Cycles* 24: GB1011. doi:10.1029/2008GB003435.

Pradhan, S. 2001. "Crop Area Estimation Using GIS, Remote Sensing and Area Frame Sampling." *International Journal of Applied Earth Observation and Geoinformation* 3 (1): 6–92. doi:10.1016/S0303-2434(01)85025-X.

Quarmby, N. A. 1992. "Towards Continental Scale Crop Area Estimation." *International Journal of Remote Sensing* 13: 981–989. doi:10.1080/01431169208904172.

Ramankutty, N., A. T. Evan, C. Monfreda, and J. A. Foley. 2008. "Farming the Planet: 1. Geographic Distribution of Global Agricultural Lands in the Year 2000." *Global Biogeochemical Cycles* 22: GB1003. doi:10.1029/2007GB002952.

Reynolds, C. A., M. Yitayew, and D. C. Slack. 2000. "Estimating Crop Yields and Production by Integrating the FAO Crop Specific Water Balance Model with Real-Time Satellite Data and

Ground-Based Ancillary Data." *International Journal of Remote Sensing* 21: 3487–3508. doi:10.1080/014311600750037516.

Song, X. P., V. P. Peter, K. Alexander, K. LeeAnn, D. B. CarlosM, H. Amy, K. Ahmad, A. Bernard, V. S. Stephen, and C. H. Matthew. 2017. "National-Scale Soybean Mapping and Area Estimation in the United States Using Medium Resolution Satellite Imagery and Field Survey." *Remote Sensing of Environment* 190: 383–395. doi:10.1016/j.rse.2017.01.008.

Stehman, S. V. 2014. "Estimating Area and Map Accuracy for Stratified Random Sampling When the Strata are Different from the Map Classes." *International Journal of Remote Sensing* 35: 4923–4939. doi:10.1080/01431161.2014.930207.

Stehman, S. V., J. D. Wickham, L. Fattorini, T. D. Wade, F. Baffetta, and J. H. Smith. 2009. "Estimating Accuracy of Land-Cover Composition from Two-Stage Cluster Sampling." *Remote Sensing of Environment* 113 (6): 1236–1249. doi:10.1016/j.rse.2009.02.011.

Tao, F. L., Y. Masayuki, and Z. Zhan. 2005. "Remote Sensing of Crop Production in China by Production Efficiency Models: Models Comparisons Estimates and Uncertainties." *Ecology Modeling* 183: 385–396. doi:10.1016/j.ecolmodel.2004.08.023.

Thenkabail, P. S., C. M. Biradar, P. Noojipady, V. Dheeravath, Y. Li, M. Velpuri, M. Gumma, et al. 2009. "Global Irrigated Area Map (GIAM), Derived from Remote Sensing, for the End of the Last Millennium." *International Journal of Remote Sensing* 30 :3679–3733. doi:10.1080/01431160802698919.

Tsiligrides, T. A. 1998. "Remote Sensing as a Tool for Agricultural Statistics: A Case Study of Area Frame Sampling Methodology in Hellas." *Computers and Electronics in Agriculture* 20: 45–47. doi:10.1016/S0168-1699(98)00011-8.

You, L., S. Wood, U. Wood-Sichra, and W. Wu. 2014. "Generating Global Crop Distribution Maps: From Census to Grid." *Agricultural. Systems* 127: 53–60. doi:10.1016/j.agsy.2014.01.002.