Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 2
CII 2023

Research on Adaptive Cluster Sampling Method based on PPS

Shaohua Wang [1], Ting Yang [1], Shengxiang Ouyang [2]

[1] University of Arts and Science, Changde 415000, China

[2] Changde Forestry Science Research Institute, Changde 415000, China

Abstract. This paper introduces the principle of PPS-based adaptive cluster sampling method and its modified HH estimator and HT estimator calculation method. It compares PPS-based adaptive cluster sampling method with SRS sampling and SRS-based adaptive group. The difference between the group sampling and the advantages and scope of the PPS adaptive cluster sampling method are analyzed. According to the case analysis, the relevant conclusions are drawn: 1) The adaptive cluster sampling method is more accurate than the SRS sampling; 2) SRS adaptive The HT estimator of the cluster sampling is more stable than the HH estimator; 3) The two estimators of the PPS adaptive cluster sampling method have little difference in the estimation of the population mean, but the HT estimator variance is smaller and more suitable; 4) PPS The HH estimator of adaptive cluster sampling is the same as the HH estimator of SRS adaptive cluster sampling, but the variance is larger and unstable.

Keywords: PPS Adaptive Cluster Sampling; Modified HH Estimation; Modified HT Estimation.

1. Introduction

Adaptive cluster sampling method is a sampling survey method first proposed by Thompson in 1990. It is a sampling survey for populations with sparse, clustered, patchy or clustered distribution characteristics. This method provides a new choice for sampling species resources in western China. During the implementation of adaptive cluster sampling method, the initial sample unit is extracted from the population at first, and then the sample is extrapolated to form an aggregation network according to the preset judgment standard, and the estimation is carried out according to the information provided by the aggregation network.[1] Traditionally, SRS method is used to extract initial sample units, but because the size difference of each unit in the population is too large or there are a few large sampling units, this method has the problems of low estimation accuracy and low sampling efficiency. PPS sampling is a sampling method with probability proportional to size. Will PPS sampling make up for the defects of SRS based adaptive cluster sampling? In order to answer this question, SRS sampling, SRS-based adaptive cluster sampling and PPS-based adaptive cluster sampling are used to analyze the simulated data in order to find a more effective and accurate sampling method.

2. Adaptive Cluster Sampling Method based on PPS

Considering the difference of population unit size, PPS sampling method is selected to obtain initial sample unit, that is, adaptive cluster sampling method based on PPS is used to investigate and estimate the population [5].

2.1. Principle of PPS Adaptive Cluster Sampling Method

PPS (Probability Proportion to Size) sampling is a method of unequal probability sampling with probability proportional to size. If the size of each unit in the population is equal in PPS sampling, the sampling probability of each unit is equal. In this case, PPS sampling is SRS sampling in the case of replacement [5]. In traditional probabilistic sampling, units of different sizes are assigned the same probability of extraction, which makes the probability of inclusion of members in a large-scale unit less than that in a small-scale unit. For example, if there are 200 members in a sample unit, the probability of one member being selected is 1/200 or 0.005, and the probability of one member being

selected in another 100 sample units is 1/100 or 0.01. In PPS sampling, different units have different sizes, and larger units have a higher probability of being selected when sampling. Usually, unit size needs to be measured with auxiliary information, and auxiliary information should have a high correlation with the target variable [5]. PPS based adaptive cluster sampling adopts PPS sampling method to select initial sample units, calculates the sampling probability of units according to the size of units, and then extracts samples from the population. PPS sampling has two methods: code method and Lahiri method, which can be selected according to specific conditions in the implementation process.

Let the overall total be Y, divide the target population into N adjacent units, and mark the flag values of N cells as $y_1, y_2, \cdots, y_N$,the size of N units is denoted by $m_1, m_2, \cdots, m_N$, the total size of the population is recorded as M_0 , $M_0 = \sum_{i=1}^{n} m_i$, so the probability that any unit i will be chosen is $Z_i = m_i / M_0$. n_1 initial sample units are extracted from N sample units, and the flag values of the n_1 initial sample units are denoted as $y_1, y_2, \cdots, y_{n_1}$, the size is denoted by $m_1, m_2, \cdots, m_{n_1}$. The probability that unit i will be selected in n_1 independent repetitions is $\pi_i = 1 - (1 - m_i / M_0)^{n_1}$.

These n_1 initial sample units are judged according to the set judgment criterion C, assuming that the judgment criterion is $C = \{x | x \geq c\}$. If the initial sample unit y_i meets the judgment criterion C, the formed aggregation network A_i includes xi sample units and ai edge units; if the initial sample unit y_i does not meet the judgment criterion C, the initial sample unit performs 0 times of sample extrapolation, and the formed aggregation network A_i includes only the initial sample unit, $x_i = 1$, $a_i = 0$.

In summary, the total size of the aggregation network A_i is $\sum_{j=1}^{x_i + a_i} m_j$, and the proportion of the total size of the units contained in the aggregation network A_i to the total size of the units is PA_i, then $P_{A_i} = \sum_{j=1}^{x_i + a_i} m_j / M_0$. Therefore, the sampling probability of the aggregation network A_i is: $\pi_i = 1 - (1 - P_{A_i})^{n_1} = 1 - (1 - \sum_{j=1}^{x_i + a_i} m_j / M_0)^{n_1}$. Because the process of adding sampling units by sample extrapolation of initial sample units is not a random process, but a process carried out step by step according to preset judgment criteria, when the initial sample units are selected, the aggregation network is fixed accordingly, and the sampling probability of the aggregation network finally obtained by adaptive cluster sampling is equal to that of the initial sample units y_i.

2.2. Estimators for PPS Adaptive Cluster Sampling

2.2.1. Revised HH Estimator

The traditional estimation method is HH estimation, which uses the three sample unit information contained in the aggregation network. H_u Dandan gave a modified HH estimator[6]. HH estimation of the total population Y is:

$$Y = t_{HH} = n_1^{-1} \sum_{i=1}^{n_1} \frac{y_i}{Z_i} \tag{1}$$

where $Z_i^{'}$>0, i=1, 2, ..., N and the sub-estimator is unbiased, but also information provided by two samples formed using only adaptive cluster sampling.

When estimating population, we obtain more sample units through sample extrapolation process of adaptive cluster sampling design, and the inference estimation based on the information of clustering network obtained by sample extrapolation process is more accurate.

$$y_i / m_i = \sum_{x_i} y_k / \sum_{x_i} m_k \tag{2}$$

From formula (1), we know:

$$Y = t_{HH} = n_1^{-1} \sum_{i=1}^{n_1} y_i / z_i^{'} = (M_0 / n_1) \sum_{i=1}^{n_i} (y_i / m_k) \tag{3}$$

Using the information provided by the aggregation network A_i to correct the above equation, the HH estimator formula corrected to the overall total Y can be obtained as:

$$\hat{Y} = t_{HH^*} = (M_0 / n_1) \sum_{i=1}^{n_1} (\sum_{x_i} y_k / \sum_{x_i} m_k) \tag{4}$$

where the mean of aggregation net A_i is $y_i / m_i = \sum_{x_i} y_k / \sum_{x_i} m_k$.

In the n_1 times of sampling by PPS sampling method, the probability of each aggregation network A_i being selected is $p_{A_i} = \sum_{x_i} m_k / M_0$. It can be seen that HH estimator of total population by PPS sampling is:

$$\hat{Y} = n_1^{-1} \sum_{i=1}^{n_1} \frac{\sum_{x_i} y_k}{\sum_{x_i} m_k / M_0} \tag{5}$$

By simplification, we can see that this formula is consistent with (6). The modified HH estimator has good properties, it is an unbiased estimate of the population total [5]. In the initial sampling, the PPS sampling process is independently repeated for n1, and the variance of the modified HH estimator can be obtained as

$$V(\hat{Y}_{HH}) = n_1^{-1} \sum_{i=1}^{N} Z_i (Y_i / Z_i - Y)^2 = n_1^{-1} \sum_{i=1}^{N} Z_{A_i} (M_0 \overline{Y}_i^* - Y)^2 \tag{6}$$

Its unbiased estimate is:

$$v(\hat{Y}_{HH}) = n_1^{-1} (n_1 - 1)^{-1} \sum_{i=1}^{n_1} (M_0 \overline{y}_i^* - \hat{Y})^2 \tag{7}$$

2.2.2. Revised HT Estimator

Traditionally, HT estimator is applied in non-replacement sampling, but HT estimator under replacement sampling is also an unbiased estimator [7]. The modified HT estimator of PPS adaptive cluster sampling is given below. The modified HT estimator of population Y is constructed according to the sampling probability of aggregation network is:

$$\hat{Y} = t_{HT^*} = \sum_{i=1}^{n'} y_i / \pi_i \tag{8}$$

where n' is the number of samples not repeated. The modified HT estimator $\hat{Y}_{HT}$ is an unbiased estimator of Y with variance:

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} / \pi_i \pi_j - 1) Y_i Y_j \tag{9}$$

In the adaptive cluster sampling process, there are two cases for calculating the second-order sampling probability π_{ij} : 1) Unit i and unit j exist in the same aggregation network, and under the influence of sample extrapolation mechanism, the case that unit i and unit j are selected is consistent with the case that they are not selected, so the second-order sampling probability of two units in the same aggregation network is the same as their respective sampling probability, $\pi_{ij} = \pi_i$ (also equal to π_j)[8]; 2) Unit i and unit j exist in different aggregation nets. In this case, complementary events are used to calculate the second-order sampling probability of simultaneous sampling, that is, the second-order sampling probability is:

$$\pi_{ij} = 1 - [(1 - \sum_{x_i} m_k / M_0)^{n_1} + (1 - \sum_{x_j} m_k / M_0)^{n_1} - (1 - \sum_{x_i + x_j} m_k / M_0)^{n_1}]$$

To sum up, we can obtain the unbiased estimate of the modified HT estimator variance:

$$v(\hat{Y}_{HT}) = \sum_{i=1}^{n'} \sum_{j=1}^{n'} \frac{1}{\pi_{ij}} (\pi_{ij} / \pi_i \pi_j - 1) y_i y_j \tag{10}$$

where n' is the number of samples that do not repeat. Note that variance estimators for $\hat{Y}_{HT}$ are not stable and may sometimes be negative[2].

3. Comparison of Adaptive Cluster Sampling Methods

3.1. Comparison of Adaptive Cluster Sampling and SRS Sampling

Adaptive cluster sampling method is proposed for the weakness of traditional sampling methods in investigating rare and cluster-like objects. Adaptive cluster sampling method has many advantages over SRS sampling in sampling surveys with such distribution characteristics. Firstly, adaptive cluster sampling and SRS sampling determine the number of sampling units differently. SRS sampling method determines the number of sampling units before sampling, while adaptive cluster sampling method finally needs to investigate the number of units depending on the criterion C set before sample extrapolation. Although the two sampling methods have different ways to determine the number of sampling units in the survey, the basic idea is the same. They both extract a part from the population and estimate the information of the population by investigating the information of this part. Secondly, the adaptive cluster sampling method has greater flexibility than SRS sampling. In sampling survey, adaptive cluster sampling method can design different initial sample extraction methods, unit number, unit size, sampling unit neighborhood configuration structure and judgment criteria or stopping rules for different populations and situations. The flexibility of this method makes the application range of adaptive cluster sampling method more extensive. Finally, the implementation cost of adaptive cluster sampling method is lower, because the average distance between two sampling units is smaller than other sampling methods[4].

Of course, perfect sampling method is our unremitting pursuit goal, but adaptive cluster sampling is not a perfect sampling method. Compared with SRS sampling method alone, it also has some disadvantages, mainly in two points: Firstly, the final number of sampling units and the number of aggregation networks formed by initial sample units are different with different judgment criteria C, and the shape and size of the aggregation networks formed are not fixed, which makes it difficult to determine the time cost, labor cost and capital cost required for sampling before investigation, and this information is often necessary before investigation. For example, the larger the value of the predetermined criterion C, the fewer and smaller the aggregation network formed, and the fewer sample units covered. For the sparsely distributed population, the sampling efficiency is greatly reduced. If the value of the predetermined criterion C is larger, the efficiency of adaptive cluster sampling can be improved, but this will inevitably lead to an infinite increase in sampling cost. Second, the information utilization rate of adaptive cluster sampling method is not high, and the information provided by sampling units is not maximized, such as edge units, only edge units appear in the initial

sample units will participate in the calculation of estimation, while most other edge units are ignored for various reasons, resulting in waste of information.

3.2. Comparison of PPS Adaptive Cluster Sampling Method and SRS Adaptive Cluster Sampling Method

PPS adaptive cluster sampling and SRS adaptive cluster sampling belong to adaptive cluster sampling method, but they are different in many aspects. Firstly, the two adaptive cluster sampling methods have different ways of extracting initial sample units. SRS adaptive cluster sampling uses simple random sampling to extract initial sample units from the population in an equal probability manner, while PPS adaptive cluster sampling uses PPS sampling to extract initial sample units from the population in an unequal probability manner. Secondly, PPS adaptive cluster sampling and SRS adaptive cluster sampling are different in application. SRS adaptive cluster sampling is suitable for the case that the size of each unit in the population is consistent or similar, and the investigation accuracy will be greatly reduced when the size of each unit in the population is greatly different. PPS adaptive cluster sampling method uses the sampling probability proportional to the size of the unit to extract the initial sample unit, which effectively solves this problem. Thirdly, PPS adaptive cluster sampling and SRS adaptive cluster sampling have different sampling mechanisms, which refer to the use of return sampling or non-return sampling. SRS adaptive cluster sampling uses simple random sampling to select initial sample units, which can be return sampling or non-return sampling, while PPS adaptive cluster sampling only returns sampling, and does not use PPS adaptive cluster sampling without return sampling. Finally, the implementation of PPS adaptive cluster sampling method is more complicated than SRS adaptive cluster sampling method. PPS adaptive cluster sampling method introduces PPS sampling into adaptive cluster sampling. Compared with SRS sampling, PPS sampling needs to consider the size difference between each unit in the population, and the sampling steps and sampling design are more complicated, which makes the implementation of PPS adaptive cluster sampling method more complicated.

4. Case Analysis

4.1. Case Study under SRS Sampling Method

Randomly select one cell from N=400 cells and include it in the sample. The probability of each cell being selected is 1/ N =1/400. This operation is repeated n times independently. In this case, the operation is to first label the 400 cells from 1 to 400, and then use R to generate 13 random numbers that may be repeated. These 13 random numbers represent the 13 selected cells. The results are shown in Figure 1. The circles with gray background represent the selected cells. Then, the information contained in these 13 cells is investigated and statistics are carried out, and 13 cells are obtained. The sum of observed values is 17. The mean value obtained from the calculation formula of the estimator is: $\bar{Y} = \sum_{i=1}^{n} Y_i / n = 1.308$, variance is:

$$v = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 / (n-1) = 10.564$$

4.2. Case Study of Adaptive Cluster Sampling Based on SRS

SRS is used to extract initial sample units. In order to ensure the comparability of estimation results, SRS sampling results are used as initial sample units. Sample extrapolation process is carried out based on initial sample units and aggregation network is finally obtained. Decision criterion C is set as $\{C \mid x \geqslant 1\}$,x is the observed value of the sample unit. The final sample obtained by sample extrapolation is a cluster formed by two irregular aggregation networks and a cluster formed by 11 initial sample units, as shown in Figure 2. The two clustering nets are formed by sample extrapolation

based on the initial sample unit extracted by SRS according to the judgment criterion C. The larger cluster in the region is formed by extrapolating 9 units from the initial sample unit with the observation value of 7, and the smaller cluster is formed by extrapolating 4 units from the initial sample unit with the observation value of 10. The black part is the edge unit. The other 11 initial sample units form a cluster separately because they do not satisfy decision criterion C. For the SRS-based adaptive cluster sampling design shown in Figure 2, the mean and variance within the study region can be calculated using the HT estimate without fallback and the HH estimate with fallback.

$$v = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 / (n-1) = 10.564 .$$

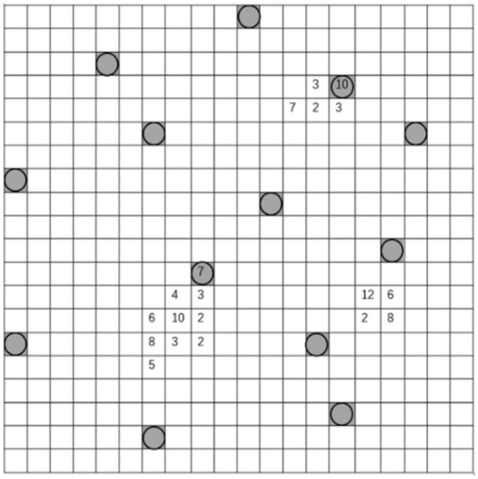

Fig 1. Sample Simulation Results Chart

SRS adaptive cluster sampling calculation estimator needs to determine the size of n=13 observed values in the cluster network and the number of units meeting the criterion C. See Table 1 for specific information. HH estimation and HT estimation performed by SRS adaptive cluster sampling method are denoted as HH estimation and HT estimation, and are calculated according to the data in the table and HH estimator calculation formula (1)(2) and HT estimator calculation formula (3)(4). The estimation result is obtained as: Mean value of HH estimator based on backward sampling is $\overline{y}_{HH} = 0.769$, variance is $V_{\overline{y}_{HH}} = 0.262$; Mean of HT Estimators Based on Non-replacement Sampling is $\overline{y}_{HT} = 0.848$, variance is $V_{\overline{y}_{HT}} = -0.023$.

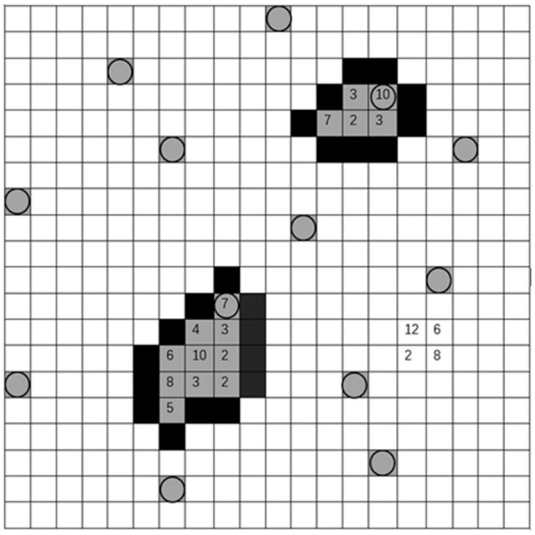

Fig 2. SRS Adaptive Cluster Sampling Simulation Results Chart

Table 1. SRS adaptive cluster simulation sampling estimation parameters

No.of network	Y values of network	Sum of Y	Unit sizes of network
1	0	0	1
2	0	0	1
3	3,10,7,2,3	25	5
4	0	0	1
5	0	0	1
6	0	0	1
7	0	0	1
8	7,4,3,6,10,2,8,3,2,5	50	10
9	0	0	1
10	0	0	1
11	0	0	1
12	0	0	1
13	0	0	1

4.3. Case Study of Adaptive Cluster Sampling Based on PPS

First, PPS sampling is used to extract the initial sample unit, the total is 400 cells, n=13 from which, PPS sampling commonly used methods are code method and Lahiri method, this paper uses Lahiri method to PPS sampling, extract 13 initial sample units, and then according to the judgment standard $C = \{C \mid x \geq 1\}$, sample extrapolation is performed to form 13 clustering nets, and the final sampling results are shown in Figure 3. Here, the size refers to the area of the sample unit, so $m_i = 100$, and the total size is $M_0 = \sum_{i=1}^{N} m_i = 40000$. The information obtained by the adaptive cluster sampling method based on PPS sampling is shown in Table 2.

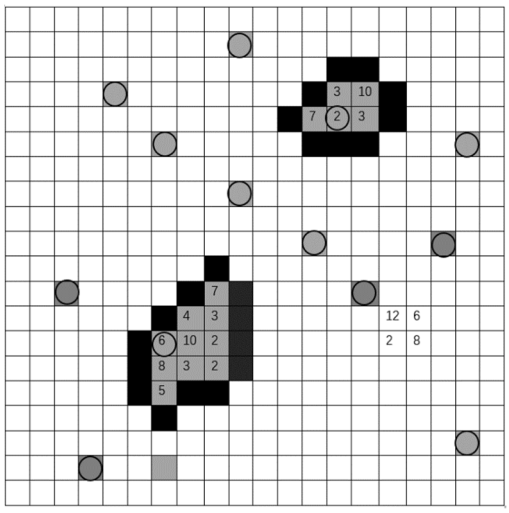

Fig 3. PPS adaptive cluster sampling simulation results

The HH estimator and HT estimator performed under the adaptive cluster sampling method based on PPS sampling are denoted as PHH estimator and PHT estimator. Based on the data in Table 2 and the modified HH estimator equation, calculate PHH estimates are generally:

$$\hat{Y} = t_{HH^*} = (M_0 / n_1)\sum_{i=1}^{n_1}(\sum_{x_i} y_k / \sum_{x_i} m_k) = 307.6293$$

The PHH estimator means: $\overline{y} = t_{HH*} / N = 0.769$

PHH estimator variance is:

$$v(\hat{Y}_{HH}) = n_1^{-1}(n_1 - 1)^{-1}\sum_{i=1}^{n_1}(M_0\overline{y}_i^* - \hat{Y})^2$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\sum_{x_i} y_k}{Z_{A_i}} - \hat{Y}\right)^2 = 43392.5$$

Table 2. PPS adaptive cluster simulation sampling estimation parameters

No.of network	Y values of network $\sum_{x_i} y_k$	Sum of Y M_{x_i}	Z_{Ai}	π_i
1	0	100	0.0025	0.03201694
2	0	100	0.0025	0.03201694
3	25	500	0.0125	0.15085402
4	0	100	0.0025	0.03201694
5	0	100	0.0025	0.03201694
6	0	100	0.0025	0.03201694
7	0	100	0.0025	0.03201694
8	50	1000	0.0250	0.28045161
9	0	100	0.0025	0.03201694
10	0	100	0.0025	0.03201694
11	0	100	0.0025	0.03201694
12	0	100	0.0025	0.03201694
13	0	100	0.0025	0.03201694

From the data in Table 2 and the modified HT estimator equations calculate:

PHT estimator population is: $\hat{Y} = t_{HT^*} = \sum_{i=1}^{n'} y_i / \pi_i = 344.007$,

PHT estimator mean is: $\overline{y}_{PHT} = t_{HT*} / N = 0.860$,

variance of PHT estimator is:

$$v(\hat{Y}_{HT}) = \sum_{i=1}^{n'}\sum_{j=1}^{n'}\frac{1}{\pi_{ij}}(\pi_{ij} / \pi_i\pi_j - 1)y_i y_j = 3826.565$$

4.4. Comparison of Case Analysis Results under Different Sampling Methods

Table 3 below contains the results of the investigation and estimation of simulated data using three sampling methods.

Table 3. Comparison of simulated sampling estimation results with different sampling methods

Parameters	SRS sampling	HH estimator	HT estimator	PHH estimator	PHT estimator
Overall	523.2	307.629	339.2	307.629	344.007
Mean	1.308	0.769	0.848	0.769	0.86
Variance	9.751	0.262	-0.023	43392.5	3826.565

To sum up: 1) For this case, the population estimation result obtained by SRS sampling method is the largest. It can be seen that the estimation effect of adaptive cluster sampling with sample extrapolation process in sampling process is obviously better than SRS sampling without this process. 2) The variance of HT estimator is smaller than that of HH estimator, which may be caused by the possibility of repeated sampling. 3) The variance of HT estimator is smaller than HH estimator, which is more suitable than HH estimator. 4) The HT estimator of PPS adaptive cluster sampling is larger than that of SRS adaptive cluster sampling, which may be due to too few valid samples in this case. 5) The HH estimator of PPS adaptive cluster sampling and the HH estimator of SRS adaptive cluster sampling have the same mean, because when the size of each unit in the population is the same, the sampling probability of each unit is equal, so that the two estimates of the mean are equal. However, the variance of SRS adaptive cluster sampling HH estimator is smaller and more stable, which is more suitable for this case.

5. Prospect

For populations with sparse, clustered, patchy or clustered distributions, the adaptive cluster sampling method has higher estimation accuracy and sampling efficiency, and PPS adaptive cluster sampling method also makes up for the defects of SRS adaptive cluster sampling to a certain extent, reducing the deviation of estimation when the cell size varies greatly. What condition does HT variance estimator of adaptive cluster sampling need to satisfy to be positive? How does PPS adaptive cluster sampling estimator change under different stopping rules? And how to apply adaptive cluster sampling to more fields, such as sociology, psychology and so on? These problems need further study.

Acknowledgments

Hunan University of Arts and Sciences Doctoral Research Initiation Project (20BSQD04); Hunan Provincial Natural Science Foundation (2023JJ60168).

References

[1] Lei Yuancai, Tang Shouzheng. Research progress on adaptive cluster sampling technique and its application[J]. Forestry science, 2009,45(3):118-127.

[2] Feng Shiyong, Ni Jiaxun, Zou Guohua. Sampling survey theory and method [M]. Beijing: China Statistics Press, 1998:181-240.

[3] Steven k. Thompson. Adaptive Cluster Sampling[J]. Journal of the American Statistical Association, 1990, 85:412, 1050-1059.

[4] Lei Yuancai, Tang Shouzheng. Application of adaptive cluster sampling technique in forest inventory [J]. Forestry Science, 2007,43(11):132-137.

[5] Ai Xiaoqing, Hu Dandan.PPS adaptive cluster sampling design and estimation [J]. Forum on Statistics and Information, 2017,32(9):3-8.

[6] Hu Dandan, Jin Yongjin, Ai Xiaoqing. Study on estimator under PPS adaptive cluster sampling [J]. Practice and understanding of mathematics, 2015,45(1):119-225.

[7] Ai Xiaoqing, Jin Yongjin. Properties and Applications of HT Estimator under Resampling [J]. Statistical Study, 2008,25(6):88-92.

[8] Hu Dandan, Jin Yongjin, Ai Xiaoqing. A New Way to Solve Small Domain Estimation: Sample Extrapolation [J]. Survey World, 2014,(9):44-46.