Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 14 (20): 2516-2522, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.2516.2522

Adaptive Cluster Sampling with Spatially Clustered Secondary Units

[1]Mena Patummasut and [2]John J. Borkowski
[1]Department of Statistics, Faculty of Science,
Kasetsart University, Lat Yao, Chatuchak, Bangkok, 10900, Thailand
[2]Department of Mathematical Sciences, Montana State University,
Bozeman, Montana, 59717, United States of America

Abstract: This study proposed the use of adaptively adding secondary units in adaptive cluster sampling when the primary units are formed by a spatial cluster of secondary units. Adaptive cluster sampling with this recommended approach of adaptively adding units is called adaptive cluster sampling with spatially clustered secondary units. The purpose of adaptive cluster sampling with spatially clustered secondary units is to provide efficient estimators and to save time and cost of travelling and observing units in the sample. For this sampling design, each primary unit is divided into four secondary units. An initial sample of primary units is selected and then secondary units are adaptively added to the sample instead of neighboring primary units. This study offers an unbiased estimator of the mean and its variance and shows the advantages and disadvantages of adaptive cluster sampling with spatially clustered secondary units in comparison to the ordinary adaptive cluster sampling based on a simple random sample.

Key words: Adaptive cluster sampling, Horvitz-Thompson estimator, spatial sampling

INTRODUCTION

Adaptive Cluster Sampling (ACS) was proposed by Thompson (1990) with the purpose of estimating the abundance of a rare and clustered population. For instance, if a researcher wishes to estimate the population size for a rare animal or plant species in a study area, then ACS can be an appropriate and efficient sampling design use (Thompson and Seber, 1996). Assume that a population study area is divided into a set of geographic units or plots of the same size. In ACS, an initial sample of units is selected by implementing a conventional sampling design, such as simple random sampling or systematic sampling. Whenever the value of the variable of interest of a unit in the sample satisfies some desirable condition determined by the researcher, its neighboring units are recursively added to the sample until there is no sampled unit satisfies the condition. The condition C is typically defined to be an interval in the range of the variable of interest. For example, if y is the observed species count for a sampled unit, then continue sampling if y>0 or, in other words, C corresponds to continued sampling if the species is present. The set of all units that satisfy the condition C in the neighborhood of each other is called a network. That is, sampling any network unit will lead to adaptively sampling every unit in the network. The units that were adaptively sampled but did not satisfy the condition are called edge units. A network and its corresponding edge units is called a cluster. Units that do not satisfy the condition, including edge units, are considered networks of size one.

It was shown that ACS will be less efficient than (non-adaptive) simple random sampling for a population having a low abundance and with individuals that are not spatially clustered (Christman, 1997; Brown, 2003). On the other hand, the advantage of ACS is its flexibility because the researcher can determine the initial sample size, the size of the sampling unit, the definition of a neighborhood and the condition defining when to adaptively sample (Turk and Borkowski, 2005). The most common neighborhood defined for a unit is the set of four adjacent units (above, below, left and right).

The researcher must also select the initial sampling plan. In addition to ACS with an initial simple random sample, systematic ACS and strip ACS were introduced by Thompson (1991). For these sampling designs, an initial sample of units is taken using systematic sampling and strip sampling before adaptively sampling. Stratification in ACS was also discussed by Thompson (1991).

Corresponding Author: Mena Patummasut, Department of Statistics, Faculty of Science, Kasetsart University, Lat Yao, Chatuchak, Bangkok, 10900, Thailand

Many scientific studies have applied ACS. For example, ACS was applied in subtidal microalgae to estimate the abundances of three algal species at three islands and two levels of wave exposure (Goldberg *et al*., 2007). Moreover, a two-stage sampling procedure was applied to estimate the total for a spatially aggregated, moving animal population in a large study are (He *et al*., 2013). For this study, ACS was considered in the first stage and capture-recapture sampling was used in the second stage.

The estimators developed for adaptive cluster sampling with an initial sample taken with or without replacement of sampled units are based on Hansen-Hurvitz and Horvitz-Thompson estimation (Thompson, 1990). Moreover, these estimators can be applied to systematic, strip and stratified ACS. Dryver and Thompson (2005) proposed an improved unbiased estimator in ACS which is derived by taking the expectation of the usual estimator conditional on a sufficient statistic which is not minimal sufficient. In addition, Dryver and Chao (2007) introduced new ratio estimators under ACS.

Note that, in ACS, when a selected unit satisfies the condition C, all units within its neighborhood are added to the sample. Notice that if a sampling unit is physically large, it can take long time to determine the response of interest, as well as be expensive in cost, when considering the neighboring network and edge units. This study proposes an adaptive strategy of adding "smaller" secondary neighboring units. That is, some part of each neighboring primary unit will be observed instead of the entire primary unit as is done in most ACS plans. ACS with this proposed approach of adaptively adding secondary units is called Adaptive Cluster Sampling with a Spatial Cluster of Secondary Units (ACS-SCSU).

ADAPTIVE CLUSTER SAMPLING WITH SPATIALLY CLUSTERED SECONDARY UNITS AND TECHNICAL NOTATION

The population is assumed to consist of N equal-sized rectangular primary units u_i (i = 1, 2, 3,…, N) and each u_i is divided into two rows and two columns forming four equal-sized rectangular secondary units u_{ij} (j = 1, 2, 3, 4). Thus, the study region contains 4N secondary units. The value of the population variable of interest corresponding to each secondary unit u_{ij} is denoted as y_{ij}. The parameter of interest in this study is the population mean:

$$\mu = \frac{1}{4N}\sum_{i=1}^{N}\sum_{j=1}^{4} y_{ij}$$

or the population total $\tau = 4N\mu$.

Adaptive cluster sampling with spatially clustered secondary units design: In ACS-SCSU, an initial sample of primary units of size n is selected by simple random sampling without replacement. Then for each sampled primary unit u_i (i = 1, 2,…, n), neighboring secondary units are added to the sample and observed whenever the variable of interest y_{ij} of a secondary unit u_{ij} satisfies a criterion C. This procedure continues until no adaptively sampled secondary units satisfy C. The set of all secondary units satisfying C as a result of u_{ij} being in the initial sample form a network while the secondary units that were adaptively sampled but did not satisfy C are edge units. Together these units form a cluster. For example, suppose a population consists of 20 primary units, each with 4 secondary units as shown in Fig. 1. To illustrate the ACS-SCSU scheme, suppose n = 2 primary units are selected by simple random sampling without replacement. These are shaded grey in Fig. 1a. Let the condition C correspond to $y_{ij}>0$ (i.e., the variable of interest being positive for secondary unit u_{ij}). Because the leftmost primary unit contains four secondary units with $y_{ij} = 0$, C is not satisfied for any secondary unit. Thus, no additional unit will be adaptively added to the sample about this primary unit. On the other hand, the rightmost primary unit does contain a secondary unit satisfying C with $y_{ij} = 103$. Thus, its two neighboring secondary units are adaptively sampled which have y_{ij}-values of 10 and 150 as shown in Fig. 1b. These two secondary units also satisfy condition C, so their neighboring secondary units are sampled as shown in Fig. 1c. The procedure continues, as shown in Fig. 1d but sampling terminates because no neighboring secondary units satisfy the condition. The final sample consists of 19 secondary units and is shown in Fig. 1e. In Fig. 1f, the network consists of the 5 dark grey secondary units with the 14 light grey secondary units being edge units or units in the original sample not satisfying C.

Estimation: For the ACS-SCSU design, the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) is used, but it is based on network inclusion probabilities which are determined post data collection and only for the networks observed in the final adaptive cluster sample. Let α_k be the probability that network k is included in the final sample. The α_k is called the inclusion probability of network k.

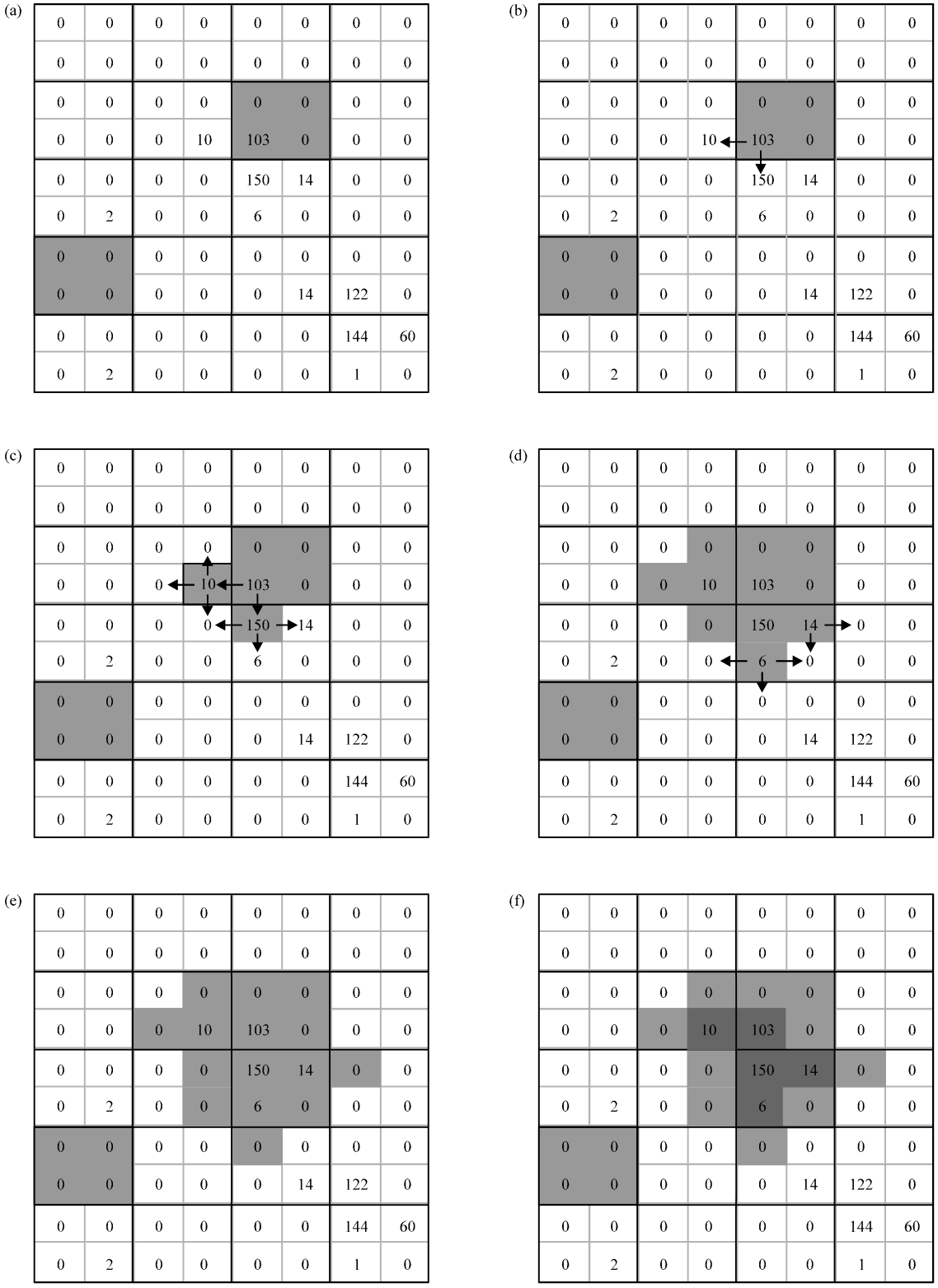

Fig. 1(a-f): Adaptive cluster sampling with spatially clustered secondary units sampling scheme

For ACS-SCSU, α_k is the probability at least one of the n primary units in the initial sample intersect network k. Let, x_k be the number of primary units in the population of n primary units which intersect network k. Using complementary probability, the probability that network k is included in the final sample is:

$$\alpha_k = 1 - \frac{\binom{N - x_k}{n}}{\binom{N}{n}} \quad (1)$$

Let, A_i be the event that network i is included in the final sample. Thus, $P(A_j) = \alpha_j$ and $P(A_k) = \alpha_k$ for network j and k, respectively. Let, α_{jk} be the probability that distinct networks j and k are both included in the final sample and it is also called the joint inclusion probability of networks j and k. It can be written as:

$$\alpha_{jk} = P(A_j \cap A_k) = P(A_j) + P(A_k) - P(A_j \cup A_k)$$

Thus:

$$\alpha_{jk} = \alpha_j + \alpha_k - P(A_j \cup A_k) \quad (2)$$

where, $P(A_j \cup A_k)$ is the probability that at least one of networks j and k is included in the final sample. Using complementary probability:

$$P(A_j \cup A_k) = 1 - \frac{\binom{N - x_j - x_k + x_{jk}}{n}}{\binom{N}{n}} \quad (3)$$

where, x_{jk} is the number of primary units in the population which intersect both networks j and k. Hence, substitution yields:

$$\alpha_{jk} = 1 - \left(\frac{\binom{N - x_j}{n} + \binom{N - x_k}{n} - \binom{N - x_j - x_k + x_{jk}}{n}}{\binom{N}{n}} \right) \quad (4)$$

Let, I_k be the indicator variable such that $I_k = 1$ if at least one primary unit in the initial simple random sample intersects network k and $I_k = 0$ otherwise. Using calculated inclusion probabilities, the Horvitz-Thompson estimator is applied in the ACS-SCSU design and the unbiased estimator of the population mean μ is:

$$\hat{\mu}_{SC} = \frac{1}{4N} \sum_{k=1}^{K} \frac{y_k I_k}{\alpha_k} \quad (5)$$

where, K is the total number of networks in the population. If network k is not included in the final sample, then $I_k = 0$ and it contributes zero to Eq. 5. Thus, the value of k does not need to be known to calculate $\hat{\mu}_{SC}$. Also note that although different primary units in the initial sample might intersect the same network, only the distinct networks observed in the final sample are utilized in the equation.

Applying the results of Horvitz and Thompson (1952), the variance of $\hat{\mu}_{SC}$ is:

$$v(\hat{\mu}_{SC}) = \frac{1}{16N^2} \sum_{k=1}^{K} \sum_{k=1}^{K} y_j y_k \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \quad (6)$$

Note that $\alpha_{kk} = \alpha_k$. An estimator of this variance is:

$$\hat{v}(\hat{\mu}_{SC}) = \frac{1}{16N^2} \sum_{k=1}^{K} \sum_{k=1}^{K} \frac{y_j y_k z_j z_k}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \quad (7)$$

and is unbiased if all joint inclusion probabilities α_{jk} are greater than zero.

SIMULATION STUDY

Data from the blue-winged teal study (Smith *et al*., 1995) was used in a simulation to investigate the performance of ACS-SCSU compared to ordinary adaptive cluster sampling of primary units. The simulation consisted of 1000 iterations. For the ith iteration, the value for the relevant estimator $\hat{\mu}_i$ and the effective (final) sample size υ_i were calculated. The formulas used to estimate an estimator variance and the estimated expected effective sample size of secondary units are:

$$\hat{v}(\hat{\mu}) = \frac{1}{1000 - 1} \sum_{i=1}^{1000} (\hat{\mu}_i - \bar{\mu})^2 \quad \text{and} \quad \upsilon = \frac{1}{1000} \sum_{i=1}^{1000} \upsilon_i \quad (8)$$

where, $\bar{\mu}$ is the sample mean of the 1000 $\hat{\mu}_i$ values (Dryver and Thompson, 2005). The blue-winged teal study area as shown in Fig. 2, consists of 50 primary units with each containing 4 secondary units. The simulation was considered for condition C to compare conventional ACS to ACS-SCSU: adaptively sample if $y_{ij} > 0$. For the 1000 simulated samples and for the condition C, $\hat{v}(\hat{\mu})$ and υ for ACS and ACS-SCSU were compared.

The simulation results were presented in Table 1. These results indicated that conventional ACS had a smaller average estimated variance than ACS-SCSU for

0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	20	4	2	12	0	0	0	0	0	10	103	0	0	0
0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	150	7144	0	0
0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	0	6393	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	122	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	114	60
0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	3	0

Fig. 2: Blue-wing teal population data (Smith *et al.*, 1995)

Table 1: Results from the simulation study on the blue-winged teal data under the condition y>0 and the estimated relative efficiencies R.E. = $\hat{v}(\hat{\mu}_{ACS}) / \hat{v}(\hat{\mu}_{SC})$

	ACS-SCSU		ACS		
n	υ	$\hat{v}(\hat{\mu}_{SC})$	υ	$\hat{v}(\hat{\mu}_{ACS})$	R.E
2	3.17	430,010.45	9.52	225,710.36	0.52
3	4.68	267,050.23	13.32	133,300.39	0.51
4	6.09	175,440.56	16.60	89,823.29	0.51
5	7.55	136,500.64	19.45	65,978.92	0.48
6	9.04	107,840.78	21.92	50,522.46	0.47
7	10.40	87,263.30	24.07	37,218.35	0.43
8	11.91	71,406.59	25.95	29,953.18	0.42
9	13.09	59,700.48	27.60	20,540.33	0.34
10	14.37	49,897.95	29.05	18,123.56	0.36
15	20.63	22,823.49	34.23	5,794.52	0.25
25	31.71	4,841.37	39.92	239.22	0.05

each initial sample size n and the estimated relative efficiencies (ratio of the estimated variances: R.E.) were less than one. Hence, ACS is more efficient than ACS-SCSU when considering the same value of the initial sample size. This, however, is misleading, because the effective sample sizes can be very different. The effective sample size υ reflects the true sampling effort incurred by the researcher (and not n). Thus, it is typical in ACS simulation studies to compare variances when υ is similar between two ACS plans (Turk and Borkowski, 2005). Note now that for the same initial sample size n, υ is smaller for ACS-SCSU than ACS. Thus, ACS-SCSU will, on average, be more efficient in terms of cost and for the time consumed travelling to observe the units in the final sample.

The average estimated variances presented in Table 1 were plotted against the estimated expected effective sample size in the graphs in Fig. 3. The graph of variances for ACS-SCSU was below the graph for ACS. This implies that estimated variances of the estimator of

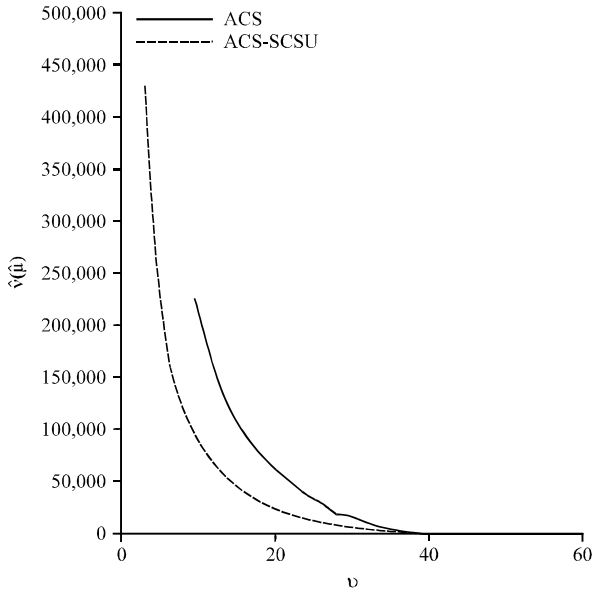

Fig. 3: Average estimated variances from the simulation study on the blue-winged teal data

the mean for ACS-SCSU are smaller, on average, than the estimated variances under ACS for the same average effective sample size. Thus, when considering the same final sample size as the important sampling requirement, ACS-SCSU is more efficient than ACS for estimating the population mean.

DISCUSSION

Adaptive cluster sampling designed with spatially clustered secondary units can be time and cost-effective when the population of interest is spatially aggregated

into relatively few networks and be more efficient that ordinary ACS with a simple random sample of primary units. Under ordinary ACS, the final sample size varies from sample to sample because units are adaptively added until no unit meets the criterion C (Thompson, 1990). Because the final sample size is not controlled, the final sample size can be quite large (Thompson, 2006). This can lead to a high cost of travelling within the study area and take much time to observe the sampled units. In comparison, ACS-SCSU will yield a smaller final sample size because secondary units are adaptively added in the sample instead of primary units and the secondary unit is one-fourth the size of a primary unit. Thus, the overall cost and time spent collecting the data is reduced. This was demonstrated by the results of the simulation study on the blue-winged teal population. ACS-SCSU has a smaller final sample size than ACS given the same initial sample size n.

The Horvitz-Thompson estimator based on partial inclusion probabilities is applied to ACS-SCSU for estimation of the population mean. Many sampling designs use the Horvitz-Thompson estimator because the inclusion probabilities can be obtained (e.g., path sampling (Patummasut and Dryver, 2012), simple latin square sampling±k designs (Borkowski, 2003), ACS (Thompson, 1990), inverse sampling (Mohammadi and Salehi, 2012), line-intercept sampling (Lucas and Seber, 1977). For ACS, the partial inclusion probabilities are used instead of actual inclusion probabilities since not all of the inclusion probabilities are known from the sample data. Similarly, under ACS-SCSU, partial inclusion probabilities are calculated and utilized in the Horvitz-Thompson estimator.

It is known that an improved unbiased estimator in adaptive cluster sampling exists and which can be derived by taking the expected value of the estimator conditional on a sufficient statistic which is not minimal sufficient (Dryver and Thompson, 2005). This idea can be used to find an improved unbiased estimator for ACS-SCSU data.

Estimation using ACS-SCSU may be improved by applying ratio estimation because, in general, the ratio estimator is often more precise (Dryver and Chao, 2007). The researcher, however, would need an appropriate auxiliary variable that is correlated with the response variable of interest. This is a problem being currently researched.

CONCLUSION

A new approach of adaptively adding units in the sample under adaptive cluster sampling was presented in this study. For ordinary adaptive cluster sampling, the whole region of a neighboring primary unit is added to the sample if the primary unit satisfies condition C. If a primary unit is large, it may take a long time and incur a high cost to observe all neighboring network units. Thus, adaptive cluster sampling of spatially clustered secondary units (ACS-SCSU) was proposed. In this sampling design, the primary unit is partitioned into four secondary units and secondary units are adaptively added to the sample instead of primary units. This leads to a smaller final sample size, so it will take less time and cost to travel and observe units in the final sample. The Horvitz-Thompson estimator based on partial inclusion probabilities was applied to estimate the population mean. In this study, ACS-SCSU and ACS were compared in a simulation study. According to the simulation results, for the same initial sample size, ACS-SCSU had a smaller expected effective sample size than that of ACS. Thus, ACS-SCSU would have lower cost and shorter time spent travelling and observing units in the final sample. ACS-SCSU estimators had smaller variances than ACS estimators for the same expected effective sample size. Thus, ACS-SCSU was more efficient than ACS for the same final sample size.

ACKNOWLEDGMENT

We are grateful to the Institute for Promotion of Science Teaching, Thailand, for financial support.

REFERENCES

Borkowski, J.J., 2003. Simple Latin square sampling K designs. Commun. Stat. Theory Methods, 32: 215-237.

Brown, J.A., 2003. Designing an efficient adaptive cluster sample. Environ. Ecol. Stat., 10: 95-105.

Christman, M.C., 1997. Efficiency of some sampling designs for spatially clustered populations. Environmetrics, 8: 145-166.

Dryver, A.L. and C.T. Chao, 2007. Ratio estimators in adaptive cluster sampling. Environmetrics, 18: 607-620.

Dryver, A.L. and S.K. Thompson, 2005. Improved unbiased estimators in adaptive cluster sampling. J. R. Stat. Soc.: Ser. B (Stat. Methodol.), 67: 157-166.

Goldberg, N.A., J.N. Heine and J.A. Brown, 2007. The application of adaptive cluster sampling for rare subtidal macroalgae. Mar. Biol., 151: 1343-1348.

He, Q. and N. Shifa, 2013. Estimating clustered population size using two stage sampling when capture probabilities vary among individuals. Int. J. Stat. Appl., 3: 39-42.

Horvitz, D.G. and D.J. Thompson, 1952. A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc., 47: 663-685.

Lucas, H.A. and G.A.F. Seber, 1977. Estimating coverage and particle density using the line intercept method. Biometrika, 64: 618-622.

Mohammadi, M. and M.M. Salehi, 2012. Horvitz-thompson estimator of population mean under inverse sampling designs. Bull. Iran. Math. Soc., 38: 333-347.

Patummasut, M. and A.L. Dryver, 2012. A new sampling design for a spatial population: Path sampling. J. Applied Sci., 12: 1355-1363.

Smith, D.R., M.J. Conroy and D.H. Brakhage, 1995. Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. Biometrics, 51: 777-788.

Thompson, S.K., 1990. Adaptive cluster sampling. J. Am. Stat. Assoc., 85: 1050-1059.

Thompson, S.K., 1991. Adaptive cluster sampling: Designs with primary and secondary units. Biometrics, 47: 1103-1115.

Thompson, S.K. and G.A.F. Seber, 1996. Adaptive Sampling. Wiley, New York, ISBN-13: 9780471558712, Pages: 265.

Thompson, S.K., 2006. Adaptive web sampling. Biometrics, 62: 1224-1234.

Turk, P. and J.J. Borkowski, 2005. A review of adaptive cluster sampling: 1990-2003. Environ. Ecol. Stat., 12: 55-94.