2012

Partial systematic adaptive cluster sampling

Arthur L. Dryver
National Institute of Development Administration, dryver@gmail.com

Urairat Netharn
Kasetsart University, ffisurnh@ku.ac.th

David R. Smith
USGS - Leetown Science Center, drsmith@usgs.gov

Research Article Environmetrics

Received: 18 November 2010, Revised: 20 March 2012, Accepted: 22 March 2012, Published online in Wiley Online Library: 4 May 2012

(wileyonlinelibrary.com) DOI: 10.1002/env.2144

Partial systematic adaptive cluster sampling

Arthur L. Dryver[a]*, Urairat Netharn[b] and David R. Smith[c]

A main benefit from taking a systematic sample is the ease of implementation when field sampling. However, it is not uncommon for a researcher to sample only one primary sampling unit (PSU) but to assume that the secondary sampling units (SSUs) were selected by simple random sampling to obtain a variance estimate. To obtain an unbiased estimator of variance for conventional or adaptive systematic sampling, it is necessary to sample >1 PSUs, and this can marginally increase cost and complicate implementation. We show that it is possible to obtain an unbiased estimate of variance if the researcher takes only a single PSU and one or more SSUs. Although this is no longer a true systematic sample, such a design retains much of the simplicity of sampling a single PSU and allows for a valid variance estimate.

This paper introduces three new sampling strategies stemming from systematic adaptive cluster sampling and the Raj estimator. The new sampling designs will be referred to as partial systematic adaptive cluster sampling. The sampling strategies are investigated in a simulation study that utilizes distance traveled as a measure of cost when comparing sampling strategies. When only a single PSU can be sampled because of cost or logistics concerns, we recommend also sampling one or more SSUs to obtain an unbiased estimate of variance.

Keywords: adaptive cluster sampling; Raj estimator; systematic sampling; unequal probability sampling

1. INTRODUCTION

Sampling for rare species in ecological studies can be costly and logistically difficult, and it is advisable to spread sampling effort throughout the study area (McDonald, 2004). Adaptive cluster sampling (ACS) has been shown to be efficient for sampling rare species and when combined with systematic sampling, effectively distributes sampling effort effectively while adapting to clustered populations (Thompson, 1991). In standard systematic ACS (SACS), at least two primary sampling units (PSU) are selected without replacement, and secondary sampling units (SSU) are "adaptively" sampled in the neighborhood of observed SSUs that meet a pre-specified condition (Thompson, 1991). However, it is not uncommon in field sampling to select only a single PSU to lower cost and increase the logistical ease of implementation (Smith *et al.*, 2003). A common practice when a single PSU is selected is to treat it as a simple random sample (SRS) for estimating the variance (e.g. Fridman and Walheim, 2000; Smith *et al.*, 2003; Pooler and Smith, 2005; Thompson, 2002). If only a single PSU is sampled by systematic sampling, the estimators of variance will be biased (Wolter, 1985).

We propose an alternative strategy that involves selection of a single large PSU with a single SSU. We call this design partial SACS (PSACS). Although this is no longer a true systematic sample, such a design retains much of the simplicity of sampling a single PSU. Importantly, an unbiased estimator of variance is available. Thus, PSACS is logistically advantageous and yields a statistically valid estimate.

The common design factors for all ACS designs are the initial sample design, the condition to adapt, and the neighborhood definition (Thompson, 1991). For SACS designs, the initial sample design is by systematic sampling with multiple random starts (Thompson, 1991). The condition to adapt is applied to the observation at the SSU level, that is, the y_i value and is commonly set at $y_i > 0$. The neighborhood specifies the SSUs that are adaptively sampled when triggered by the condition to adapt. A neighborhood must be symmetric in the sense that if unit i is in the neighborhood of unit j, then j is in the neighborhood of unit i, but beyond this restriction, the definition has considerable flexibility. Commonly and for this paper, the neighborhood includes the four adjacent units. Adaptive sampling continues until no additional SSUs meet the condition. All units within the neighborhood of one another meeting the condition form a network, and the adaptively added units that do not meet the condition are called edge units. A network and its associated edge units form a cluster. An initially sampled unit that does not meet the condition is a network of size 1.

The objective of this paper is to present the PSACS design and to evaluate its performance compared with ACS designs. Section 2 covers the advantages and flexibility of PSACS designs versus SACS or strip ACS. Section 3 presents the PSACS estimators notation. Then, Sections 4 and 5 contain an illustrative example and simulation results, respectively. Finally, the conclusions are in section 6.

* *Correspondence to: Arthur L. Dryver, National Institute of Development Administration, Bangkapi, Bangkok, Thailand 10240. E-mail: dryver@gmail.com*

a *National Institute of Development Administration, Bangkapi, Bangkok, Thailand 10240*

b *Department of Fishery Management, Faculty of Fisheries, Kasetsart University, Chatuchak, Bangkok, Thailand 10900*

c *USGS - Leetown Science Center, Aquatic Ecology Lab, 11649 Leetown Road, Kearneysville, WV 25430, U.S.A.*

2. PARTIAL SYSTEMATIC ADAPTIVE CLUSTER SAMPLING DESIGNS

As an example, consider a population of 200 SSUs and three types of SACS designs (Figure 1). The numbers in the figures represent the count of the object of interest in the SSU, such as a rare plant (Philippi, 2005). Blank units mean the count is zero in that SSU. In this illustration, units are adaptively sampled when the observed count >0. As a result, there are two networks of size >1.

A sample could be taken with only one PSU selected (Figure 1(A)). The selected PSU (shaded gray) consists of 20 SSUs, two of which have y_i values that meet the condition to adapt. Thus, neighboring units were adaptively added, and the units in the bold outline make up a single network of size 5. The adjacent empty units in the dotted lines make up the edge units, and the network with its edge units make up a cluster. In this case, variance is not estimable.

Alternatively, a sample could be taken with >1 PSU selected, which follows the standard SACS design (Figure 1(B)). In this example, two PSUs are selected and are half the size of the single PSU selected in Figure 1(A) to keep the number of SSUs in the initial sample consistent. Thus, the initial sample consists of two PSUs with 10 SSUs each. An unbiased estimate of variance can be calculated.

Finally, a sample could be taken with one PSU and a single SSU, which follows the proposed PSACS design (Figure 1(C)). The PSACS design is an alternative to selecting >1 PSUs. The additional SSU can be selected using three different methods:

1. From all the SSUs excluding the SSUs in PSU selected, gray color, or
2. From all the SSUs excluding the PSU selected and SSUs in the network intersected in the dark lines, or
3. From all the SSUs excluding the PSU selected and the cluster. The cluster consists of all units within the dashed lines, edge units plus units in the network.

The three different methods for selecting the additional SSU create three options for PSACS: PSACS without replacement of units, PSACS without replacement of networks, and PSACS without replacement of clusters, respectively. The new sampling strategies, designs and estimators are based on Salehi and Seber (1997) and Dryver and Thompson (2007). Ultimately, the fundamental concepts behind the strategies stem from Raj (1956) and Thompson (1990). All three proposed sampling designs can yield unbiased estimates of the population mean and variance.

3. PARTIAL SYSTEMATIC ADAPTIVE CLUSTER SAMPLING ESTIMATORS

The number of PSUs in the population is denoted by N, and the number of SSUs in PSU k is M_k. The number of PSUs selected is assumed to be 1 for the following formulas, and it is assumed to be the first unit sampled. The number of SSUs selected in addition to the single PSU is $m-1$. Thus, the initial sample consists m sample selections, 1 PSU and, $(m-1)$ SSUs. It is also assumed that the number of SSUs

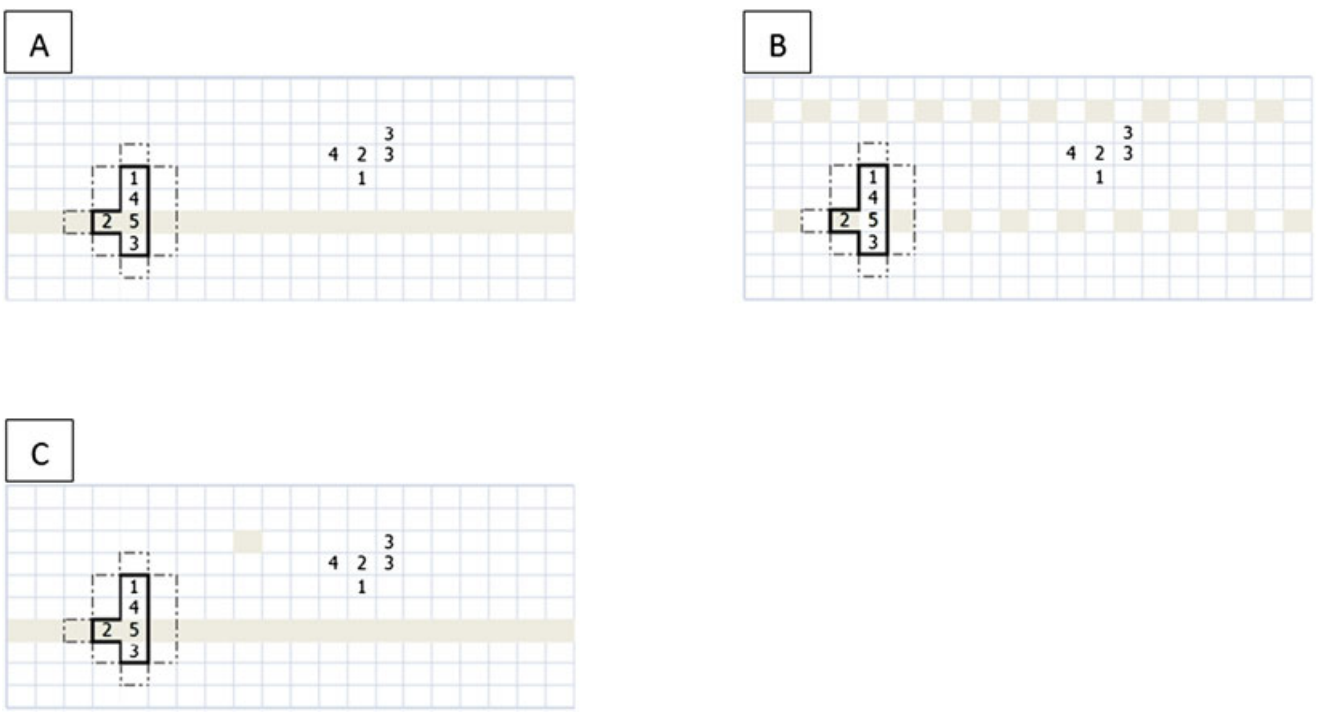

Figure 1. A single example population with three different sampling designs sampled. The shaded units are in the initial sample. In (A), a systematic adaptive cluster sample is taken with one large primary sampling unit (PSU) sampled. In (B), a systematic adaptive cluster sample is taken with two smaller PSUs sampled. In (C), a partial systematic adaptive cluster sample taken with one large PSU and one secondary sampling unit

is the same for all PSUs and there are $\sum_{k=1}^{N} M_k = H$ total SSUs in the population. In the case where all the PSUs are of the same size, $M_k = M \ \forall k$, and then the total number of SSUs is simply NM. The population vector, $y = (y_1, y_2, \ldots, y_H)$, is considered fixed and unknown constants. The parameter of interest for this paper is the population mean for SSUs, $\mu = \frac{1}{H}\sum_{i=1}^{H} y_i$. Let Ψ_i denote a collection of SSUs of which unit i is a member. Let w_i be defined for every SSU and represent the average SSU value of the network to which it is contained in Ψ_i. That is,

$$w_i = \frac{1}{x_i}\sum_{j\in\Psi_i} y_j$$

where x_i is the number of SSUs in the network to which unit i belongs.

We present formula for the three PSACS design options. The formulas that apply to all three options have subscript "*." The formula specific to an option has subscripts "u," "n," and "c" to denote PSACS without replacement of units (u), networks (n), or clusters (c), respectively. An unbiased estimator for the population mean is a weighted sum of unbiased estimates of the population total, z_{*i}, divided by the number of secondary units in the population (Raj, 1956) and is denoted as

$$\hat{\mu}_{w*} = \frac{1}{H}\sum_{i=1}^{m} c_i z_{*i} \qquad \sum_{i=1}^{m} c_i = 1 \tag{1}$$

where c_i are the weights and must add up to 1. The variance is

$$\mathrm{var}(\hat{\mu}_{w*}) = \frac{1}{H}\sum_{i=1}^{m} c_i^2 \mathrm{Var}(z_{*i})$$

An unbiased estimator of the variance is

$$\widehat{\mathrm{var}}(\hat{\mu}_{w*}) = \hat{\mu}_{w*}^2 - 2\frac{\sum_{i<j} z_{*i}z_{*j}}{H^2 m(m-1)}$$

where $\sum_{i<j}$ is the sum of all the $\binom{m}{2}$ pairs.

For the special case where the weights are all equal, $c_i = \frac{1}{m}\ \forall i$, then the aforementioned equations reduce to the following (Raj, 1956):

$$\hat{\mu}_{*} = \frac{1}{Hm}\sum_{i=1}^{m} z_{*i} \tag{2}$$

with variance

$$\mathrm{var}(\hat{\mu}_{*}) = \frac{1}{(Hm)^2}\sum_{i=1}^{m} Var(z_{*i})$$

An unbiased estimator of the variance is

$$\widehat{\mathrm{var}}(\hat{\mu}_{*}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{z_{*i}}{H} - \hat{\mu}_{*}\right)^2$$

The z_{*1} is the same for all three sampling designs,

$$z_{*1} = \frac{\sum_{j\in s_{u0}} w_j}{p_1}$$

where $p_1 = \frac{M_1}{H}$, the probability of selecting a PSU of size M_1, and s_{u0} consist of all SSUs in the PSU selected on the first draw. The z_{*1} is an unbiased estimator of the population total as estimated from the first selected PSU.

The three PSACS design options require that different sets of SSUs be defined. Let $s_{u(i-1)}$ represent the set of units observed up to the $i-1$ selection, excluding the adaptively added units. Let $s_{n(i-1)}$ represent the set of units observed up to the $i-1$ selection, including the adaptively added units within networks but excluding edge units. Let $s_{c(i-1)}$ represent the set of units observed up to the $i-1$ selection, including the adaptively added units within clusters (i.e., networks plus edge units). Then, z_{*i} for $i = 2, \ldots, m$ is defined as follows:

$$z_{*i} = \sum_{j\in s_{*(i-1)}} w_j + \frac{w_i}{p_{*i}}$$

where

$$p_{*i} = 1/\left(H - \sum_{j \in s_{*(i-1)}} 1\right)$$

or simply the probability of selecting a single SSU from the remaining SSUs. The calculation for z_{*i} differs in terms of $s_{*(i-1)}$ and equals

$$z_{ui} = \sum_{j \in s_{u(i-1)}} w_j + \frac{w_i}{1/(H - \sum_{j \in s_{u(i-1)}} 1)}$$

$$z_{ni} = \sum_{j \in s_{n(i-1)}} w_j + \frac{w_i}{1/(H - \sum_{j \in s_{n(i-1)}} 1)}$$

$$z_{ci} = \sum_{j \in s_{c(i-1)}} w_j + \frac{w_i}{1/(H - \sum_{j \in s_{c(i-1)}} 1)}$$

The z_{*i} are unbiased estimators of the population total. The SSU y-values are substituted by their network averages, w_i. Sampling may be performed without replacement of units, networks, or clusters, and this determines the remaining units to sample for future selections and what units to include in the estimator z_{*i}. For example, z_{ni} is the sum of all SSU w_j values for units selected prior to the ith selection plus w_i divided that unit probability of selection. The probability of selection for the ith unit is one divided by the number of SSUs whose networks have not already been selected in the $i-1$ prior selections, $(H - \sum_{j \in s_{n(i-1)}} 1)$.

3.1. Theory behind the estimators

For simplicity, this section focuses on PSACS without replacement of clusters, but the concepts easily follow to the other PSACS design options. Because of the definition of a neighborhood, if you observe any unit in the network, then you will observe the entire network. The $\mathbf{w}_i$-values are the average network values. Thus, by using w_i as opposed to the specific y_i selected, the variability among the y-values within the same network is eliminated reducing the variance of the estimator (Thompson, 2002).

The concept of the estimators of the population mean and its variance originally follows from Raj (1956) as the proposed sampling designs can be viewed as types of unequal probability sampling designs from the transformed population of w_i's. $E[z_{c1}] = E\left[\frac{\sum_{j \in s_{u0}} w_j}{p_1}\right]$ equals the expectation of the sum of the w_i selected on the first selection divided by the probability of selection on the first draw. Thus, the $E[z_{c1}] = \tau$, the population total. The $E\left[z_{ci} = \sum_{j \in s_{c(i-1)}} w_j + \frac{w_i}{p_{ci}}\right]$ for $i \geq 2$, the z_{ci} can be broken into 2 parts, $E[z_{ci}] = E\left[\sum_{j \in s_{c(i-1)}} w_j\right]$ and $E\left[\frac{w_i}{p_{ci}}\right]$. The $E\left[\sum_{j \in s_{c(i-1)}} w_j\right]$ equals the $\sum_{j \in s_{c(i-1)}} w_j$ is the sum of all y_i observed for the $i-1$ selections. The $E\left[\frac{w_i}{p_{ci}}\right]$ is the expectation of w_i divided by the probability of selecting that unit from the remaining unobserved units. Thus, $E\left[\frac{w_i}{p_{ci}}\right] = \tau - \sum_{j \in s_{c(i-1)}} w_j$, the population total minus the sum of the previously observed y-values, $\tau - \sum_{j \in s_{c(i-1)}} w_j$. Thus, $E[z_{ci}] = \sum_{j \in s_{c(i-1)}} w_j + E\left[\frac{w_i}{p_{ci}}\right] = \tau$. The estimator $\hat{\mu}_c = \frac{1}{Hm} \sum_{i=1}^{m} z_{ci}$ is the average of the z_{ci} divided by the number of secondary units, and thus, the $E[\hat{\mu}_c] = \mu$, the population mean.

3.2. Theory behind using the unequal weights

The weighted version of the proposed estimators, Equation 1, using unequal weights can yield lower variance than using equal weights. Unfortunately, the formula for estimating variance when using the unequal weights can yield negative values. For the latter reasons, the authors cover both estimators in Equations 1 and 2. As mentioned in the previous section, 3.1, each z_{*i} is an unbiased estimate of the population total. The z_{*1} utilizes a single PSU of size M_1 SSUs plus some information from adaptive added unit estimate τ, whereas z_{*2} conditions on some to all of the previously observed SSUs depending on without replacement of units to without replacement of clusters and a single SSU to estimate the remaining units in the population. For this reason, the var(z_{*1}) is expected to be much smaller than the variance var(z_{*2}). The latter can be thought of in a similar manner to comparing the var$(\bar{x})$ and var(x). In Sections 4 and 5 for $m = 2$, the weights are $c_1 = \frac{M_1}{M_1+1}$ and $c_2 = \frac{1}{M_1+1}$. These weights yield significantly lower variances than $c_1 = c_2 = \frac{1}{2}$, as used in equation 2.

4. ILLUSTRATIVE EXAMPLE

We use the population in Table 1 to illustrate the proposed designs and estimators. The population consists of $H = 12$ SSUs and $N = 3$ PSUs. Each row makes up a PSU, and each PSU consists of $M = 4$ SSUs. The PSACS design begins with an initial selection a single PSU. A single SSU is then selected from the remaining units available for selection, where the remaining units depend upon the sampling design. For illustration, let $m = 2$ and the condition to adapt be $y_i \geq 50$. Table 1 contains both the y_i-values, that is, population of interest, and the corresponding w_i-values, that is, the transformed population (Thompson, 2002; Dryver and Chao, 2007).

Table 1. A fictitious population, y_i, with $\mu = 29.0$, and the transformed version of the population, w_i. The condition to adaptively sample is $y_i \geq 50$

Population, y_i				$\Rightarrow$	Corresponding w_i			
60	70	0	1		65	65	0	1
2	5	6	80	$\Rightarrow$	2	5	6	90
7	8	9	100		7	8	9	90

4.1. Partial systematic adaptive cluster sampling examples

Suppose the initial PSU selected is the third row, consisting of units with values $y = 7, 8, 9, 100$ and the associated $w = 7, 8, 9, 90$. Then,

$$z_{*1} = (7 + 8 + 9 + 90)/(1/3) = 342$$

and is the same for all three sampling designs.

The selection of the SSU and calculation of the z_{*2} are different for all three PSACS design options.

1. For PSACS without replacement of units, the SSU will be selected from the eight SSUs comprised of the SSUs in the population excluding the SSUs in the PSU already selected. The z_{u2} will be calculated by adding up the w_i values for all the units in the PSU selected plus the w_i-value of the SSU selected divided by its probability of selection $\frac{1}{8}$.

Table 2. All possible samples from applying partial systematic adaptive cluster sampling without replacement of units to the population in Table 1

Final sample	P(S)	ν	z_{u1}	z_{u2}	$\hat{\mu}_u$	$\widehat{\text{var}}(\hat{\mu}_u)$	$\hat{\mu}_{wu}$	$\widehat{\text{var}}(\hat{\mu}_{wu})$
60,70,0,1,2;5	1/24	6	393	147	22.50	105.06	28.65	419.64
60,70,0,1,5;2	1/24	6	393	171	23.50	85.56	29.05	377.22
60,70,0,1,6;2,5	1/24	7	393	179	23.83	79.51	29.18	363.15
60,70,0,1,80;2,5,100,6,9	1/24	10	393	851	51.83	364.17	40.38	−691.71
60,70,0,1,7;2,5	1/24	7	393	187	24.17	73.67	29.32	349.11
60,70,0,1,8;2,5	1/24	7	393	195	24.50	68.06	29.45	335.12
60,70,0,1,9;2,5	1/24	7	393	203	24.83	62.67	29.58	321.15
60,70,0,1,100;2,5,80,6,9	1/24	10	393	851	51.83	364.17	40.38	−691.71
2,5,6,80,60;70,0,100,1,9	1/24	10	309	623	38.83	171.17	30.98	−376.89
2,5,6,80,70;60,0,100,1,9	1/24	10	309	623	38.83	171.17	30.98	−376.89
2,5,6,80,0;100,1,9	1/24	8	309	103	17.17	73.67	22.32	277.01
2,5,6,80,1;100,9	1/24	7	309	111	17.50	68.06	22.45	265.82
2,5,6,80,7;100,1,9	1/24	8	309	159	19.50	39.06	23.25	199.38
2,5,6,80,8;100,1,9	1/24	8	309	167	19.83	35.01	23.38	188.43
2,5,6,80,9;100,1	1/24	7	309	175	20.17	31.17	23.52	177.51
2,5,6,80,100;1,9	1/24	7	309	823	47.17	458.67	34.32	−588.39
7,8,9,100,60;70,0,2,5,80,1,6	1/24	12	342	634	40.67	148.03	33.37	−392.42
7,8,9,100,70;60,0,2,5,80,1,6	1/24	12	342	634	40.67	148.03	33.37	−392.42
7,8,9,100,0;80,1,6	1/24	8	342	114	19.00	90.25	24.70	339.34
7,8,9,100,1;80,6	1/24	7	342	122	19.33	84.03	24.83	326.94
7,8,9,100,2;80,1,6	1/24	8	342	130	19.67	78.03	24.97	314.58
7,8,9,100,5;80,1,6	1/24	8	342	154	20.67	61.36	25.37	277.72
7,8,9,100,6;80,1	1/24	7	342	162	21.00	56.25	25.50	265.50
7,8,9,100,80;1,6	1/24	7	342	834	49.00	420.25	36.70	−633.86
Expectation =		8.08			29.00	139.05	29.00	27.22
Bias =					0.00	0.00	0.00	0.00

Final sample lists all units in each initial sample followed by adaptively added units placed after the semicolon. The final sample size is denoted ν. The final rows contain the expectation and bias of the estimator for mean and variance.

2. For PSACS without replacement of networks, the SSU will be selected from the seven SSUs comprised of the SSUs in the population excluding the SSUs in the networks intersected by the PSU already selected. The z_{n2} will be calculated by adding up the w_i values for all the units in the networks intersected by the PSU selected plus the w_i-value of the SSU selected divided by its probability of selection $\frac{1}{7}$.
3. For PSACS without replacement of clusters, the SSU will be selected from the five SSUs comprised of the SSUs in the population excluding the SSUs in the cluster and networks intersected by the PSU already selected. The z_{c2} will be calculated by adding up the w_i values for all the units in the cluster and networks intersected by the PSU selected plus the w_i-value of the SSU selected divided by its probability of selection $\frac{1}{5}$.

Now suppose, regardless of design option, that the SSU selected has the value $y_2 = 5$. Then the z_{*2} would depend on the PSACS design option as follows:

$$z_{u2} = 7+8+9+90+\frac{5}{(1/(12-4))} = 114+\frac{5}{(1/8)} = 154$$

$$z_{n2} = 7+8+9+90+90+\frac{5}{(1/(12-5))} = 204+\frac{5}{(1/7)} = 239$$

$$z_{c2} = 7+8+9+90+90+1+6+\frac{5}{(1/(12-7))} = 211+\frac{5}{(1/5)} = 236$$

The number of SSUs in the first PSU sampled is four and in the second selection is one to make a total of five nonadaptively sampled SSUs. The weights are then $c_1 = \frac{4}{5}$ and $c_2 = \frac{1}{5}$ for the $\hat{\mu}_{w*}$. By using the z_{i*} and Equation 1, the corresponding estimates for the population

Table 3. All possible samples from applying partial systematic adaptive cluster sampling without replacement of networks to the population in Table 1

Final sample	P(S)	ν	z_{n1}	z_{n2}	$\hat{\mu}_n$	$\widehat{\text{var}}(\hat{\mu}_n)$	$\hat{\mu}_{wn}$	$\widehat{\text{var}}(\hat{\mu}_{wn})$
60,70,0,1,2;5	1/24	6	393	147	22.50	105.06	28.65	419.64
60,70,0,1,5;2	1/24	6	393	171	23.50	85.56	29.05	377.22
60,70,0,1,6;2,5	1/24	7	393	179	23.83	79.51	29.18	363.15
60,70,0,1,80;2,5,100,6,9	1/24	10	393	851	51.83	364.17	40.38	−691.71
60,70,0,1,7;2,5	1/24	7	393	187	24.17	73.67	29.32	349.11
60,70,0,1,8;2,5	1/24	7	393	195	24.50	68.06	29.45	335.12
60,70,0,1,9;2,5	1/24	7	393	203	24.83	62.67	29.58	321.15
60,70,0,1,100;2,5,80,6,9	1/24	10	393	851	51.83	364.17	40.38	−691.71
2,5,6,80,60;70,0,100,1,9	1/21	10	309	648	39.88	199.52	31.40	−404.54
2,5,6,80,70;60,0,100,1,9	1/21	10	309	648	39.88	199.52	31.40	−404.54
2,5,6,80,0;100,1,9	1/21	8	309	193	20.92	23.36	23.82	153.09
2,5,6,80,1;100,9	1/21	7	309	200	21.21	20.63	23.93	143.64
2,5,6,80,7;100,1,9	1/21	8	309	242	22.96	7.79	24.63	87.51
2,5,6,80,8;100,1,9	1/21	8	309	249	23.25	6.25	24.75	78.25
2,5,6,80,9;100,1	1/21	7	309	256	23.54	4.88	24.87	69.02
2,5,6,80,100;1,9	0							
7,8,9,100,60;70,0,2,5,80,1,6	1/21	12	342	659	41.71	174.46	33.78	−423.81
7,8,9,100,70;60,0,2,5,80,1,6	1/21	12	342	659	41.71	174.46	33.78	−423.81
7,8,9,100,0;80,1,6	1/21	8	342	204	22.75	33.06	26.20	201.94
7,8,9,100,1;80,6	1/21	7	342	211	23.04	29.79	26.32	191.44
7,8,9,100,2;80,1,6	1/21	8	342	218	23.33	26.69	26.43	180.97
7,8,9,100,5;80,1,6	1/21	8	342	239	24.21	18.42	26.78	149.72
7,8,9,100,6;80,1	1/21	7	342	246	24.50	16.00	26.90	139.36
7,8,9,100,80;1,6	0							
Expectation =		8.21			29.00	94.64	29.00	20.12
Bias =					0.00	0.00	0.00	0.00

Final sample lists all units in each initial sample followed by adaptively added units placed after the semicolon. The final sample size is denoted ν. The final rows contain the expectation and bias of the estimator for mean and variance.

mean can be calculated as follows:

$$\begin{aligned}\hat{\mu}_{wu} &= \tfrac{1}{12}\left(342\times\tfrac{4}{5}+154\times\tfrac{1}{5}\right) &= 25.37\\ \hat{\mu}_{wn} &= \tfrac{1}{12}\left(342\times\tfrac{4}{5}+239\times\tfrac{1}{5}\right) &= 26.78\\ \hat{\mu}_{wc} &= \tfrac{1}{12}\left(342\times\tfrac{4}{5}+236\times\tfrac{1}{5}\right) &= 26.73\end{aligned}$$

The following is how to calculate estimates for the population mean by using equal weights for all z_{i*} Equation 2:

$$\begin{aligned}\hat{\mu}_{u} &= \tfrac{1}{12}\times\tfrac{1}{2}(342+154) &= 20.67\\ \hat{\mu}_{n} &= \tfrac{1}{12}\times\tfrac{1}{2}(342+239) &= 24.21\\ \hat{\mu}_{c} &= \tfrac{1}{12}\times\tfrac{1}{2}(342+236) &= 24.08\end{aligned}$$

Because the sampling design must be chosen before sampling begins, only one estimate would be calculated as only one of the three would be appropriate.

As confirmed by examination of estimates from all possible samples, the estimators for mean and variance are unbiased Tables 2–4. For the example population presented in Table 1, PSACS without replacement of clusters is the most efficient design.

4.2. Systematic adaptive cluster sampling example

Suppose a SACS is taken from Table 1, but only a single PSU is selected, and a PSU is defined as the entire row. Each sample will consist of $M_k = 4$ SSUs. In this case where only a single PSU is taken, a researcher might estimate the variance pretending the SSUs came from a

Table 4. All possible samples from applying partial systematic adaptive cluster sampling without replacement of clusters to the population in Table 1

Final sample	P(S)	ν	z_{c1}	z_{c2}	$\hat{\mu}_c$	$\widehat{\text{var}}(\hat{\mu}_c)$	$\hat{\mu}_{wc}$	$\widehat{\text{var}}(\hat{\mu}_{wc})$
60,70,0,1,2;5	0							
60,70,0,1,5;2	0							
60,70,0,1,6;2,5	1/18	7	393	174	23.63	83.27	29.10	371.94
60,70,0,1,80;2,5,100,6,9	1/18	10	393	678	44.63	141.02	37.50	−444.13
60,70,0,1,7;2,5	1/18	7	393	180	23.88	78.77	29.20	361.39
60,70,0,1,8;2,5	1/18	7	393	186	24.13	74.39	29.30	350.87
60,70,0,1,9;2,5	1/18	7	393	192	24.38	70.14	29.40	340.36
60,70,0,1,100;2,5,80,6,9	1/18	10	393	678	44.63	141.02	37.50	−444.13
2,5,6,80,60;70,0,100,1,9	1/15	10	309	528	34.88	83.27	29.40	−268.64
2,5,6,80,70;60,0,100,1,9	1/15	10	309	528	34.88	83.27	29.40	−268.64
2,5,6,80,0;100,1,9	1/15	8	309	203	21.33	19.51	23.98	139.60
2,5,6,80,1;100,9	0							
2,5,6,80,7;100,1,9	1/15	8	309	238	22.79	8.75	24.57	92.81
2,5,6,80,8;100,1,9	1/15	8	309	243	23.00	7.56	24.65	86.19
2,5,6,80,9;100,1	0							
2,5,6,80,100;1,9	0							
7,8,9,100,60;70,0,2,5,80,1,6	1/15	12	342	536	36.58	65.34	31.73	−266.00
7,8,9,100,70;60,0,2,5,80,1,6	1/15	12	342	536	36.58	65.34	31.73	−266.00
7,8,9,100,0;80,1,6	1/15	8	342	211	23.04	29.79	26.32	191.44
7,8,9,100,1;80,6	0							
7,8,9,100,2;80,1,6	1/15	8	342	221	23.46	25.42	26.48	176.49
7,8,9,100,5;80,1,6	1/15	8	342	236	24.08	19.51	26.73	154.17
7,8,9,100,6;80,1	0							
7,8,9,100,80;1,6	0							
Expectation =		8.80			29.00	59.88	29.00	14.56
Bias =					0.00	0.00	0.00	0.00

Final sample lists all units in each initial sample followed by adaptively added units placed after the semicolon. The final sample size is denoted ν. The final rows contain the expectation and bias of the estimator for mean and variance.

SRS as opposed to a systematic sample. An estimator of the Hansen–Hurwitz type for SACS (Thompson, 2002) is simply the average of the w_i in the PSU selected. The estimator for our example is calculated as

$$\hat{\mu}_{\text{SACS}} = \frac{1}{4}\sum_{i=1}^{4} w_i$$

A finite population correction factor (Thompson, 2002) can be used, $\frac{12-4}{12}$. Thus, the variance can be estimated,

$$\widehat{\text{var}}(\hat{\mu}_{\text{SACS}}) = \left(\frac{12-4}{12}\right)\frac{1}{4(4-1)}\sum_{i=1}^{4}(w_i - \hat{\mu}_{\text{SACS}})^2$$

Estimates are biased when an inappropriate estimator is applied. A primary motivation for ACS is the ability to obtain unbiased estimates while sampling nearby units that meet the prespecified condition. As is illustrated in Table 5, when the units nearby are expected to be similar, treating the SSUs in the single PSU as a SRS tends to overestimate the variance (Thompson, 2002). The estimate of variance SACS with a single PSU treated as a SRS overestimated the true variance by about approximately 30-fold, $\frac{272.56}{8.29}$.

5. EMPIRICAL STUDY

In this section, we conduct a simulation to compare the three proposed PSACS design options to existing ACS strategies, SACS and ACS. In all designs, the initial sample is taken by SRS without replacement. Table 6 shows the empirical population consisting of counts of blue winged teals in 252-km units as observed from aircraft Smith *et al.* (1995). In addition, the data from Smith *et al.* (1995) was augmented with 20 rows of zeros at the bottom of the population to create a larger population for comparison. The number of SSUs in Table 6 is 200, whereas in augmented population the number of SSUs is 600. The condition to adaptively add units is $y_i \geqslant 1$. The neighborhood is the four adjacent units.

Tables 7 and 8 contain the definition of the PSUs used for PSACS and SACS with the population in Table 6. The definition of the PSUs for the augmented data was consistent with the definitions in Tables 7 and 8 with the number of PSU augmented accordingly for the additional 20 rows. The PSUs for PSACS designs and SACS design with single PSU selected are defined as an entire strip/row. For PSACS designs, only a single PSU was be selected. The PSUs for SACS design when two PSUs are selected are defined as every other unit. The reason for

Table 5. All possible samples from applying systematic adaptive cluster sampling (SACS) with a single primary sampling unit selected from the population in Table 1

Final sample	P(S)	ν	$\hat{\mu}_{\text{SACS}}$	$\widehat{var}(\hat{\mu}_{\text{SACS}})$	$(\hat{\mu}_{\text{SACS}} - \mu)^2$
60, 70, 0, 1; 2,5	$\frac{1}{3}$	6	32.75	231.15	14.06
2, 5, 6, 80; 1,9, 100	$\frac{1}{3}$	7	25.75	306.26	10.56
7, 8, 9, 100; 6,1, 80	$\frac{1}{3}$	7	28.50	280.28	0.25
Expectation =		6.67	29.00	272.56	8.29

Final sample lists all units in each initial sample followed by adaptively added units placed after the semi-colon. The final sample size is denoted ν. The final rows contain the expectation of the estimator of the mean and variance.

Table 6. Blue wing teal data

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	3	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	1	0
0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	6	6339	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

Table 7. Primary sampling unit layout for partial systematic adaptive cluster sampling designs and systematic adaptive cluster sampling design with a single primary sampling unit selected

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

Table 8. Primary sampling unit layout for systematic adaptive cluster sampling design with two primary sampling units in the sample

1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11	1	11
2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12	2	12
3	13	3	13	3	13	3	13	3	13	3	13	3	13	3	13	3	13	3	13
4	14	4	14	4	14	4	14	4	14	4	14	4	14	4	14	4	14	4	14
5	15	5	15	5	15	5	15	5	15	5	15	5	15	5	15	5	15	5	15
6	16	6	16	6	16	6	16	6	16	6	16	6	16	6	16	6	16	6	16
7	17	7	17	7	17	7	17	7	17	7	17	7	17	7	17	7	17	7	17
8	18	8	18	8	18	8	18	8	18	8	18	8	18	8	18	8	18	8	18
9	19	9	19	9	19	9	19	9	19	9	19	9	19	9	19	9	19	9	19
10	20	10	20	10	20	10	20	10	20	10	20	10	20	10	20	10	20	10	20

the difference is that for SACS, if two PSUs will be selected, then each PSU must be half the size to maintain a similar number of units in the final sample. This follows the concept shown in Section 2 in Figure 1.

One of the benefits of SACS and PSACS is the cost savings over ACS, and cost is often highly correlated with the distance traveled (Morrison *et al.*, 2008). For this reason, distance traveled is included in the simulation results. To calculate distance, data collection starts from the upper-most left-hand corner unit and then goes to the nearest SSU in the initial sample, continuing until the path connects all SSUs in the initial sample. The total path for calculating distance is the path connecting the initial sample plus the path among the adaptively added units. Note that the adaptively added units are adjacent to the units in the initial sample. The calculated path is not always optimal but is a logical approach that mimics movements in field research.

An example from start to finish, including the path assumed to be taken, is given in Table 9. In this example, the five units are traversed from the upper left corner in the first column straight down to the sixth row where the single PSU is sampled, indicated by italicized 1s. Next, the sixth row is in bold 1s for the PSU sampled. Then, in the last (20th) column, on the other end of the sampled PSU, the path continues down three rows to the ninth row containing the SSU sampled. Finally, the path continues to the SSU sampled in the fourth column ninth row, adding another 14 units. The sample yielded 18 adaptively added units. Note that all adaptively units are adjacent. Thus, there are 39 SSUs in the sample, 20 SSUs in the PSU, plus 1 SSU and 18 SSUs adaptively added. The additional units traveled to "connect the dots" is 22 units, $5 + 3 + 14$.Thus, the total distance traveled is the 61 units that the path traverses or equivalently the sum of all the 1s in Table 9.

The simulation results for sampling a single PSU under the SACS design confirmed substantial bias under the assumption that the sample comes from an SRS (Table 10). The $E[\widehat{\text{var}}(\hat{\mu})]$ assuming a SRS was calculated for the two populations as there are only $10 = \binom{10}{1}$ and $30 = \binom{30}{1}$ combinations for the populations investigated. The expected values of the biased variance estimator were 6247.24 and 2082.41 for the blue teal data and the augmented population, respectively. The average estimate from SRS-type estimators of variance were 57% and 42% lower than the true variance, respectively.

The PSACS unequal weighted estimators performed significantly better than their equally weighted counterparts with variances that were over 50% less. Given the latter fact, the next statements on PSACS in this section will be focused on the unequally weighted estimators, Equation 1. The unequally weighted estimators used the weights, $c_1 = \frac{20}{21}$ and $c_1 = \frac{1}{21}$. Either the without replacement of networks or cluster options had the lowest variance compared with without replacement of units. This is to be expected as there is a lower probability and no probability of repeat observations for without replacement of networks and without replacement of clusters, respectively (Salehi and Seber, 1997; Dryver and Thompson, 2007). As the weight on z_{2*}, which utilizes the second observation, is less than 5% in the simulation a large difference is not expected for the unequal weighted estimators. PSACS was much more efficient, lower variance, when taking into

Table 9. The final sample and path assumed taken to travel between selected units

1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	1	1	1
1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1	0	*1*
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*1*
0	0	0	0	**1**	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Travel starts at the upper left corner. Bold '1s' indicate initial sample, regular '1s' indicate adaptively added units, and italicized '1s' indicate units only in travel path. Note that the order of units sampled is not considered, although it technically should be considered for partial systematic adaptive cluster sampling without replacement of networks and partial systematic adaptive cluster sampling without replacement of clusters.

Table 10. Simulation results with 20 000 iterations on the population data sets given in the Table 6 and the augmented data

	Population in Table 6			Augmented population		
	var($\hat{\mu}$)	ν	Distance	var($\hat{\mu}$)	ν	Distance
PSACS $\hat{\mu}_u$ $m = 2$	29192.1	30.3	43.1	11780.4	24.4	51.9
PSACS $\hat{\mu}_n$ $m = 2$	23995.6	30.3	43.2	10654.6	24.4	51.8
PSACS $\hat{\mu}_c$ $m = 2$	23023.3	30.4	43.5	10648.1	24.3	51.9
PSACS $\hat{\mu}_{wu}$ $m = 2$	10647.9	30.3	43.1	4740.5	24.4	51.9
PSACS $\hat{\mu}_{wn}$ $m = 2$	10592.2	30.3	43.2	4645.9	24.4	51.8
PSACS $\hat{\mu}_{wc}$ $m = 2$	10570.4	30.4	43.5	4689.4	24.3	51.9
SACS $n = 1$ $\hat{\mu}_{hh}$	11204.3	28.2	32.7	5053.0	23.0	37.4
SACS $n = 1$ $\hat{\mu}_{ht}$	10668.7	28.2	32.7	4869.8	23.0	37.4
SACS $n = 2$ $\hat{\mu}_{hh}$	5846.3	36.3	54.0	2605.9	26.3	61.4
SACS $n = 2$ $\hat{\mu}_{ht}$	4308.1	36.3	54.0	2225.3	26.3	61.4
ACS $\hat{\mu}_{hh}$ $n = 7$	17983.3	14.8	44.8	6168.7	9.8	70.1
ACS $\hat{\mu}_{ht}$ $n = 7$	16233.4	14.8	44.8	5952.2	9.8	70.1
ACS $\hat{\mu}_{hh}$ $n = 8$	15944.5	16.9	48.5	5454.9	11.2	75.5
ACS $\hat{\mu}_{ht}$ $n = 8$	14088.0	16.9	48.5	5188.1	11.2	75.5
ACS $\hat{\mu}_{hh}$ $n = 9$	14018.0	18.8	51.6	4836.8	12.7	80.5
ACS $\hat{\mu}_{ht}$ $n = 9$	12156.7	18.8	51.6	4709.0	12.7	80.5
ACS $\hat{\mu}_{hh}$ $n = 10$	12364.9	20.7	54.6	4180.0	14.0	85.0
ACS $\hat{\mu}_{ht}$ $n = 10$	10688.8	20.7	54.6	4071.3	14.0	85.0

The variance estimates for the weighted estimator were negative approximately 4% of the time for the nonaugmented population and less than 1% for the augmented population. ACS, adaptive cluster sampling; SACS, systematic ACS; PSACS, partial SACS.

distance traveled compared with ACS. PSACS had a lower variance than SACS with a single sample unit but required a larger distance traveled to obtain the data. Finally, PSACS had a considerably higher variance than SACS with 2 units, but the distance traveled for SACS with two units was considerably higher.

6. DISCUSSIONS AND CONCLUSIONS

In situations where only a single PSU is sampled using SACS, using SRS estimators for the estimating the variance can lead to severely biased variance estimates. For example in Section 4.2, the true variance was approximately 3% of the expected variance estimator when treating the SACS with 1 PSU as an SRS. In the simulation, Section 5, the reverse was found to what was expected and in the example. The expected sample variance calculated treating the single PSU as coming from a SRS was approximately 50% of the true variance. Thus, this method of estimating the variance for when a single PSU is selected for SACS can be significantly biased in either direction, over or under estimate.

The authors can envision a couple of situations when only a single PSU might be advantageous. A researcher does not have sufficient time to sample >1 PSU at a site. Also, when conducting a pilot survey, it might not be desirable to sample >1 PSU. Thus, PSACS sampling only a single PSU and a single SSU may be very desirable. If the researcher encounters a SSU that meets the condition to adapt, then the researcher would need to add neighboring units just as in ACS.

The unequally weighted estimators, Equation 1, in the simulation, Table 10, offered slightly lower variance than a single PSU but higher distance traveled. The additional distance traveled was lower than that of taking two smaller PSUs. PSACS is not optimal in comparison with SACS with two smaller PSUs should funding not be a major constraint or if cost is not a function of distance traveled. Thus, PSACS can be viewed as compromise between taking two smaller PSUs and a single larger PSU when cost is a function of distance traveled. When only a single PSU can be sampled, the PSACS design can be followed by supplementing the single PSU with one or more SSUs. Under this circumstance, the newly proposed sampling strategies are not only worth considering but possibly the only option at present that offers an unbiased estimate of variance. The authors propose weights that are logical but may not be optimal. Alternatively, the second estimate z_{*2}, on the basis of the single SSU, could receive a minuscule weight to obtain an estimator whose variance should be very similar to that of a single PSU. The optimality of the unequally weighted estimator could be topic of further investigation.

Partial systematic adaptive cluster sampling with a single PSU and a single SSU is probably the most optimal tradeoff between cost and an unbiased variance estimator. The researchers do not suggest PSACS with multiple SSUs as the cost of this procedure would quickly be equivalent to SACS with two smaller PSUs. In addition, the unequally weighted estimator uses weights in proportion to the number of the SSUs in each sample selection. Thus, the researcher would have to sample a significant number of SSUs in relation to the size of the single PSU selected to impact the variance. The case where PSACS with multiple secondary units would be of more interest is when no units sampled met the condition. This suggests development of a PSACS design with a stopping rule where sampling continues until selection of a unit that meets the condition to adapt.

Acknowledgements

The authors would like to thank the Graduate School of Business Administration, NIDA, Bangkok, Thailand for their support. The authors would also like to thank the associate editor and referees for their helpful comments and suggestions.

REFERENCES

Dryver AL, Chao C-T. 2007. Ratio estimators in adaptive cluster sampling. *Environmetrics* **18**(6): 607–620.
Dryver A, Thompson S. 2007. Adaptive sampling without replacement of clusters. *Statistical Methodology* **4**(1): 35–43.
Fridman J, Walheim M. 2000. Amount, structure, and dynamics of dead wood on managed forestland in Sweden. *Forest Ecology and Management* **131**(1-3): 23–36.
McDonald L. 2004. *Sampling Rare Populations*. Island Press: Washington, DC.
Morrison LW, Smith DR, Young CC, Nichols DW. 2008. Evaluating sampling designs by computer simulation: a case study with the missouri bladderpod. *Population Ecology* **50**(4): 417–425.
Philippi T. 2005. Adaptive cluster sampling for estimation of abundances within local populations of low-abundance plants. *Ecology* **86**(5): 1091–1100.
Pooler P, Smith D. 2005. Optimal sampling design for estimating spatial distribution and abundance of a freshwater mussel population. *Journal of the North American Benthological Society* **24**(3): 525–537.
Raj D. 1956. Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association* **51**: 269–284.
Salehi M, Seber G. 1997. Adaptive cluster sampling with networks selected without replacement. *Biometrika* **84**: 209–219.
Smith D, Conroy M, Brakhage D. 1995. Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics* **51**(2): 777–788.
Smith DR, Villella RF, Lemarie DP. 2003. Application of adaptive cluster sampling to low-density populations of freshwater mussels. *Environmental and Ecological Statistics* **10**: 7–15.
Thompson S. 1990. Adaptive cluster sampling. *Journal of the American Statistical Association* **85**: 1050–1059.
Thompson S. 1991. Adaptive cluster sampling: designs with primary and secondary units. *Biometrics* **47**(3): 1103–1115.
Thompson S. 2002. *Sampling: Second Edition*. John Wiley & Sons, Inc.: New York, New York.
Wolter KM. 1985. *Introduction to Variance Estimation*. Springer-Verlag: New York, New York.

