19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Multi-species attributes as the condition for adaptive sampling of rare species using two-stage sequential sampling with an auxiliary variable

B.Panahbehagh[a], D. R. Smith[b], M.Salehi M.[c], D. J. Hornbach[d] and J. A. Brown[e]

[a]Department of Mathematics, Isfahan University of Technology, Isfahan, Iran, bardia_p@math.iut.ac.ir

[b]U.S. Geological Survey, Kearneysville, West Virginia, USA, drsmith@usgs.gov

[c]Department of Mathematics, Statistics and Physics, Qatar University Doha, Qatar, salehi@qu.edu.qa

[d]Department of Environmental Studies, Macalester College, St. Paul, Minnesota, USA, hornbach@macalester.edu

[e]Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, j.brown@math.canterbury.ac.nz

Abstract: Assessing populations of rare species is challenging because of the large effort required to locate patches of occupied habitat and achieve precise estimates of density and abundance. The presence of a rare species has been shown to be correlated with presence or abundance of more common species. Thus, ecological community richness or abundance can be used to inform sampling of rare species. Adaptive sampling designs have been developed specifically for rare and clustered populations and have been applied to a wide range of rare species. However, adaptive sampling can be logistically challenging, in part, because variation in final sample size introduces uncertainty in survey planning. Two-stage sequential sampling (TSS), a recently developed design, allows for adaptive sampling, but avoids edge units and has an upper bound on final sample size. In this paper we present an extension of two-stage sequential sampling that incorporates an auxiliary variable (TSSAV), such as community attributes, as the condition for adaptive sampling. We develop a set of simulations to approximate sampling of endangered freshwater mussels to evaluate the performance of the TSSAV design. The performance measures that we are interested in are efficiency and probability of sampling a unit occupied by the rare species. Efficiency measures the precision of population estimate from the TSSAV design relative to a standard design, such as simple random sampling (SRS). The simulations indicate that the density and distribution of the auxiliary population is the most important determinant of the performance of the TSSAV design. Of the design factors, such as sample size, the fraction of the primary units sampled was most important. For the best scenarios, the odds of sampling the rare species was approximately 1.5 times higher for TSSAV compared to SRS and efficiency was as high as 2 (i.e., variance from TSSAV was half that of SRS). We have found that design performance, especially for adaptive designs, is often case-specific. Efficiency of adaptive designs is especially sensitive to spatial distribution. We recommend that simulations tailored to the application of interest are highly useful for evaluating designs in preparation for sampling rare and clustered populations.

Keywords: *Adaptive sampling, sequential sampling, endangered species, environmental monitoring, population estimation*

1. INTRODUCTION

Assessing populations of rare species is challenging because of the large effort required to achieve precise estimates of density and abundance. Adaptive sampling designs have been developed specifically for rare and clustered populations (Thompson 1990) and have been applied to a wide range of rare species including Pacific hake larvae (Lo et al. 1997), rare trees (Magnussen et al. 2005), terrestrial herpetofauna (Noon et al. 2006), subtidal macroalgae (Goldberg et al. 2007), plant disease organisms (Ojiambo and Scherm 2008), red sea urchin (Skibo et al. 2008), and freshwater mussels (Smith et al. 2003, Outiero et al. 2008, Horbach et al. 2010). Compared to conventional sampling designs, adaptive sampling can result in higher efficiency (i.e., higher precision for fixed cost) and higher rates of encountering occupied habitat and detecting rare species (Brown 2003, Smith et al. 2004). Efficiency depends on spatial distribution (Brown 2003), but encounter and detection rates tend to be relatively high with only a modest degree of clustering (Smith et al. 2004). A criticism of adaptive cluster sampling is that the final sample size is random and can be large compared to time allotted to complete the survey.

Two-stage sequential (TSS) sampling (Salehi and Smith 2005, Brown et al. 2008) was developed, in part, to solve one of adaptive cluster sampling's shortcomings – that the final sample size is random and uncontrolled. The TSS design does not require a neighborhood, but is adaptive in the sense that it does allow sampling effort to increase when the target species is encountered. A predetermined value C determines if the condition for sequential sampling is met. The condition is typically based on an attribute of the target species, such as, the count of individuals within a sampling unit.

Conventional and adaptive sampling designs are available that make use of auxiliary information, which can be useful when the target and auxiliary populations are correlated (Thompson 2002). The presence of a rare species, at least in some communities, has been shown to be associated with overall community density (Myers et al. 2000, Hornbach et al. 2010). Community attributes are more readily observable than the presence of a rare species. Thus, the overall community richness or abundance can be used as an auxiliary variable to inform sampling of rare species. The TSS design can incorporate auxiliary information simply by setting the condition C as a function of an auxiliary variable.

In this paper we present an extension of two-stage sequential sampling that incorporates an auxiliary variable (TSSAV), such as the count of the overall community within a sampling unit, as the condition to sequentially increase sampling. We report on a simulation study to examine the general performance of the TSS with auxiliary variables design. The parameters of the simulation study were set to be similar to a study of endangered freshwater mussels in Minnesota USA (Hornbach et al. 2010).

2. METHODS

2.1. TSSAV Design

In the TSS design (Salehi and Smith 2005), sampling occurs sequentially, but only within those areas (i.e., primary sampling units) where the target species was encountered at a level sufficient to satisfy a predetermined condition. Because the TSS design is two-stage, the study area comprised of N units is divided into M large primary units. Each primary unit contains N_i units. At the first stage of sampling, a sample (m) is selected probabilistically. (When $m = M$, then the design is equivalent to a stratified sequential sampling design.) At the second stage, an initial probability sample is taken within each primary unit (n_1), and if the target species is encountered then a sequential sample of pre-determined size (n_2) is taken. Here we consider only the designs where sample sizes are constant within each primary unit, i.e., $n_{i1} = n_1$ and $n_{i2} = n_2$ for all i from 1 to M, although the design can allow sample size to vary among primary units (Salehi and Smith 2005).Thus, the final sample size (η) is random, but it is controlled to be within the range of $m^* n_1 \leq \eta \leq m^*(n_1+ n_2)$.

Let the count of individuals of the target population in the i^{th} primary unit and j^{th} sampling unit be denoted by y_{ij}. Similarly, let the count of individuals of the auxiliary population in the i^{th} primary unit and j^{th} sampling unit be denoted by x_{ij}. To incorporate auxiliary information into the design (TSSAV), the condition to sequentially sample, C, can be a function of the auxiliary variable x_{ij}. For example, the condition C could be satisfied when $x_{ij} \geq C$ for any sampling unit within the i^{th} primary unit where C is a positive integer. If at least one sampling unit satisfies the condition then a sequential sample of size n_2 is taken – this is determined within each primary unit independently. Let l_i be the number of sampling units in the final sample within the i^{th} primary unit. The estimators for TSS presented in Salehi and Smith (2005) also hold for TSSAV with the exception that l_i is a function the auxiliary variable rather than the target variable.

2.2. Estimators

Salehi and Smith (2005) presented an unbiased estimator for the two-stage sequential sampling based on Murthy's estimator (Murthy 1957, Salehi 2001). The estimator for population total in i^{th} primary unit (τ_i) is

$$\hat{\tau}_i = \sum_{j \in s_i} \frac{P(s_i|j)}{P(s_i)} y_{ij}$$

where s_i is the set of units selected in the i^{th} primary unit, $P(s_i)$ is the probability of selecting the sample s_i, $P(s_i|j)$ is the conditional probability of selecting the sample s_i given the j^{th} unit was selected first in the i^{th} primary unit, and y_{ij} is the count of individuals of the target population in the i^{th} primary unit and j^{th} sampling unit, The probabilities of obtaining s_i depends on the number of units that satisfy the condition within a primary unit. Let l_i denote the number of units in the i^{th} primary unit that satisfy the condition C. In the TSSAV design, l_i is a function of the auxiliary variable. The ratio of $P(s_i|j)/P(s_i)$ is given by

$$\frac{P(s_i|j)}{P(s_i)} = \begin{cases} \dfrac{N_i}{n_1} & n_2 = 0 \\ \dfrac{N_i}{n_1 + n_2} & n_2 > 0 \text{ and } l_i < n_2 \\ \dfrac{N_i(n_1+n_2-1)!}{(n_1+n_2)! - \frac{n_2!(n_1+n_2-l_i)!}{(n_2-l_i)!}} & n_2 > 0 \text{ and } l_i \le n_2 \text{and } j \text{ satisfies } C \\ \dfrac{N_i\{(n_1+n_2-1)! - n_2!\,(n_1+n_2-1-l_i)!/(n_2-l_i)!\}}{(n_1+n_2)! - \frac{n_2!(n_1+n_2-l_i)!}{(n_2-l_i)!}} & n_2 > 0 \text{ and } l_i \le n_2 \text{and } j \text{ not satisfy } C \end{cases}$$

Variance estimators depend on joint probabilities of selecting units j and j' in the first draw (Salehi and Smith 2005). Because selection of primary units and subsampling are conducted independently, population total for the study area is estimated using standard estimators, e.g., Horvitz-Thompson estimators (Thompson 2002, Salehi and Smith 2005).

2.3. Simulation Study

We set up a series of simulation case studies using estimates of endangered freshwater mussels reported by Hornbach et al. (2010) as a basis for target population densities (Table 1). Endangered species of freshwater mussels are often found at low density and spatially clustered (Smith et al. 2003, Hornbach et al. 2010). Rarity is thought to be ≤0.10 m^{-2} (Green and Young 1993). Hornbach et al. (2010) found densities of endangered species to be in the order of 0.01 m^{-2}. Expected density for the auxiliary population was 1 and 3 m^{-2}, and correlation between counts of target and auxiliary populations was 0.33 and 0.67 (Table 1).Figure 1 shows the populations for E(ρ) = 0.67. Primary unit sampling fraction (*m/M*) ranged from 0.25 to 1.0 by 0.25. Initial and sequential sampling fractions (i.e., n_1/N and n_2/N, respectively) ranged from 0.1 to 0.4 by 0.1. Condition to sequentially sample (*C*) ranged from 1 to 5 by 1. Each case was replicated 1,000 times, and average estimates were considered to be estimates of expected values.

We used the software program SAMPLE that was designed for simulating and

Table **1**. Parameters for the simulation case studies. Density (no. m^{-2}) is denoted by μ, coefficient of variation by CV, occupancy (i.e., proportion of units occupied) by ψ, and expected value by E(•). The target population parameters are denoted by subscript Y, and the auxiliary population parameters by subscript X. The correlation between the target and auxiliary populations is denoted by ρ.

$E(\mu_Y)$	$E(CV_Y)$	$E(\psi_Y)$	$E(\mu_X)$	$E(CV_X)$	$E(\psi_X)$	ρ
0.01	11.04	0.011	1.0	1.98	0.40	0.33
0.01	11.04	0.011	1.0	1.98	0.40	0.67
0.01	11.04	0.011	3.0	1.31	0.79	0.33
0.01	11.04	0.011	3.0	1.31	0.79	0.67
0.10	3.93	0.085	1.0	1.98	0.40	0.33
0.10	3.93	0.085	1.0	1.98	0.40	0.67
0.10	3.93	0.085	3.0	1.31	0.79	0.33
0.10	3.93	0.085	3.0	1.31	0.79	0.67

comparing sampling designs, particularly adaptive designs (Morrison et al. 2008, Smith et al. 2011). Each design was simulated 1,000 times. SAMPLE accepts sample or population data, depending on whether a simulation or an analysis is desired. The software can be downloaded at https://profile.usgs.gov/drsmith. Correlated counts variables were generated using the *corcounts* package for *R.* available on the CRAN home page (*http://lib.stat.cmu.edu/R/CRAN/*). We generated the spatial pattern following the Poisson cluster process (Brown, 2003). The number of clusters was selected from a Poisson distribution, and cluster centers were randomly located throughout the site. Individuals within the cluster were located around the cluster center at a random distance following an exponential distribution and a random direction following a uniform distribution. Expected cluster size was 100 with an expected radius of 1.

To evaluate design performance, we used measures of design efficiency, probability of sampling an occupied unit, and robustness to model inaccuracy. Efficiency is the ratio of variance from a simple random sampling design to variance from the candidate design with final sample size equal among the two designs. Final sample size is fixed for conventional designs, but is random in adaptive designs. Thus, for adaptive designs the expected sample size was the average of final sample sizes over the 1,000 simulations. Coefficient of variation (CV) was another measure of efficiency and precision. CV is the ratio of standard error (SE) over density, where SE was the standard deviation of the density estimate.

The probability of sampling an occupied unit was measured by relative risk and odds ratio. Relative risk is the ratio of the proportion of occupied units in the final sample (p_1) relative to that of the population (p_2). Thus, relative risk of sampling an occupied unit = p_1/ p_2. The success of encountering occupied units can also be measured using odds ratios (Agresti 1990), where $\text{odds ratio} = \text{relative risk} \times \frac{(1-p_2)}{(1-p_1)}$.

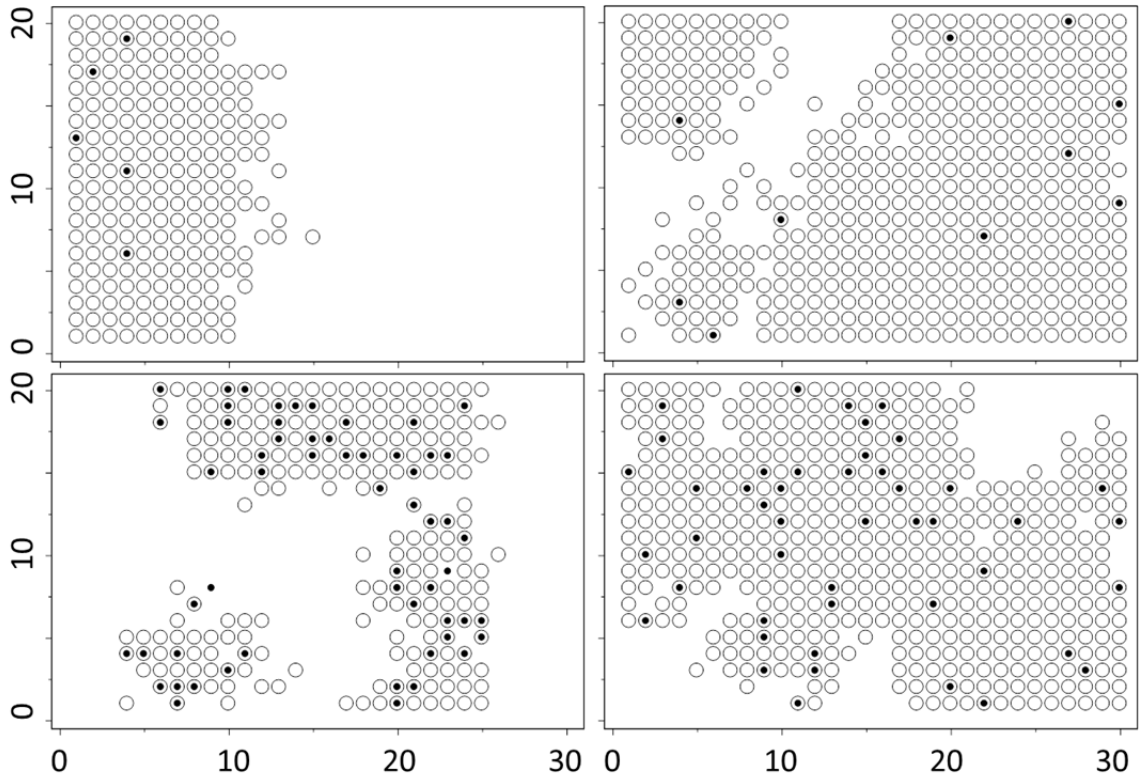

Figure 1. Four of the 8 populations used as simulation case studies. Empty circles denote units occupied by members of the auxiliary population. Filled circles denote units occupied by members of the target population. Count is ≥ 1 in an occupied unit. Populations on the left are for moderate auxiliary population density ($E[\mu_X]$= 1.0), and populations on the right are for high auxiliary population density ($E[\mu_X]$= 3.0). Populations on the top are for very rare target population density ($E[\mu_Y]$= 0.01), and populations on the bottom are for moderately rare target population density ($E[\mu_Y]$= 0.10). Expected correlation between target and auxiliary populations was 0.67 for all 4 populations.

We used Random Forests (Breiman 2001) to evaluate the relative importance of population characteristics and design factors on estimator efficiency and probability that a sampling a unit is occupied by the target population. Population characteristics included density of the target and auxiliary populations and correlation between those populations. Design factors included n_1, n_2, m_1, and criteria to adapt. Probability of sampling an occupied unit was measured by odds ratio. Factors were ranked based on increased node purity, a measure of factor importance, as reported from the R package 'randomForest' (Liaw and Wiener 2002).

3. RESULTS

Efficiency tended to be > 1 consistently across sampling fraction only for moderate auxiliary population density ($E[\mu_X]$ = 1 m^{-2}) and low target population density ($E[\mu_Y]$ = 0.01 m^{-2}; Figure 2). When expected auxiliary population density was 1 m^{-2}, but expected target population density was 0.10 m^{-2}, then efficiency tended to remain ≤ 1, except for final sample size > 300 (sampling fraction > 0.50). Correlation had a modest effect; the correlation = 0.67 tended to be associated with higher efficiency than for correlation = 0.33. When expected auxiliary population density was 3m^{-2}, tended to remain ≤ 1 regardless of target population density or correlation.

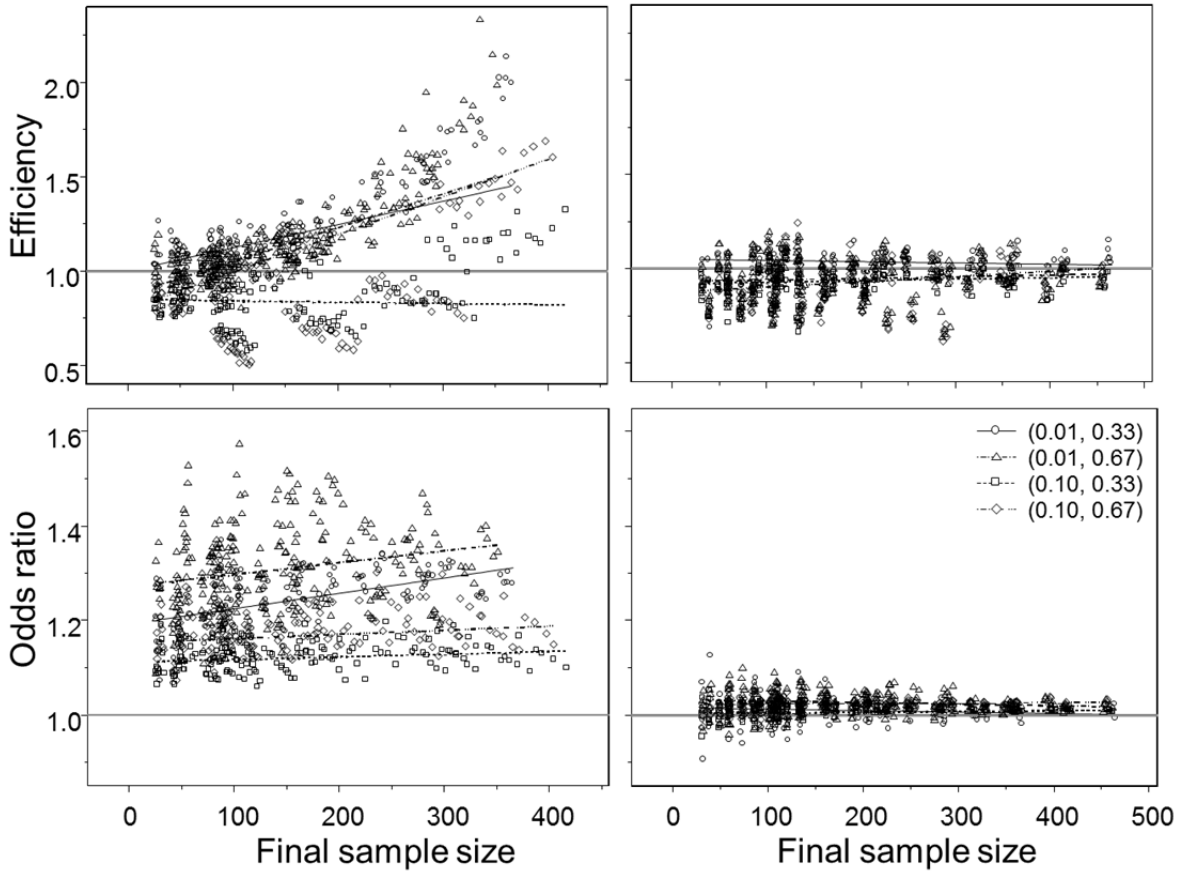

Figure 2.Efficiency and odds ratio across final sample size for the simulation cases.Efficiency is the variance from TSSAV sampling relative to variance from simple random sampling. Odds ratio is for the odds of sampling an occupied unit by TSSAV sampling compared to simple random sampling. Graphs on the left are for moderate auxiliary population density and occupancy ($E[\mu_X]$= 1.0; $E[\psi_X]$ = 0.40), and graphs on the right are for high auxiliary population density and occupancy ($E[\mu_X]$= 3.0; $E[\psi_X]$ = 0.79). Symbols and lines differ by target population density and correlation between the auxiliary and target population. (Density and correlation are shown in parentheses in the lower-right graph legend.)

The probability of encountering units occupied by the target population (as measured by odds ratio) was uniformly > 1 for moderate auxiliary population density (Figure 2). Odds ratio was higher for expected target density of 0.01 m^{-2} than for 0.10 m^{-2} and for correlation of 0.67 than for 0.33. When expected auxiliary population density was 3 m^{-2}, odds ratio tended to be at or slightly above 1 and tended to remain above 1 as final sample size increased.

Auxiliary population density was the most important factor for determining both efficiency and probability of sampling an occupied unit (Table 2). Target population density and primary unit sample size were nearly equally important factors for determining efficiency. Initial sample size was moderately important for determining efficiency. However, auxiliary population density was overwhelming the most important factor determining probability of sampling an occupied unit.

Table **2**. Node purity, which is a direct measure of factor importance, from a Random Forests analysis implemented using the R package 'randomForest' (Liaw and Wiener 2002).

Factor	Efficiency	Odds ratio
Auxiliary population density (μ_X)	13.45	16.28
Target population density (μ_Y)	12.66	2.38
Correlation (ρ)	1.46	0.68
Primary sampling unit sample size (m_1)	12.64	0.21
Initial sample size (n_1)	8.55	0.50
Sequential sample size (n_2)	3.55	1.16
Criteria to adapt	1.05	0.26

4. CONCLUSIONS AND RECOMMENDATIONS

Sample designs, such as probability proportional to size and stratified sampling, have been available for incorporating an auxiliary variable that is correlated with the variable of interest to inform sampling (Thompson, 2002; LeLay et al., 2010; Smith et al., 2011). However, these designs assume *a priori* knowledge of the auxiliary variable. The design that we present and evaluate, TSSAV, incorporates observations of the auxiliary variable to inform and adapt sampling as the survey progresses. This can be an advantage when prior information is not available, but there is reason to believe that the auxiliary and target variables are correlated. The auxiliary variable in the TSSAV design should be more readily or cheaply observed than the target variable. Otherwise there would be no advantage over using the target variable itself for adaptively adding units. In the case of sampling freshwater mussels, using a community attribute (e.g., count of all species) ensures that the auxiliary variable will be readily observed in the initial sample size while the rare species would be infrequently observed among the $n_1$units. However, the probability of detecting the rare species increases with the addition of the sequential sample (n_1+n_2).

We developed a set of simulation case studies to approximate sampling of endangered species of freshwater mussels (Hornbach et al., 2010), and results are conditional on this set. However, the simulation results should apply to any populations with similar densities, abundances, and correlation. The results of the simulation indicate the following:

- The density and distribution of the auxiliary population is the major determinant of the performance of TSSAV in terms of efficiency and probability of sampling units occupied by the target population. TSSAV is a good candidate design when the auxiliary population density and distribution are moderate.
- The density and distribution of the target population is an important determinant of the efficiency of TSSAV. As with other adaptive designs, efficiency is linked to rarity and spatial clustering. Efficiency can be expected to tend higher for rarer target populations, all else being equal.
- Correlation between auxiliary and target populations in the range of $0.33 \leq \rho \leq 0.67$ was less important than expected at determining the performance of TSSAV. However, higher correlation is expected to increase efficiency and probability of sampling occupied units.
- Of the design factors, the most important was the sample size for the primary units. As with conventional two-stage sampling, efficiency improved as the sampling fraction for primary units increased.
- A broader set of recommendations for the performance of TSSAV will require a wider evaluation of population characteristics and design factors than presented here. We have found that design performance, especially for adaptive designs, is often case-specific. Efficiency is especially sensitive to spatial distribution. We recommend that simulations tailored to the application of interest are highly useful for evaluating designs in preparation for sampling rare and clustered populations.

ACKNOWLEDGMENTS

We thank Doug Nichols for programming support. John Young and Arthur Dryver provided helpful comments on an earlier draft of the manuscript.

REFERENCES

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*,16, 199-231.

Breiman, L., Friedman,J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California.

Brown, J.A., and Manly, B.J.F. (1998). Restricted adaptive cluster sampling. *Environmental and Ecological Statistics*, 5, 49-63.

Brown, J. (2003). Designing an efficient adaptive cluster sample. *Environmental and Ecological Statistics*, 10, 95-105.

Brown, J.A., Salehi, M.M., Moradi, M., Bell, G., and Smith, D.R. (2008). Adaptive two-stage sequential sampling. *Population Ecology*, 50, 239–245.

Goldberg, N.A., Heine, J.N., and Brown, J.A. (2007). The application of adaptive cluster sampling for rare subtidal macroalgae. *Marine Biology*, 151, 1343-1348.

Green, R.H., and Young, R.C. (1993). Sampling to detect rare species. *Ecological Applications*, 3, 351-356.

Hornbach, D.J., Hove, M.C., Dickinson, B.D., MacGregor, K.R., and Medland, J.R. (2010). Estimating population size and habitat associations of two federally endangered mussels in St. Croix River, Minnesota and Wisconsin, USA. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 20, 250-260.

LeLay, G., Engler, R., Franc, E., and Guisan, A. (2010). Prospective sampling based on model ensembles improves the detection of rare species. *Ecography*,33, 1015-1027.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R News 2(3), 18-22.

Lo, N.C.H., Griffith, D., and Hunter, J.R. (1997). Using a restricted adaptive cluster sampling to estimate Pacific hake larval abundance.*California Cooperative Oceanic Fisheries Investigations Reports*,38, 103-113.

Magnussen, S., Kurz, W., Leckie, D.G., and Paradine, D. (2005).Adaptive cluster sampling for estimation of deforestation rates. *European Journal of Forest Research*,124, 207-220.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement, *Sankhya*, 18, 379–390.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonsesca, G.A.B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403, 853-858.

Noon, B.R., Ishwar, N.M., and Vaudevan, K. (2006). Efficiency of adaptive cluster and random sampling in detecting terrestrial herpetofauna in a tropical rainforest. *Wildlife Society Bulletin*,34, 59-68.

Ojiambo, P., and Scherm, H. (2010). Efficiency of adaptive cluster sampling for estimating plant disease incidence. *Phytopathology*, 100, 663-670.

Outeiro, A., Ondina, P., Fernandez,C., Amaro,R., and Miguel.E.S. (2008). Population density and age structure of the freshwater pearl mussel, *Margaritifera margaritifera*, in two Iberian rivers. *Freshwater Biology,*53, 485-496.

Salehi, M. M. (2001), A new proof of Murthy's estimator which applies to sequential sampling, *Australian & New Zealand Journal of Statistics*, 43, 901–906.

Salehi, M. M., and Smith, D.R.(2005). Two-stage sequential sampling: A neighborhood-free adaptive sampling procedure.*Journal of Agricultural, Biological, and Environmental Statistics*,10, 84-103.

Skibo, K.M., Schwarz, C.J., and Peterman, R.M. (2008). Evaluation of sampling designs for red sea urchins *Strongylocentrotus franciscanus* in British Columbia. *North American Journal of Fisheries Management,* 28, 219-230.

Smith, D.R., Brown, J.A., and Lo, N.C.H. (2004). Application of adaptive cluster sampling to biological populations. Pages 77-122 In Thompson, W.L. editor. Sampling rare or elusive species: Concepts, designs, and techniques for estimating population parameters. Island Press, Covelo, CA.

Smith, D.R., Lei, Y., Walter, C.A., and Young, J.A. (2011). Incorporating predicted species distribution in conventional and adaptive sampling designs. Pages __-__ In Gitzen, R.A., editor, Design and analysis of long-term ecological monitoring studies. Cambridge, UK

Smith, D.R., Villella, R.F., and Lemarié.D.P. (2003). Application of adaptive cluster sampling to low-density populations of freshwater mussels. *Environmental and Ecological Statistics*, 10, 7-15.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of American Statistical Association*, 85, 1050-1059.

Thompson, S.K. (2002). Sampling. Wiley: New York.