70 Simulation of long-term monitoring sample designs in Denali National Park

TRENT MCDONALD, West, Inc., 2003 Central Avenue, Cheyenne, Wyoming 82001; tmcdonald@west-inc.com

CAROL ROLAND, Denali National Park and Preserve, P.O. Box 9, Denali Park, Alaska 99755-0009

JESSICA FRIED, West, Inc., 2003 Central Avenue, Cheyenne, Wyoming 82001

SARA GOEKING, Denali National Park and Preserve, P.O. Box 9, Denali Park, Alaska 99755-0009

KAREN OAKLEY, U.S. Geological Survey Biological Resources Division, Alaska Biological Science Center, 1011 East Tudor Road, Anchorage, Alaska 99503

Introduction

A park-wide sampling strategy for the long-term ecological monitoring (LTEM) program for Denali National Park and Preserve is presently in the design stages. The goal of this monitoring program is to watch various ecological resources in the park to better understanding the current status of the resource and direction of trends. Part of the design process for the Denali monitoring project involves discussions of appropriate sampling designs for various resources. For vegetation resources, several sample designs have been discussed among project participants, including stratified random sampling and systematic sampling. In the course of these discussions, questions have arisen regarding the feasibility and accuracy of the systematic sampling approach applied over an area as large as Denali.

In this paper we explore some of the questions surrounding systematic sampling of the park by presenting the results of a computer exercise designed to mimic vegetation sampling in the park. In this exercise, we construct a realistic representation of a vegetation parameter (basal area of white spruce, *Picea glauca)* over the entire park and sample that parameter using a systematic grid of points. Once this hypothetical population of basal area is sampled, statistical estimators for the total and mean of the parameter are computed, and the sampling is repeated. In the end, this computer exercise allows us to draw conclusions about the statistical validity and accuracy of the systematic sample design by summarizing the variance and bias of our estimators. We also investigate grid spacing and its effects on variance and bias of the estimators.

Methods

Our goal was to construct a reasonably realistic representation of a sampling scheme, and replicate it a large number of times to verify and assess the design's statistical properties. Long-run averages, bias, and variances of proposed estimators could be assessed in this simulation because true underlying quantities were known. We choose to focus our simulation efforts on a single vegetation resource, basal area of white spruce, because it is an important attribute of the vegetation structure in the park, and because some information on basal area in Denali was known. Basal area was also indicative of a large number of variables present in the park because it exhibited typical patterns in its distribution.

Our simulation to assess properties of a systematic sample design can be outlined as follows: (1) a reasonably realistic map of basal area for the entire park was constructed, (2) the map of basal area was sampled using a randomly placed systematic

From **Crossing Boundaries in Park Management: Proceedings of the 11th Conference on Research and Resource Management in Parks and on Public Lands,** edited by David Harmon (Hancock, Michigan: The George Wright Society, 2001).

grid, (3) sample estimates were computed and stored, (4) steps 2 and 3 were repeated a large number of times, and (5) bias, variance, and confidence interval coverage was computed. Details of each step follow.

Step 1 — Construction of the basal area map

A grid of 2,355,882 points, spaced 100 m x 100 m apart and large enough to span the park, was defined. For purposes of the simulation, artificial basal area values were assigned to each point in this 100-m x 100-m grid. Artificial basal area values were assigned to each location by randomly sampling from a *mixture distribution* that was chosen to approximate the perceived distribution of basal area in the park. These mixture distributions appear in Figure 70.1. The general mixture distribution shape was bi-modal, with one mode near zero and another at larger values. The relative size and placement of each mode varied according to elevation and slope. Once generated, this grid of 2.35 million locations and associated basal area values was viewed as the sample *universe* or "truth." Expected values of sample quantities (computed during the simulation) were compared to "true" basal area quantities of this map.

Step 2 — Simulated sampling

Sample grids of various sizes were defined and randomly placed over the larger 100-m x 100-m grid of basal area constructed in step 1. At each sample location, basal area was noted and the resulting list of basal areas values from all grid points constituted one sample. Sample grid spacings were 20 km, 17.5 km, 15 km, 12.5 km, 10 km, 7.5 km, 5 km, 3.5 km, and 2 km. Due to irregularities in the border of Denali National Park, the number of sample grid points inside the park border varied across random placements of the sample grid.

Step 3 — Sample calculations

Sample quantities of interest were calculated for each random placement of the sample grid. Let the number of points in the 100-m x 100-m basal area map defined in step 1 be *N* (*population size*). Let the number of points that fell inside the boundary of the park from the ***i***-th random placement of the sample grid defined in step 2 be n_i *(sample size)*. Let x_{ij} be the ***j***-th basal area value of the grid sample obtained from the ***i***-th random placement of the sample grid. For each random grid placement (***i*** = 1, …, 500), mean basal area was estimated as:

$$\bar{x}_i = \sum_{j=1}^{n_i} x_{ij} / n_i \, .$$

The estimated standard error of mean basal area was computed as:

$$s_i = \sqrt{\left(\frac{N - n_i}{N} \right) \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i - 1}} \, .$$

A 95% confidence interval for the true basal area was computed as:

$$\bar{x}_i \pm 1.96 s_i \,.$$

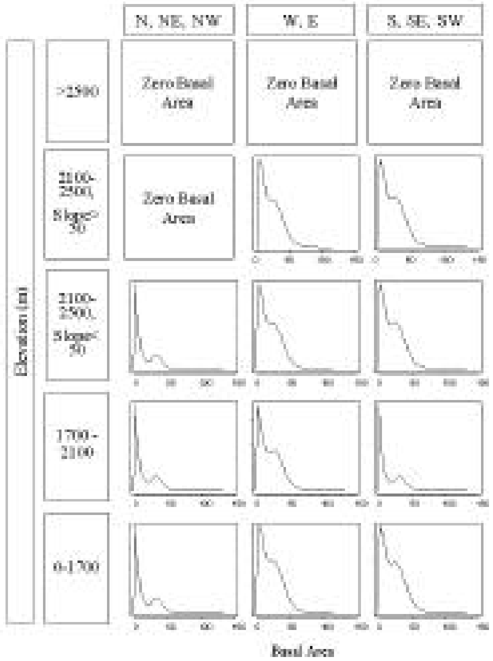

Figure 70.1. Mixture distribution used to generate basal area values in the Denali sampling simulation. Vertical axes plot relative frequency.

Step 4 — Iteration

Steps 2 and 3 were repeated 500 times. Each repeat of steps 2 and 3 defined a single *iteration* of the simulation. Sample quantities from each iteration (step 3) were stored for later summarization.

Step 5 — Summarization

Bias, variance, and *root mean squared error (RMSE)* were assessed for all sample quantities of interest. Let μ represent the true average basal area on the 100-m x 100-m map defined during step 1. Let

$$\bar{x}_{\cdot} = \sum_{i=1}^{500} \bar{x}_i / 500$$

and

$$s_{\bar{x}}{}^2 = \sum_{i=1}^{500} (\bar{x}_i - \bar{x}_.)^2 / 499$$

be the simulated expected value and variance of the mean basal area estimator. The expected value and standard error of the variance estimate were:

$$s_. = \sum_{i=1}^{500} s_i / 500$$

and

$$s_s{}^2 = \sum_{i=1}^{500} (s_i - s_.)^2 / 499 \, .$$

Bias in the estimator of mean basal area was computed as:

$$b = \bar{x}_. - \mu \, .$$

Estimated *RMSE* of mean basal area was calculated as:

$$RMSE_x = \sqrt{b^2 + s_{\bar{x}}{}^2} \, .$$

Bias in the estimate of variance was computed as:

$$b_s = s_. - \sqrt{s_{\bar{x}}{}^2} \, ,$$

and the *RMSE* of the standard error estimator was:

$$RMSE_s = \sqrt{b_s{}^2 + s_s{}^2} \, .$$

Coefficients of variation (CV) for both the estimators and observed sample size were computed as the standard deviation divided by expected value. For example, CV of the mean estimator was computed as:

$$\sqrt{s_{\bar{x}}{}^2} / \bar{x}_. \, .$$

Coverage of the sample confidence interval was computed as the proportion of confidence intervals (out of 500) that contained the true mean basal area. Coverage of the confidence interval was:

$$c = \sum_{i=1}^{500} I_i / 500$$

where I_i was an indicator function that took on a value of 1 if the confidence interval from iteration *i* contained the true mean, and 0 otherwise. Theory holds that *c* should equal 0.95 for confidence intervals with nominal coverage of 95%.

Lower values of *RMSE* were considered better than higher values of because *RMSE* is a function of both variance and bias. For example, an unbiased estimator with large variance might have *RMSE* equal to a biased estimator with small variance. Prior to simulation, it was acknowledged that *RMSE* generally decreases as sample size increases, but it was of particular interest to note whether a large gain in *RMSE* was obtained by any one grid spacing. If so, this grid spacing would be considered for implementation. Confidence interval coverage was assessed the same way as *RMSE*. It was of interest to note whether or not a large improvement in confidence interval coverage was obtained by a single grid spacing.

Results

The average number of grid points inside Denali was 53.8 for the 20-km grid, 74.4 for the 17.5-km grid, 102.5 for the 15-km grid , 143.7 for the 12.5-km grid, 227.2 for the 10-km grid, 414.8 for the 7.5-km grid, 930.5 for the 5-km grid, 1915.9 for the 3.5-km grid, and 5868.7 for the 2.5-km grid. The standard error of sample size as a function of grid size is plotted in Figure 70.2. Variability in sample size ranged from 15 for the 2.5-km grid to 2 for the 20-km grid. The CV of sample size (i.e., average sample size divided by its standard error) ranged from 0.2% for the 2.5-km grid to 3.7% for the 20-km grid.

Bias and standard error of both the mean and standard error estimator is plotted in Figure 70.3. Bias in both the mean and standard error estimator was small for all sizes of grids. *RMSE* for both the mean and standard error estimators are plotted in Figure 70.4. *RMSE* of both estimators increased as grid size increased and as sample size decreased. No large gains, or "jumps," in performance of either estimator were apparent as grid size increased. CV of the mean estimator was remarkably small even for smaller sample sizes. CV of the mean estimator for the 20-km grid was 12.3%. CVs for denser grids were all less than 11%.

Coverage of the 95% confidence intervals is plotted as a function of grid size in Figure 70.5. Coverage of the confidence intervals ranged from 0.93 to 0.97, with average coverage across grids equal to 0.948.

Conclusions

The systematic sample design proved to be a useful design for sampling the artificial population constructed here. Bias in the estimate of mean basal area was negligible and variation in the estimator was relatively small for all grids. Coverage of the sample confidence intervals was adequate for all grid sizes. The resources required to sample the park were relatively constant and predictable because variation in the number of sample points was less than 4% for all grid sizes. We hypothesize that use of a systematic grid in a real study of Denali National Park and Preserve would yield highly accurate and precise estimates of white spruce basal area and other parameters that behave similarly to basal area.

No obvious "jumps" in precision of the mean estimator were evident that might aid choice of a particular grid size. The plot displayed in Figure 70.4(b) shows a very slight "break" in the *RMSE* of the standard error estimator between the 5- and 7.5-km grids, and between the 12.5- and 15-km grids; however, these breaks are not prominent enough to influence large-scale management decisions. Choice of a particular grid spacing for a real study of Denali will likely rely heavily on logistic and budgetary considerations.

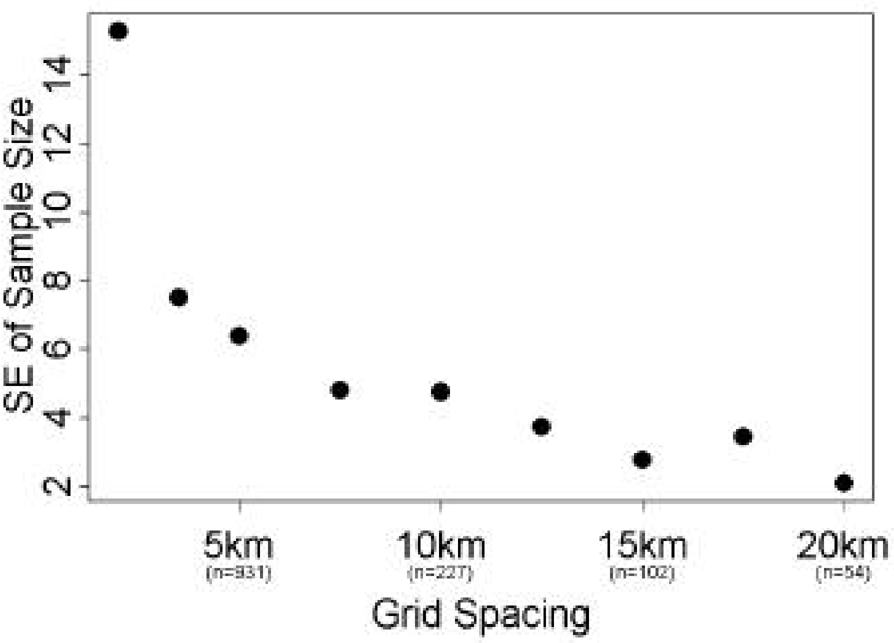

Figure 70.2. Variability of observed sample size (n) as a function of grid size.

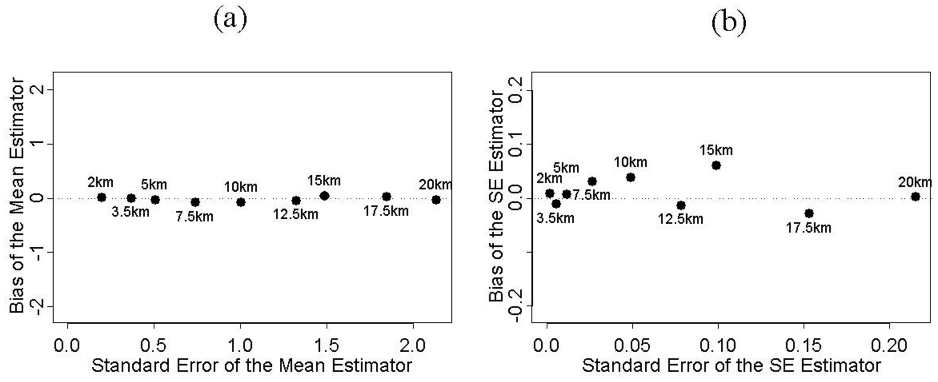

Figure 70.3. Bias and standard error of the mean (a) and standard error (b) estimators. Point labels denote sample grid spacing. RSME is distance from the origin to each point.

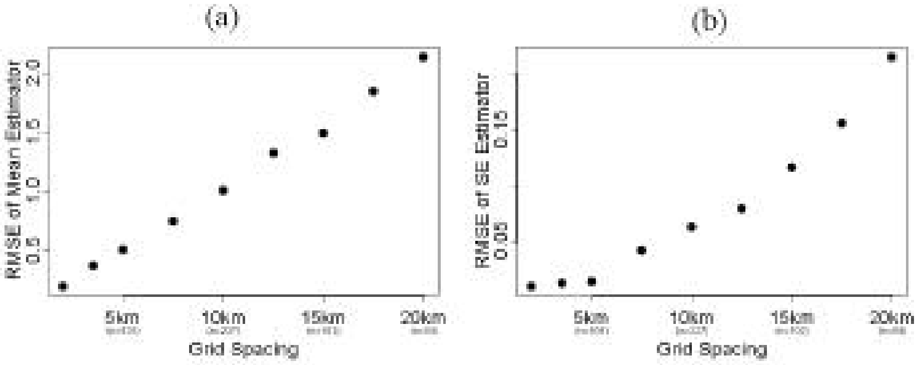

Figure 70.4. RMSE of the mean (a) and standard error (b) estimators as a function of grid size.

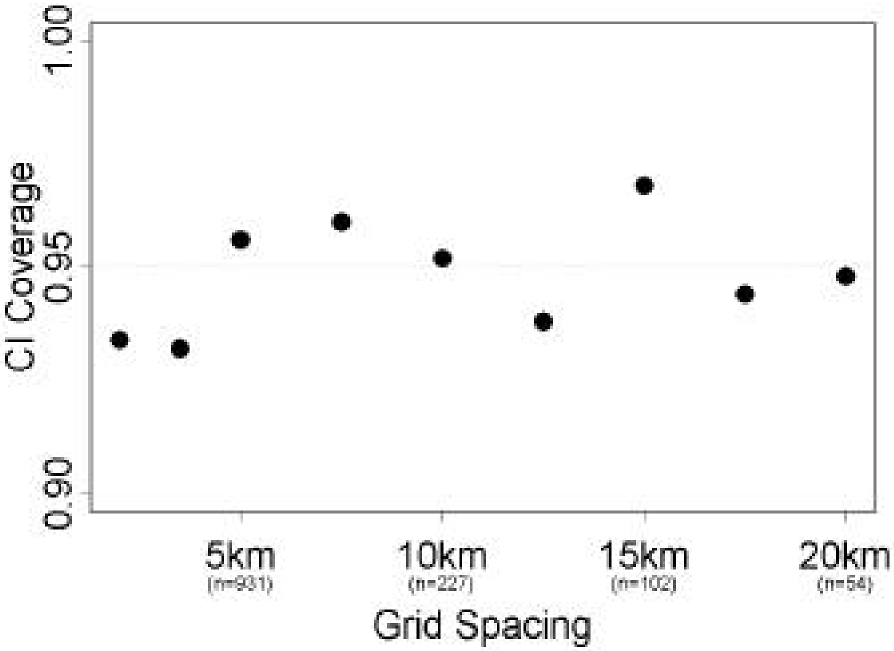

Figure 70.5. Estimates of confidence interval coverage as a function of grid size.

The results of this simulation apply to estimation of the parkwide mean of a parameter that behaves like basal area. Performance of a systematic grid for other parameters that do not behave like basal area remains unknown. Performance of the grid is also unknown for non-mean estimators such as regression, analysis of variance, principal components, etc. In addition, if estimates of the mean are sought for subsections of the park, precision (i.e., variance) will likely suffer due to reduced sample sizes in those regions.