Evaluating probability sampling strategies for estimating redd counts: an example with Chinook salmon (*Oncorhynchus tshawytscha*)

Jean-Yves Courbois, Stephen L. Katz, Daniel J. Isaak, E. Ashley Steel, Russell F. Thurow, A. Michelle Wargo Rub, Tony Olsen, and Chris E. Jordan

Abstract: Precise, unbiased estimates of population size are an essential tool for fisheries management. For a wide variety of salmonid fishes, redd counts from a sample of reaches are commonly used to monitor annual trends in abundance. Using a 9-year time series of georeferenced censuses of Chinook salmon (*Oncorhynchus tshawytscha*) redds from central Idaho, USA, we evaluated a wide range of common sampling strategies for estimating the total abundance of redds. We evaluated two sampling-unit sizes (200 and 1000 m reaches), three sample proportions (0.05, 0.10, and 0.29), and six sampling strategies (index sampling, simple random sampling, systematic sampling, stratified sampling, adaptive cluster sampling, and a spatially balanced design). We evaluated the strategies based on their accuracy (confidence interval coverage), precision (relative standard error), and cost (based on travel time). Accuracy increased with increasing number of redds, increasing sample size, and smaller sampling units. The total number of redds in the watershed and budgetary constraints influenced which strategies were most precise and effective. For years with very few redds (<0.15 redds·km^{-1}), a stratified sampling strategy and inexpensive strategies were most efficient, whereas for years with more redds (0.15–2.9 redds·km^{-1}), either of two more expensive systematic strategies were most precise.

Résumé : La gestion des pêches requiert comme outils essentiels des estimations précises et non faussées de la taille des populations. Chez une grande diversité de poissons salmonidés, le décompte des frayères dans un échantillon de sections de cours d'eau sert couramment à suivre les tendances annuelles d'abondance. En utilisant une série chronologique de 9 années d'inventaires géoréférencés de frayères de saumons chinook (*Oncorhynchus tshawytscha*) dans le centre de l'Idaho, É.-U., nous testons une gamme étendue de stratégies courantes d'inventaire pour l'estimation de l'abondance totale des frayères. Nous évaluons deux tailles d'unités d'échantillonnage (sections de 200 et de 1000 m), trois proportions échantillonnées (0,05, 0,10 et 0,29), ainsi que six stratégies d'échantillonnage (échantillonnage par indices, échantillonnage aléatoire simple, échantillonnage systématique, échantillonnage stratifié, échantillonnage par groupements adaptatifs et un plan équilibré spatialement). Nous évaluons les stratégies d'après leur exactitude (couverture de l'intervalle de confiance), leur précision (erreur type relative) et leur coût (temps de déplacement). L'exactitude augmente en fonction directe du nombre de frayères et de la taille des échantillons et en fonction inverse de la taille des unités d'échantillonnage. Le nombre total de frayères dans le bassin versant et les contraintes budgétaires déterminent quelles stratégies sont les plus précises et les plus efficaces. Les années avec très peu de frayères (<0,15 frayère·km^{-1}), la stratégie d'échantillonnage stratifié et les stratégies moins chères s'avèrent être les plus efficaces, alors que les années avec plus de frayères (0,15–2,9 frayères·km^{-1}), l'une ou l'autre de deux stratégies systématiques plus coûteuses sont plus précises.

[Traduit par la Rédaction]

Introduction

Chinook salmon (*Oncorhynchus tshawytscha*) populations are important for the economy, recreation, and culture of the Pacific Northwest (Lichatowich 1999). Over the last century, these populations have experienced substantial declines (Lichatowich 1999). Nehlsen et al. (1991) listed 64 Chinook salmon stocks that are at high-to-moderate risk of extinction

Received 22 February 2007. Accepted 10 October 2007. Published on the NRC Research Press Web site at cjfas.nrc.ca on 7 August 2008.
J19846

J.-Y. Courbois,[1] S.L. Katz,[2] E.A. Steel, A.M. Wargo Rub, and C.E. Jordan.[3,4] NOAA Fisheries, Northwest Fisheries Sciences Center, 2725 Montlake Boulevard E, Seattle, WA 98112, USA.
D.J. Isaak and R.F. Thurow. US Forest Service, Rocky Mountain Research Station, Boise Aquatic Sciences Laboratory, 322 E Front Street, Suite 401, Boise, ID 83702, USA.
T. Olsen. US Environmental Protection Agency, National Health and Environmental Effects Research Laboratory (NHEERL), Western Ecology Division, 200 SW 35th Street, Corvallis, OR 97333, USA.

[1]Present address: c/o Amazon.com Inc., 705 5th Avenue S, Suite 400, Seattle, WA 98104 USA.
[2]Present address: Channel Islands National Marine Sanctuary, 113 Harbor Way, Suite 150, Santa Barbara, CA 93109 USA.
[3]Corresponding author (e-mail: chris.jordan@noaa.gov).
[4]Present address: NOAA Fisheries, c/o US EPA, 200 SW 35th Street, Corvallis, OR 97333, USA.

or are of special concern in the region. Ten of these stocks are listed under the United States Endangered Species Act (ESA) (Myers et al. 1998). Hundreds of millions of dollars are spent annually in the Columbia River basin alone to conserve and restore Chinook salmon and other salmonid species (\$122 million per year in early 1990s; Lichatowich 1999).

Given the lack of existing experimental design applied to the distribution of management actions, estimation of population status is key to assessing the net impact of habitat degradation, climate change, hatchery practices, harvest, and hydropower. Improving management practices also requires feedback on salmon population response to past management actions (e.g., adaptive management; Bellman 1961; Hilborn 1979; Halbert 1993). Scientific monitoring aims to provide accurate and precise abundance assessments that enable tracking of changes in population attributes for a particular region of interest (Stehman and Overton 1994; Conquest and Ralph 1998; Conquest 2002). Measures of population condition can include both the size of the population at a particular time (status) and the change in population size over time (trend). In this paper, we focus on status. For salmonid fishes, these regions of interest may be delimited by the boundaries of evolutionarily significant units (ESU, Waples 1991), hydrological units, watersheds, or other political entities. Censusing every fish in such large areas is prohibitively expensive, even if it were possible. Therefore, a necessary component of monitoring is sampling in a statistically valid and representative manner (Williams 1978). There are numerous sampling strategies, each with a unique set of strengths and weaknesses that can affect the accuracy and precision of parameter estimates. Therefore, determining which strategy is optimal for a given application is important, because it affects the power of the results and project costs (Cochran 1977; Lohr 1999; Larsen et al. 2001).

Most modern monitoring programs, like the US Environmental Protection Agency's (EPA) Environmental Assessment and Monitoring Program (EMAP), are based on probability sampling, which requires that every element, for example, locations or individuals, within the sampling universe possesses a nonzero probability of being included in a given sample (Thompson 1992). Probability designs have several advantages over nonprobability designs. They allow for probability-based inferences about the entire sampling universe, include a measure of uncertainty, and objectively select study elements, thereby removing possible selection bias. Importantly, the specific feature of probability samples that allows for inferences to be extended to the population is the use of randomization for choosing study elements (Fisher 1935).

Historically, monitoring of many salmon populations was based on data from annual commercial harvest of adult fish from the ocean (Lichatowich 1999), but recent monitoring efforts have utilized sport harvest, dam and weir passage counts, stream surveys for juveniles and spawning adults, and counts of nests, often referred to as redds (Myers et al. 1998). In fact, the complex life history of salmon offers many potential indicators of salmon abundance, the merits of each depending on the specific question being addressed.

Redd counts have long been used to monitor salmon populations (Hassemer 1993). Not only do they provide indirect measures of spawner abundance and reproductive effort, but they are relatively easy to observe (especially for species that spawn during low flows) for extended time periods. Despite the ubiquity of redd counts, however, little guidance exists for designing efficient sampling strategies in stream networks. Furthermore, the sampling universe itself has not always been chosen based on best representation of the overall spawning habitat. As a result, estimates of spawning based on redd counts can suffer from a loss of both accuracy and precision because of inefficient sampling designs and less than optimal choice of the sampling universe. Early agency monitoring efforts commonly consisted of establishing index reaches where redds were counted in accessible portions of stream networks with high fish densities (Hassemer 1993). Index reaches are portions of a river or stream where monitoring metrics, such as redd density, are regularly generated as the basis for extrapolating population estimates over a much larger unstudied area. Unfortunately, these reaches are rarely representative of the population as a whole because they were based on a nonprobabilistic sample and can only provide broader inference if untenable assumptions are made (e.g., temporal consistency in spawning distributions). Despite inherent limitations, index sampling remains valuable because of the temporal extent of data at these sites and the ability to examine long-term population dynamics (Isaak et al. 2003).

Our goal was to compare several probability sampling strategies for estimating the total abundance of redds in a large river basin and to demonstrate a process of selecting a sampling strategy. A sampling strategy is a sampling design along with an estimator for the parameter of interest. A unique database consisting of multiple censuses of Chinook salmon redds was used as a baseline for comparison of estimators derived from each strategy. We describe the statistical performance of the different sampling strategies and estimate the economic costs for each design. We conclude with a discussion of the trade-offs associated with sample strategy selection.

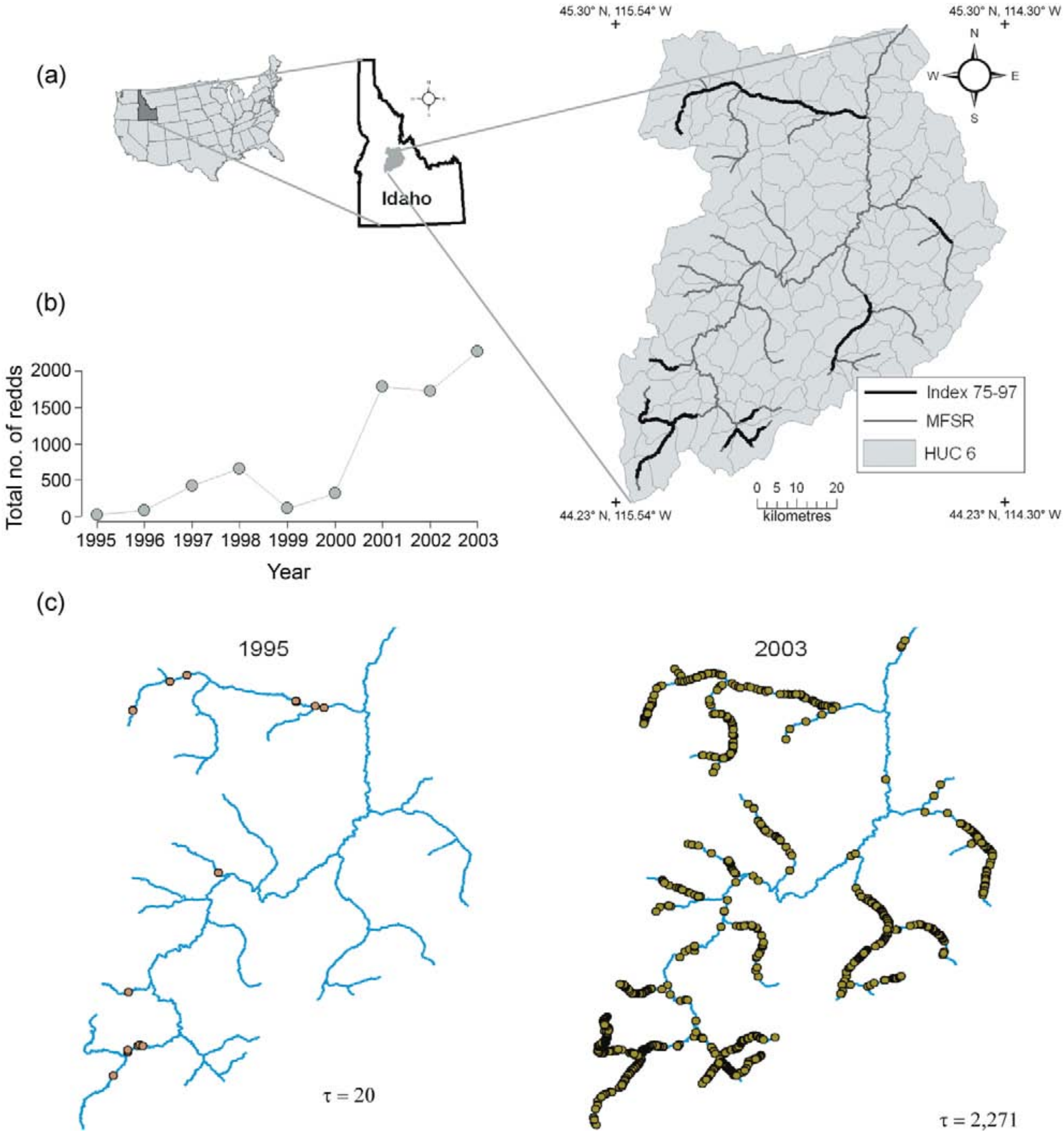
Materials and methods

Study site

The Middle Fork Salmon River (MFSR), a National Wild and Scenic River, drains 7330 km^2 of central Idaho (Fig. 1). For most of its length, the river flows through the Frank Church River of No Return Wilderness. The watershed consists of two level 4 USGS hydraulic unit codes (HUCs, ICBEMP 1999) or 126 level 6 HUCs. HUCs are accounting and cataloging units for describing four levels of successively smaller waterways in the United States. From its origin at the confluence of Bear Valley and Marsh Creeks, the MFSR flows north-northeast for 171 km through the Salmon River Mountains and joins the Salmon River 92 km downstream from Salmon, Idaho (Thurow 2000).

Chinook salmon populations in the upper MFSR are composed of wild, indigenous fish referred to as both "spring" and "summer" Chinook salmon based on the timing of adult migration past Bonneville Dam in the lower Columbia River (Matthews and Waples 1991). They are categorized as stream-type Chinook salmon because they rear for at least 1

Fig. 1. (*a*) Map of Middle Fork Salmon River (MFSR, Idaho, USA). (*b*) Number of redds for each year. (*c*) Two extreme examples, low- and high-run years, of the census data to be sampled.



year in fresh water (Healey 1991). Adult Chinook salmon enter the MFSR in early summer, migrate toward natal areas in larger tributaries, and stage before spawning. Redd construction usually begins during the last week of July and is completed by mid-September. Females typically deposit eggs in large redds (~ 3 m^2) constructed in riffle crests or areas with similar hydraulic and substrate characteristics (Healey 1991).

Chinook salmon redd censuses

Since 1995, United States Forest Service (USFS) biologists with the Rocky Mountain Research Station have conducted annual, spatially continuous surveys for Chinook salmon redds within the portion of the MFSR accessible to Chinook salmon. Surveys were conducted using low-altitude helicopter flights at the end of the spawning period (mid-September). Ground observers surveyed stream reaches with closed canopies that were difficult to count aerially (i.e., dense canopy). When a redd was observed, a global positioning system (GPS) was used to geo-reference the location. Both ground and aerial coordinates were differentially corrected and assembled into a geographic information system (GIS) database. For more details, see Isaak and Thurow (2006). We used the data from nine of these surveys (1995 through 2003) in our analysis; the total number of redds ranged from 20 to 2271 (Fig. 1).

Sampling

The sampling universe was that portion of the stream network surveyed by USFS biologists (Fig. 1). The sampling units were stream segments that were either 200 or 1000 m long. The river network was approximately 680 km of stream and comprised 3400 two-hundred-metre segments and 693 one-thousand-metre segments. Not all of the units were exactly 200 and 1000 m owing to confluences and headwaters. Although a preliminary analysis demonstrated that shorter units result in higher precision, we included the longer units because they were less expensive to sample and were more characteristic of lengths surveyed by agency biologists. Three different percentages of the stream network were sampled: 29%, 10%, and 5% (respectively, n = 1000, 340, and 170 for 200 m segments; n = 204, 69, and 35 for 1000 m segments).

Assumptions and notation

We assumed that redd counts and locations occurred without error, the classical design-based approach to statistical inference (Thompson 1992). Under this assumption, the population values are fixed and the estimators' statistical properties depend solely on the variability caused by the different possible samples. Violation of this assumption would have a minimal effect on our results because all sampling strategies were subject to the same errors.

Let y_i, $i = 1,2, \ldots, N$, be the population of interest, where y_i is the redd count in river segment i, N = 693 or 3400. Furthermore, let U be the universe of sampling segments, $U = \{1,2, \ldots, N\}$ and s represent a sample, $s \subseteq U$. Also, define S to be the set of all samples: $s \in S \;\forall\; s \subseteq U$. The parameter of interest is the status, the total number of redds in the watershed, denoted by $\tau = \sum_{i=1}^{N} y_i$. Let $\hat{\tau}$ represent an estimator for τ. The total of the responses in the sample will be denoted by $t = \sum_{i \in s} y_i$. Finally, let $\bar{y} = t/n$ and $\bar{y}_U = \tau/N$ be the sample mean and population mean, respectively.

For design-based inference, the properties of an estimator are based on the sampling strategy rather than an assumed distribution. For example, the expected value of $\hat{\tau}$ is $E[\hat{\tau}] = \sum_{s \in S} p(s)\ \hat{\tau}(s)$, where $s \in S$ indicates the summation over all possible samples, $p(s)$ is the probability that sample s is selected, and $\hat{\tau}(s)$ is the estimate resulting when sample s is selected. The root mean squared error (RMSE) of the estimator $\hat{\tau}$, or standard error if $\hat{\tau}$ is unbiased, is the square root of the variance of its sampling distribution:

$$(1) \qquad \mathrm{SE}(\hat{\tau}) = \sqrt{\mathrm{var}\ (\hat{\tau})} = \sqrt{\sum_{s \in S} p(s)(\hat{\tau}_s - \tau)^2}$$

This standard error is often called the planning standard error because, although it is unknown, it is used to compare different sampling strategies (Lohr 1999). These properties of $\hat{\tau}$ are functions of the entire population and thus are usually unknown to the researcher (see below). The estimated standard error for $\hat{\tau}$, $\widehat{\mathrm{SE}}(\hat{\tau})$, is an estimator for the planning standard error and is a function of the data (see below).

Sampling strategies

Index sampling

Index sampling is a nonrandom sampling design where units are often selected based on stream access and high fish densities. For purposes of this study, we used the 1995–1997 spawning ground index areas established by the Idaho Department of Fish and Game (Elms-Cockrum 1998). These surveys are taken annually at 22 index reaches — some of which have been surveyed for 50 years (Fig. 1). The total length of the index reaches was 198.8 km or about 29.2% of the stream network suitable for Chinook salmon spawning.

To develop status estimators from the nonprobability index sample, we made two assumptions. First, the index reaches covered the entire Chinook salmon spawning grounds, so there were no redds outside this area. By making this assumption, the estimator becomes the simple sum of the sample, $\hat{\tau} = t$. Second, the index sites were representative of the entire network, in which case the resulting estimator inflates the redd count by the fraction of area sampled: $\hat{\tau} = \frac{1}{0.292}\ t = 3.42\ t$.

The two assumptions represent the limits for a family of estimators based on index sampling. We also evaluated estimators intermediate to these that assume index sites represented from 40% to 90% (in increments of 10%) of the total redds in the watershed.

Estimators from index samples do not have a survey design variance associated with them because there is only one possible sample. This is a major limitation of index sampling — no estimate of precision is possible.

Simple random sampling without replacement (SRS)

SRS selects units independently and with equal probability. The estimate for the total number of redds from SRS is $\hat{\tau} = \frac{N}{n}\ t = N\ \bar{y}$, where $\bar{y}$ is the sample mean. This estimator has the variance

$$\mathrm{var}(\hat{\tau}_{\mathrm{SRS}}) = N^2\left(1 - \frac{n}{N}\right)\ \frac{S^2}{n}$$

where $S^2 = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2/(N-1)$ is the population variance and $\bar{y}_U$ is the mean of the response over the universe. The estimator for the planning standard error is

$$(2) \qquad \widehat{\mathrm{SE}}(\hat{\tau}_{\mathrm{SRS}}) = \sqrt{N^2\left(1 - \frac{n}{N}\right)\frac{\sum_{i \in s} (y_i - \bar{y})^2}{n-1}}$$

The square of this is unbiased for its respective population variance (Thompson 1992).

Systematic sampling (SYS)

In SYS, every rth unit is selected systematically along a line, for a one-dimensional population, or through space, for a two-dimensional population. An easy way to systematically sample a river network is to connect all tributaries into one line and systematically sample along this line in one dimension. To do this, we calculated the closest integer to N/n, which we denote as K, and chose a random number r between 1 and K, then sampled the rth unit and every Kth unit after that. Note that K is the total number of possible systematic samples, where one of these is chosen by the random start. For example, if 5% of the network was sampled, there were only 1/0.05 = 20 possible systematic samples; thus the standard error could be exactly calculated. However, constructing a single line from a linear network is not uniquely possible, so we had to decide on an ordering for the tributaries. We decided to order the network's tributaries in a random order for every simulation. This added one extra level of randomization to the design. Certainly, if there was good information about the correlation structure between sampling units, the ordering could be optimized for SYS.

The estimator for the number of redds from SYS is $\widehat{\tau} = Kt$. The planning variance is intractable in the sense that it is too expensive to calculate, owing to the random ordering of tributaries. Estimating this standard error is in turn complicated by the restrictions imposed by the systematic samples within the tributaries (Bellhouse 1988). The planning variance reflects only cluster-to-cluster variation within tributaries, and since only one cluster is selected for each, an unbiased estimator does not exist. We evaluated three estimators that rely on population assumptions for unbiasedness. The first assumes that the systematic sample is equivalent to an SRS or that the population is in random order:

$$(3) \qquad \widehat{\mathrm{SE}}_1(\widehat{\tau}_{\mathrm{SYS}}) = \sqrt{N^2\left(1-\frac{n}{N}\right)\frac{\sum\limits_{i\in s}(y_i-\overline{y})^2}{n}}$$

The second, the successive difference estimator, is suggested when the population exhibits a trend (Cochran 1977) within the tributaries:

$$(4) \qquad \widehat{\mathrm{SE}}_2(\widehat{\tau}_{\mathrm{SYS}}) = \sqrt{N^2\left(1-\frac{n}{N}\right)\frac{\sum\limits_{i\in s}(y_i-y_{i-1})^2}{2n(n-1)}}$$

A final estimator for the variance of the systematic estimator is the neighborhood variance estimator (Stevens and Olsen 2003). This estimator is a weighted average of local estimates of variance, one at each sampled point. The sampled units surrounding each point contribute to each local variance estimator, and their contribution depends on their inclusion probability and the rank (compared with other sampled units in the neighborhood) of their distance from the sampled unit in question. The neighborhoods are not defined in terms of distance but rather by the closest four sampled points (with some adjustment to make the neighborhood function symmetric). The form of the estimator is

$$(5) \qquad \widehat{\mathrm{SE}}_3(\widehat{\tau}_{\mathrm{SYS}}) = \sqrt{\sum_{i\in s}\sum_{j\in D_i} w_{ij}\left(N\frac{y_j}{n}-\overline{y}_{D_i}\right)^2}$$

Where D_i is the set of sampled units that are the neighbors to unit i, w_{ij} is the weight that decreases as a function of the relative distance of sampled unit j to sampled unit i, and $\overline{y}_{D_i}$ is the weighted average of the units in the neighborhood of unit i.

All three of these standard error estimators are motivated by assumptions about the underlying population autocorrelation structure. They do not specifically estimate the randomization produced by the design nor in particular the effect of the randomizing tributaries to the line along which we sample. The validity of the underlying assumptions on which these estimators are based is unknown, and thus their model-based theoretical properties, such as unbiasedness, are unknown.

Spatially balanced design (GRTS)

The generalized random-tessellation stratified design (GRTS; Stevens and Olsen 1999, 2003) emphasizes spatial balance of the sample in the sense that every sample reflects the spatial balance of the population. It provides better coverage than SRS and more randomness than SYS. For the GRTS design, every unit in the finite population is assigned a random address that was used to order the population randomly, then a systematic design was used to select units along this random ordering as one would do for an SYS sample. The random addresses and the random ordering, however, are closely tied to the spatial location of the units so that the sampled units are spread out in space. The resulting sampling design is similar to a spatially stratified design with one unit selected per stratum; however, the strata are random sets, so the second-order inclusion probabilities are nonzero.

GRTS is an equal probability design, so the estimator for the total is the same as used for SRS, $\widehat{\tau} = \frac{N}{n}\, t = N\,\overline{y}$. The estimator for the standard error of this estimator is the neighborhood estimator (eq. 5).

Stratified sampling (STR)

In STR, the sampling universe is split up into mutually exclusive but exhaustive strata, and a simple random sample is independently selected from each stratum. There are two reasons to use STR. (*i*) In the case when estimates are required for subsets of the population, STR allows for control over these estimators at the design stage. (*ii*) STR can result in more precise estimators when variability within strata is smaller than variability between strata. We stratified the sample units into two strata based on whether or not they occurred in an index reach. This stratification allows for explicit control over the sampling design within the index reaches and thus control the precision of estimates within the reaches. This advantage is explored further in the Discussion. Furthermore, because the index area is identified as good spawning habitat, we might expect the variance to be smaller within than between the two strata. There are many other possible ways to stratify the MFSR, by tributary, for example, that we did not consider. As in predictive model fitting, we do not want our designs optimized for our training data (the populations we have on the MFSR); rather, we want our results to apply more generally.

We allocated the sample to strata using optimal allocation, which minimizes the standard error of the estimator for the total. Because redd distributions varied through time (Isaak and Thurow 2006), the allocations varied accordingly. Consequently, we calculated the optimal allocation for each year and used the median of these values to allocate the sample to all populations. The resulting allocation was 0.49

of the sample allocated to the index reaches; the range of the annual optimal allocation was 0.41–0.76.

The estimator for the total number of redds from STR is the sum of the SRS estimators for the strata totals, $\widehat{\tau}_{\mathrm{STR}} = \sum_{h=1}^{H} \widehat{\tau}_h$, where H is the number of strata and $\widehat{\tau}_h$ is the SRS estimator for the total within stratum h. Because sampling is independent in each stratum, the variance of this estimator is the sum of the variances of the estimators for each stratum:

$$\mathrm{var}(\widehat{\tau}_{\mathrm{STR}}) = \sum_{h=1}^{H} \mathrm{var}(\widehat{\tau}_h) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h}$$

where the second equality holds under SRS in each stratum, and n_h and N_h are the sample and population sizes for stratum h, respectively. The estimator for this variance is the same function calculated on the sample.

Adaptive cluster sampling (ACS)

ACS begins with SRS. Whenever a sampled unit contains one or more redds, adjacent units are also sampled. If these units also contain one or more redds, their adjacent units are also sampled. Sampling continues until no redds are encountered.

We used the draw-by-draw estimator to estimate the status (Thompson 1992). The population is made up of a series of networks of units, where a network is a set of units that contain redds and would result in the entire network being in the sample if any one of the units is sampled. Let Φ_i denote the network which includes unit i, and let m_i be the number of units in that network. Let ω_i represent the average of the observations in the network, which includes the ith unit of the initial sample, that is, $\omega_i = \frac{1}{m_i} \sum_{j \in \Phi_i} y_j$. The estimator for the total becomes $\widehat{\tau}_{\mathrm{ACS}} = \frac{N}{n} \sum_{i \in s} \omega_i$, where N and n are still the population and sample size, respectively. The variance of this estimator is

$$\mathrm{var}(\widehat{\tau}_{\mathrm{ACS}}) = \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} (\omega_i - \overline{y})^2$$

The estimate of this variance is the same equation but is calculated on the sample.

For the ACS, the sample size might reflect the size of the original SRS or the size of the final sample that includes elements in the networks. We evaluated both of these approaches; the first was denoted adaptive-n, indicating that the original sample sizes were equal to n above, while the second was denoted adaptive-En, indicating that the original sample sizes were the sample sizes such that the expected final sample sizes were equal to n above. For adaptive-En, the original sample sizes were different for each year and sampling proportion. For example, when 5% of the MFSR was sampled using 200 m units, setting the expected final sample size to 170 resulted in the initial sample sizes of 168, 162, 126, 109, 160, 141, 66, 71, and 41 for redd numbers of 20, 83, 424, 661, 110, 318, 1789, 1730, and 2271, respectively. The resulting final sample sizes varied with number of redds, from 168 to 170 for the fewest (20) and from 41 to 241 for the most (2271). Thompson (1992) describes how to calculate the expected final sample size given a population and an original sample size.

Simulations and statistical evaluation

We evaluated the sampling strategies using Monte Carlo simulations. For each strategy and population, the simulations worked identically; we selected 1000 independent samples from the MFSR segments and calculated $\widehat{\tau}$, $\widehat{\mathrm{SE}}(\widehat{\tau})$, and cost (see below). The resulting collection of 1000 redd population estimates approximated the sampling distribution of the strategy's estimator for τ so the properties of the estimator could be approximated from this collection of $\widehat{\tau}$ values. For example, the standard deviation of this collection approximated the true standard error of the estimator. The simulations were done in the R statistical package (R Foundation for Statistical Computing, Vienna, Austria; code is available from the first author).

The statistical criteria used to compare the strategies were their accuracy and precision. The accuracy of each strategy was evaluated with the empirical coverage of a 95% confidence interval for τ. This empirical probability indicates the fraction of the time that the true total is within the 95% confidence interval around the estimated total and thus is a measure of accuracy. This value should be close to 0.95. Confidence intervals were calculated using the central limit theorem:

$$(6) \qquad \widehat{\tau} \pm 1.96 \times \mathrm{SE}(\widehat{\tau})$$

There are three elements to this equation; (*i*) a good estimate for the total number of redds (unbiased); (*ii*) a good estimate for the standard error of the estimator for the total (unbiased); and (*iii*) the sampling distribution for the estimator of the total should be well approximated by the normal distribution (the 1.96 constant in eq. 6). Empirical coverage probabilities away from 95% may be due to failure of any of these elements.

Even for accurate strategies, the resulting empirical coverage is unlikely to equal 95% because we are evaluating 1000 replicate samples instead of all possible samples. This is simulation error and it was assessed analytically. If the true coverage probability for a strategy is 0.95, then we would expect 95% of simulated empirical coverages to be within $1.96 \times \mathrm{SE(cover)}$ of 0.95, where cover is the expected nominal coverage, 0.95. This SE was $\sqrt{0.95(1-0.95)/1000} = 0.00689$, so a "hypothesis-test" interval half-width will be approximately 0.014, and we might expect results, when the true coverage probability is 0.95, between 0.936 and 0.964. If the coverage probability falls outside this range, then we conclude that the strategy is not accurate. Precision is defined as the inverse of the standard error; thus, a small variance implies a large precision. We standardized the estimators' planning standard errors (eq. 1) because population sizes (number of redds) varied widely, which because the strategies are unbiased, results in a coefficient of variation (CV):

$$(7) \qquad \mathrm{CV} = \frac{\mathrm{SE}(\widehat{\tau})}{\tau}$$

This is the CV for the sampling distribution of the estimator sometimes known as the relative standard error; we shall

Fig. 2. Location and type of access points (landing strip or trailhead) in the Middle Fork Salmon River (MFSR). Each point along the river system is assigned an access point based on its proximity.

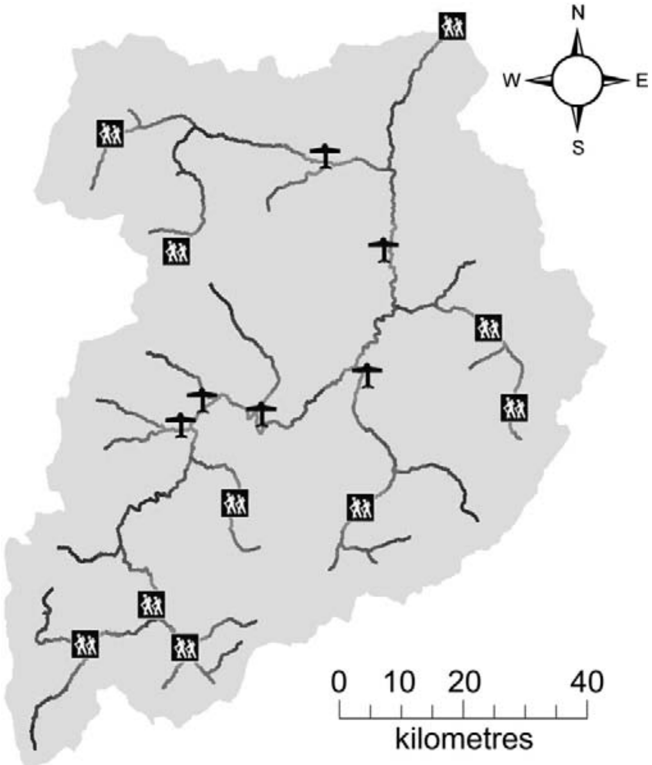
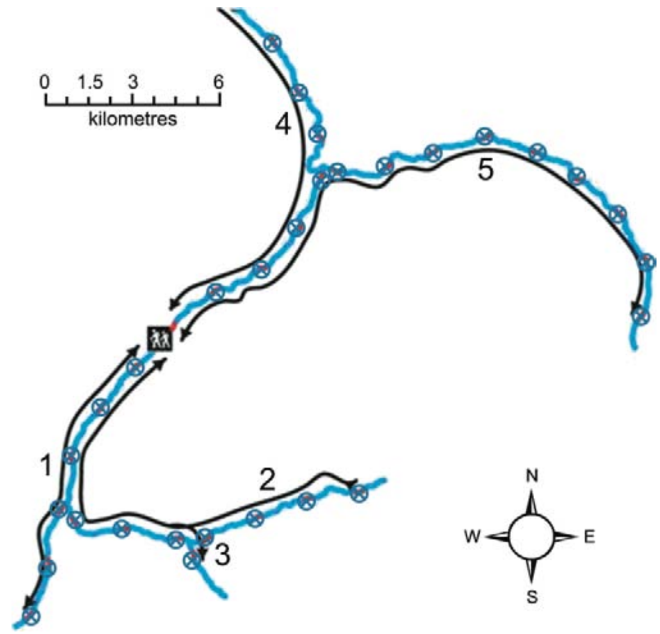


Fig. 3. An example of the cost function. The segments marked with $\otimes$ are sampled in this systematic random sample. To access these points, the crew makes five trips, marked 1–5, from the access site up each tributary along the black arrows. Notice that trips 2 and 3 are separate distances from the access site. The cost of sampling this region is the total of the crew's travel distance (i.e., the sum of the maximum distance in each of the five river tributary directions).



continue to refer to it as the CV. Small values of CV are desirable. Most of these were calculated analytically and did not suffer from simulation error. For the other cases, we approximated the CV through simulation by calculating the standard deviation of the 1000 estimates for τ. We assessed simulation error here by placing the 1000 replicates into groups of 100. Then we approximated the standard error for $\widehat{\tau}$ from each group of 100. The resulting 10 standard errors were used to assess the simulation error for the standard error approximation, and a confidence interval was calculated for the true unknown standard error.

Cost

The MFSR flows primarily through a roadless area with limited vehicle access. Although the censuses that produced the data were collected via aerial surveys, ground-based redd counts are used in many areas, and we wanted to define a cost function that reflects alternative modes of transportation. To begin, we identified 15 access sites (Fig. 2) that signified points past which crews must continue on foot to sample.

To estimate the cost of foot travel, we assigned an access point to each sampling unit and calculated the river distance from the access point to each segment (Fig. 2). We assumed once a crew is dropped off, it can walk up a tributary to the farthest selected site (maximum distance 33 km, median of 10.3 km) and back in one trip counting redds on any additional selected segment on the way. The cost to sample all of the sites along a tributary thus was equivalent to the farthest distance traveled along that tributary, $\max_j\{d_{ij}\}$, where d_{ij} is the distance from sample site j to the access site to which it belongs, i (Fig. 3).

Two other costs were additively included in the cost function: the cost of access and the sampling cost. To estimate the access cost, we assumed that the costs associated with dropping off a crew at any one of the access points are equal. For sampling cost, we assumed it would take five times longer to conduct a redd survey than it would to walk a comparable length of stream. Therefore the complete cost function was

$$(8) \qquad C = \sum_{i=1}^{15}(C_i + \max_j\{d_{ij}\}) + 5(l \times n)$$

where the sum is over the 15 access sites, C_i is the cost for access point i, and l is the length of each n sampling segment in kilometres. The cost function has the units kilometres traveled. Because in our case, $C_i = C_j \; \forall \; i, j$, this part of C is ignored.

The costs of the designs were compared, but we also determined which designs were most cost-effective in terms of precision. We evaluated cost and precision together by fixing cost and finding the design that was most precise. Because the index sampling strategy is currently being used, its cost (1254.4 km) was used as a baseline. Because costs varied over the 1000 samples drawn from each design, we chose sample sizes that insured 95% of our samples had costs less than 1254.4 km. Thus, sampling can be accomplished at or under the current index sampling budget with 95% certainty.

Fig. 4. Simple random sampling with n = 170 200 m units. (*a*) An example of one sample with an enlargement showing the small-scale distribution of sampled segments. (*b*) The approximate sampling distribution for the estimator for status. The mean of these 1000 estimates is 665.4, while the actual number of redds (661) in 1998 is indicated by the vertical line. (*c*) The approximate sampling distribution for the estimator for the standard error. The true standard error for the estimator is 154.66, as indicated by the vertical line. The mean of these 1000 estimators of standard error is 153.6. (*d*) Several examples of confidence intervals; the vertical line indicates the true number of redds (661). Of all the intervals, 939 out of 1000 contain the value 661. Here, iteration number 4 does not contain the truth.

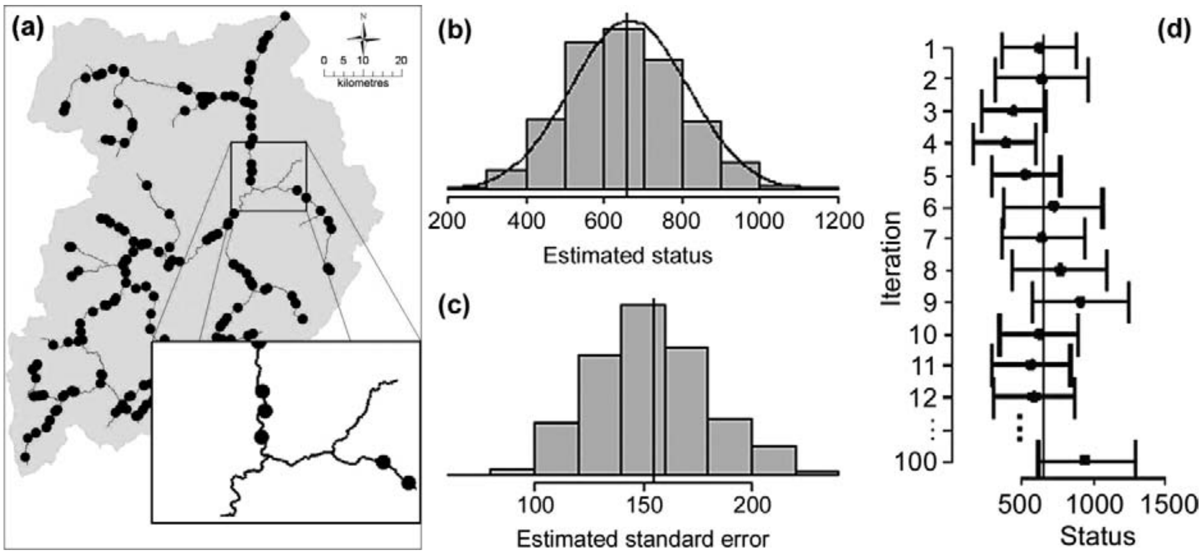
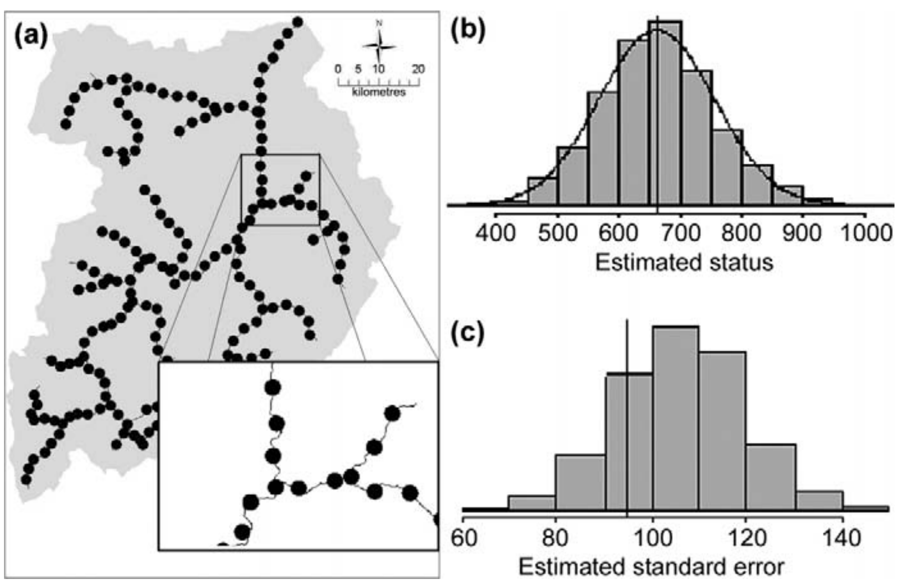


Fig. 5. Systematic sampling with n = 170 200 m units. (*a*) An example of one sample with an enlargement showing the small-scale distribution of sampled segments. (*b*) The approximate sampling distribution for the estimator for status. The mean of these 1000 estimates is 664.35, while the actual number of redds (661) for 1998 is indicated by the vertical line. (*c*) The approximate sampling distribution for the estimator for the standard error that assumes a simple random sample (eq. 2). The true standard error for the estimator is 94.58, as indicated by the vertical line. The mean of these 1000 estimators of standard error is 105.97.



A second method used to evaluate cost and precision was to look at precision per cost, which balances the need for the most informative answer (highest precision) with lowest cost. We defined the standardized precision as the inverse of the CV and measured the precision per cost by

$$\frac{\text{precision}}{\text{cost}} = \frac{\tau}{\text{SE}(\hat{\tau}) \times C} \tag{9}$$

which is in units of 1/(kilometres traveled).

Results

To introduce the analytical protocol and performance measures, we work through two examples that include SRS and SYS for 1998. In both cases, the sampling designs used 200 m units and a 0.1 sampling proportion (Fig. 4). There were τ = 661 redds in the MFSR in 1998 (Fig. 1).

For the SRS (Fig. 4), the average of the 1000 estimates (one from each randomly selected sample) was 658.03, close enough to 661 to suggest that this estimator was accurate (unbiased). Since this estimator is unbiased analytically, the difference between 658.03 and 661 was due to simulation error. Furthermore, the sampling distribution for this estimator is approximately normal (Fig. 4*b*). The standard deviation of these 1000 estimates for τ, 107.99, approximated the true standard error of the estimator, which in this case can be calculated exactly as 106.44. The discrepancy between the standard deviation and the true standard error was due to simulation error. The CV, $\mathrm{CV}(\widehat{\tau}) = 106.44/661 \times 100 = 16.1$, measures the precision of SRS. The average of the 1000 estimates for the standard error was 105.36, suggesting that this estimator is also unbiased (Fig. 4*c*). We demonstrate how for each sample, we can use $\widehat{\tau}$ and $\widehat{\mathrm{SE}}(\widehat{\tau})$ to construct a confidence interval and count how many of these intervals include $\tau = 661$ to determine the empirical coverage probability (Fig. 4*d*). In the SRS case, 924 out of 1000, or 92.4%, include the true value. This value is not within our testing bounds (93.6 to 96.4), indicating that the true coverage was slightly less than 0.95. The reason for this failure is explored later.

For SYS (Fig. 5), $\tau = 661.78$, indicating an unbiased estimator, and the sampling distribution also was approximately normal (Fig. 5*b*). The standard deviation of these 1000 estimates for the total was 94.58 (Fig. 5*c*), resulting in a precision of 14.3. For SYS, we do not calculate the standard error exactly, but depend on the approximation from the simulation. Consequently, this approximation is affected by simulation error. We estimated the magnitude of the simulation error by splitting the 1000 outcomes into 10 groups of 100 each and calculating intervals by using a normal approximation. The 95% interval was 13.9–14.8, and the SYS strategy appeared to be more precise than SRS (14.8 < 16.1).

The average of the 1000 estimates for the standard error, 105.97, was larger than the approximate true standard error of 94.58, indicating a bias in this estimator. In fact, although $\widehat{\tau}_{\mathrm{SYS}}$ is unbiased, the estimator for the standard error is biased because we used the estimator based on an SRS (eq. 3). This positive bias results in wide confidence intervals and a slightly inflated empirical coverage: 95.9 > 95%. However, this lies within our accuracy bounds (93.6 to 96.4), suggesting that SYS was reasonable.

These two examples indicate the characteristics of accuracy and precision that will be referred to when evaluating the relative performance of the various strategies. We now turn to an evaluation of all the strategies.

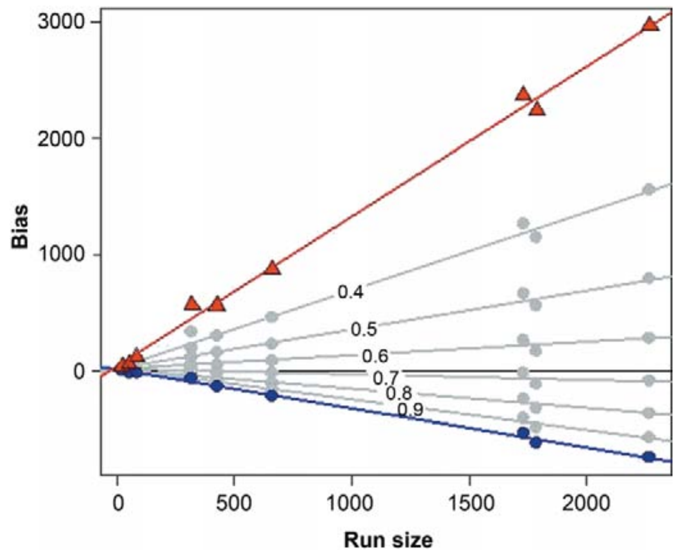
The index strategy

Estimators for the index strategy were biased (Fig. 6). This was not unexpected because the assumptions that had to be made to derive estimates were incorrect. However, approximate estimates could be obtained by assuming that 70% of the redds were in the index reaches regardless of run size. Even with this assumption, however, inaccuracy was still substantial and ranged from 9% (–4.3 redds) in 1999 to 8% (–107.6 redds) in 2001. Because of this large bias and no means of calculating a standard error, additional comparisons between the index strategy and other designs were impossible.

Accuracy

Larger sampling fractions and larger redd numbers resulted in better accuracy (Table 1). Strategies that sampled 200 m segments were more accurate than those sampling 1000 m segments (see bold data vs. shaded cells in Table 1). None of the designs were accurate when redd abundance was low (Table 1). However, as the sampled proportion increased, accuracy improved for intermediate numbers of redds. Only when sampling 29% of the network with 200 m units and during years with large redd numbers could we be confident that nominal 95% confidence intervals are actually 95% (Table 1*c*).

Fig. 6. Actual bias (number of redds) of the estimators from index sampling. The values 0.4–0.9 are the assumed proportion of redds that are in the index reaches (counted) used to construct estimators. All assumptions result in biased estimators, with the bias depending on the run size.



Precision

As redd numbers increased, precision improved for all designs (Tables 2 and 3). Precision also improved as the sampling fraction increased and with 200 m reaches relative to 1000 m reaches (Tables 2 and 3).

The SYS and GRTS designs were most precise, with STR also performing well with lower numbers of redds. SRS and the adaptive designs were generally less precise.

Cost

As expected, designs that select fewer, larger units were less expensive (Figs. 7*a*–7*c*). However, cost differences decreased for larger sampling fractions (Fig. 7*c*). The index design was the least expensive strategy (Fig. 7*c*) when the sampling proportion was large (29%), because the index reaches are clustered near access points; however, the index design was the most expensive for smaller sampling proportions. The least expensive probabilistic strategy was adaptive-En because of its small initial sample sizes; the most expensive was adaptive-n, because of its larger final sample sizes. These two strategies were also the most variable (Figs. 7*a*–7*c*) because the final sample sizes depend on whether the initial SRSs contain any redds and thus were unknown. Cost differences among remaining designs were trivial.

Table 1. Accuracy as measured by the empirical coverage of the 95% confidence interval.

	Run size								
Strategy	20	83	110	318	424	661	1730	1789	2271
(*a*) Sampling fraction = 0.05 (*n* = 170).									
SRS	57.5	88.9	88.3	89.2	91.5	92.3	92.8	91.1	93.2
SYS									
1	56.8	90.2	87.5	**95.2**	96.8	**95.7**	98.2	97.8	98.1
2	56.8	89.0	87.2	**94.2**	**95.6**	**94.2**	97.8	97.2	97.0
3	56.8	89.0	86.2	93.0	**93.6**	92.0	**96.4**	**95.6**	**95.3**
GRTS	65.7	89.3	83.2	90.2	93.3	88.4	93.3	32.1	**95.3**
STR	75.3	88.1	87.9	91.4	91.3	91.1	90.7	89.4	92.7
Adaptive-n	58.4	89.5	92.8	90.5	93.2	**94.7**	**93.6**	93.2	**94.1**
Adaptive-En	56.4	89.2	86.2	91.8	91.8	92.0	92.8	92.8	92.1
(*b*) Sampling fraction = 0.1 (*n* = 340).									
SRS	81.7	92.1	92.0	92.5	93.0	92.4	92.9	92.7	**94.0**
SYS									
1	86.5	90.1	**95.2**	97.5	96.6	**95.9**	97.7	98.8	99.5
2	86.5	89.5	**95.2**	96.6	**94.5**	92.7	**96.1**	97.8	97.5
3	87.3	80.7	92.4	93.2	91.4	91.3	92.4	**95.7**	**95.4**
GRTS	86.7	86.4	89.1	91.9	**96.3**	**95.2**	**94.0**	**95.7**	**94.0**
STR	82.8	91.0	91.7	92.9	**94.6**	92.9	93.3	91.9	**94.6**
Adaptive-n	82.9	92.6	92.0	93.1	**94.0**	**95.0**	92.5	92.6	**95.6**
Adaptive-En	84.2	91.9	86.2	92.2	92.7	**95.1**	**94.3**	93.5	**95.5**
(*c*) Sampling fraction = 0.29 (*n* = 1000).									
SRS	84.8	**94.6**	92.6	**95.3**	**94.4**	**94.5**	**94.9**	**94.3**	**96.2**
SYS									
1	**96.1**	**95.8**	99.7	97.9	96.0	96.4	97.8	99.9	99.7
2	**96.1**	**95.3**	99.1	**95.2**	91.9	91.5	92.6	98.3	97.6
3	98.4	**95.9**	99.5	96.9	94.3	94.7	96.0	99.5	99.2
GRTS	91.4	92.8	**93.9**	**95.5**	**96.1**	**96.2**	**94.2**	**95.1**	97.8
STR	92.4	**95.1**	92.9	**94.6**	92.6	93.3	93.2	93.4	93.0
Adaptive-n	87.9	**94.2**	93.4	**94.4**	**95.0**	**93.8**	**95.3**	**94.8**	**95.8**
Adaptive-En	92.6	**94.3**	**94.3**	**94.3**	**94.2**	**96.0**	**95.4**	**95.1**	**94.2**

Note: The data indicate the approximate accuracies for the strategies when sampling 200 m units. A strategy is accurate if its empirical coverage is between 93.6% and 96.4%. Accurate strategies are indicated in bold for 200 m units or shaded for 1000 m units (values not given). SRS, simple random sampling; SYS, systematic sampling; GRTS, spatially balanced design; STR, stratified sampling; Adaptive-n, adaptive cluster sampling when n = the original sample size; Adaptive-En, adaptive cluster sampling when n = the expected final sample size. The three rows for SYS refer to the three estimators for the standard error: 1 = assuming simple random sampling (eq. 2); 2 = successive difference estimator (eq. 3); and 3 = neighborhood estimator (eq. 4).

Precision per unit cost

When costs were held constant, the general trend was for inexpensive designs, such as STR, to have larger sample sizes. As a consequence of more samples, these inexpensive designs were more precise when there were small numbers of redds. Conversely, the more expensive strategies, such as SYS and GRTS, were most precise during large runs, when sparsely distributed redds was less of an issue (Table 4).

Precision per cost also demonstrated that the choice of an optimal design and sample unit size depended on the sample size and run size. For small sample proportions (5%), the inexpensive, larger sampling units (1000 m) provided better value than the smaller 200 m units (Table 5*a*). In this case, the STR strategy was best for small runs and the GRTS strategy was best for large runs (Table 5*a*). However, for the intermediate sample size (10% sampling fraction), during small run years, sampling 1000 m units was most cost effective, while for large-run years, the more expensive SYS design with 200 m units was most cost effective (Table 5*b*). Finally, for large samples, the expensive but precise SYS strategy with 200 m units was most cost effective (Table 5*c*).

Discussion

We examined the performance of different redd sampling strategies for a large river network. Our results illustrate design tradeoffs and can provide guidance for selecting optimal monitoring strategies.

We did not exhaust the set of possible probability sampling strategies that can be used to count redds. We concentrated on standard designs and designs that have been suggested in the literature. Our goal was not to optimize a strategy for the MFSR, but rather to suggest designs that can be used more generally. Likewise, the estimators we considered were also standard, and there may be ways to

Table 2. Precision for strategies using 200 m units as measured by coefficient of variation × 100 (eq. 7).

Strategy	Run size 20	83	110	318	424	661	1730	1789	2271
(*a*) Sampling fraction = 0.05 (*n* = 170).									
SRS	110.9	52.8	50.8	33.9	28.3	23.4	20.2	20.8	18.4
SYS*	107.6	51.9	50.6	26.8	22.9	20.5	14.8	15.0	13.5
GRTS*	95.8	50.1	49.4	34.0	25.2	23.2	16.7	17.7	13.7
STR	86.7	45.8	42.4	27.3	25.3	20.9	17.9	19.3	15.8
Adaptive-n	110.9	52.0	48.7	32.6	26.5	21.6	17.1	17.6	15.5
Adaptive-En	111.6	53.3	50.3	36.0	30.9	27.2	26.8	28.7	32.3
(*b*) Sampling fraction = 0.1 (*n* = 340).									
SRS	76.3	36.3	35.0	23.3	19.5	16.1	13.8	14.3	12.6
SYS*	69.2	39.1	28.1	16.7	16.1	14.3	10.5	10.2	9.1
GRTS*	65.3	35.1	34.6	22.0	15.2	13.3	11.4	11.0	10.0
STR	58.6	31.3	28.9	18.6	17.3	14.3	12.2	13.3	10.8
Adaptive-n	76.3	35.8	33.5	22.5	18.2	14.8	11.7	12.1	10.7
Adaptive-En	76.7	36.7	33.5	24.7	21.1	18.4	17.9	19.0	19.7
(*c*) Sampling fraction = 0.29 (*n* = 1000).									
SRS	39.4	18.8	18.1	12.1	10.1	8.3	7.1	7.4	6.5
SYS*	23.5	16.4	11.7	9.6	9.3	7.6	5.8	4.5	4.0
GRTS*	39.9	19.1	17.8	10.4	8.0	6.2	5.9	5.6	4.3
STR	27.3	15.4	14.0	8.9	8.7	7.2	6.2	6.8	5.4
Adaptive-n	39.4	18.5	17.3	11.6	9.4	7.7	6.1	6.3	5.5
Adaptive-En	39.6	18.9	17.8	12.5	10.6	9.1	8.0	8.8	8.2

Note: The most precise strategies for each run size (column) are shaded; when more than one strategy is highlighted within a column, the difference in precision is unclear because of simulation error. SRS, simple random sampling; SYS, systematic sampling; GRTS, spatially balanced design; STR, stratified sampling; Adaptive-n, adaptive cluster sampling when n = the original sample size; Adaptive-En, adaptive cluster sampling when n = the expected final sample size.

*Precision is evaluated through simulation.

improve them that we did not consider. In particular, we did not consider any modern variance estimators, for example, the bootstrap may result in more accurate estimators.

Given that the ecological processes that determine the redd distributions we observed are likely to be similar in other locations, our results will be fairly robust. However, given that we do not know this to be true, the relative performance of these sampling designs may not be similar in all other cases.

Poor accuracies

It may be confusing that theoretically accurate strategies, those with unbiased estimators and unbiased estimators of variance, SRS and STR for example, were not consistently accurate. There are two reasons for this failure: small counts and the sensitivity of the normal approximation in the confidence intervals (eq. 6).

During small runs, many of the samples contained zero redds. In these cases, the estimate of the standard error and the confidence interval were zero (Fig. 8*a*). When this occurred, our confidence intervals did not contain the true value and the accuracy suffered. In 1995 (20 redds), for the 0.05 sampling fraction and 200 m units, 437 of the 1000 SRSs had zero redds. These zeros cause a loss of precision during small runs.

The surprising performances of all the designs, but the adaptive sampling design in particular, during small run sizes illustrate an interesting dynamic where the spatial structure of the sampling units interacts with the spatial structure of the sampled populations. In this case, the interaction results in numerous samples that have no redds. This poses a problem for estimation of clustered and rare populations and illuminates the need for a solution. One possible solution is to recruit a design-based likelihood estimator when a large number of zero redd counts are observed (Nicholson and Barry 1995). One could estimate the likelihood of observing zero redds in a sample given a range of very small population sizes pre hoc, and then identify the 95% confidence intervals for the estimated likelihood. Such design-based estimators could explicitly prevent estimating zero redds (Kincaid 1997). Solutions to this interesting design problem may become increasingly important for those designing monitoring for other more rare or endangered populations.

For the larger runs, there were few zeros; however, the accuracies still consistently deviated from nominal. In this case, there appears to be a failure in the normal approximation. Although histograms of the estimates appear normal on inspection, the estimator for the total and the estimator for its standard error were highly correlated (Fig. 8*b*). This correlation existed for most of the strategies regardless of the number of redds or the sample size. For SRS with 200 m units and 0.1 sampling proportion, these correlations ranged from 0.77 to 0.93. This positive correlation would be absent in data that is normally distributed and where the variance

Table 3. Precision for strategies using 1000 m units as measured by coefficient of variation × 100 (eq. 7).

Strategy	Run size 20	83	110	318	424	661	1730	1789	2271
(*a*) Sampling fraction = 0.05 (*n* = 35).									
SRS	129.1	63.8	66.8	50.8	43.0	35.4	33.8	35.8	32.2
SYS*	115.6	59.6	62.8	42.9	39.3	29.6	27.8	31.9	25.0
GRTS*	65.4	47.4	172.0	21.3	35.6	24.5	24.0	25.2	18.2
STR	94.4	49.9	49.6	36.0	35.2	29.3	27.4	30.3	25.4
Adaptive-n	127.7	58.5	59.7	44.5	38.1	29.8	25.0	24.4	23.5
Adaptive-En	131.8	65.7	69.7	62.8	61.3	56.9	61.6	66.0	63.7
(*b*) Sampling fraction = 0.1 (*n* = 69).									
SRS	85.6	44.3	46.3	35.2	29.8	24.6	23.4	24.9	22.3
SYS*	81.7	38.9	38.0	23.6	22.2	19.6	16.9	19.1	17.0
GRTS*	38.9	29.9	47.8	21.6	23.8	18.4	17.0	15.4	14.9
STR	64.4	34.2	33.9	24.6	24.2	20.2	18.9	20.9	17.5
Adaptive-n	88.6	40.6	41.4	30.9	26.5	20.7	17.3	16.9	16.3
Adaptive-En	91.6	45.5	48.3	42.6	42.1	38.9	41.6	44.3	42.8
(*c*) Sampling fraction = 0.29 (*n* = 204).									
SRS	46.1	22.8	23.8	18.1	15.4	12.6	12.1	12.8	11.5
SYS*	50.1	21.3	16.3	12.1	10.2	7.6	7.3	6.7	6.8
GRTS*	40.2	19.7	21.1	13.4	13.2	10.6	11.0	8.4	8.6
STR	28.5	16.1	15.5	11.2	11.9	10.1	9.4	10.5	8.6
Adaptive-n	45.6	20.9	21.3	15.9	13.6	10.7	8.9	8.7	8.4
Adaptive-En	46.8	23.2	24.3	20.8	19.7	18.3	19.8	20.9	21.7

Note: The most precise strategies for each run size (column) are shaded; when more than one strategy is highlighted within a column, the difference in precision is unclear because of simulation error. SRS, simple random sampling; SYS, systematic sampling; GRTS, spatially balanced design; STR, stratified sampling; Adaptive-n, adaptive cluster sampling when *n* = the original sample size; Adaptive-En, adaptive cluster sampling when *n* = the expected final sample size.

*Precision was evaluated through simulation.

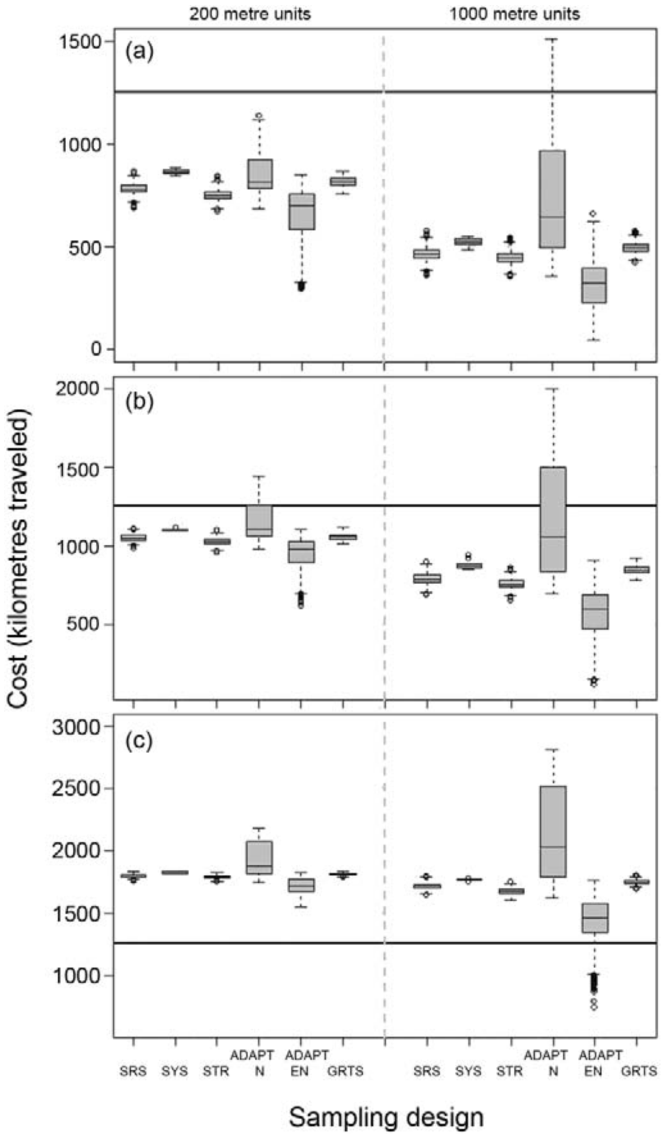
and the mean are independent. However, in the MFSR redd data, when one underestimates the number of redds, one also underestimates the variance of the estimator, resulting in a loss of accuracy. This effect was most certainly due to the fact that we were analyzing counts that are bounded by zero, making estimators based on the central limit theorem slow to converge in this way (one needs large counts or very large samples). Various transformations of the data were unable to eliminate this consequence of zero counts.

Accuracy of the SYS strategy

The trends in accuracy for the SYS design differed from the other strategies. This strategy was more accurate at intermediate run sizes than at large run sizes, and its confidence intervals appeared too wide and contained the true value too often, resulting in empirical levels above 95%. This situation results because the estimators for the standard error were biased high (Fig. 9); thus, this overestimated estimator counteracts the zero and mean–variance correlation problems discussed earlier. The interaction seemed to work in our favor for the intermediate run sizes, but the overestimated variance dominated for larger run sizes (accuracies >95%), while the zero problem dominated for the smaller run sizes (accuracies <95%).

The reasons for the failure of these estimators were not intuitive. Systematic designs perform well because their implicit spatial balance avoids bad samples that occur under many of the other designs — bad samples being those that have all the sampled locations occurring in some small or limited portion of the watershed. These bad samples with poor spatial balance across the entire sampled universe are possible in designs like SRS. Since the naïve standard error estimator based on SRS (eq. 3) assumes that every sample is possible, including these bad samples, it tends to overestimate the standard error of the systematic samples and is therefore biased for these designs. The properties of the neighborhood estimator, on the other hand, vary depending on the sampling proportion and the stochastic nature of the population response. These factors interact because the neighborhoods are not defined by the responses of points within some fixed radius in space, but rather the population responses measured at the four closest points. For high sampling fractions, the estimator for the variance will depend on an average of sampled points very close together, which one might expect to be relatively well correlated, all else being equal. However, when sampling a small proportion of the watershed, the neighborhood is made up of observations farther apart and is therefore expected to be less correlated. Additionally, as the response becomes increasingly stochastic, spatial correlation will fall faster for a given increase in neighborhood size, and as a result the standard error estimator will increase even more quickly. Furthermore, Stevens and Olsen (2003) report that if the population has no spatial structure, the local estimates of variance estimate the plan-

Fig. 7. Distribution of costs (eq. 8) for the sampling strategies: (*a*) 0.05 sampling fraction, (*b*) 0.1 sampling fraction, and (*c*) 0.29 sampling fraction. The solid horizontal line is the cost of index sampling (1254.4 km).



ning variance and hence the average of these is a good estimator; however, it may overestimate the variance if the population is clustered (Stevens and Olsen 2003).

The bias in these estimators suggests that when employing an SYS design, use the successive difference estimator (eq. 4) during small runs and the neighborhood estimator (eq. 5) during large runs. However, even when implementing such a complicated strategy, the benefit is marginal, since there is still a substantial relative bias that consistently exceeds 0.05 (Fig. 9).

STR has other benefits

Stratifying by index performed well in terms of its precision and accuracy and has the added benefit of providing estimates for the number of redds in the index reaches with a precision controlled by the sampler (by allocating the samples intelligently). This advantage is extremely important for many managers, as the resulting estimate can be used to continue the long time series of index samples that may exist (50+ years in the MFSR). For stratified designs, however, allocation of sampling units to different strata will be problematic. Optimal allocation, as used here, is not equal to proportional allocation and is only coincidentally close to equal allocation. Equal allocation is not recommended, but optimal allocation requires information that is typically unknown. In areas where there are few index reaches, an allocation that allows a census of the index stratum seems logical. However, in areas like the MFSR, where the index sites make up almost 30% of the watershed, a census will be difficult if not impossible, and it is not clear how to decide on allocation.

ACS

It is surprising that the ACS strategy did not perform better. ACS is expected to perform well in the case of a rare, clustered population (Thompson 1983), which seems to describe the redd distributions we observed. There are several possible explanations for this failure. The first is that the scale of the sampling unit and the clustering do not correspond. For example, if the scale of the clustering is much larger than the sampling unit, finding a redd in a unit from the original sample does not indicate that the adjacent units are likely to contain redds. A similar problem occurs when the opposite mismatch occurs. The idea of designing an ACS based on characteristics (scale) of the point process has been explored elsewhere (Thompson 1983). A second possible reason that ACS did poorly is that in small-run years, when the clustering is most apparent, the initial sample rarely contains redds and does not meet the criteria for sampling the adjacent units; while in the large-run years the redds are no longer strongly clustered, overestimating the number of redds. ACS also suffers from more practical issues, such as the design makes planning difficult because the final sample size is unknown. This complicates predicting costs, and it makes it difficult for crews to plan trips when they have no idea how long a region will need to be sampled on any one day. Furthermore, the estimator's and sample's complexity may make trend estimation and sharing the data with others difficult.

Segments vs. points

We sampled segments of streams rather than points on the stream network. To sample segments requires frame construction in advance of sampling, which might be prohibitive. However, segmenting the network beforehand leads to the relative simplicity of a finite population and a situation in which many of the ambiguities that can arise during frame construction (such as how to deal with confluences and edge effects) being addressed before sampling begins. In contrast, the EPA recommends an alternative approach, whereby points are sampled from an infinite set along the stream network, and then measurement units (the stream reaches) are defined around each selected point. The point strategy does not require initial frame development, but it adds complications such as the possibility for an infinite number of sampling units, overlapping measurement units, ambiguity at confluences, and measurement units extending beyond the edge of the universe or strata (Stevens and Urqu-

Table 4. Precision for strategies keeping costs fixed so that 95% of the samples have cost less than or equal to 1253.4 m (i.e., the cost of the index strategy).

		Run size								
Strategy	n	20	83	110	318	424	661	1730	1789	2271
200 m units										
SRS	476	63.1	30.0	28.9	19.3	16.1	13.3	11.4	11.8	10.4
SYS*	458	59.4	29.6	30.4	14.8	13.8	11.3	7.8	7.5	6.9
GRTS *	467	60.0	30.5	28.0	15.4	12.1	9.5	9.1	7.7	7.1
STR	486	47.0	25.3	23.3	15.0	14.1	11.6	10.0	10.8	8.8
Adaptive-n	258–478	63.0	29.7	27.8	19.2	16.1	13.6	12.5	13.0	12.4
1000 m units										
SRS	123	64.1	31.7	33.2	25.2	21.4	17.6	16.8	17.8	16.0
SYS*	116	61.1	33.8	31.7	20.7	14.2	10.8	11.8	13.8	10.5
GRTS *	122	48.5	23.3	31.1	18.4	21.5	13.1	14.4	12.2	10.8
STR	130	41.9	22.8	22.4	16.2	16.4	13.8	12.9	14.3	11.9
Adaptive-n	31–122	63.7	89.8	29.6	25.3	24.2	22.3	23.3	25.2	25.1

Note: Precision is measured by the coefficient of variation × 100 (eq. 7). The most precise strategies for each column (run size) are shaded; when more than one strategy is highlighted within a column, the difference in precision is unclear because of simulation error. SRS, simple random sampling; SYS, systematic sampling; GRTS, spatially balanced design; STR, stratified sampling; Adaptive-n, adaptive cluster sampling when n = the original sample size.

*Precision is evaluated through simulation.

hart 2000). Some of these complications can lead to inadvertent unequal probability designs, although many of these issues have been resolved (see for example Stevens and Urquhart 2000). If a particular watershed will be sampled every year, then spending the effort to create a finite sampling frame in advance might be worth it, but if the sampling is to occur only once, the additional work is unlikely to be cost effective.

The cost function

In constructing a cost function, we made a number of simplifications to the reality of conducting redd surveys in remote locations. The cost function we used is necessarily simple and particular to one implementation in the MFSR. Indeed, given that the MFSR is censused with helicopter surveys, our choices were de facto arbitrary. There are myriad cost functions that account for diverse modes of transport and eccentricities of specific watersheds; however, we needed a cost function that was relatively realistic in accounting for the choices researchers were likely to make in other watersheds — primarily balancing numbers of access points and travel by foot with total number of samples. The cost function we used did not include or assume equal travel to all access sites; because some sections of the MFSR are remote, these costs might be substantial or differ substantially. The cost function did not account for crew trips. When there are many sample sites at a particular access point, crews might need to camp or visit the site more than once. Crews cannot remain on site indefinitely and additional costs accrue if the crews have to camp overnight. Additionally, this cost function is specific to a watershed where roads are not common. In watersheds where roads are common, crews might hike from the road to individual sites and back again, essentially driving from site to site. Road distance and the distance from the road to the sites rather than distance along the network would be important in that case. However, in spite of the various specifics of the cost function we used, analyzing sampling effort was useful in demonstrating the approach to balancing statistical precision with allocated resources in the design process.

Sampling larger areas

It is possible that these designs will not perform competitively when scaled up to larger areas. Although there is no evidence that these designs are scale-dependant, most large-scale designs will be more complex, having two or several stages of sampling. Currently there are several large-scale, multistage monitoring programs implemented in the Pacific Northwest to assess the health of endangered anadromous fish populations and (or) their habitat. These programs involve numerous federal, state, and local management authorities that have diverse mandates to monitor stream networks. An important feature of these sampling designs is that in each case, sampling is a two-step process. In the first step, large sampling units (e.g., level 6 HUCs) are chosen with some form of randomization scheme. In the second step, stream reaches within the selected HUCs are selected. This second stage is where our results can be applied.

Recommendations

Sampling objectives and resources (budget) should first dictate which strategy is appropriate. For example, if precise estimates of status within the index region are necessary, then the STR strategy should be used. After these considerations, this study suggests sampling as many small units as possible and that designs that implicitly spread the sample out in space are most precise (SYS or GRTS), but there is an advantage to stratifying by the index reaches during small runs. Stratifying by index reaches demonstrated consistently competitive performance under the range of observed conditions, even if performance was not outstanding under any circumstance; this suggests that some form of stratification is a reasonable basic design template. This generally good alternative will also benefit from reduced costs (although modestly) when oversampling the easy-to-access index stratum. However, while some form of stratification

Table 5. Precision per average cost in 1/(km traveled) × 1000 (eq. 9).

	Run size								
Strategy	20	83	110	318	424	661	1730	1789	2271
(*a*) Sampling fraction = 0.05.									
200 m units (*n* = 170)									
SRS	1.16	2.43	2.52	3.79	4.54	5.49	6.41	6.19	7.00
SYS*	1.07	2.23	2.29	4.31	5.06	5.64	7.84	7.70	8.58
GRTS *	1.28	2.45	2.48	3.60	4.87	5.27	7.36	6.93	8.93
STS	1.54	2.90	3.14	4.87	5.26	6.38	7.46	6.88	8.42
Adaptive-n	1.16	2.46	2.62	3.85	4.67	5.57	6.28	6.04	6.28
Adaptive-En	1.15	2.45	2.61	3.80	4.60	5.51	6.42	6.26	6.55
1000 m units (*n* = 35)									
SRS	1.67	3.37	3.22	4.24	6.01	6.08	6.37	6.01	6.70
SYS*	1.66	3.22	3.06	4.48	4.89	6.48	6.92	6.03	7.69
GRTS *	3.10	4.28	2.81	9.52	5.68	8.26	8.43	8.04	11.15
STR	2.38	4.50	4.55	6.22	6.37	7.66	8.19	7.41	8.84
Adaptive-n	1.67	3.56	3.40	4.03	4.04	4.33	4.13	3.81	3.73
Adaptive-En	1.71	3.86	3.76	5.05	5.37	6.50	7.34	6.68	7.11
(*b*) Sampling fraction = 0.1.									
200 m units (*n* = 340)									
SRS	1.24	2.61	2.72	4.06	4.87	5.89	6.88	6.64	7.51
SYS*	1.31	2.32	3.23	5.42	5.62	6.33	8.67	8.84	9.96
GRTS *	1.44	2.68	2.72	4.29	6.20	7.11	8.27	8.54	9.46
STR	1.66	3.10	3.36	5.23	5.61	6.80	7.94	7.31	8.98
Adaptive-n	1.24	2.65	2.82	4.13	4.98	5.95	6.70	6.47	6.93
Adaptive-En	1.24	2.63	2.86	4.04	4.84	5.71	6.25	6.02	6.08
1000 m units (*n* = 69)									
SRS	1.40	2.84	2.72	3.57	4.22	5.12	5.36	5.06	5.61
SYS*	1.40	2.92	2.99	4.81	5.12	5.79	6.73	5.94	6.67
GRTS *	3.04	3.94	2.47	5.47	4.95	6.40	6.92	7.66	6.65
STR	2.04	3.86	3.88	5.35	5.44	6.54	6.99	6.32	7.55
Adaptive-n	1.41	3.00	2.87	3.48	3.59	3.94	3.88	3.70	3.62
Adaptive-En	1.43	3.15	3.05	3.96	4.22	1.93	5.38	4.91	5.20
(*c*) Sampling fraction = 0.29.									
200 m units (*n* = 1000)									
SRS	1.41	2.96	3.07	4.60	5.52	6.67	7.79	7.52	8.51
SYS*	2.34	3.35	4.68	5.72	5.94	7.22	9.50	12.24	13.65
GRTS *	1.49	2.88	3.11	5.31	6.89	8.88	9.31	9.80	12.64
STR	2.05	3.62	3.98	6.24	6.42	7.76	9.00	8.20	10.29
Adaptive-n	1.41	2.99	3.19	4.68	5.66	6.80	7.95	7.70	8.48
Adaptive-En	1.41	2.98	3.17	4.61	5.53	6.56	7.37	7.02	7.50
1000 m units (*n* = 204)									
SRS	1.26	2.56	2.44	3.21	3.80	4.61	4.83	4.56	5.06
SYS*	1.13	2.66	3.48	4.67	5.55	7.42	7.81	8.42	8.33
GRTS *	1.42	2.90	2.71	4.28	4.35	5.41	5.19	6.77	6.64
STR	2.10	3.70	3.85	5.35	5.04	5.94	6.35	5.71	6.95
Adaptive-n	1.27	2.71	2.61	3.30	3.62	4.26	4.47	4.49	4.48
Adaptive-En	1.27	2.72	2.61	3.27	2.49	3.91	3.90	3.70	3.66

Note: Larger numbers indicate the most cost-effective strategies. SRS, simple random sampling; SYS, systematic sampling; GRTS, spatially balanced design; STR, stratified sampling; Adaptive-n, adaptive cluster sampling when *n* = the original sample size; Adaptive-En, adaptive cluster sampling when *n* = the expected final sample size.

*Precision evaluated through simulation.

is recommended, there are still a number of design decisions within and between strata. What one does within strata is another matter. Given that the SYS and GRTS strategies performed consistently better than SRS, either of these designs should be used within each stratum. As for allocation of sampling resources to the two strata, oversampling the index reaches is appropriate for three reasons: (*i*) the estimators for the status within the index region will be more precise so that the long time series can be continued; (*ii*) the status estimators for the entire watershed will be more

Fig. 8. Problems with accuracy for all designs. (*a*). The sampling distribution of the estimate of status for simple random sampling (SRS) of 170 200 m units during 1995. 43.7% of the samples contained no redds, resulting in estimates of status and standard error equal to zero. When this happened, the confidence interval was also zero. The true value, 20 redds, is marked by the vertical line. (*b*) Estimates of the total number of redds and its standard error from SRS 170 200 m units during 2002. These display a positive correlation of 0.83. Solid points indicate iterations whose confidence interval contained the true value, while open point iterations did not; only 93.2% of the points are solid, indicating that this strategy is not accurate despite theoretical considerations. The open triangle locates the true values: 1789 redds and standard error 342.1.

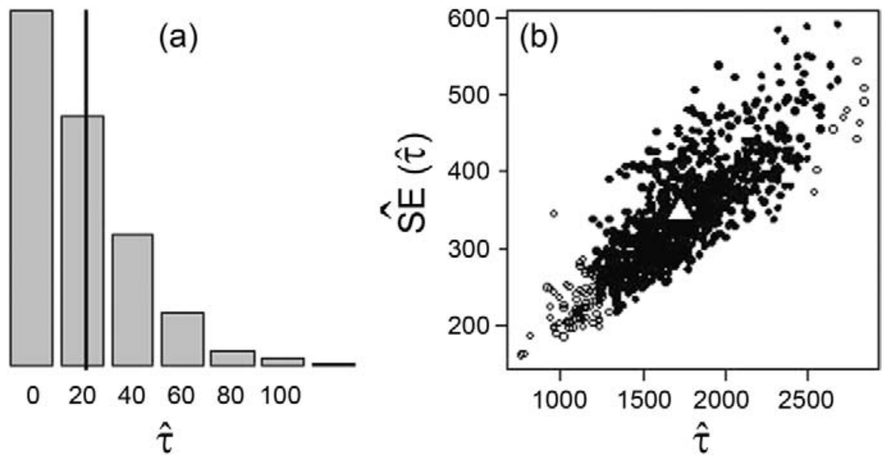
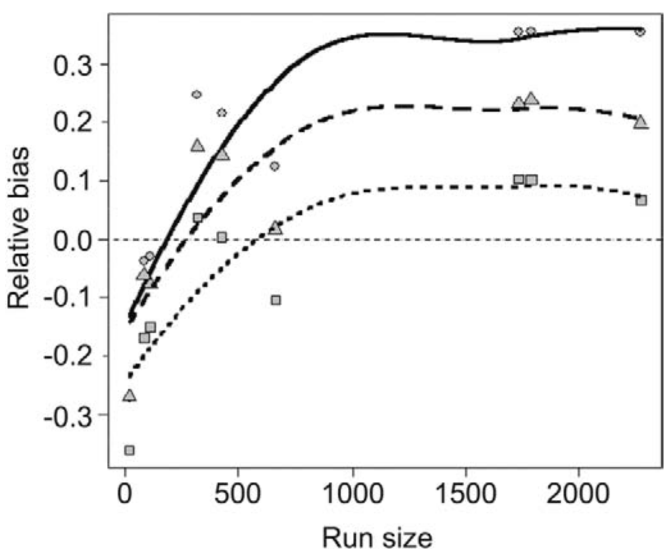


Fig. 9. Relative bias for the standard error for systematic sampling with n = 170 (sampling proportion 0.05 and 200 m units). The lines are smooth lines (LOESS, Cleveland 1979) to show the trends. The bias of the estimators depends on the run size, with all three showing negative bias during small runs and positive bias during large runs.



precise; and (*iii*) costs will be lower. Finally, it should be noted that while the focus of this work was on the design of sampling schemes for aquatic resources on river networks, Chinook salmon redds in particular, it would not have been possible to do these analysis were it not for the time series of censuses in the MFSR. As such, for large-scale monitoring program design, we recommend one census-based data collection if possible to serve in sampling design guidance and assessment.

Acknowledgements

We thank Steven Smith, George Pess, Martin Liermann, and two anonymous reviewers for constructive critiques of the manuscript and Blake Feist for expert assistance on the figures. This work was funded in part by the Bonneville Power Administration (Project 2003-017 to CEJ).

References

Bellhouse, D.R. 1988. Systematic sampling. *In* Handbook of statistics. *Edited by* P.R. Krishnajah and C.R. Rao. North Holland, Amsterdam. pp. 125–144.

Bellman, R.E. 1961. Adaptive control processes: a guided tour. Princeton University Press, Princeton, N.J.

Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. **74**: 829–836. doi:10.2307/2286407.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. Wiley, New York.

Conquest, L.L. 2002. Biomonitoring. *In* Encyclopedia of environmetrics. Vol. 1. *Edited by* A.H. El-Shaarawi and W.W. Piegorsch. John Wiley & Sons, Ltd., Chichester, UK. pp. 199–205.

Conquest, L.L., and Ralph, S.C. 1998. Statistical design and analysis considerations for monitoring and assessment. *In* River ecology and management: lessons from the Pacific coastal ecoregion. *Edited by* R.J. Naiman and R.E. Bilby. Springer, New York. pp. 455–475.

Elms-Cockrum, T.J. 1998. Salmon spawning ground surveys, 1997. Idaho Department of Fish and Game, Pacific Salmon Treaty Program Technical Report. Available from Idaho Department of Fish and Game (IDFG), Boise, Idaho, USA.

Fisher, R.A. 1935. The design of experiments. Oliver and Boyd, Edinburgh, UK.

Halbert, C. 1993. How adaptive is adaptive management? Implementing adaptive management in Washington State and British Columbia. Rev. Fish. Sci. **1**: 261–283.

Hassemer, P.F. 1993. Salmon spawning ground surveys, 1989–92. Idaho Department of Fish and Game, Boise, Idaho. Proj. F-73-R-15.

Healey, M.C. 1991. Life history of Chinook salmon. *In* Pacific salmon life histories. *Edited by* C. Groot and L. Margolis. UBC Press, Vancouver, B.C. pp. 313–393.

Hilborn, R. 1979. Some failures and successes in applying systems analysis to ecological systems. Journal of Applied Systems Analysis, **6**: 25–31.

ICBEMP. 1999. Interior Columbia River Ecosystem Management Project. Spatial data, October 1999. ICBEMP, Portland, Ore. Available from www.icbemp.gov/spatial/expnm/hyd.shtml.

Isaak, D.J., and Thurow, R.F. 2006. Network-scale spatial and temporal variation in Chinook salmon (*Oncorhynchus tshawytscha*) redd distributions: patterns inferred from spatially continuous replicate surveys. Can. J. Fish. Aquat. Sci. **63**: 285–296. doi:10.1139/f05-214.

Isaak, D.J., Thurow, R.F., Rieman, B.E., and Dunham, J.B. 2003. Temporal variation in synchrony among chinook salmon (*Oncorhynchus tshawytscha*) redd counts from a wilderness area in central Idaho. Can. J. Fish. Aquat. Sci. **60**: 840–848. doi:10.1139/f03-073.

Kincaid, T.M. 1997. Estimating absence. Ph.D. thesis, Oregon State University, Corvallis, Ore.

Larsen, D.P., Kincaid, T.M., Jacobs, S.E., and Urquhart, N.S. 2001. Designs for evaluating local and regional scale trends. Bioscience, **51**: 1069–1078. doi:10.1641/0006-3568(2001)051[1069:DFELAR]2.0.CO;2.

Lichatowich, J.A. 1999. Salmon without rivers: a history of the Pacific salmon crisis. Island Press, Washington, D.C.

Lohr, S.L. 1999. Sampling: design and analysis. Duxbury Press, Pacific Grove, Calif.

Matthews, G., and Waples, R. 1991. Status review for Snake River spring and summer Chinook salmon. Department of Commerce, National Oceanic and Atmospheric Administration, Northwest Fisheries Science Center, Seattle, Wash. NOAA Fisheries Tech. Memo. No. NMFS-NWFSC- 200, June 1991. Available from www.nwfsc.noaa.gov/publications/techmemos/tm200/tm200.htm.

Myers, J.M., Kope, R.G., Bryant, G.J., Teel, D., Lierheimer, L.J., Wainwright, T.C., Grant, W.S., Waknitz, F.W., Neely, K., Lindley, S.T., and Waples, R.S. 1998. Status review of Chinook salmon from Washington, Idaho, Oregon, and California. Department of Commerce, National Oceanic and Atmospheric Administration, Northwest Fisheries Science Center, Seattle, Wash. NOAA Fisheries Tech. Memo. No. NMFS-NWFSC-35. Available from www.nwfsc.noaa.gov/publications/techmemos/tm35/.

Nehlsen, W., Williams, J.E., and Lichatowich, J.A. 1991. Pacific salmon at the crossroads: stocks at risk from California, Oregon, Idaho, and Washington. Fisheries, **16**: 4–21.

Nicholson, M., and Barry, J. 1995. Inferences from spatial surveys about the presence of an unobserved species. Oikos, **72**: 74–78. doi:10.2307/3546040.

Stehman, S.V., and Overton, S.W. 1994. Environmental sampling and monitoring. *In* Handbook of statistics. Vol. 12. *Edited by* G.P. Patil and C.R. Rao. Elsevier Science B.V., Amsterdam. pp. 263–306.

Stevens, D.L., Jr., and Olsen, A.R. 1999. Spatially restricted surveys over time for aquatic resources. J. Agric. Biol. Environ. Stat. **4**: 415–428. doi:10.2307/1400499.

Stevens, D.L., Jr., and Olsen, A.R. 2003. Variance estimation for spatially balanced samples of environmental resources. Environmetrics, **14**: 593–610. doi:10.1002/env.606.

Stevens, D.L., Jr., and Urquhart, N.S. 2000. Response designs and support regions in sampling continuous domains. Environmetrics, **11**: 13–41. doi:10.1002/(SICI)1099-095X(200001/02)11:1<13::AID-ENV379>3.0.CO;2-8.

Thompson, S.K. 1983. Adaptive sampling of spatial point processes. Ph.D. thesis, Oregon State University, Corvallis, Ore.

Thompson, S.K. 1992. Sampling. Wiley, New York.

Thurow, R.F. 2000. Dynamics of Chinook salmon populations within Idaho's Frank Church wilderness: implications for persistence. *In* Wilderness Science in a Time of Change Conference. Vol. 3. Wilderness as a Place for Scientific Inquiry, 23–27 May 1999, Missoula, Montana. *Edited by* S.F. McCool, D.N. Cole, W.T. Borrie, and J. O'Loughlin. USDA Forest Service, Rocky Mountain Research Station, Ogden, Utah. Proceedings, RMRS-P-15-VOL-3. pp. 143–151.

Waples, R.S. 1991. Pacific salmon, *Oncorhynchus* spp., and the definition of "species" under the Endangered Species Act. Mar. Fish. Rev. **53**: 11–22.

Williams, B. 1978. A sampler on sampling. John Wiley & Sons, New York.