PERSPECTIVE:
FISH HABITAT

John G. Williams

Williams is an independent consultant in Davis, California.He can be contacted at jgwill@dcn.org.

Sampling for Environmental Flow Assessments

ABSTRACT: Environmental flow studies normally involve sampling. Except for hydrologically based methods, the studies usually are conducted only on part of the reach of stream to which the study results will be applied. This assumes that the study sites, or the samples, are representative of the stream as a whole. There is no way to be sure of this without studying the whole stream, but it is possible to estimate how close the study results and the "true" results are likely to be, if the study uses a probability (random) sampling design that allows for estimating confidence intervals. Sampling designs that allow for interval estimates of statistics are normal in science, but, strangely, not in environmental flow studies. Here, I describe and discuss some appropriate sampling designs and related considerations in the context of environmental flow studies. I then describe and critique the sampling in some recent studies using the Physical Habitat Simulation System (PHABSIM), the method most commonly used in the United States, which I describe briefly for readers not familiar with it. Finally, I explain why deliberate selection of samples gives biased results, and describe the experience that persuaded statisticians of this fact early in the twentieth century.

Muestreo para valoración de caudales ambientales

RESUMEN: la evaluación de caudales ambientales comúnmente implica actividades de muestreo. Con excepción de los métodos hidrológicos, los estudios normalmente se realizan sólo en una parte de la extensión total de un río, sobre la cual los resultados serán aplicables. Esto supone que los sitios de estudio, o las muestras obtenidas, son representativas del cauce entero. No hay forma de tener seguridad acerca de la veracidad de esta práctica más que estudiando un río en su totalidad, pero es posible estimar la certidumbre de dichos resultados si el estudio se basa en un diseño de muestreo (aleatorio) probabilístico que permita estimar intervalos de confianza. Los diseños de muestreo que permiten estimar la confiabilidad estadística son comunes en la ciencia, pero curiosamente no lo son en estudios de caudales ambientales. En la presente contribución se describen y discuten algunos diseños de muestreo de este tipo, y consideraciones relacionadas, en el contexto de estudios de causales ambientales. Después se describe y critica el muestreo realizado en estudios recientes que utilizan el sistema de simulación de Hábitat Físico (PHABSIM), considerado como el método más comúnmente utilizado en los Estados Unidos de Norteamérica, el cual se explica brevemente para los lectores no familiarizados. Finalmente se menciona por qué la selección deliberada de muestras produce sesgo en los resultados y se comenta la experiencia que persuadió de este hecho a los estadísticos a principios del siglo veintiuno.

DUDLEY REISER, R2 RESSOURCE CONSULTANTS

Introduction

In 2008, the American Fisheries Society adopted a strong resolution calling for funding of instream flow protection programs, and making good recommendations regarding attributes of instream flow recommendations or the studies that produce them (AFS 2008). The resolution is silent, however, about one aspect of instream or environmental flow studies that often falls short of normal scientific practice: sampling.

Environmental (or instream) flow assessments, studies to inform decisions about how much water to leave in a river or stream to maintain environmental resources, generally involve sampling. Except for hydrologically based methods such as the Tenant or the range of variation methods (Tenant 1976; Richter et al. 1997), the environmental flow methods (EFMs) normally are applied only to part of the reach of stream that will be affected by the decision, and the results are extrapolated to the rest of it. This procedure assumes that the sample is representative of the whole.

Sampling is ordinary in science and is a well-developed part of statistics, but has received surprisingly little attention in environmental flow assessments (Williams 2010). For example, Annear et al. (2004), in an otherwise comprehensive review, do not discuss sampling at all. Unfortunately, most applications of EFMs use deliberate or purposive methods rather than random (probability) methods to select samples, so the assumption that the sample represents the stream lacks a solid foundation, and it is not possible to assess quantitatively how good the representation is likely to be. There are good reasons, rooted in practical experience as well as in theory, why random or probability sampling is standard practice in science and in other fields such as opinion polling that rely on survey sampling: "Without a census, a statistical survey with the incorporation of probability sampling is the only way to assure the selection of a representative sample from which can be drawn unbiased conclusions about the population as a whole" (Stevens et al. 2007).

Sampling for EFMs is essentially similar to sampling for other kinds of studies, so textbooks such as Cochran (1977), Jesson (1978), or Thompson (2002) provide comprehensive treatments that are applicable to EFMs. However, the lack of attention to proper sampling approaches in environmental flow assessments indicates a need for a brief review focused on EFMs. Here, I provide such a review, using examples from applications of the Physical Habitat Simulation System (PHABSIM), probably the best known and widely used EFM in the United States, which calculates an index of habitat called weighted usable area (WUA) as a function of discharge (See Box 1 for a brief description of PHABSIM). I refer primarily to PHABSIM because many readers will be familiar with it, but the discussion also applies to other methods that rely on samples, such as the numerical habitat models described by Guay et al. (2000) or Railsback and Harvey (2001), or the demonstration flow assessment (DFA) method described by Railsback and Kadvany (2008).

Environmental flow methods generally estimate an index of habitat value at a given discharge from samples of transects or short reaches of stream, which begs the question, how good are the estimates? We would like to know how close the estimated value of the index is to the unknown "true" value of the index for the river reach of interest at that discharge, the value that could be calculated by someone with perfect knowledge of the river. Without knowing the "true" value, this question cannot be answered. However, if an appropriate sampling design is used, it is possible to estimate how similar the estimate and the "true" value are likely to be. That is, one can calculate an interval estimate, an interval within which the unknown "true" value will fall with a given probability, i.e., a confidence interval. Statisticians regard this approach to estimation as basic, and it is conventional in science. Castleberry et al. (1996) urged that it be used in applications of PHABSIM. However, PHABSIM and other EFMs generally do not use this approach. Instead, sample sites or transect locations usually are selected deliberately, and the results of EFM studies are presented as point estimates at a given discharge, rather than as interval estimates.

Developing interval estimates requires an appropriate sampling design, and my main objective is to provide guidance regarding such designs for EFMs. Following a general exposition, I present four case studies to illustrate the main points. I also touch on aspects of study planning that are related to sampling designs, and the empirical evidence that convinced statisticians of the need for probability sampling.

Aspects of study design

Study objectives

A good sampling design is only part of a good study design (Box 2). Specifying the problem to be addressed is always a critical step in any study, since it is otherwise impossible to tell whether the study will actually address the problem. As noted by Cochran (1977), "A lucid statement of the objectives is most helpful. Without this, it is easy in a complex survey to forget the objectives when engrossed in the details of planning, and to make decisions that are at variance with the objectives." Typically, the objective in environmental flow studies will be to define a relationship between discharge and habitat value for a given reach of stream. Usually, this will not be done for its own sake, but rather to serve some management purpose that provides context for the study.

One factor that should be considered in environmental flow studies, but typically is not, is how accurate the estimates of habitat value resulting from the study need to be to meet the purpose of the study. This could be visualized in terms of an accuracy/utility function (Figure 1). (By accuracy, I mean both precision and lack of bias.) If the study is intended to inform a decision about instream flow standards, it may be useful to consider the flow increments that the decision-maker is likely to consider. If such increments are about 10%, say from 10 to 11 m^3/s, there may not be much point to developing an estimate that is precise enough to support a decision between 10.0 and 10.5 m^3/s. On the other hand, a study that cannot usefully distinguish the habitat values at 10 and 12 m^3/s would be less accurate than would be desirable. Accuracy must be balanced against effort and expense, and very likely a study that is as accurate as is desirable will not be practicable, but if the results are expected to be too inaccurate, then the utility of the study should be questioned.

Sampling universe

A clear definition of what is to be sampled is important. Defining this "sampling universe" may be easy for some environmental flow assessments, since flow decisions generally apply to a well defined reach of stream, but it may be a major challenge in some cases. Even when the reach of interest is well defined, access to parts of it may not be available for practical or legal reasons. Then, the actual sampling universe is the accessible parts of the reach, to which any conclusions from the study should be restricted. If possible, the actual

Box 1. Summary description of PHABSIM.

PHABSIM is a collection of hydraulic and biological models used to assess the value of habitat in a stream as a function of discharge for a particular species or life stage. PHABSIM operationally defines and estimates a suitability S for a species or life stage, and uses S ($0 \le S \le 1$) to weight the area of the stream, yielding a statistic called weighted usable area (WUA). Conceptually, S varies continuously over the surface of a stream, and is defined operationally in terms of "microhabitat" variables; usually these are water depth, velocity, and substrate size or cover, but sometimes other variables such as distance to cover or velocity gradient. Substrate and cover are estimated by field surveys, usually as categorical variables, but water depth and velocity are estimated over a range of discharge with a hydraulic model. The biological models normally used to calculate S for depth and velocity are curves called habitat suitability criteria (HSC), that vary between 0 and 1 as a function of one of the microhabitat variable at points or small areas in the stream. For categorical variables such as substrate, a suitability value is assigned to each category (Figure B1). The sampling and modeling issues involved with developing habitat selection models are well covered in Manly et al. (2002), and the discussion applies to HSC as well.

The hydraulic models used in PHABIM are usually one-dimensional (1-D), with which depth and velocity are estimated at points along transects perpendicular to the flow. Increasingly, however, two-dimensional (2-D) models are being used, with which areas of the stream are divided into many small cells or tiles, and depth and velocity are estimated for each. At a given discharge, the values of the suitability curves for each point or cell are combined, often by simple multiplication, to calculate WUA.

Details are complicated by the large number of options available in PHABSIM, but in the general case, with the 1-D models, the river reach of interest is represented by a set of transects, and each transect is assumed (e.g., through "weighting factors") to represent some fraction of the total reach. One of several hydraulic models is used to estimate water depth and velocity at usually 20 or more points along each transect. The values at the points are assigned to "cells" with nominal areas calculated from the weighting factors and the distance between the points on the transect. WUA for the transect at a given discharge is estimated by:

$$\widehat{WUA}_t = \sum_{i=1}^{n} a_i \hat{S}_i$$

where the summation is over the n cells on the transect, a_i is the nominal area of the ith cell on the transect, and, for the default option,

$$\hat{S}_i = \hat{s}_{v,i}\hat{s}_{d,i}\hat{s}_{s,i}$$

and $s_{v,i}$, $s_{d,i}$, and $s_{s,i}$ are the values of the HSC for the species or life stage in question for velocity, depth, and substrate/cover for the ith cell, again at the given discharge; the "hats" indicate that the terms are estimated from samples. Repeating this process at each transect over a range of discharge and interpolating results in curves of WUA over discharge, and summing over these gives a composite curve of WUA over discharge for the reach (e.g., Figure B2). Usually, the sum is normalized by stream length, so the final results are reported as WUA per length, which has a dimension of length. The process is similar with 2-D models, except in this case the a_i are real areas that are determined by the grid of the hydraulic model.

The composite curves are the basic product of PHABSIM, although they can be used to assess flow regimes rather than specific levels of flow by combining the curves with a hydrograph to produce times series of WUA. As stated by Annear et al. (2004:149), "…the primary value of PHABSIM is its ability to identify trade-offs between streamflow and hydraulic habitat …" Bovee et al. (1998) and Waddle (2001) describe PHABSIM in detail. Recently, some investigators (e.g., Guay at al. 2000) have estimated the $\hat{S}_i$ using logistic regression with microhabitat values as independent variables, rather than using the usual HSI, but the basic process is similar.

Figure B2. WUA for juvenile Chinook salmon at the Seiad study site, Klamath River, California. Redrawn from Hardy et al. (2006).

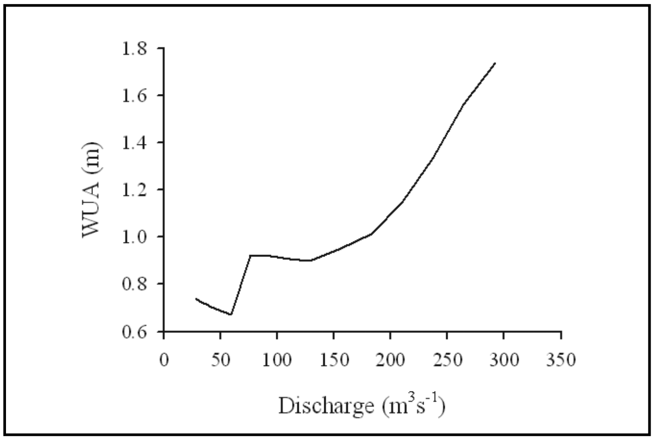
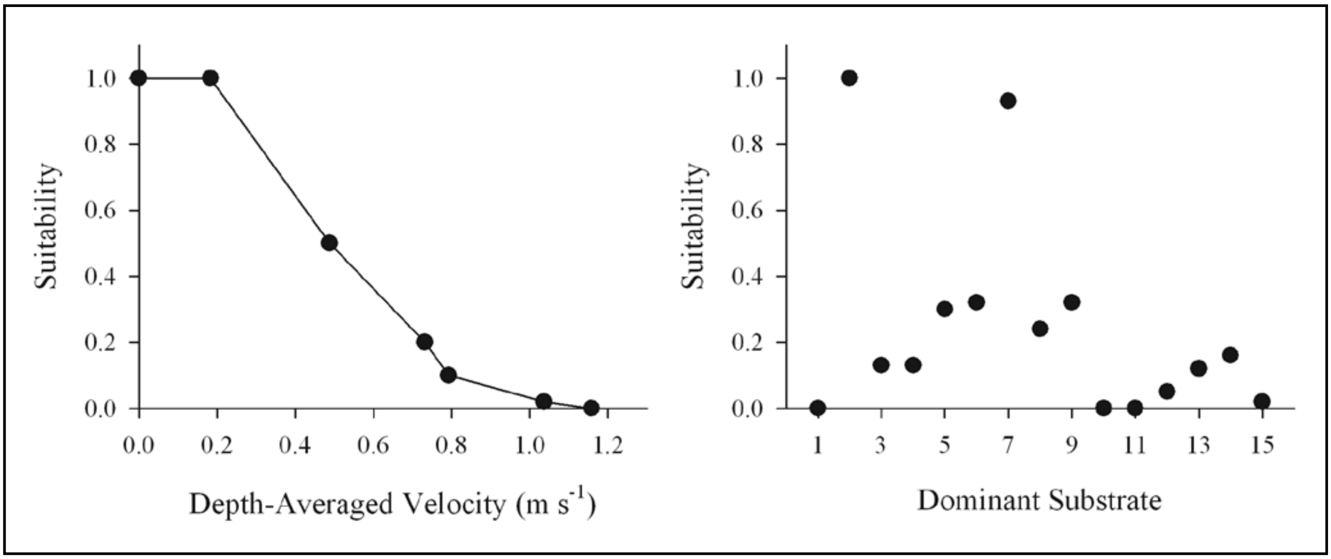


Figure B1. Suitability criteria for water velocity and dominant substrate type for juvenile Chinook salmon in the Klamath River, re-drawn from Hardin et al. 2005. Although suitability criteria are estimated from data, confidence intervals for the criteria were not calculated; presenting only point estimates is the normal practice.



sampling universe should be determined during the planning phase of the project; if it is too small a fraction of the reach of stream that will be affected by the flow decision, then the utility of the study should be reconsidered.

In other cases, the sampling universe may be restricted to certain types of habitat; for example, Gallagher and Gard (1999) applied PHABSIM to "optimal [Chinook salmon] spawning areas based on substrate size and past redd distributions," for an assessment of how well WUA predicts spawning site selection. Similarly, if the concern is for a riffle-dwelling or a pool-dwelling fish, then restricting the sampling to the relevant habitat probably is appropriate. Or, suppose that the habitat index to be used will return a habitat value of zero for sand or mud substrates, and that these can be identified reliably on aerial photographs of the stream. Effort sampling such areas would be wasted, unless the objective is to test the index. When the sampling universe is restricted to part of the stream, however, some operational definition will be needed for the areas to be included or excluded.

Sampling units

The sampling units, the things that make up the sample and determine the sample size, must also be defined. This is easy for transect-based methods, such as 1-D PHABSIM, where the transects make natural units. For area-based methods, such as DFA or 2-D PHABSIM, the study sites are the sampling units, and decisions must be made about how big the sites should be and how their boundaries will be determined. For methods incorporating 2-D hydraulic models, the sites should provide for sensible boundary conditions for the hydraulic models, so the hydraulic modeler should be involved in defining the units.

In EFMs such as PHABSIM, sampling also occurs when data on microhabitat variables are collected. When depth and velocity data are collected at transects for calibrating the 1-D models, for example, the depth is usually measured at particular points, and these values are applied to some interval across the transect. Similarly, the vertically averaged water velocity at the same points is estimated by measurements measured at one or two depths, and applied to the same interval across the transect. Finally, WUA is calculated interval by interval across the stream at a given discharge, and summed to produce an estimate for the transect. It might seem that the intervals are the basic sampling unit, so that the sample size would be the total number of intervals on the all the transects.

To see why the intervals are not the basic sampling unit from a statistical point of view, and why the total number of intervals on all transects is not the sample size, consider the consequences of making the size of the intervals across a single transect smaller and smaller. In the limit, if there were infinitely many intervals, the sample size would be infinite, but there would still be information only about that single transect. Similarly, with 2-D PHABSIM, no matter how fine the computational grid in the 2-D model, there would still be information only about the site. In terms of sampling theory, the basic point is that whenever any of the cells or tiles in a site is selected, all of them are selected, and similarly, whenever any interval on the transect is selected, all of them are selected (Thompson 2002). The reasoning is the same for systematic samples.

For 1-D methods, the number of potential sampling units is effectively infinite, but for 2-D methods, there can be only a limited number of non-overlapping units. Accordingly, as described in the textbooks cited above, a finite population correction factor should be included in the variance estimates, unless the fraction of the sampling universe contained in the samples is small.

Sampling frame

The method by which the sample will be selected—some method by which random numbers will be associated with sampling units on the stream—must also be defined. In survey sampling, this is normally a list, such as a list of telephone numbers used to select a sample for a political opinion poll. For an area-based EFM, the stream of interest might be divided into potential study sites, with the sample selected from a list of the sites. For EFMs, however, the sampling frame probably will be some representation of the stream such as US. Geological Survey (USGS) topographic maps or aerial photographs. Then, for example, a line could be drawn down the center of the channel, and distances could be measured along this line. For transect-based methods, samples can be selected by placing transects at distances along the line corresponding to the random numbers. For area-based methods, the upstream boundary of the study sites might be specified by lines across the channel at the points corresponding the selected random numbers, or, for numerical habitat models, by the first locations upstream from the points that provide reasonable boundary conditions for the hydraulic model. Division of the stream into potential study sites could also be done before the sampling, with sites being selected if the

Box 2. Steps in developing the sampling plan for an EFM.

1. Clearly define the purpose of the flow assessment.
2. Determine how accurate the estimates resulting from the study must be, in order to satisfy the purpose of the study.
3. Determine the actual sampling universe for the EFM.
4. Develop the conceptual plan for the sampling, including the sample sizes and strata to be used.
5. Estimate the accuracy/utility function for the EFM.
6. Tentatively assess whether the proposed method can satisfy the purpose of the study.
7. Develop the details of the sampling plan, and conduct the study.
8. Test whether the results meet the accuracy criterion identified in Step 2, and if not, collect more samples.

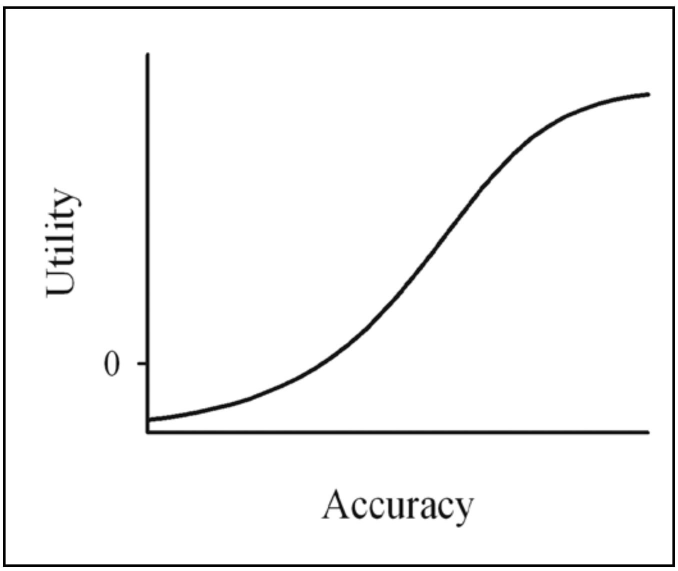
random numbers correspond to points along the stream center line that fall within the sites.

Sampling designs for PHABSIM and other environmental flow assessment methods

The basic requirement for probability (random) samples is that all elements in the sampling universe—for example all possible transect locations on the study stream in 1-D PHABSIM studies—have knowable, non-zero probability of being selected for the sample, so that the probability of the actual sample can be calculated. Much effort has gone into developing sampling designs that meet this requirement but also provide for economy of effort and spatial balance. Designs may also provide for variation in the probability that potential elements of the sample will be included in the selected sample, i.e., variation in the sampling intensity, so that more effort can go into areas thought to be important. Probability sampling does not require that knowledge or information about the study stream be ignored, but rather that it be incorporated into the study and sampling design, or into the estimator of the index of habitat used by the EFM, and not into choosing the actual sample.

The objective of the sampling design for a PHABSIM study should be to produce an estimate of the unknown "true" WUA curve for stream in question that is as accurate as possible, given the resources at hand and practical complications such as lack of access to parts of the stream, and to allow the accuracy of the estimate to be assessed. Probability sampling avoids sampling bias, but sampling bias cannot be eliminated if parts of the stream are unavailable for sampling, as is often the case. Precision is determined partly by the variation in the thing being sampled, which the investigator cannot control, but also by the sample size and the sampling design.

Figure 1. Conceptual accuracy/utility for a hypothetical EFM study. Estimating such curves for the method to be used would be a useful step in planning the project, even if the curve is based only on opinion. Accuracy could be quantified as the width of the expected confidence intervals for the index to be estimated. Williams (2010) describes an approach for anticipating the width of confidence intervals for a given sample size, for 1-D PHABSIM studies. The utility could be negative if the study is inaccurate enough to be misleading.



Economy of effort

Stratified and multi-stage sampling are common approaches to reducing the effort and cost of the sampling required to achieve a given level of accuracy. The basic idea of stratified sampling is that the sampling will be more efficient if information about the stream can be used to divide it into strata, such that sites within the strata generally are more like other sites within the same strata than sites within other strata. Using habitat types as strata is an obvious possibility, although this is not always effective (Williams et al. 2004); whether to use habitat types for strata, or the particular set of habitat types to be used, should depend on the morphology of the stream under study.

Stratification also allows for more intensive sampling in areas thought to be important. When investigators want the sample to include particular habitats, such as the shallow riffles that might cause passage problems, such habitats could be made a separate stratum. It may be preferable, however, to make the assessment of passage a separate part of the overall study. Transects at hydraulic controls that are required by step-backwater hydraulic models could similarly be put in a separate stratum, so that they are not given too much weight in the habitat modeling, or simply be left out of the habitat modeling.

With 1-D modeling, it will often be possible to model more transects on a given budget if the transects are clustered in space, so that the time spent getting people and equipment to the transects is reduced. For example, transects can be clustered within habitat types, which can be clustered within sites. This is multi-stage sampling. However, when estimating the variance of the resulting habitat-index estimates, the variance among the study sites and among the habitat types must be added to the variance among the transects within the habitat types. Thus, developing an interval estimate of the habitat index requires that there be replicate study sites containing replicate habitat types with replicate transects.

Some sort of multi-stage, stratified sampling design is common in PHABSIM studies, even if probability sampling is not. Examples are described below. Particularly if the stream is long, it is often divided into segments based mainly on discharge, especially junctions with major tributaries, and possibly also on slope, geology, etc. (Bovee et al. 1998). Study sites are then distributed among the segments (e.g., Hardy et al. 2006). This is a form of spatial stratification. Typically, sampling for 1-D PHABSIM is stratified by "mesohabitat types" such as pools, riffles, runs, etc. Bovee et al. (1998) also suggest that "representative reaches" can be selected within the segments. Transects are then placed within the representative reaches or the habitat types. Provided that they are implemented within a probabilistic sampling design, these approaches may be appropriate and effective.

Spatial balance

A common concern with simple random sampling, especially for natural resources, is that the sample may be clustered, so that parts of the area to be sampled are represented more heavily than other parts. Natural resources typically exhibit spatial patterns, and if these differ from the patterns in the sampling, the sampling will be inefficient at best and perhaps misleading. Increasing the sample size will tend to remedy this, but using a spatially balanced sampling design is more effective, and some investigators in various fields have favored using a systematic sample with a randomly selected starting point for this reason (e.g., Mier and Picquelle 2008). In a systematic sample with 1-D PHABSIM, transects would be located at a fixed intervals from a randomly chosen point. One practical drawback with systematic samples for environmental flow studies results from periodic variation in the sampling universe. For example, pool-riffle sequences in alluvial streams may have a period of about 7 bankfull channel widths, in which case transects located every 7 channel widths would likely give misleading results, as would transects located every 3.5 or 14 channel widths, etc. Also, estimating confidence intervals from systematic samples raises complications. Statisticians treat systematic samples as single samples (Thompson 2002), so for 100 transects divided into 10 systematic samples each with a random start, $n = 10$. Although standard methods for calculating confidence intervals are often used with single systematic samples (Manly 2002), this will tend to overestimate the width of the intervals if adjacent transects tend to be similar (Thompson 2002).

Spatial stratification avoids the problems associated with picking samples at a fixed interval. The stream could be divided into intervals, and one or more samples could be drawn randomly from each. But suppose, however, that after the sampling was completed and confidence intervals were calculated around the resulting WUA curves, the confidence intervals were wider than what was thought necessary to meet the objectives of the study. Increasing the sample size to increase the precision of the WUA estimates would be awkward. The same problem can occur with systematic samples.

An alternative for achieving spatial balance that avoids such problems is available. Called GRTS, for the "generalized random tessellation stratified" design, this is a flexible probabilistic sampling approach that gives good spatial balance (Stevens and Olsen 2004) and is increasingly used in monitoring programs (Stevens et al. 2007). The details of GRTS are complex, but essentially it can be regarded as an extension of spatial stratification which retains efficiency in the face of the practical complications, such as access problems, that field work on streams often entails. A Windows-based program for selecting GRTS samples, S-Draw, is available at www.west-inc.com, and a GRTS library for the R statistical program is available from the EPA at www.epa.gov/nheerl/arm/analysispages/software.htm.

With multi-stage sampling, spatial balance may be an issue at each stage. For example, if primary sites are stratified by habitat types such as pools and riffles, it would seem desirable that the primary sites be well distributed over the sampling universe, and that the pool transects have a roughly even distribution over the length of the pools.

There is obvious tension between achieving spatial balance and increasing efficiency of effort by clustering samples. There is no general resolution for this problem, but information about the likely spatial pattern of habitat value in the stream could be used to develop a sampling design that strikes a reasonable balance between the two. A related question with 2-D studies concerns the trade-off between more study sites and larger sites. The best balance may depend on the EFM used. With a demonstration flow assessment method, there would probably be less cost associated with having more but smaller study sites than would be the case with numerical habitat models involving hydraulic models. Practical experience is probably needed for useful guidance on the trade-offs involved, but simulations may also be useful.

Case studies

Four recent case studies illustrate the main points made above. Two were conducted in the context of major controversies, one of which has been the subject of two National Research Council (NRC) reports, and two are more "run-of-the-mill" studies conducted by scientists from the U. S. Geological Survey. These studies involved more than just applications of PHABSIM, but my review deals only with the sampling aspect of that component of the study.

Salmon River Basin: U.S. Geological Survey scientists used 1-D PHABSIM "to identify streamflow needs from July to September to provide fish passage and support various life stages of bull trout, Chinook salmon and steelhead trout" (*Salvelinus confluentus*, *Oncorhynchus tshawytscha*, *O. mykiss*) in selected streams in the upper Salmon River Basin in Idaho, with 51 transects in 8 study sites on 6 streams (Maret et al. 2006). That is, the study aimed to elucidate the relationships between flow and both habitat and passage. The study sites were selected deliberately, based partly on access, and "one to two transects were selected to represent each major mesohabitat in the reach," apparently with care to place one or more transects in shallow riffles that might cause passage problems at low flow.

Rather than support instream flow recommendations for particular streams, the study results are presented more as preliminary information that might guide further work, as suggested by language at p. 12: "Once an adequate number of sites have been characterized using PHABSIM, it may be feasible to develop habitat/discharge relations for streams with similar basin characteristics in specific geographic locations." Results were presented site by site, with no attempt in this report to apply the results beyond the sites. Accordingly, the only sampling issues immediately raised in this study involve how well the transects represent the sites, discussed below. However, the larger project suggested by the language cited above would require a proper sampling design and probability sampling for the "adequate number of sites" in order to allow for interval estimates.

Maret et al. (2006) stratified their sites by mesohabitat types, but did not describe the types used or how they were defined. The variance in estimates of WUA for the sites then depends on the variance within and among the mesohabitat types. With only two transects per type, however, the variance within the mesohabitat types can be estimated only poorly, and with one transect it cannot be estimated at all, so the stratification would not be effective even if the transects had not been deliberately selected. More useful results probably could have been obtained with the same number of transects by using a spatially balanced probability sampling design without stratification, or by collecting data for and modeling more transects within the mesohabitat types, again with a probability design. If collecting the data and modeling the transects is a small part of the overall work, as seems likely, then the latter alternative will likely be preferable. Similarly, probably little extra effort would have been involved in collecting data for hydraulic modeling on a few deliberately selected transects that could address passage separately from habitat. Trying to address both flow and habitat within the same sampling design will tend to bias one or both assessments.

North Fork Shenandoah River: U.S. Geological Survey scientists applied RHABSIM, a variant of PHABSIM, to three segments of the North Fork Shenandoah River during the low flow season (Kristolic et al. 2006). The stated objectives of the study were "to enhance understanding of summer low-flow conditions in the North Fork Shenandoah River, relate water availability to physical habitat needs of fish, and develop a relation for the availability of suitable fish habitat and instream flows." The segments (the sampling universes) were defined in terms of physical characteristics of the basin and stream gage location, so that there was a stream gage near the lower end of each site. Kristolic et al. (2006) mapped the river into 9 habitat types, and deliberately placed 35 transects in 6 study sites 45 to 610 m in length. The 80.6 km lower segment included 21 transects in 4 sites, the 28.3 km middle segment included 5 transects in 1 site, and the 63.2 km upper segment included 9 transects in 1 site. Kristolic et al. (2006) placed transects in "sections of the river where the mesohabitat was homogenous across the channel." They chose accessible sites that included relevant habitat types, and developed WUA curves for each segment from the transect curves for each habitat type and the proportion of the segment occupied by the habitat type.

Like most applications of PHABMSIM, the Kristolic et al. (2006) study set out to characterize flow/habitat relations for lengthy reaches of a river, not just sites along it. As such, the objective of the study was more demanding than the objective of Maret et al. (2006), and required a more careful sampling design. Unfortunately, the sampling design fell short. The location of the study sites was spatially unbalanced, apparently because the study was expanded midcourse: "The original scope of the project focused on the Seven Bends area…, therefore, five of the six reaches are in the northern end of the basin, in the middle and lower sections of the river. As the study progressed, the Plains Mill reach…was added to incorporate the upper section…" The actual sampling universe was limited to accessible parts of the river where the mesohabitat homogenous across the channel, and so was biased in the statistical sense. All but one of the sites included habitat types represented by only one transect, so the variance of the transect data within the actual sampling universe could not have been estimated, even if the transects had been placed randomly.

Snake River Basin: R2 Resource Consultants (2004) developed flow recommendations for the "usual and accustomed" fishing places of the Nez Perce Tribe in Idaho, for a large number of streams in a ~65,000 km^2 project area. The recommendations were used in stream adjudication, a legal process to quantify water rights in the Snake River Basin. Sampling was necessary not just to develop relations between flow and habitat in particular streams, but also to select streams for study, since it was impractical to study all the streams that would be affected by the adjudication.

For developing the sampling frame for streams, the project area was divided into 1,145 subbasins stratified by the Strahler order of the associated stream segment, using 1:100,000 scale maps. These subbasins were the sampling universe. Groups of similar subbasins were defined based on 14 physical characteristics, using a cluster analysis following a principal components analysis. Thirty-four groups of subbasins were defined, following some ad-hoc adjustments of the cluster analysis results to deal with outliers and small clusters. At least three stream segments in each group were randomly selected for PHABSIM analysis, and PHABSIM analyses were also conducted on four other streams identified as important by the Nez Perce Tribe.

On each stream segment, study sites were selected randomly, based on distance along the stream from the downstream boundary of the subbasin (the site sampling frame), although in some cases parts of the stream were excluded from sampling because access was too difficult. The 208 sites were approximately 25 channel widths long. Sites were stratified by six habitat types: pool, run, riffle, cascade, simple island, and complex island. Three transects were then placed randomly in one habitat unit of each habitat type that made up more than 10% of the site, although in some cases transects were also placed randomly in habitats that were judged to be important, even if they made up less than 10% of the site. In addition, for pool habitats, a transect was placed at the hydraulic control, as required by the hydraulic model used.

Flow recommendations were developed separately for each of the 34 groups of subbasins, so the study sampled stream segments within the groups, study sites within the stream segments, habitat types within the stream segments, and transects within the habitat types. With some minor exceptions, the sampling was random, so bias in the recommendations is minor. However, the recommendations are point estimates, and because only one site was selected per stream segment, interval estimates cannot be calculated—there is no way to estimate the within-segment variance.

Interval estimates for the relation between flow and habitat could have been developed for the sites, but they were not. However, the general approach used here would also be suitable for the Maret et al. (2006) study discussed above.

Klamath River: "Water is for fighting over," and nowhere more so than the Klamath River basin in Oregon and California, the site of well publicized conflict between agricultural, hydropower, and fisheries interests. In a study for the Department of the Interior, Hardy et al. (2006) used several methods, including 2-D PHABSIM, to develop recommendations for "instream flows on a monthly basis for specific reaches of the main stem Klamath River below Iron Gate Dam by different water year types," to "provide for the long-term protection, enhancement, and recovery of the aquatic resources" of the river, especially anadromous fishes. Chapter 5 of NRC (2008) reviews the complex approach taken to combine PHABSIM results for several species and life stages with each other and with other considerations, but I focus on the sampling for the PHABSIM modeling.

Hardy et al. (2006) divided the 310 km of Klamath River between Iron Gate Dam and the estuary into five segments of unequal length, based on tributary junctions. Eight study sites were deliberately selected, with one site in the most upstream and downstream segments, and two sites in the remaining segments. The hydraulic model was applied to each site, but results were unsatisfactory in the most downstream site. Results from modeling the seven remaining sites, which totaled 8.4 km in length, were extended to the segments using a mapping of the entire river into five habitat types: pool, run, low slope, moderate slope, and steep slope; thus, if a given habitat type occupied 25% of the study site but only 20% of the segment, the model results for that habitat type would be weighted by a factor of 0.8.

For reasons already discussed, interval estimates could not have been developed for two of the segments with this study design, even if the study sites had been randomly chosen, and the estimates would have been poor for the other three. Only about 3% of the sampling universe was sampled in this study, so a finite population correction would have little effect on the estimates of variance.

Of the recent PHABSIM studies described above, only the R2 Resources Consultants study used random sampling to select study sites or transect locations. Even this study fell short of what would be needed to calculate confidence intervals for the WUA curves that were developed. The other studies selected sampling sites deliberately, apparently based on the judgment of the investigators and on access. Besides the bias inherent in the deliberate selection of samples, the sample sizes in the studies reviewed seem small, given the evident variability of rivers and simulation studies of the relationship between transect numbers and the width of confidence intervals (Williams 1996, 2010), although the sample sizes are ordinary or large for PHABSIM studies (Payne et al. 2004). Given the likely magnitude of the sampling error in PHABSIM studies and other sample based EFMs, careful attention to estimating the error is needed to meet ordinary norms for scientific practice.

Although I have criticized particular studies, my criticisms are aimed not at the authors, but at the standard of statistical practice in environmental flow assessment, which the studies reflect. Hardy et al. (2006) did not select their study sites by themselves, but rather, in consultation with an inter-agency group. I chose USGS studies for this article partly because the USGS has an internal review process, which presumably approved them, and both the Salmon River Basin and Shenandoah River studies involved inter-agency collaborations.

Why deliberate selection doesn't work

To understand the problems with deliberate selection of transect locations or study sites, consider how one might try to select a "representative" sample of transects that would generate something close to the "true" WUA curve for the reach of interest. Somehow, the investigator would have to inspect the entire reach, make judgments about how the conditions in the river that determine WUA would change with discharge along the entire reach, and then select sites that would, in the aggregate, reasonably represent the way the conditions that determine WUA change with discharge. This is a tall order, and, if the investigator could fill it, what would be the need for the PHABSIM study?

I should note that a recent National Research Council study reviewed Hardy et al. (2006), and with minor caveats approved the deliberate selection of study sites in that study. However, the committee that wrote the report included no statisticians (although it did include one of the developers of PHABSIM), and cited no statistical authorities to support its approval of deliberately selected sampling sites. The report (NRC 2008:193) noted that "…the representativeness of the study sites was determined by inter-agency group agreement and was not statistically assessed," but the report also implied that "basic channel properties" might be used to assess whether the Klamath River samples sites used in Hardy et al. (2006) were reasonably representative of the river. How might this work? Stream channel width and slope are basic channel properties. Suppose that aerial photographs taken over a range of discharge were available, so that the width (W), change in width with discharge ($\Delta W/\Delta Q$), and slope (S) could be determined photogrammetrically for the 310 km whole reach. Then, following the suggestion in NRC (2008), one might determine whether the means of W, $\Delta W/\Delta Q$, and S were approximately the same for sites and for the river. If so, one might argue that the sites were reasonably representative.

The idea that information about the whole could be used for purposive selection of a representative sample was advocated by some statisticians early in the twentieth century, but was abandoned following a critical review in a seminal paper by Neyman (1934) that laid the basis for modern survey sampling. This study addressed the very practical problem of getting results quickly from census data, at a time when compiling and reducing the data took years. Some statisticians supposed that information known for all the census districts might be used to select representative districts, for which the complete census data could be compiled and reduced to give preliminary results, upon which public policy could be based until the full results were in. Others advocated selecting random samples for the same purpose, and this was a situation in which, over time, the two approaches could be tested. Neyman (1934) showed that stratified random samples gave reasonably accurate estimates, whereas the deliberately selected samples did not. More fundamentally, Neyman (1934) rigorously considered what it means for a sample to be representative, and concluded that the most that can be done is to define a "representative method of sampling and a consistent method of estimation." Essentially, these are methods that allow for calculating interval estimates. Much subsequent experience has confirmed Neyman's argument (Converse 1987). There are three reasons for random sampling: to allow estimates of variance, so that the quality of the estimate can be assessed; to get more accurate estimates (unless the sample is very small); and to avoid bias, either deliberate or unconscious (Jesson 1978). Information about the stream should not be ignored, but it should be used in developing the sampling plan, not in choosing the sample itself.

Conclusions

To be scientifically credible, EFMs based on samples should incorporate probability sampling, and report interval estimates for whatever index of habitat the method produces. The sampling design should provide for a balance between sampling efficiency and spatial balance. The GRTS design, which is increasingly used in water quality monitoring, seems well suited for EFMs as well. All streams are unique, so knowledge of the stream and of the assessment method to be used are critical for developing a good sampling design, and results will be most satisfactory if a statistician familiar with sampling is involved in planning the study.

Acknowledgments

This work was completed with funding by the Public Interest Energy Research of the California Energy Commission through the Instream Flow Assessment Program of the Center for Aquatic Biology and Aquaculture of the University of California, Davis. Suggestions by Peter Moyle, Joe Kiernan, several reviewers, and especially Ken Newman, improved the text

References

American Fisheries Society (AFS). 2008. Resolution on the development of instream flow programs. Fisheries 33(11):563.

Annear, T., and 15 others. 2004. Instream flows for riverine resource stewardship, revised edition. Instream Flow Council, Cheyenne, Wyoming.

Bovee, K. D., B. L. Lamb, J. M. Bartholow, C. B. Stalnaker, J. Taylor, and **J. Henriksen.** 1998. Stream habitat analysis using the instream flow incremental methodology. U.S. Geological Survey, Biological Resources Division Information and Technical Report USGS/BRD-1998-0004. Available at: www.mesc.usgs.gov/products/publications/3910/3910.pdf (May 2007).

Castleberry, D. T., J. J. Cech Jr., D.C. Erman, D. Hankin, M. Healey, G. M. Kondolf, M. Mangel, M. Mohr, P. B. Moyle, J. Nielsen, T. P. Speed, and **J. G. Williams.** 1996. Uncertainty and instream flow standards. Fisheries 21(8):20-21.

Cochran, W. G. 1977. Sampling techniques, third edition. John Wiley and Sons, New York.

Converse, J. M. 1987. Survey research in the United States: roots and emergence 1890-1960. University of California Press, Berkeley.

Gallagher, S. P., and **M. F. Gard.** 1999. Relationship between Chinook salmon (*Oncorhynchus tshawytscha*) redd densities and PHABSIM-predicted habitat in the Merced and Lower American rivers, California. Canadian Journal of Fisheries and Aquatic Sciences 56:570-577.

Guay, J. C., D. Boisclair, D. Rioux, M. Leclerc, M. Lapointe, and **P. Legrendre.** 2000. Development and validation of numerical habitat models for juveniles of Atlantic salmon (*Salmo salar*). Canadian Journal of Fisheries and Aquatic Sciences 57:2065-2075.

Hardin, T. S., R. T. Grost, M. B. Ward, and **G. E. Smith.** 2005. Habitat suitability criteria for anadromous salmonids in the Klamath River, Iron Gate Dam to Scott River, California. California Department of Fish and Game Stream Evaluation Report 05-1, Sacramento.

Hardy, T. B., R. C. Addley, and **E. Saraeva.** 2006. Evaluation of instream flow needs in the lower Klamath River, Phase II, Final Report. Institute for Natural Systems Engineering, Utah State University, Logan.

Jesson, R. J. 1978. Statistical survey techniques. John Wiley and Sons, New York.

Kristolic, J. L.; D. C. Hayes, and **P. M. Ruhl.** 2006. Physical habitat classification and instream flow modeling to determine habitat availability during low-flow periods, North Fork Shenandoah River, Virginia. U.S. Geological Survey Scientific Investigations Report 2006-5025.

Manly, B. F. J., L. L. McDonald, D. L. Thomas, T. L. McDonald, and **W. P. Erickson.** 2002. Resource selection by animals, second edition. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Maret, T. R., J. E. Hortness, and **D. S. Ott.** 2006. Instream flow characterization of upper Salmon River Basin streams, central Idaho, 2005. U.S. Geological Survey Scientific Investigations Report 2006-5230.

Mier, K. L., and **S. J. Picquelle.** 2008. Estimating abundance of spatially aggregated populations: comparing adaptive sampling with other study designs. Canadian Journal of Fisheries and Aquatic Sciences 65(2):176-197.

National Research Council (NRC). 2008. Hydrology, ecology, and fishes of the Klamath River Basin. National Academies Press, Washington, D.C.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society A 97:558-606.

Payne, T. R., S. D. Eggers, and **D. B. Parkinson.** 2004. The number of transects required to compute a robust PHABSIM habitat index. Hydroécologie Appliquée 14: 27-53.

R2 Resource Consultants. 2004. Instream flow claims for the Snake River Basin, Volume 1, Part 1, Main Report. R2 Resource Consultants, Redmond, Washington.

Railsback, S. F., and **B. C. Harvey.** 2001. Individual-based model formulation for cutthroat trout, Little Jones Creek, California. General Technical Report PSW-GTR-182, Pacific Southwest Research Station, Forest Service, U. S. Department of Agriculture.

Railsback, S. F., and **J. Kadvany.** 2008. Demonstration flow assessment: judgment and visual observation in instream flow studies. Fisheries 33(5):217-236.

Richter, B. D., J.V. Baumgartner, R. Wigington, and **D. P. Braun.** 1997. How much water does a river need? Freshwater Biology 37:231-249.

Stevens, D. L., Jr., and **A. R. Olsen.** 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99:262-278.

Stevens, D. L. Jr., D. P. Larsen, and **A. R. Olsen.** 2007. The role of sample surveys: why should practitioners consider using a statistical sampling design. Pages 11-24 *in* D. H. Johnson, B. M. Shrier, J. S. O'Neil, J. A. Knutzen, X. Augerot, T. A. O'Neil, and T. N. Parsons, eds. Salmonid field protocols handbook: techniques for assessing status and trends in salmon and trout populations. American Fisheries Society, Bethesda, Maryland.

Tenent, D. L. 1976. Instream flow regimens for fish, wildlife, recreation and related environmental resources. Fisheries 1(4):6-10.

Thompson, S. K. 2002. Sampling, 2nd edition. John Wiley and Sons, New York.

Waddle, T. J. (editor). 2001. PHABSIM for Windows: user's manual and exercises. U.S. Geological Survey Open-File Report 2001-340.

Williams, J. G. 1996. Lost in space: minimum confidence intervals for idealized PHABSIM studies. Transactions of the American Fisheries Society 125:458-465.

_____. 2010. Lost in space, the sequel: spatial sampling issues with 1-D PHABSIM. River Research and Applications 26:341-352. Published online March 2009, DOI: 10.1002/rra.1258.

Williams, L. R., M. L. Warren, Jr., S. B. Adams, J. L. Arvai, and **C. M. Taylor. 2004.** Basin visual estimation technique (CVET) and representative reach approaches to wadeable stream surveys: methodological limitations and future directions. Fisheries 29(8):12-22.