Regional Estimation of Juvenile Coho Abundance in Streams

FINAL REPORT

prepared by

Scott W. Overton
Department of Statistics (Emeritus)
Oregon State University
Kidder Hall #44
Corvallis, OR 97331
overton@stat.orst.edu

and

Trent L. McDonald
Western EcoSystems Technology, Inc.
2003 Central Ave
Cheyenne, WY 82001
tmcdonald@west-inc.com

West Technical Report #98-5

November 6, 1998

Contents

1. Introduction

Estimation of the number of fish in a small stream is a difficult problem. Fish are highly mobile, difficult to see, difficult to catch, and live in three dimensions. For sampling fish in a single stream, the Hankin and Reeves method (Hankin and Reeves 1988) is a generally accepted method which provides an index to fish abundance. In addition to the difficulties inherent in estimating numbers of fish on a single stream, attempting to estimate the number of fish over a large area only adds difficulties. Assuming sufficient resources are available to overcome all the difficulties, this document presents a method for regional estimation of fish abundance by making recommendations for frame construction, sample selection, and statistical estimators. Our use of the term "regional" is intended to imply areas which are larger than can be feasibly sampled by a single stream protocol and may include large watersheds, large basins within a watershed, multiple watersheds, etc.

In making our recommendations, we assume that techniques for estimation of abundance on a single stream (stream segment) are known and focus on statistical procedures for regional estimation. We focus on regional statistical considerations for two reasons. First and foremost, estimation of fish numbers on a single stream and selection of which streams to sample are separate topics. From a regional perspective, the single stream sampling protocol is a device which measures the number of fish in any particular stream segment. So long as the same measurement (i.e., protocol) is applied to each segment, or as long as each of several protocols are unbiasedly measuring the number of fish in a segment, the regional estimation routine does not rely on any details about how the single stream protocol was implemented. Under this approach methodological and technological advances in single stream methods do not invalidate the regional estimation protocol. It would be undesirable for regional estimation procedures to depend upon current field methods because changes in the single stream protocol would necessitate changes in the regional methods. Secondly, we say nothing about single stream methods because we recognize that the region for which estimates are sought could be very large (e.g., multi-state) and that different land owners and management agencies may have different resources. Assuming unbiasedness, our focus on regional estimation allows managers to implement different protocols based on their individual resources.

We assume that an abundance estimate is desired for a single species of fish but accept the fact that data may be collected on multiple species during a single survey. We caution that collecting data on two or more species of fish using a single survey may yield undesirable estimates for one or both species because optimal sample frame coverages and optimal single stream sampling techniques may be different for different species. Furthermore, sample sizes adequate for one species may not be adequate for another.

Our methods target estimation of abundance of juvenile coho salmon (*Oncorhynchus kisutch*) which reside in coastal streams of Northern California and Southern Oregon during summer.

Straightforward modifications of the methods can be made to sample other salmonids such as steelhead (*Oncorhynchus mykiss*) or other trout species such as cutthroat trout (*Oncorhyncus clarkii*). The three-year life cycle of coho make it desirable to conduct surveys in three successive years so that estimates encompass three independent age classes. For multiple year samples discussed in Section 6, we propose a three-year rotating systematic design and advocate estimation of a three-year moving average to summarize the population.

A general outline for the remainder of this paper is as follows: notation and a short glossary of terms are given in Section 2; frame development is covered in Section 3; sample selection is outlined in Section 4; equations for yearly estimates are given in Section 5; and sampling over short time periods is covered in Section 6. Statistical calibration at the single stream level, although not part of the regional estimation procedure, is covered in the Appendix. Statistical calibration is covered in the Appendix because it is a popular technique and has the potential to vastly improve single stream estimates and consequently the regional estimates.

2. Notation and Definitions

We employ the following definitions:

- *stream segment*: a designated length of a watercourse.

- *sample unit*: the fundamental entities on which measurements are taken. In this paper, sample units are stream segments.

- *universe:* the group of sample units to which we wish to make inference. For example, the universe may be all stream segments in a river basin or watershed. When necessary, we denote the universe by the symbol U.

- *sample frame*: a list of sample unit identifiers which will be used to draw the sample. The frame is a description of the universe of interest and is structured in a manner consistent with the realities of that universe. The frame must also be compatible with the intended sample design in that it should identify units and any structure used by the sample design. In this paper, the frame is a list of stream segment identifiers. Ideally, the sample frame lists all sample units in the universe.

- *basin*: a river basin or watershed. We anticipate keen interest in estimates for basins such that usually the universe will be either a single river basin or a group of two or more contiguous basins; however, the statistical validity of estimates does not depend on the universe containing whole basins.

- *sub-basin*: a subset of a basin. At some places in the protocol, it will be convenient to divide large basins into smaller basins. Sub-basins will be the smallest entities for which estimates are made.

- *inclusion probability*: the statistical probability that a sample unit is included in the selected sample. In statistical literature, a distinction is made between first-order and second-order inclusion probabilities. We do not need to make this distinction here because under the proposed design and variance estimation approach these probabilities are treated as constant across all units.

- *stream system*: a main stream and all its smaller tributaries. This definition is somewhat ambiguous because one investigators might call a stream a tributary while another calls the stream a main stream. Nonetheless, we find use of the term to be convenient and the ambiguity in its definition irrelevant.

We have endeavored to keep mathematical notation to a minimum. When necessary we use the following notation to represent components of the finite population, design, and estimates. The following list of symbols is not complete but contains the key definitions for our protocol.

- U = the population
- u = a single element of the population (i.e., a sample unit)
- S = a set of u selected from U as the sample ($S \subseteq U$)
- $\hat{T}_u$ = the estimated number of fish on steam segment u.
- $\hat{V}(\hat{T}_u)$ = estimated variance of the estimated number of fish on stream segment u.
- x_u = some measured characteristic of stream segment labeled u (e.g., length, depth, etc.)
- N = the number of stream segments in the population (i.e., size of U)
- n = actual sample size (i.e., size of S)

3. Frame Development

The general approach of this protocol is to: 1) construct a sampling frame to represent the stream system or universe of interest; 2) draw a sample of stream segments from the frame; 3) conduct physical sampling of the selected stream segments to produce counts on each segment; and 4) estimate abundance using the formulas in Section 5. This section gives our recommendations for construction of the sampling frame.

Construction of a high quality sampling frame is critical to the success of the survey. Agencies and managers should recognize that, while it may seem unproductive initially to invest heavily in frame development, investment in frame development is one of the best ways to assure long term success of the project. On the other hand, we have seen frame development efforts become mired in detail and the field aspect of the survey never reach full fruition. A high quality frame achieves a balance between need for survey accuracy and acceptance of trivial frame inconsistencies.

In general, there are two major tasks which must be completed for sample frame construction, universe delineation and segment identification. Universe delineation involves the demarcation of streams which are in the universe and which could potentially be sampled. Segment identification comes after universe delineation and breaks the previously defined universe of fish habitat into manageable segments. A sample of these manageable units is then drawn and measurements are taken on each. Once universe delineation and segment identification are complete, these documents become working definition of the universe.

3.1 Map materials

This sub-section contains a discussion of map resolution and proper perspective for frame construction. Detailed steps for frame construction are given in the next sub-section.

During universe delineation, project managers will need to make decisions about whether or not to include certain stretches of stream based on an overall regional perspective. While the concept of "regional perspective" is difficult to define and convey, decisions at this stage to include or exclude certain areas from the universe should be based on things like study objectives and whether or not variables like elevation, slope, aspect, watershed size, rainfall patterns, etc. indicate poor (or favorable) fish habitat. For example, if researchers are confident that a certain stream is too high, too steep, too small, and has too little water to support fish, it should be excluded from the universe. Conversely, a stream should be included in the universe if environmental conditions look favorable to fish. Major fish blockages, such as dams or falls, should also be considered at this stage, as well as auxiliary information such as temperature if it is available. Although decisions at this stage are sometimes difficult, universe delineation should be a relatively simple process and less tedious than segment identification where a more detailed consideration of the presence of fish bearing habitat can take place. Pilot data and previous investigator experience can aid decision making here.

To convey proper regional perspective during universe delineation, we propose that relatively low resolution paper maps (e.g. 1:250,000) be used. This resolution map is advantageous because typically only substantial streams are displayed while very small streams with assumably low numbers of fish are excluded. Appropriately scaled electronic maps or digital line graphs (DLG) can also be used; however, we caution project managers that the level of detail present in many DLG files is greater than that needed for universe delineation, and it is possible to become mired in this detail. We emphasize that universe delineation should be relatively simple and should include all watercourses ("blue lines") present on the map which are not obviously too high, too steep, too small, etc., or lie behind an obstruction. We acknowledge that if frame construction is to be automated or strictly standardized across investigators, some form of DLG map and geographic information system programing will have to be used. Automated frame construction may also be attractive from an efficiency and cost point of view.

Once the universe is delineated, we propose that segment identification be carried out using relatively high resolution maps (e.g., 1:24,000) or digital line graphs (DLG). Aerial photographs are not absolutely necessary but can be a useful as a supplement to the maps or DLG files. While low resolution maps provide proper regional perspective for universe delineation, high resolution maps provide the detail required to identify physical features necessary to establish segment boundaries. These details include cliffs, falls, bridges, roads, meadows, confluences, etc. Inevitably, some stretches of stream, previously included in the universe, will need to be excluded from the universe at the segment identification level due to perceived low fish abundance. For example, a segment might be excluded if falls are shown at high resolution but not at low resolution. Segment identification can be easy in some situations but it is often tedious and requires iteration between universe delineation and segment identification.

Both universe delineation and segment identification (and hence frame construction) can be performed using either paper maps of any resolution or DLG files and a valid sampling frame will result. The main question surrounding which maps to use centers on which is most efficient and whether or not the frame built with one kind of map can be judged better than another. The concept of better in this case should be based on regional perspective and geographic coverage. A relatively simple pilot study which generates a frame for a medium sized watershed using two different map types would answer many questions regarding the efficiency of various resolution maps to produce sample frames. Regardless of whether paper maps or DLG files are used, the maps or DLG files, together with all indexing materials, must be archived.

3.2 Frame construction procedure

Details regarding frame construction are given in the list below. In chronological order, the general steps required for frame construction are:

1. *Conceptual definition*. The conceptual geographic region of interest should be defined. This conceptual definition is simple and reflects study objectives. Examples of a conceptual definition are “All waters in management unit A” or “All waters in the Mad River watershed”.

2. *Universe delineation*. Using low resolution maps (e.g., 1:250,000 USGS maps), all watercourses present on the map which appear favorable for fish productions should be marked as potential fish habitat and included in the universe. Streams which do not fit the conceptual definition of the universe should be excluded. Streams should be excluded if project managers and biologists are certain no fish are present. A overall regional perspective should influence decisions to include or exclude a stream and should be based on whether or not factors such as slope, aspect, elevation, water volume, etc. indicate

favorable habitat for fish. See the preceding sub-section for more on our notion of regional perspective.

3. *Segment identification*. Using high resolution maps (e.g., 1:24,000 USGS or DLG), all watercourses delineated during the previous step should be subdivided into manageable stream segments. Each stream segment should be approximately the same length and manageable by the intended single stream sampling technique; however, stream segments can differ substantially in length without effecting the statistical validity of estimates. Each segment's endpoints should be well defined and relatively easy to identify. Permanent landmarks such as confluences, bridges, cliffs, falls, meadows, etc. are easy to identify in the field and serve well as segment endpoint markers. Places where the maps do not agree with the physical environment are accounted for at the sample level without modification of the frame. If a tributary of a segment is present at a sampled segment which is not represented on maps, investigators can either exclude the tributary if it is small enough or include the tributary as a regular part of the segment which is sampled. Known tributaries, forks, and bifurcations should be treated as separate segments if they are large enough and potentially contain a significant number of fish. The target length of each segment is arbitrary and should be compatible with the intended single stream sampling technique. If the Hankin and Reeves (Hankin and Reeves 1988) procedure is to be used on each segment, segments of approximately 1 mile are suggested.

   Based on maps, investigator knowledge of the river system, and perceived presence of fish, a decision to include or exclude certain stream segments can be made at this stage. If investigators are certain that no fish are present in a segment or that the number of fish in a segment is trivially small, the segment should be excluded from the universe. If investigators feel there is some likelihood of finding fish in a segment but are not certain as to the magnitude of fish numbers, perhaps pre-survey pilot work at the segment should be conducted. Otherwise, a segment should be included if investigators are unsure about the presence of fish. To exclude large numbers of fish, even by accident, runs counter to study objectives; therefore, it would be better to include uncertain segments than it would be to exclude them provided the potential costs associated with their inclusion in the sample are not too high.

4. *Identification of estimation domains.* Groups of segments in the universe for which separate estimates are sought should be identified. For convenience, we call these groups of segments basins or sub-basins, where basins contain one or more sub-basins. Basins could be multiple watersheds, single or grouped management units, or other area containing more than say 100 miles of stream. Sub-basins could be watersheds or tributary systems comprised of say 50-100 miles of stream. Basins and sub-basins can vary tremendously in size without effecting the statistical validity of resulting estimates. Sub-basins, in addition to being interesting as estimation domains, will be used as devices to provide spatial coverage of the sample. Certain basins will be natural (e.g., the Klamath

Basin or the Mad River Basin) while others will be composites of smaller river-systems assembled together for convenience. Small coastal streams near the mouth of a large stream should probably be included in the same basin as the large stream. The primary consideration when constructing basins should be the integrity of the populations of Coho that occupy the basin. It is understood that precision of sub-basin (and small basin) annual estimates will be poor, owing to small sample size, and that greater utility will come from estimates at the basin and universe level.

5. *Assign segment identifiers*. Once stream segments, basins, and sub-basins are identified, each segment in the universe should be given a unique identifier. It is likely that the tasks of assignment of identifiers, identification of segments, and identification of estimation domains will by necessity be completed concurrently and therefore are not completely separated tasks. We list them as separate tasks for clarity. Standardization of stream identifiers will prove valuable in coordinating the work of different researchers. For identifiers, we suggest twelve digits (more or less as needed) of the form ‘bbbbnnnnxxxx' where ‘bbbb' is a four letter abbreviation or code designating the basin containing the stream segment, ‘nnnn' is a four letter abbreviation or code designating the sub-basin of the stream, and ‘xxxx' is a sequential number assigned to each segment of the ‘bbbbnnnn' stream. The ‘bbbb' codes should generally be defined so that, when sorted, the lowest elevation waters are placed before higher waters. The same condition should hold for the ‘nnnn' and ‘xxxx’ codes. Sorting on ‘nnnn’ within ‘bbbb’ or ‘xxxx’ within ‘nnnn’ should place lower elevation waters before higher waters. Placing low elevation waters together and before higher elevation waters assures sample coverage of all elevations. Coverage of all elevations is desirable because elevation is a known proxy for coho abundance. Once assigned, changes to the identifier of a segment should be resisted. The stream segment definitions and identifiers should be recorded on archived maps.

6. *Assemblage into computer system*. The list of stream segment identifiers should be assembled into a computer file in a format such that sorting, sample selection, and other manipulations of the frame are convenient. The list of stream segment identifiers is the sample frame.

4. Sample Selection

The sampling scheme recommended here is designed to provide maximum information for estimation of population status. Sample selection for estimation of population trends is a large and separate topic and consequently, will not be treated here. Section 6 contains comments and recommendations regarding establishing status over short time periods.

The recommended design for selection of stream segments is a general systematic sample (GSS). In other contexts, GSS designs are variable probability designs, meaning that units (segments) are included in the sample with unequal probability; however, for reasons which will be made clear later, we advocate an equiprobable GSS design in which all segments have the same (first order) inclusion probability. We will not consider the variable probability version of a GSS and will only consider the special case of equal probabilities.

There are two tasks required in order to draw an equiprobable GSS. First, the frame needs to be ordered in such a way that the population is advantageously sampled (see Section 4.1 for our suggested ordering). Second, the sample needs to be drawn systematically across the frame. This section begins with some general comments on the merits of a GSS. Considerations when setting sampling rates (sample size) are covered next. Details of the basic tasks required to draw a GSS are given in separate sub-sections below followed by a discussion of inaccessible segments.

The primary desirable characteristic of a GSS is the ability to control and ensure spatial coverage of the sample. In establishing status, broad spatial coverage is likely to be the single most important characteristic of the design because spatial variation in fish numbers is likely the biggest source of variation in the population. Spatial coverage is achieved when a sample unit is selected from all parts of the study area and a wide range of environmental conditions are encountered. This type of broad spatial coverage is achieved in the GSS sample through judicious ordering of the frame prior to sample selection. Although variances are sometimes difficult to compute, the spatial coverage of a GSS generally increases the precision of estimates by reducing variance relative to a completely random sample. Variance of a completely random sample is larger than that of a GSS because the completely random design occasionally provides poor spatial coverage.

A secondary desirable characteristic of an equiprobable GSS is that the yearly sample will achieve approximate proportional allocation on any definable grouping of segments. That is, if we assume interest lies in a particular sub-basin which has N_h segments in it where N_h is known from the frame, the realized sample size from that sub-basin will be approximately proportional to N_h. If segments are approximately the same length, proportional allocation means that the number of miles of stream sampled each year in a any particular sub-basin will be approximately proportional to the total number of miles in that sub-basin. Approximate proportional allocation in all areas is desirable because sub-basin or basin boundaries can be redrawn after the survey is complete and the number of sampled units in the redefined areas will remain approximately proportional to size of the area. Consequently, there is no need to anticipate all possible sub-basin or basin definitions prior to sample selection.

Other desirable characteristics of a GSS include: 1) simple estimation formulas; and 2) fixed size samples. Simple estimation formulas improve survey utility. Fixed size samples may be important for planning purposes.

Before a sample can be drawn, sample size must be determined. Determination of an appropriate sample size is a difficult task. From a pure scientific point of view, sample size should be set to achieve a certain precision in the estimates. In reality, sample size is almost always a direct function of budgetary and time constraints. A detailed accounting of the time, personnel, and equipment required to sample an individual stream segment is needed to set a feasible sample size. In addition, personnel availability and reporting deadlines need to be considered. At least in the initial years of study, we advocate devoting resources to determining as accurately as possible the dollars, time, personnel, equipment, etc. required to sample a single stream segment rather than devoting resources to theoretical statistical exercises which attempt to determine sample size which achieve a certain precision. In fact, it is only after the exact requirements of the single stream sampling method are known that a statistical exercise to determine sample size becomes useful. If done correctly, the statistical exercise will incorporate cost constraints to calculate an "optimum" sample size. As a general guide, we recommend construction of stream segments which can be surveyed in one day by one crew. While sampling a segment in one day is not a requirement, this characteristic will make certain planning calculations easier. We also suggest pilot studies be conducted on selected streams and/or basins in order to collect hard data regarding time requirements and to construct initial variance estimates.

For illustrative purposes, a pilot frame for the Lower Klamath basin is given in Table 1. The pilot frame in Table 1 was developed by Mr. Ross Taylor for Simpson Timber Company. This table contains a concrete example of frame materials which will be referenced later to illustrate details of sample selection. An important piece of information not presented in Table 1 is the fixed map showing exact locations of all stream segment.

Once the sample is drawn, we suggest that the selected stream segments be represented on high resolution maps (e.g., 1:24,000) or aerial photographs highlighting all visible features. These maps or photos should be carried in the field by the sampling crew to assist in conduct of the survey. Use of the maps or photos will aid field crews in identifying permanent features which define the segment.

4.1 Order of the frame

It should be emphasized that the order of the frame is left to the discretion of the researcher, but that the frame's order is key to the spatial pattern of the sample and consequently the efficiency of the survey. Changing the frame’s order will not change the fundamental validity of the sample. When ordering the frame, our objective is to ensure a high degree of spatial dispersion in the selected segments. While statistically valid, it would be inefficient and result in high variance if we obtained a sample of segments which were very close to one another.

Sub-basins which are generally lower in elevation should be placed in the frame before higher elevation sub-basins. A strict linear ordering from lowest sub-basin to highest sub-basin is not always possible in topologically varied and diverse systems. Luckily, the ordering of large blocks of stream segments (the sub-basins) is not critical to ensuring spatial coverage over the entire region. Within sub-basin, segments should be roughly ordered from lowest elevation waters to highest elevation waters. Within sub-basin, tributaries to a main stream can be inserted into the sequence of the main stream if they are short, say less than three segments in lengths. Long tributaries should be appended at the end of the list of segments from the main stream.

If the segment identification scheme mentioned in Section 3 is adopted, the ordering from lowest to highest waters that we advocate can be accomplished by first sorting on the 'bbbb' portion of the id string, then sorting on the 'nnnn' portion within 'bbbb', and finally on the 'xxxx' portion within 'nnnn'.

4.2 Selection of the systematic sample

This sub-section describes an algorithm to select a equiprobable GSS.

To draw an equiprobable GSS, set the real number k = N/n, where N is size of the universe or length of the frame and n is the desired sample size. In general, k is a real number and not an integer. Choose a random number, say m, between 0 and k, where m is a real number and not necessarily an integer. Generate an *indicator sequence* as a vector containing the numbers {m, m+k, m+2k, . . . m+(n-1)k}. The final step is to include the i-th element of the sorted frame in the GSS if any number, say b, of the indicator sequence falls between i-1 and i (i.e., segment i is included if $(i-1)<b<i$). The S-Plus® (MathSoft Inc., Seattle WA) computer code to carry out the GSS algorithm appears in Table 2.

As an example of the sample selection process, an equiprobable GSS sample of size n=13 was drawn from the N=150 segments listed in the pilot frame of Table 1. For this example, k = 150/13 = 11.53. A random start between 0 and k was drawn and was m=6.106. The indicator sequence was then constructed to be {6.102, 17.644, 28.183, ..., 144.56 }. Segment identifiers of units selected in this particular sample were {0B004, 0C004, 0D005, 0E004, 0E016, 0E027, 0F005, 0G004, 0I006, 0J007, 0K009, 0L008, 0M011}. Note that this sample selected at least one segment from every sub-basin except Wilson Creek and Surpur Creek which where small sub-basins containing three and four segments, respectively.

4.3 Inaccessible stream segments

Lack of investigator access to sample units is a common problem which plagues many environmental surveys. The problem arises because investigators do not usually know which sample units they can get to prior to sample selection and inspection. For example, investigator access can be barred due to lack of navigatable waters (if access is by boat), lack of passable roads, or lack of permission to use private lands. If investigators know for certain that some streams are inaccessible to humans (and/or fish) prior to sample selection, these segments should be removed from the frame prior to sample selection. Statistical inferences will not apply to areas removed from the frame. In some situations, investigators may be willing to make an *assumption* that data from the accessible portion of the universe are essentially the same as data from the inaccessible portion. In these cases, inference to the entire population (accessible and inaccessible) is by assumption and is not a statistical inference.

A distinction should be made between accessible to humans and accessible to fish. This section deals only with denied access to humans. If, after sample selection, investigators find that fish access to a selected segment is completely blocked, this knowledge constitutes a valid zero for the number of fish in the unit and it should be recorded as such. In this case, investigators may wish to estimate the proportion of segments which are blocked to fish using sample data. If humans cannot get to a segment and are unsure whether fish are present, the segment is inaccessible.

It is not essential to the statistical validity of the survey that all inaccessible units be removed from the frame prior to sample selection. We recommend an approach which estimates and reports the proportion of inaccessible stream segments in the population. Conclusions can then be qualified by stating that estimates only apply a certain proportion of the population. Estimation of the proportion of inaccessible sample segments in the universe can be performed the same way any other proportion is estimated. We view the fact that a segment is inaccessible as "data" on that segment. The "data" collected from an inaccessible unit is the fact that researchers could not take any measurements there and this "data" could be coded as a 0-1 indicator variable; 0 if the segment was inaccessible, 1 if the segment was accessible. If a single equiprobably sample of size n was taken and n_a units were accessible, n_a/n is the estimated proportion of units in the population which were accessible.

When a segment selected in the sample turns out to be inaccessible to investigators, this fact should be recorded in the study's database but the entire sample should not, in general, be redrawn, nor should additional samples be taken. There are certainly exceptions to this rule. If for example, a large portion (say 20% or more) of the sample turns out to be inaccessible, an additional sample could be drawn. This additional sample, if drawn, should be the same type as the original (i.e, equiprobable GSS) and independently drawn. Often, the additional sample complicates data analysis if an appropriate accounting for the design under which data were gathered is to made. For example, if two samples are drawn and the resulting data is analyzed as one sample, the correct first-order inclusion probability is the probability of being included in the first sample or the second sample. If two independent equiprobable GSS samples are drawn and units in the first sample is not excluded from selection by the second sample, the probability that a unit is included in the composite sample is $(n_1 + n_2 - n_1n_2)/N$. Our main point with this example is that inclusion probabilities, both first and second order, can be complicated when two samples are drawn. Other sample schemes which exclude previously drawn units may yield simpler inclusion probabilities.

In general, the frame should not be updated or changed once it is fixed unless inaccessiblity or other problems become chronic in the study over multiple years. Because of the complications it causes at the analysis stage, changes to the frame (e.g., removal of inaccessible segments) should be resisted and made only when absolutely necessary. When inaccessiblity becomes a problem, we recommend drawing a second sample to supplement the first in such a way that the inclusion probabilities are simple to compute.

5. Yearly Abundance Estimation

This section gives estimators for total number of fish and variance of the total fish estimate assuming data were collected from an equiprobably GSS sample of stream segments. This section starts by defining the information required from each sampled segment then gives the estimators for total and variance assuming a single population is of interest and no calibration is performed. Next, we assume estimates from different populations are to be combined into a single estimate for a multi-population area. Last we consider estimates of total fish when auxiliary information is available.

For regional estimation, two pieces of information are required from each sampled stream segment; the estimated total number of fish in each segment and the estimated variance of the total number of fish in each segment. Let $\hat{T}_u$ represent the estimated total number of fish in segment *u*. Let $\hat{V}(\hat{T}_u)$ represent the estimated variance of $\hat{T}_u$. Although we state that $\hat{V}(\hat{T}_u)$ is required, it is not absolutely necessary in all cases. $\hat{V}(\hat{T}_u)$ may not be required in cases where the variation in fish numbers from segment to segment dominates variation within segments. In such cases, incorporation of $\hat{V}(\hat{T}_u)$ is likely inconsequential and the variance estimates given below without terms involving $\hat{V}(\hat{T}_u)$ will be approximately correct.

Incorporation of $\hat{V}(\hat{T}_u)$ into the regional variance estimate allows highly accurate and precise single stream estimates to be combined with less accurate and imprecise single stream estimates provided all single stream protocols are unbiasedly estimating total number of fish. An advantage of this feature is illustrated by the following example. Suppose field crews could only sample one fourth of a segment before logistical problems forced them to stop. Logistical problems might include adverse weather, mechanical breakdowns, unanticipated cliffs or falls, etc. In this situation, we advocate a relatively ad hoc inflation of the total fish estimate from the portion of the segment sampled to the entire segment and an appropriate adjustment in $\hat{V}(\hat{T}_u)$. In this example, the estimate of total fish in the portion sampled would be multiplied by four and the estimated variance of total fish in the portion sampled would be multiplied by 16. This ad hoc approach is approximate but allows use of all the data and could likely be improved if the portion sampled could be thought of as a probability sample from the entire segment and if a more accurate inflation technique could be identified.

5.1 Single population estimates

This section gives the estimator of total number of fish and an estimator for variance of the population estimate assuming an equiprobable GSS or simple random sample (SRS) is taken.

Let u represent a stream segment in the universe and let S represent the sample of u's which were drawn. Recall that the population is of size N (i.e., there exist N u's in the universe) and the sample size is n (i.e., there exist n u's in S). For every $u \in S$ (the symbol '$\in$' means 'element of' or 'in'), we have $\hat{T}_u$ and $\hat{V}(\hat{T}_u)$. The estimate of total number of fish in the population is,

$$\hat{T}_P = \frac{N}{n} \sum_{u \in S} \hat{T}_u \ . \tag{1}$$

An estimate of the approximate variance of $\hat{T}_P$ is,

$$\hat{V}(\hat{T}_P) = \frac{N(N-n)}{n} s_u^2 + \left(\frac{N}{n} \right) \sum_{u \in S} \hat{V}(\hat{T}_u) \tag{2}$$

where

$$s_u^2 = \frac{1}{n-1} \left[\sum_{u \in S} \hat{T}_u^{\,2} - \frac{\left(\sum_{u \in S} \hat{T}_u \right)^2}{n} \right]$$

is the common sample variance of $\hat{T}_u$ (Cochran 1977, equations 11.3 and 11.24; Särndal et al. 1992, chapter 16). Equation 2 is the sum of the regular simple random sample variance estimate and an additional term to account for variation within segments.

Under an equiprobable GSS design, the variance estimator in Equation 2 is approximate because it does not account for the systematic nature of the design and was derived assuming simple random sampling. Ignoring the systematic nature of the GSS and using the simple random sample estimator is a common practice which usually leads to satisfactory variance estimates. The approximate variance estimate is usually satisfactory because it is usually too large when sampling from a population such as stream segments; however, there are situations where this variance estimator is not satisfactory and estimators specifically designed for use with systematic samples should be used. Situations where Equation 2 does not provide a satisfactory variance estimate include those in which the response of interest is cyclic in the population and the systematic sample happens to have the same frequency as the underlying cycle in responses. Use of estimators specifically designed for systematic samples is an unnecessary complication unless circumstances prove they are required. Some verification of the applicability of Equation 2 may be done once data are available.

An approximate 95% confidence interval for true regional total is $\hat{T}_P \pm 2\sqrt{\hat{V}(\hat{T}_P)}$.

5.2 Combination of multiple universe estimates

This sub-section provides formula for combining several abundance estimates into one estimate applicable to a multi-universe area. This situation is commonly called stratified analysis or stratified sampling.

We assume that H populations of known size are of interest and that $\hat{T}_P$ and $\hat{V}(\hat{T}_P)$ are available from each population. We also assume that *independent* probability samples (e.g., equiprobable GSS's) are taken from each universe. We will add a subscript to our notation to accommodate designation of the population. N_h will denote size of the h-th population, $\hat{T}_{P_h}$ will denote the abundance estimate from the h-th population, and $\hat{V}(\hat{T}_{P_h})$ will denote estimated variance of abundance from the h-th population.

Total fish in all H populations can be estimated as

$$\hat{T}_H = \sum_{h=1}^{H} \hat{T}_{P_h} .$$

An unbiased estimator for variance is,

$$\hat{V}(\hat{T}_H) = \sum_{h=1}^{H} \hat{V}(\hat{T}_{P_h})$$

assuming $\hat{V}(\hat{T}_{P_h})$ is unbiased for the variance in every population (Särndal et. al. 1992, Result 3.7.1).

5.3 Incorporation of auxiliary information

This sub-section describes incorporation of auxiliary information which is available at the stream segment level. The techniques of this section apply when the auxiliary information is a single number for each segment and is known on all sampled and un-sampled segments. We consider this situation because we hope that the correlation between auxiliary information and total fish numbers is sufficient to cause a gain in accuracy and precision of the fish numbers. Contrary to the previous section, measurement errors inherent in the $\hat{T}_u$ are not accounted for in this section. We assume the $\hat{V}(\hat{T}_u)$ are small relative to segment-to-segment variation and are overshadowed by the gain in precision afforded by the auxiliary information. We cover estimation when a single auxiliary variable is available. Estimation with more than one auxiliary variable is possible and the interested reader is referred to Särndal et al. (1992) for more details on that topic.

Statistical calibration is another technique which incorporates auxiliary information. In the context of this report we will reserve the term 'statistical calibration' to mean incorporation of information which is available on units *within* whole stream segments (like pools). See the Appendix for a detailed description of statistical calibration and incorporation of information available on units within single stream segments.

Let the value of the auxiliary variable be denoted by x_u. Provided an equiprobable GSS or SRS of stream segments is taken, the *regression estimator* of $\hat{T}_P$ is,

$$\hat{T}_P = N[\bar{T}_s + \hat{B}(\bar{x}_U - \bar{x}_s)] ,$$

where

$$\bar{T}_s = \sum_{u \in S} \hat{T}_u / n,$$

$$\bar{x}_U = \sum_{u \in U} x_u / N,$$

$$\bar{x}_S = \sum_{u \in S} x_u / n,$$

$$\hat{B} = \frac{\sum_{u \in S} (x_u - \bar{x}_s)(\hat{T}_u - \bar{T}_s)}{\sum_{u \in S} (x_u - \bar{x}_s)^2}$$

(Särndal et. al. 1992, equation 7.8.9). Note that to construct this estimator, only the totals of the auxiliary variable on the universe and sample, $\sum_{u \in U} x_u$ and $\sum_{u \in S} x_u$, are needed and not every individual x_u. An estimator for the approximate variance of $\hat{T}_P$ is,

$$\hat{V}(\hat{T}_P) = \frac{N^2(1 - \frac{n}{N})}{n(n-1)} \sum_{u \in S} [1 + a_s(x_u - \bar{x}_s)]^2 e_u^2 , \qquad \textbf{(3)}$$

where

$$e_u = \hat{T}_u - T^*_u ,$$
$$T^*_u = \bar{T}_s + \hat{B}(x_u - \bar{x}_s),$$
$$a_s = n(\bar{x}_U - \bar{x}_s) / \sum_{u \in S} (x_u - \bar{x}_s)^2$$

(Särndal et. al. 1992, page 274). Särndal et. al. (1992) report that the ratio of the approximate variance of the regression estimator (Equation 3) to the variance of the expansion estimator

(Equation 1) is $1\text{-}r^2$, where r is the true finite population correlation coefficient between x_u and $\hat{T}_u$. Consequently, the regression estimator is an improvement over the simple expansion estimator when $r \neq 0$. If $r > 0.7$, improvement in the regression estimator is substantial.

6. Implementation Over Several Years

This section describes a slight modification of the GSS sample selection procedure which can be used to sample a universe of stream segments over several years. Although the original intent of this document and regional estimation protocol was to produce the best possible estimate of abundance in a single year, focus has migrated slightly to include estimation over relatively short time periods. We retain our focus on juvenile coho populations and assume surveys will be run in each of three successive years. Modification of our GSS procedure to allow for sample selection over 1, 2, 4, or more years is straightforward.

Of particular relevance in deciding what design to use over multiple years is whether or not population trends or population status is emphasized as an estimation objective. In making our recommendations in this section, we assume that the primary objective of the surveys are to establish population status (i.e., size) as accurately as possible during a short period of time. While the design we describe is powerful for establishing status because it assures a high degree of spatial coverage, it is perhaps not as powerful as other designs for detecting trends. We note the work and sample designs advocated by Dr. Tony Olsen and Dr. Don Stevens of the Environmental Protection Agency (Corvallis, OR, pers. com.) which put relatively more emphasis on detection of trends at the cost of some power to determine status. We anticipate that Olsen and Stevens will publish their work sometime in 1998 or 1999. Our sample scheme could be repeated over multiple short time periods but its power to detect trends should be considered.

In this section, we recommend a three-year rotating systematic sample. Another name for this design is an interpenetrating systematic sample. With this simple approach, all areas of the population are sampled and desirable geographic coverage will be achieved. If summarization of all three years into a single number is necessary, we recommend a moving average which incorporates previous year's estimates into the current year's estimate.

6.1 Sample selection over multiple years

We assume that the survey will be conducted over three years. Three year corresponds to the life cycle of the anadromous coho salmon. Surveys for non-anadromous fish species may wish to consider surveys spanning one or more years.

Recall that there are N stream segments in the sampling frame and that we desire a sample of n segments each year. For the rotating systematic sample, let $k = N/(3n)$ which is real number and not necessarily an integer. We now view the ordered frame as $3n$ groups of stream segments, where each group is of size k, and we draw a GSS sample of size $3n$. Once drawn, we only visit a systematic sample of n units from the $3n$ in any given year in such a way that all $3n$ units are eventually visited over three years.

The original sample of $3n$ units is drawn as a GSS with step size equal to k. The $3n$ segments in the sample are then divided into three groups systematically. The first group of segments consists of the 1st, 4th, 7th, ..., etc. segments in the sample. The second groups consists of the 2nd, 5th, 8th, ..., etc. segments in the sample. The third group consists of the 3rd, 6th, 9th, ..., etc. segments in the sample. If the sample is drawn in this way, the three year sample is a GSS with high resolution (i.e., step size k) while the sample from any given year is a GSS with one third the resolution of the overall sample (i.e., step size $3k$).

We randomly determine which group of segments to visit each year by constructing a random ordering of the integers 1, 2, and 3 and visiting the corresponding group. For example, if a random shuffling of the first three integers produces 2, 3, 1, we visit the 2nd group of segments the first year, the 3rd group of segments the second year, and the 1st group of segments the third year. Conceptually, the random ordering of the numbers 1, 2, 3 can be constructed by assigning the numbers 1, 2, and 3 to cards and then shuffling the cards.

The inputs to the computer code appearing in Table 2 can be easily modified to accommodate selection of the high resolution GSS sample. Assuming the function F.equal.gss of Table 2 is available and the segment identifiers are 1 through N, the segments to visit each year can be determined using the following code;

```
s <- F.equal.gss( 1:N, 3*n )
ord <-sample(1:3, replace=F)
ind <- c(F,F,F); ind[ ord[1] ] <- T
s.1 <- s[ ind ]
ind <- c(F,F,F); ind[ ord[2] ] <- T
s.2 <- s[ ind ]
ind <- c(F,F,F); ind[ ord[3] ] <- T
s.3 <- s[ ind ]
```

The segments listed in `s.1` would be visited the first year. The segments listed in `s.2` would be visited the second year. The segments listed in `s.3` would be visited the third year.

6.2 Moving average estimates

Yearly estimates of fish abundance do not necessarily need to be summarized across years; however, an average (over 3 years) will capture the full life history of one cohort of coho and potentially reflect status more accurately than individual single year estimates. If a large proportion of an anadromous fish cohort do not return in a single year, surveys should be run until all or most of the cohort has returned.

Each year of a three year study will yield an estimate of total fish in the region, labeled $\hat{T}_P$ and an estimate of variance, labeled $\hat{V}(\hat{T}_P)$. An additional subscript is needed in our notation to describe the moving average. Let $\hat{T}_{P_i}$ represent the estimate of total fish in year i and let $\hat{V}(\hat{T}_{P_i})$ represent its variance. At the end of any 3 year period, the moving average estimate is,

$$\hat{T}_{\bar{P}} = \frac{\sum_{i \in \zeta} \hat{T}_{P_i}}{3} .$$

where ζ is the set of years over which to average. For example, after three years of surveys $\zeta = \{1,2,3\}$, after five years of surveys $\zeta=\{3,4,5\}$. Consistent with our treatment of a systematic sample as a simple random sample, we estimate the variance of the moving average as,

$$\hat{V}(\hat{T}_{\bar{P}}) = \frac{1}{3^2} \sum_{i = \zeta} \hat{V}(\hat{T}_{P_i}) .$$

References

Cochran, W.G., (1977) *Sampling Techniques*, New York: John Wiley and Sons.

Hankin, D.G., and G.H. Reeves, (1988) “Estimating total fish abundance and total habitat area in small streams based on visual estimation methods”. *Canadian journal of Fisheries and Aquatic Sciences* **45**: 834-844.

Särndal, C.E., B. Swensson, and J. Wretman, (1992) *Model assisted survey sampling,* New York: Springer-Verlag. 694 pages.

Appendix: Statistical Calibration

In this appendix, we describe statistical calibration at the single stream level as one method of improving abundance estimates for a segment. Although statistical calibration is not necessarily a regional estimation procedure, we include this appendix because calibration is a popular and widely applicable technique which has the potential to improve precision and accuracy of regional estimates. We stress that information for calibration is available on units within stream segments such as pools, riffles, and runs. If information is available on segments as a whole, see the regression estimator advocated in the main body of this report.

Statistical calibration relies on an auxiliary variable that is correlated with fish numbers to inject information into the system and thereby improve estimates. While improved estimates are always desirable, identification of a correlated auxiliary variable sufficient to cause the improvement is often difficult and full of conjecture prior to data collection. In the absence of a known and strong correlations, we recommend a simple approach to variable collection at the single stream level which keeps the number of measured auxiliary variable to a minimum. In fact, we recommend collecting either one variable or no auxiliary variables if a obvious correlated variable cannot be found.

As an example where calibration has been considered, we point out that in the single stream sampling technique currently favored by Hankin (Dr. Dave Hankin, Humbolt State University, pers. com.) both diver and electro-fishing counts are obtained from a sample of the segment's pools and runs. The diver counts are thought to be less accurate than the electro-fishing counts but are much less expensive logistically to collect and do not harm the fish. Consequently, diver counts are usually carried out on a relatively large number of pools and runs in the segment while electro-fishing is conducted on a relatively small sample of the pools and runs which also received diver counts. Thus, there are two phases to sampling: a large first phase sample is selected and diver counts are made, then a smaller second phase sample (drawn from the first phase sample) is selected and electro-fishing is conducted on the second phase sample. In the end, it is hoped that the correlation between diver counts and electro-fishing counts can be exploited to correct the diver counts so that they reflect a count which electro-fishing would have produced had electro-fishing been carried out on the larger sample of pools and runs.

Calibration is facilitated by estimation of a regression equation which relates the variable of interest (e.g., electro-fishing count) to the auxiliary variable (e.g., diver count). This regression equation is called the *calibration relationship.* Once estimated, the calibration relationship is used to predict the variable of interest at all places where only the auxiliary estimate is known. The resulting predictions of the variable of interest are then summed to obtain an estimate for the entire segment.

Calibration information can be combined across segments (or even across years) to estimate the calibration relationship without effecting the statistical validity of regional abundance estimates. Furthermore, it is not necessary for the calibration relationship be the same in all areas which are combined in order for the resulting estimates to be legitimate. Variance of the resulting estimates can be improved if differences in the calibration relationship can be identified and modeled, but such differences do not invalidate the calibrated estimate. We mention combining of information from different areas because we anticipate that relatively little calibration data will be available on a single segment.

The calibration estimator is now defined. Let u_c represent the *calibration units*. Calibration units are different than the sample units (u) because calibration units (u_c) are defined within a single stream segment. Calibration units may, for example, consist of pools, riffles, or runs. Let the variable of interest be denoted y_{u_c} (e.g., electro-fishing count) and the auxiliary variable be denoted x_{u_c} (e.g., diver count). Let S_1 represent the sample of calibration units on which x_{u_c} is known. Let S_2 represent the sample of u_c from S_1 on which y_{u_c} is known ($S_2 \subseteq S_1$). Let n_1 be the number of u_c in S_1 and n_2 be the number of u_c in S_2. We assume that S_1 was selected from among all units in the segment (or multiple segments if combining across segments) with equal probability. We assume S_2 was selected from S_1 with equal probability.

It is not necessary that x_{u_c} be a single variable. If there exists substantial variation in the calibration relationship, it may be possible to improve the relationship and subsequent estimator efficiency by incorporating additional environmental variables. Incorporation of more than one variable into the calibration relationship is relatively straight forward once a single variable has been used. For simplicity, we assume a single variable, x_{u_c}, will be used.

The calibrated estimate of $\hat{T}_u$ is,

$$\hat{T}_u = \frac{N}{n_1} \sum_{u_c \in S_1} \hat{y}_{u_c} + \frac{N}{n_2} \sum_{u_c \in S_2} (y_{u_c} - \hat{y}_{u_c})$$

(Särndal, et. al. 1992, equation 9.7.25) where $\hat{y}_{u_c}$ is obtained from the ordinary least squares (OLS) regression of y_{u_c} on x_{u_c} in S_2. The ordinary least squares regression estimate of $\hat{y}_{u_c}$ is,

$$\hat{y}_{u_c} = \hat{\alpha} + \hat{\beta} x_{u_c},$$
$$\hat{\beta} = \frac{\sum_{u_c \in S_2} (y_{u_c} - \bar{y}_{u_c})(x_{u_c} - \bar{x}_{u_c})}{\sum_{u_c \in S_2} (x_{u_c} - \bar{x}_{u_c})^2},$$
$$\hat{\alpha} = \bar{y}_{u_c} - \hat{\beta} \bar{x}_{u_c},$$
$$\bar{y}_{u_c} = \sum_{u_c \in S_2} y_{u_c} / n_2,$$
$$\bar{x}_{u_c} = \sum_{u_c \in S_2} x_{u_c} / n_2 .$$

An estimate of the approximate variance of $\hat{T}_u$ under equiprobable designs is,

$$\hat{V}(\hat{T}_u) \approx \frac{N(N-n_1)}{n_1} s_{y_c}{}^2 + \frac{N^2(n_1-n_2)}{n_1 n_2} s_{d_c}{}^2$$

where

$$s_{y_c}{}^2 = \sum_{u_c \in S_2} (y_{u_c} - \bar{y}_{u_c})^2/(n_2-1) \quad \text{and} \quad s_{d_c}{}^2 = \sum_{u_c \in S_2} (y_{u_c} - \hat{y}_{u_c})^2/(n_2-1) .$$

Table 1: Pilot sample frame developed for stream systems in the Lover Klamath basin. Pilot frame materials provide by Mr. Ross Taylor. For illustration, a sample of n=13 is drawn from the N=150 segments with a random start of m=6.106. Note k=150/13=11.53. For brevity, the basin code recommended in Section 3 has been left off the segment id.

Sub-basin	Segment Id	Frame Position	Indicator Sequence	Sub-basin	Segment Id	Frame Position	Indicator Sequence
Wilson Creek	0A001	1		Ah Pah Creek	0F001	72	
	0A002	2			0F002	73	
	0A003	3			0F003	74	
Hunter Creek	0B001	4			0F004	75	
	0B002	5			0F005	76	75.33691
	0B003	6			0F006	77	
	0B004	7	6.10614		0F007	78	
	0B005	8			0F008	79	
	0B006	9			0F009	80	
	0B007	10			0F010	81	
	0B008	11			0F011	82	
	0B009	12			0F012	83	
	0B010	13		Bear Creek	0G001	84	
	0B011	14			0G002	85	
Turwar Creek	0C001	15			0G003	86	
	0C002	16			0G004	87	86.87537
	0C003	17			0G005	88	
	0C004	18	17.6446		0G006	89	
	0C005	19		Surpur Creek	0H001	90	
	0C006	20			0H002	91	
	0C007	21			0H003	92	
	0C008	22			0H004	93	
	0C009	23		Tectah Creek	0I001	94	
	0C010	24			0I002	95	
	0C011	25			0I003	96	
MaGarvey	0D001	26			0I004	97	
	0D002	27			0I005	98	
	0D003	28			0I006	99	98.41383
	0D004	29			0I007	100	
	0D005	30	29.18306		0I008	101	

	0D006	31	
	0D007	32	
	0D008	33	
	0D009	34	
	0D010	35	
	0D011	36	
	0D012	37	
Blue Creek	0E001	38	
	0E002	39	
	0E003	40	
	0E004	41	40.72152
	0E005	42	
	0E006	43	
	0E007	44	
	0E008	45	
	0E009	46	
	0E010	47	
	0E011	48	
	0E012	49	
	0E013	50	
	0E014	51	
	0E015	52	
	0E016	53	52.25999
	0E017	54	
	0E018	55	
	0E019	56	
	0E020	57	
	0E021	58	
	0E022	59	
	0E023	60	
	0E024	61	
	0E025	62	
	0E026	63	
	0E027	64	63.79845
	0E028	65	
	0E029	66	
	0E030	67	
	0E031	68	
	0E032	69	
	0E033	70	
	0I009	102	
	0I010	103	
Mettah Creek	0J001	104	
	0J002	105	
	0J003	106	
	0J004	107	
	0J005	108	
	0J006	109	
	0J007	110	109.9523
	0J008	111	
	0J009	112	
	0J010	113	
Roach Creek	0K001	114	
	0K002	115	
	0K003	116	
	0K004	117	
	0K005	118	
	0K006	119	
	0K007	120	
	0K008	121	
	0K009	122	121.4908
	0K010	123	
	0K011	124	
	0K012	125	
Tully Creek	0L001	126	
	0L002	127	
	0L003	128	
	0L004	129	
	0L005	130	
	0L006	131	
	0L007	132	
	0L008	133	133.0292
	0L009	134	
Pine Creek	0M001	135	
	0M002	136	
	0M003	137	
	0M004	138	
	0M005	139	
	0M006	140	
	0M007	141	

	0E034	71			0M008	142	
					0M009	143	
					0M010	144	
					0M011	145	144.5677
					0M012	146	
					0M013	147	
					0M014	148	
					0M015	149	
					0M016	150	

Table 2: S-Plus® code implementing the equiprobable generalized systematic sample (GSS) algorithm. Example: to draw a GSS of size 3 from the numbers 1 to 10, execute s<-F.equal.gss(1:10, 3). Resultant sample is stored in s.

```
F.equal.gss _ function( f, n ){
#	Purpose: Return an equiprobable GSS sample of size n from frame f.
#	Inputs:
#	f = NX1 vector of sample unit identifiers
#	n = desired number of units in the sample (a scalar)
#	Output:
#	A nX1 vector of the sample unit identifiers which are in the sample.
N <- length(f); k <- N/n; m <- runif(1) * k
s <- ceiling( seq(m, by=k, length=n) )
return( f[s] )}
```