A Comparison of Variance Estimates Of Stream Network Resources

Sarah Williams
Department of Statistics

In partial fulfillment of the requirements
for the degree of Master of Science
Colorado State University
Fort Collins, Colorado
Fall 2006

Table of Contents

Appendix

Abstract

Many studies of environmental resources are designed to make statements about the status of the resource. A second objective often relates to detecting change in the resource. Changes may be practically undetectable at rates as low as 1- to 2-percent over a 10-15 year period. One way of collecting data to ensure detection of change is through a probability sampling design. In any sampling there will be variance and there already exists a variety of methods to choose from to compute the estimated variance of the status of a resource. When there exists directional trend of a response variable there will also be an underlying linear trend, so we consider the use of a linear model. For an environmental indicator the linear model could be as simple as an overall mean plus slope by time with random factors of site and visit year.

Such a linear model can be interpreted in terms of components of variance of a mixed linear model. By contrast, Stevens and Olsen (2003) proposed a local area neighborhood method for estimating variance. This design-based estimate of variance lacks consideration of time. Instead, with a linear model we can examine what aspects of the study design contribute to the variance of an environmental indicator. We relate these two methods of estimating variance using an extensive set of data.

The Oregon Coastal Range coho salmon habitat data, consisting of 1535 field visits to 1055 distinct sites over a period of eight years, has 35 response variables over which to compare the two methods of estimating variance. Its analysis shows that the local estimate decreases as the site variance increases. Conversely, the local estimate increases as the residual variance increases. The local neighborhood estimate also can be described as a linear combination of these components of variance; the local neighborhood estimate increases as either of the components of variance increases. We conclude that, although related, these two methods of estimating variance should be used for different purposes: the local estimate for status and the components of variance for evaluating trend detection.

Acknowledgements

The work reported here was developed under the STAR Research Assistance Agreements CR-829095 awarded to Colorado State University and CR-829096 awarded to Oregon State University by the U.S. Environmental Protection Agency (EPA). This presentation has not been formally reviewed by the EPA. The work done and views expressed here are solely those of the author and STARMAP. EPA does not endorse any products or commercial services mentioned in this report.

I wish to thank the following individuals and organizations for their contributions to preceding research of the design and analysis of aquatic studies:

- Don Stevens of the Department of Statistics, Oregon State University, for the developing the local area neighborhood estimate of variance, the vehicle of comparison herein.
- Thomas Kincaid of Dynamac Corporation, for developing the R software programs used and valuable computational input through the analysis stages.
- Kim Jones, project leader of the Oregon Dept. of Fish & Wildlife's Aquatic Inventories for valuable input of the individual response variables analyzed.
- The Oregon Department of Fisheries and Wildlife for use of the coho habitat data that has been collected through their teaming efforts.

And gratitude to N. Scott Urquhart of the Department of Statistics, Colorado State University, director of STARMAP, for his establishment and continued interest of this project, and without whose guidance this project would have not been possible.

Section 1: Introduction

Human interactions with the earth directly impact environmental resources. Forests have become lumber for houses. Open spaces have become neighborhoods and shopping malls. And, once free-flowing rivers have been dammed to supply power for everyday activities. In the past few decades the global community has seen the need to protect and monitor resources that affect the larger natural balance. We act not only for the concern of our own but also the creatures of the earth. State and local agencies have environmental groups in working order to administer efforts on the localized scale. The Environmental Protection Agency (EPA), formed in 1970, is an independent federal agency established to coordinate programs aimed at reducing pollution and protecting the environment on the national level. Portions of funds available to such organizations have been allocated to resource monitoring. Interest may lie in an isolated area or encompass an entire region covering many states and land types. Thus, researchers are constantly refining their methods for collection and analysis of the data associated with monitoring efforts.

For this discussion we are concerned with studies that cover a large geographic region and variables of interest address aquatic resources. The EPA-funded program, Space-Time Aquatic Resources Modeling and Analysis Program (STARMAP) is one such research effort. Thus measurements of interest have been taken in lakes or along streams. The end result of the monitoring is to be a statement of how the region has changed with respect to these measurements.

The questions of how to estimate trend for data that has spatial dependencies are constantly being examined and methods refined. Improved designs for data collection of aquatic resources have been formulated and are being implemented in new and existing studies. The shortcomings of a simple random sample over large geographic areas are acknowledged. The EPA's Office of Water, Oceans and Wetlands and affiliated state agencies frequently use a Generalized Random Tessellation Stratified (GRTS) design to form a sampling list accounting for spatial balance. In the quest to describe how a resource has changed over time, there must be in place a scheme for revisiting sites. Due to financial limitations, all sites of interest usually cannot be visited at every interval over the course of the study and researchers will choose or uniquely define an appropriate revisit design.

There are many design-based estimates of variance. For instance, Horvitz and Thompson (1952) introduced a method of estimation of population totals and other parameters utilizing inclusion probability. This estimator, referred to as the Horvitz-Thompson estimate, or HT, has many desirable properties including that there exists an unbiased estimate of the total variance. However, the HT has problems as it is sensitive to observations with low inclusion probabilities that happen to make it into the sample. The design-based estimate compared herein to the linear-model approach is referred to as the local area neighborhood variance estimator. This form is a derivative of the HT, to be explained further in Section 3.

The end product of resource monitoring should ultimately include a statement of how the resource has changed over time. Models should incorporate a time component in order to achieve this objective. Due to limited time and money, sampling designs, such as GRTS, have been formulated and scrutinized, on the basis of the most effective data collection. Design-based estimates provide a 'snap-shot' view of the resource's finite population variance. These estimates do not have any consideration of time in their computation. Thus, even if we have visits during different years at a single sample site, it is as if this site was sampled twice at the same time. The panel revisit design was chosen as a means of monitoring change. Panels are sets of sites such that if one site in a panel is revisited in a time interval, all other sites in the panel will be revisited in that interval. It only makes sense to have time in the estimate of a resource's variance. The linear model leads to providing estimates of variance of trend estimates across a region for purposes of hierarchical modeling.

Section 2: The Sampling Design

In any study it is difficult to determine the best way to go about collecting data. In truth, many designs probably would be suitable for providing the investigator the information desired. In fact, what would be a good design in the collection stage may not be knowable until after the data have been collected and analyzed. What may be a good design for one response may not be very good for another. It is with experience with similar studies that investigators are able to formulate the most appropriate way to go about collecting certain information.

A primary criterion is to avoid missing the valuable information while concurrently avoiding having too much information of marginal value. It is the job of the consulting statistician to find the middle ground so the investigator collects data that are as useful and informative as possible. Further, sampling of aquatic resources poses additional complexities.

Streams and rivers have varying and interesting forms; they follow the landscape in infinite patterns. Compare, for instance, a study of the carbon content of all types of ground cover in a remote area versus the carbon content of aquatic plants. In the former case, a simple sampling scheme might be to construct a grid at whose intersecting points a sample of the ground cover is obtained and tested. If one carried out such a sample for an aquatic resource many areas would be underrepresented as the lake or stream may not exist at many of the grid intersections. As for surveying resources, what happens at one point directly may affect the waters downstream and have little to no effect on waters upstream. The opposite may be true as well. The unique shapes and flow of these bodies of water must also be of interest and taken into account when designing a sampling frame.

The Environmental Monitoring and Analysis Program (EMAP) is a research program of the EPA. EMAP has as its goal "…to develop the scientific understanding for translating environmental monitoring data from multiple spatial and temporal scales into assessments of current ecological condition and forecasts of future risks to our natural resources." (www.epa.gov/emap) Through their many research projects and grants, EMAP has aimed to improve the gathering of aquatic data.

Here we will discuss the two important components of planning an aquatic study: spatial distribution and providing a revisit plan aiming to detect trend. It is also of interest to keep in mind that a design must be flexible enough to accommodate unforeseen obstacles. As any gathering of data will show variability, a good design will include a way for estimating this variance. This variance may be the natural variance of the resource of interest or perhaps due to perturbations during the collection stage.

2.1 The Spatial Design: Planning for spatial distribution and the GRTS sample

An aquatic study needs to capture all of the variability across the region. We want to ensure that we are not convenience sampling where a large portion of survey sites are in close proximity. Nor is there much information gained by having "clumps" of data in the center or at the perimeter. The ideal situation is where there is continuity in the points of interest such that points are equally spaced, or nearly so. Systematic designs would provide for this but do not provide completely defensible estimates of variance. This is why the construction of spatially structured list frames has become widely studied to ensure spatial distribution. There are many ways of constructing spatially structured list frames. Graph theoretic, Latin square and Markov processes are a few methods that have been studied recently. See Stevens and Olsen (1999) for a more complete listing of available methods.

One way of drawing a list frame for an aquatic study is by way of Generalized Random Tessellation Stratified (GRTS) sampling. This method is structured so that the sample of an aquatic resource is well dispersed as well as proximity-preserved over the region of interest. See Stevens and Olsen (1999) for a complete development. Briefly, the method recursively partitions the geographic region of interest into sampling regions such that within each partition there is one site to be sampled. Hierarchically, the region is divided into tessellations randomly assigned an address such that points close in space have related addresses. The region is divided into finer and finer tessellations; at each stage of this division, addresses are randomly permuted. The addresses are then placed into an ordered list which will preserve the randomness achieved in the address assignment. Finally, a systematic sample of the list is taken, assuring a spatially balanced randomized sample.

In a monitoring study on a stream the region of interest is already one-dimensional. After the hierarchical random assignment of addresses inclusion densities can be varied to accommodate objectives. For example, they have been varied based on the Strahler order (Strahler, 1957) of the stream (1^{st}, 2^{nd}, 3^{rd}, etc.), or based on an agency determining the relative monitoring importance of the various streams.

2.2 The Temporal Design: Revisiting sites for trend detection by way of panel plans

Eventually, interests will include how the responses over the region are changing over time and there needs to be a plan for revisiting sites. This is the temporal design chosen to make a reasonable schedule for obtaining measurements over the duration of the study. It would be statistically convenient to have a balanced design; data would be collected at every site during every observation interval in the study. This is rarely feasible, possible or even economically efficient. The large scale of a study will often prevent such a feat as time and funds rarely allow complete revisits.

Various revisit plans have been developed and been shown to have desirable properties. Those considered here are in the form of panel plans. A panel is a subset of sites that are all visited in the same year, the time interval considered here. By taking panels, it is possible to achieve an indication of the changes that may be taking place over years. This also provides for an estimate of variability that may exist within a given year. Urquhart and Kincaid (1999) provide an overview and compare the schematics for five planned revisit panel plans. For example, the schematics below show a rotating panel and an augmented serially alternating design.

Figure 1: Examples of Panel Designs.

Rotating Panel

year

panel	1	2	3	4	5	6	7	8	9	10	11	12
-3	x											
-2	x	x										
-1	x	x	x									
0	x	x	x	x								
1	x	x	x	x	x							
2		x	x	x	x	x						
3			x	x	x	x	x					
4				x	x	x	x	x				
⋮					⋱	⋱	⋱	⋱	⋱			

Augmented Serially Alternating

year

panel	1	2	3	4	5	6	7	8	9	10	11	12	
1	x				x				x				…
2		x				x				x			…
3			x				x				x		…
4				x				x				x	…
5	x	x	x	x	x	x	x	x	x	x	x	x	…

In the rotating panel design, in any year five panels are visited and the same panel is visited for five consecutive years. This is a good design for status, however a weak design for trend, but often is implemented. After the five-year block of visits, the site is never observed again.

EMAP projects often use an augmented serially alternating design. This is appropriate when there is some cyclical behavior to be monitored. Above, the panel is visited every five years providing a well temporally spaced sample. Then there is the panel that is visited every year providing a good look at how the resource is changing from year to year. Having a set of sites that are visited each year provides additional benefits in the analysis stage. Any subset of the data, determined by years, is going to have the property of connectedness. A connected design assures that we can estimate population parameters of interest by way of linear model theory because in the two-factor, no-interaction model, connectedness is equivalent to estimability. Thus, we can easily estimate the model parameters $\hat{\mathbf{ß}} = (\mathbf{X}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{S}^{-1}\overline{\mathbf{Y}}$ and $\mathrm{cov}(\hat{\mathbf{ß}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{X})^{-1}$ with no complications.

Many panel revisit designs have been compared as far as power to detect temporal changes. It is shown in Urquhart and Kincaid (1999) that panel plans having revisits across the duration of the study will have higher power for trend detection. Those that use random revisits close in time will most likely miss valuable information since environmental resources may be slowly changing or fluctuating through time. It should also be noted at this point that the optimal design will lie somewhere between the extremes of never revisit and always revisit.

Section 3: Local Area Estimate of Variance

The study of finite population parameter estimation is not a new endeavor. Survey sampling has an extensive body of theory and methodology and is practiced everyday in varying capacity across many disciplines. As mentioned, Horvitz and Thompson (1952) published an influential work on the subject under the title 'A generalization of sampling without replacement from a finite universe'. In a probability sampling design, each element of the finite population is assigned a probability of being included in the sample (referred to as the first order inclusion probability). Then each pair, triplet, etc., of elements can be assigned a probability of being included in the sample (referred to as the second, third, etc., order inclusion probability).

The complexity of probability sample designs can vary greatly which leads to the estimators of population parameters being equally complex. But, for a probability sampling design where all elements of interest have positive inclusion probability, one can use a Horvitz-Thomson estimate equivalence. The Horvitz-Thomson estimate of variance is

$$S_p^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 \frac{Z_i}{p_i} - \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \frac{Z_i Z_j}{p_i p_j} \right).$$

The Z_i are indicators of sample inclusion, p_i are first order inclusion probabilities and y_i are the observations for a population of size N. This estimator is unbiased; however it can possibly take negative values.

We can compare this form to the simple random sample estimate of variance,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i \right)^2 \right).$$

One may be tempted to use this estimator if it is not known that the sample was collected under some predetermined probability sampling design. The main difference, of course, is that the HT estimator has embedded in it the inclusion probabilities that were used to choose the sample. Thus, elements with small probability of being sampled will contribute more to the variance and elements with a larger inclusion probability will contribute less to the estimate. This is the reason why HT has some flaws; the observations that make it into the sample with small probability will heavily influence the variance estimate.

Of course, it may not be appropriate to use the HT estimator in all studies where one can obtain the sampling unit inclusion probabilities. As more complex sampling designs arise researchers have been carefully reformulating the theory to fit the individual scenarios. It depends on the type of study how one should go about modifying an estimator. For the studies of environmental resources one has to consider how the sample behaves in space. The GRTS design preserves spatial proximity. And as in most spatial samples units close in space often will yield similar response values. So, it would also be of use to incorporate this spatial component of the sample in variance estimation.

Stevens (2003) developed an estimator for variance of finite population environmental resources. This estimator is referred to as the Local Area Neighborhood (NBH) variance estimator and is appropriate for a study that is designed with properties such as that of a GRTS sample. This estimator has very desirable properties. It has been shown to be unbiased and stable through simulations. And as a comparison, the Horvitz-Thomson estimate, assuming independent random sampling, may be two times the magnitude of NBH (Stevens, 2003).

Neighborhoods are chosen based on some simple criteria. The neighborhood $D(s_i)$ is the point s_i and, as a minimum, its three closest neighbors. Neighborhoods can include more than these four points, however local variance tends to increase with the addition of more neighbors. The local neighborhood estimate of variance is,

$$S^2_{NBH}(\hat{Y}_T) = \sum_{s_i \in R} \sum_{s_i \in D(s_j)} w_{ij} \left(\frac{y(s_j)}{p(s_j)} - \bar{y}_{D(s_i)} \right)^2$$

The weights, w_{ij}, should vary inversely with $p(s_i)$ and decrease with increasing distance between sites s_i and s_j. Also, $\sum_i w_{ij} = \sum_j w_{ij} = 1$ assuring that neighborhood totals are averages over the neighborhoods and the NBH totals will sum to the total estimate. Here $y(s_i)$ denotes the observation of resource y at site i and $\bar{y}_{D(s_i)} = \sum_{s_j \in D(s_j)} w_{ij} \frac{y(s_j)}{p(s_j)}$ the neighborhood total. See Stevens (1997, 1999, and 2003) for further theoretical development.

Section 4: Linear Model Estimate of Variance

Now consider a second approach to estimating population variance. We will discuss the linear model that has been chosen for this type of data and justify why it accurately portrays the data. The sampling design, as presented in Section 2, incorporates features to deal with varying locations and times across the geographic region of interest. In design-based estimates of variance, as discussed in the previous section, the computation is as if all the data were collected at one time since there are no factors of time or location, other than nearness to neighbors, in the formulation of the estimator.

Agencies and individuals conducting the research want to have some way to monitor trend. The sampling schemes for monitoring that are being used by agencies such as EMAP and affiliates seek to sample across many sites in a region and also over time. Since great amounts of money and time are invested in these studies, the analysis must ultimately include these components. We can incorporate this information into an analysis in the form of a linear model.

Whether or not the data exhibit a linear trend, it always can be described in terms of a linear model. Here we turn our focus to the mixed linear model, mixed in the sense that it includes random and fixed factors in the model. In the most compact form, we write the model as **y** = **Xß** + **Z?** + **e**. As is standard notation of the mixed linear model **y** denotes the vector of observed univariate data, **ß** is a vector of unknown fixed-effects parameters with known design matrix **X,** g is a vector of unknown random effects with known design matrix **Z** and all remaining noise is collected in e. Of course if the model is lacking either fixed effects or random effects, the corresponding terms are omitted.

The linear model naturally leads into a statement of how the resource is changing over time. In the case of a single slope, the above model becomes **y** $= a + bX_i +$ **Z?** + **e**; the value of beta is how the mean value of **y** is changing across a one-unit change in X (time) when all other variables are held constant. All of the information in **Z**g + e gathers the random variation of the observations, **y**. The linear model supports an estimate of variance of the variables of interest.

4.1 Some basic mixed linear model theory

Hocking (2003) gives a complete theoretical development for mixed models of balanced data in his chapter 13. Studies with panel designs will, most likely, not be balanced. The same theory applies, but we only have to note that the structure of the variance-covariance matrix will be slightly different. Once again, the mixed linear model is denoted as $\mathbf{y} = \mathbf{Xß} + \mathbf{Z?} + \mathbf{e}$, and the expected value of the model is $E[\mathbf{y}] = \mathbf{Xß}$.

Let $\Re$ denote the set of random effects. Each component of $\Re$ has an associated f_t, the variance component of the random variable $t \in \Re$. Then all of the randomness of the model is contained in the term $\mathbf{Z?} + \mathbf{e}$ where $\mathbf{Z}$ is the design matrix of the random factors. It can be partitioned into matrices $\mathbf{Z_t}$, corresponding to random term $t \in \Re$. Each of the $\mathbf{Z_t}$ are $n \times n_t$ where n is the total number of observations and n_t the number of distinct classes of variable t. Then we can represent $\mathbf{V}_t$, the incidence matrix associated to random effect t as,

$$\mathbf{V}_t = \mathbf{Z}_t \mathbf{Z}_t^{\mathrm{T}}.$$

Therefore, each of the $\mathbf{V_t}$ is $n \times n$. The covariance matrix is then the sum of these matrices weighted by their associated variance components, also an $n \times n$ matrix:

$$\mathrm{var}[\mathbf{y}] = \mathbf{S} = \sum_{t \in \Re} f_t \mathbf{V}_t .$$

In the typical way of linear models, we will also need to compute the expected mean squares in order to estimate the variance components under Method 3. When the data are balanced the expected mean square for $t \in \Re$ is $y_t = \sum_{\substack{m \in \Re \\ m \supseteq t}} \frac{n}{a_m} f_m$ where a_m is the number of levels of random effect m. The unbalanced case is more complicated: If the mean square for variable t is denoted by $\mathbf{y}^{\mathrm{T}} \mathbf{A}_t \mathbf{y}$, we can express its expected mean square in terms of the trace of the matrices that correspond to each of the random effects: $y_t = \sum_{\substack{m \in \Re \\ m \supseteq t}} f_m \mathrm{tr}[\mathbf{A}_t \mathbf{Z}_m \mathbf{Z}_m^{\mathrm{T}}]$. PROC MIXED computes and solves these systems of equations for each analysis.

4.2 Development of the linear mixed model

Due to site access costs, it does not make sense that the researcher would only be interested in one indicator during the data collection. In most studies there will be multiple responses of interest taken at the site. For instance, when the site is visited, not only will the field scientists take measurements of water chemistry, they may also observe the general state of the site indicating habitat status.

Even though time is generally a continuous variable, we assume a discrete time structure and that observations are made in a small time window. For analysis purposes, time is recorded as simply the year of visit (i.e., 1998, 1999, etc.) and can be translated onto the scale of: 1, 2, 3, etc. So all observations made in the first year of the study have time = 1, observations in the second year have time = 2, etc. The effect of time could manifest itself as a trend across the ordered variable of time, or as random perturbations in each time interval, or both. Consider the following preliminary model with S_i indicating site, T_j indicating time and E_{ij} the residual term:

$$\mathrm{Y}_{ij} = S_i + T_j + E_{ij}, \text{ with}$$

$$\begin{aligned} S_i &\sim (0, s^2_{site}) \\ T_j &\sim (a + b(t_j), s^2_{year}) \\ E_{ij} &\sim (0, s^2_{resid}) \end{aligned}$$

Existing software does not support such analysis. Thus, to adequately model time, its trend must be separated from its random effect by entering the model in two different ways: 1. As a continuous variable with a regression slope, and 2. As a classification variable to accommodate random contributions of sampling period.

The factor of site is also random because there was randomness in the selection of sites to survey the entire population of those eligible. Also, both time and site are considered as random since these are only one possibility of many sets of site and time that the study could have been performed. Since they are random in the model, these two terms have a variance component associated with each as indicated in the model above. They are referred to as s^2_{year} for the variable of 'Time' and s^2_{site} for the variable of 'Site'. This model is strictly additive and there is no assumption of interaction or nesting of any of the factors; of course, more complex situations should include such terms. Any error due to possible interaction effect and variance of the response is absorbed into the residual term, R_{ij}, and has variance component s^2_{resid}.

The resulting model is then,

$$\mathrm{Y}_{ij} = a + b(t_j) + S_i + T_j + E_{ij}\text{, with}$$

$$S_i \sim (0, s^2_{site})$$
$$T_j \sim (0, s^2_{year})$$
$$E_{ij} \sim (0, s^2_{resid})$$

where i = 1,…, s indexes site, j = 1,…, t indexes time or year. By this design we understand that there can be more than one observation at any site/time combination. However, for this study we simplify to the case where researchers only take one set of observations and there is no need to account for replication error. Y_{ij} denotes the response variable of interest at the given site, i, and time, j, combination. Let a denote the overall average at the beginning of the study. Any of the variables that are collected can be modeled in this fashion. The expected value of the above mixed linear model is $\mathrm{E}[\mathrm{Y}_{ij}] = a + b(t_j)$ since each of the random effects have a mean of zero. The variance of any observation is $\mathrm{Var}[\mathrm{Y}_{ij}] = s^2_{site} + s^2_{year} + s^2_{resid}$, and the covariance will be determined by the temporal design.

4.3 A small example

If the design were balanced, where all sites are visited in all years, then we have a very simple parameter estimation problem. In the balanced case, the covariance matrix is highly and regularly patterned, leading to the simple formulas for the expected mean squares given in section 4.1. If the data are not balanced, then it is still quite simple to describe the covariance matrix using the notation above.

Consider a small example of a subset of data consisting of 8 observations having the following revisit structure:

		year			
		1	3	4	7
site	1	x	x	x	x
	2	x		x	
	3			x	x

This example exhibits the properties which one hopes for in the larger and full-sample analysis. This sample is connected and so we will be able to estimate all parameters of the linear model. Connected visually means that in the above diagram, if we start at any of the observations and restrict movement to only vertical and horizontal movement, and changing direction only at the location of an observation, then it is possible to trace a path to any of the other observations. If it is desired to carry out analysis on only a subset of the data then the sub-sample should be chosen such that the cells in a table are connected.

If we sort on year and arrange into tabular notation then the data become:

Obs.	Year	Site
1	1	1
2	1	2
3	3	1
4	4	1
5	4	2
6	4	3
7	7	1
8	7	3

And will have the following $\mathbf{Z}$ and $\mathbf{V}$ matrices:

$$\mathbf{Z}_{year} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{V}_{year} = \mathbf{Z}_{year}\mathbf{Z}^{t}_{year} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{Z}_{site} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{V}_{site} = \mathbf{Z}_{site}\mathbf{Z}^{t}_{site} = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{Z}_{resid} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{V}_{resid} = \mathbf{Z}_{resid}\mathbf{Z}^{t}_{resid} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Then we have, $\mathbf{S} = s_{year}^{2}\mathbf{V}_{year} + s_{site}^{2}\mathbf{V}_{site} + s_{resid}^{2}\mathbf{V}_{resid}$:

$$\mathbf{S} = \begin{bmatrix} s_{R}^{2} & s_{year}^{2} & s_{site}^{2} & s_{site}^{2} & 0 & 0 & s_{site}^{2} & 0 \\ s_{year}^{2} & s_{R}^{2} & 0 & 0 & s_{site}^{2} & 0 & 0 & 0 \\ s_{site}^{2} & 0 & s_{R}^{2} & s_{site}^{2} & 0 & 0 & s_{site}^{2} & 0 \\ s_{site}^{2} & 0 & s_{site}^{2} & s_{R}^{2} & s_{year}^{2} & s_{year}^{2} & s_{site}^{2} & 0 \\ 0 & s_{site}^{2} & 0 & s_{year}^{2} & s_{R}^{2} & s_{year}^{2} & 0 & 0 \\ 0 & 0 & 0 & s_{year}^{2} & s_{year}^{2} & s_{R}^{2} & 0 & s_{site}^{2} \\ s_{site}^{2} & 0 & s_{site}^{2} & s_{site}^{2} & 0 & 0 & s_{R}^{2} & s_{year}^{2} \\ 0 & 0 & 0 & 0 & 0 & s_{site}^{2} & s_{year}^{2} & s_{R}^{2} \end{bmatrix}.$$

Noting that, $s_{R}^{2} = s_{site}^{2} + s_{year}^{2} + s_{resid}^{2}$.

Section 5: A Method of Comparison

Coho Salmon in the Pacific Northwest have been affected by human activity. Land use, most notably area timber harvest, has reduced suitable spawning grounds of the species. This has lead to diminished population and the coho salmon that reside along the Oregon coast are being considered to be listed as threatened under the Endangered Species Act. Realization of this has lead to efforts to monitor this population and their habitat conditions by the Oregon Department of Fisheries and Wildlife (ODFW). The ODFW has been monitoring Oregon’s streams for over fifty years and a more structured program was imposed in 1990 (Stevens 2003). Further, in 1998 the Oregon Plan for Salmon and Watersheds was launched with its mission stated: “To restore the watersheds of Oregon and to recover the fish and wildlife populations of those watersheds to productive and sustainable levels in a manner that provides substantial environmental, cultural, and economic benefits.” (Biennial Report, 2005)

An Evolutionary Significant Unit or ESU is used to define separate populations of salmon. These are large geographic regions where the fish are evolutionarily similar including lifespan, population density and genetic traits. One such region is the Oregon Coastal ESU. ODFW has classified five of its own separate reporting units or Monitoring Areas (MA) that fall within the Oregon Coastal ESU. These are more similar geographically and each MA is treated as its own stratum when sampling. ODWF has concentrated its efforts within these five smaller reporting regions as seen in Figure 2. These regions and their abbreviations are: North Coast (NC), Mid-Coast (MC), Mid-South Coast (MS), Umpqua River Basin (UMP) and South Coast (SC). The MAs were defined prior to the 1998 launch of the plan and are based on river drainage basins.

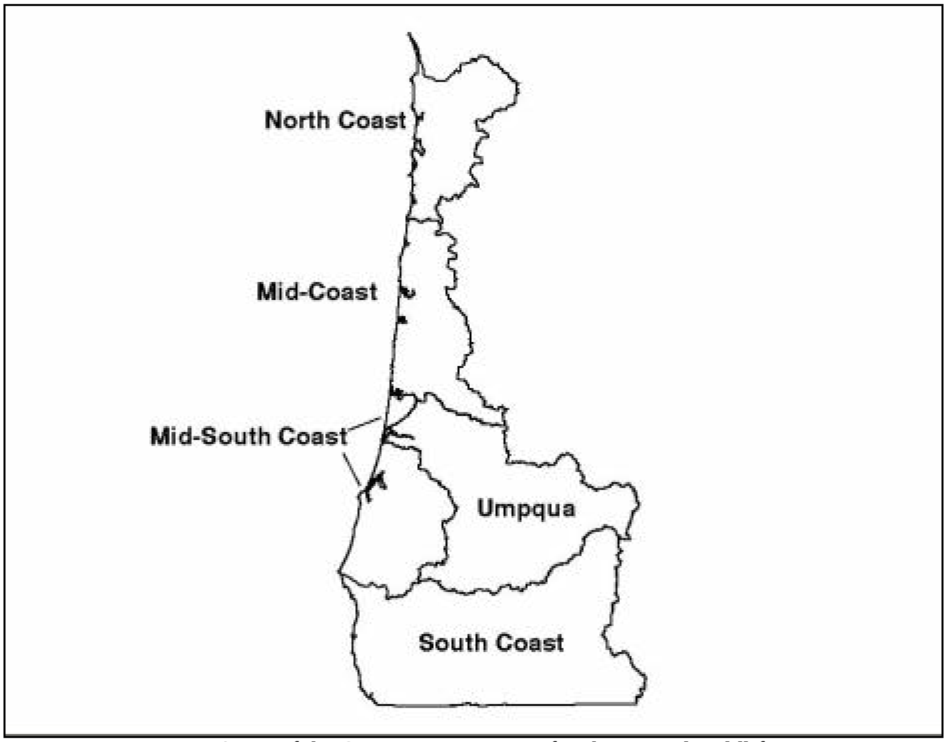


Figure 2: Monitoring Areas of the Oregon Department of Fisheries and Wildlife

The Oregon Department of Fisheries and Wildlife followed a GRTS design in the Oregon Plan for Salmon and Watersheds to assure that within every MA-defined stratum there is a spatially balanced sample obtained independently of the other MAs. They used a variation on a serially alternating panel design consisting of 40 equally site-allotted panels: one panel of sites is visited every year, 3 panels are visited on a 3-year cycle, 9 are visited on a 9-year cycle and the remaining 27 panels define sites visited on a 27-year cycle. So, in any year, 4 panels are visited and are expected to represent an approximately equally-sized sample each year. This was developed on the notion that the salmon follow a 3-year life cycle and the revisit schedule should be able to resemble the salmons' naturally occurring cohorts. A schematic representation of this is presented in the Appendix Section C. Thus, the ODFW has assured that the data collected can be analyzed using the local area neighborhood estimators for describing regional status.

In each of the Monitoring Areas the length, in kilometers, of each of the three types of salmon habitat of interest were obtained. These three domains are identified as spawning, juvenile and habitat. Then it was determined that within each MA that the target sample sizes were 164 spawning sites, 56 juvenile sites and 52 habitat sites. Thus, these targets were used to establish sampling intensities on the frame in order to have a spatially well-balanced sample and to calculate inclusion probabilities.

As mentioned, the goal of the Oregon Plan for Salmon and Watersheds is to monitor the condition of the salmon habitat. At each of the sample sites, when they are visited, the researcher is charged with collecting information about the site that reflects the condition of the stream at that sampling site. This includes the surrounding land and waters coming into and flowing out of the site. Environmental indicators of interest to the ODFW include information such as the condition of the water, stream bed and surrounding landscape.

We are using this monitoring data of the ODFW as an example to compare and contrast the local area neighborhood and linear model estimates of variance. This data set provides observations for 35 quantitative variables that are of interest for the stream monitoring endeavor. The data collected at each site are fairly dense. Measurements include features of the channel and water, surrounding land conditions and other indicators of habitat condition. All measurements collected may have large effects on the status of future salmon populations and the general stream health. At the time of this comparison study there are available data for eight years of monitoring from the period 1998-2005. The temporal design prescribed revisits in every year, however, for reasons unknown to the author, none were made in 2004.

We have computed the simple random sample estimate of variance, s^2, the local neighborhood area estimate of variance and the time, site and residual variance components (s^2_{site}, s^2_{year}, s^2_{resid}) of the mixed linear model. The scale of measurements differs greatly across the variables of interest. Some are percentages (expressed in decimal form) of the sampled area that exhibits some trait of interest; while other measurements are made on a larger interval scale, such as, the number of large woody debris present, LWDPIECE1. By standardizing the estimates to the estimate of the simple random sample variance, we can make generalizations and comparisons across all 35 variables of interest.

Data were analyzed both as the complete Oregon Coast ESU and the five separate ODFW defined Monitoring areas. The complete data set consists of 1535 observations. The MA with the largest number of observations had 323 while the smallest had 299 observations. Thus, we have 6 differently sized regions to observe the likenesses and differences of the variance estimates for each of the 35 variables.

For the local area neighborhood estimates, the R package `psurvey.analysis`, contains functions that will correctly compute the NBH estimates. For the linear model estimate, the SAS function PROC MIXED was called with the proper parameters to achieve the model discussed in Section 4. The complete results are displayed in Appendix A. For clarification, the model that would correspond to the variable WIDTH, the average channel width in meters of the wetted portion of the channel, would be analyzed in SAS PROC MIXED as:

$$\text{WIDTH}_{ij} = ß_0 + ß_1(\text{time}_i) + \text{SITE}_j + \text{TIME}_i + R_{ij}$$

Below is a small example table of the variance estimate results for two variables: LWDPIECE1 and SECCHNAREA. First, consider the column labeled 'Count'. Notice that these counts differ between the two variables for a given region. This number is how many observations were used in the final computation of the variances since each of the two programs removed missing values for analysis. It was verified that these matched in both the R and SAS analyses. The next four columns give the linear model estimates of the variance components. We also have included the value of $s^2_{site} + s^2_{resid}$, denoted 'Site + Res'. In PROC MIXED, there is the possibility to obtain variance components less than 0, which occurred for s^2_{year}. This has positive probability of occurring when the true variance is zero or nearly zero. The last column shows the simple random sample estimate of variance, denoted 'SRS Var.'

Table 1: Variance estimates obtained for two response variables

Response	Region	Count	Linear model estimates: Site	Year	Residual	Site+Res	Local	SRS Var
LWDPIECE1	1-NC	329	77.73	13.54	149.59	227.32	138.47	234.91
	2-MC	308	81.11	-0.43	14.43	95.54	52.19	96.13
	3-MS	294	178.85	9.30	21.48	200.34	113.16	200.97
	4-UMP	299	76.14	0.95	13.96	90.10	40.00	89.75
	5-SC	321	63.51	2.55	21.93	85.44	33.58	86.88
	ALL	1551	99.54	2.47	48.12	147.65	75.46	148.72
SECCHNAREA	1-NC	314	247805.84	-9.22	92810.91	340616.75	225939.60	342010.30
	2-MC	304	191217.93	-3020.04	73495.43	264713.36	77529.41	262645.77
	3-MS	286	48221.19	-154.31	21056.20	69277.39	44262.48	69658.16
	4-UMP	272	48744.10	143.67	7334.76	56078.86	37507.50	57134.55
	5-SC	304	257152.82	9281.23	154926.41	412079.24	164045.10	405046.59
	ALL	1480	163931.28	871.82	74237.91	238169.19	113843.40	238242.17

For purpose of discussion, Table 1 displays 12 rows of the 210 rows of output displayed in Appendix A. LWDPIECE1 records the number of pieces of large woody debris per 100 meters of primary stream length. This variable ranged from 0 to 149.1 with a mean value of 30.7 across all regions. SECCHNAREA is the area of secondary channels in square meters. Contrasted with LWDPIECE1, this variable ranged in continuous values from 0 to 5730 with a mean value of 195. It can be seen that the scale of the computed variances differ greatly and it would be difficult to make any meaningful comparisons both within one variable across the regions or if we wanted to compare the variances of two or more variables. Of course this is true simply because of the differing scales of measurement.

To remedy this, we have taken each of the first five variance measures and divided by its SRS variance, as displayed in Table 2. This puts everything, except for the sum of the site and residual components, on the same zero-to-one scale. Again, since we are not working with a balanced design, it is possible that the linear model variance could be noticeably different from the simple random sample variance. This difference is such that the explained ratio will still be close to one and this is exemplified in the 'Site + Res/SRS' column.

Table 2: Estimates of table 1 standardized to the simple random sample estimate of variance.

Response	Region	Count	Linear model estimates: Site/SRS	Year/SRS	Res/SRS	Site+Res/SRS	Local/SRS
LWDPIECE1	1-NC	329	0.331	0.058	0.637	0.968	0.589
	2-MC	308	0.844	-0.004	0.150	0.994	0.543
	3-MS	294	0.890	0.046	0.107	0.997	0.563
	4-UMP	299	0.848	0.011	0.156	1.004	0.446
	5-SC	321	0.731	0.029	0.252	0.983	0.386
	ALL	1551	0.669	0.017	0.324	0.993	0.507
SECCHNAREA	1-NC	314	0.725	0.000	0.271	0.996	0.661
	2-MC	304	0.728	-0.011	0.280	1.008	0.295
	3-MS	286	0.692	-0.002	0.302	0.995	0.635
	4-UMP	272	0.853	0.003	0.128	0.982	0.656
	5-SC	304	0.635	0.023	0.382	1.017	0.405
	ALL	1480	0.688	0.004	0.312	1.000	0.478

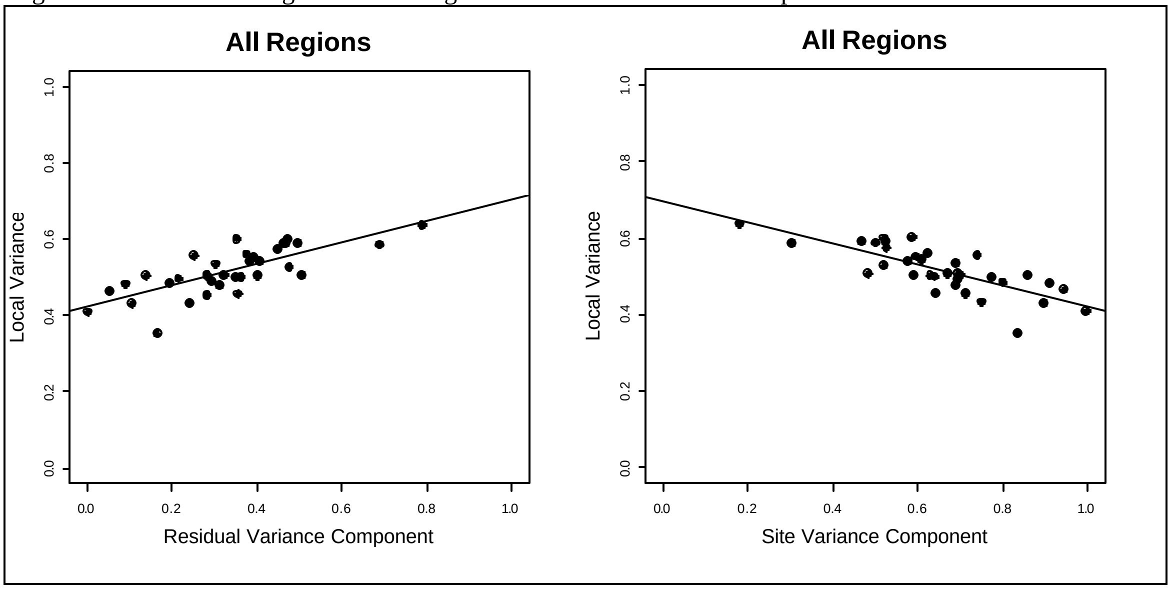
From Table 2, we can make some preliminary observations without computation. It should be noted that these observations hold for all variables:

- Site variance is most often greater than the residual variance
- The sum of site and residual variance components is “close” to the simple random sample variance.
- The variance due to year effect is quite small compared to the site and residual effects.

Some further observations that have been made by way of plotting the data:

- As residual variance increases so does the local estimate
- As site variance increases the local estimate decreases

Figure 3: Plots of the Oregon Coastal Region linear model variance components estimates



The plots in Figure 3 are for the collective larger coastal region and each of the points represents one of the 35 response variables. Similar plots were observed in all instances of the 5 smaller Monitoring Areas.

Results

The following results are based on the data from the monitoring of coho salmon by the Oregon Department of Fisheries and Wildlife. In this effort they have collected many environmental indicators of the condition of the sampling sites and, for this analysis, we utilized 35 numerical variables. They are described in the Appendix, Section A. Upon completing the preliminary examination of the mixed linear model and the collective local area estimates of variance, all estimates were assembled into one large data set of the simple random sample estimate scaled figures. With this data, graphs were constructed and some rough observations were made of the estimates having to do with the complete data. It should be noted that at first the computations were carried out for the complete data and that, as a confirmation of these observations, the same analyses were done on the data partitioned into the five smaller monitoring areas of the ODFW. Similar results were observed when computations were done for the 5 smaller regions defined by the monitoring areas of the ODFW.

There is the general observation that as residual component of variance increases, the local area estimate of variance increases as well. The average ratio, computed as $\frac{1}{35}\sum_{i=1}^{35}\frac{s_{resid,i}^2}{s_{NBH,i}^2}$, was found to be 0.63. Indicating that, on average, the residual component of variance is 0.63 times the local area estimate. Equivalently, we can say that the local area estimate is 1.59 times the residual component of variance estimate on average. There is another important observation that as the site component of variance increases the local area estimate of variance decreases. The same relationships described above were carried out and the average ratio, computed as $\frac{1}{35}\sum_{i=1}^{35}\frac{s_{site,i}^2}{s_{NBH,i}^2}$, and is equal to 1.32. Again, indicating that on average the site component of variance is 1.32 times the local area estimate. Or, we can say that the local area estimate is 0.76 times the site component of variance estimate on average.

It is also of interest to express the local estimate as a linear combination of the site and residual variance components. A multiple linear regression was performed and it was found that we can express the local estimate of response variable *i* as,

$$\hat{s}^2_{NBH,i} = 0.42 * s^2_{site,i} + 0.72 * s^2_{resid,i} .$$

Noting that this expression was forced through the origin we can get a measure similar to the standard R^2 by computing $\frac{SS_{Reg} - n\bar{y}^2}{SS_{Tot} - n\bar{y}^2} = 0.515$. We could also consider using the measure as proposed by Hocking (2002). That is, the square of the correlation between the computed local estimate and the predicted local estimates, 0.517. Neither measure shows a strong relationship but does give some indication that the above equation does try to explain some of the variance in the local estimate across the responses considered.

When there is a larger estimate of the time component of variance there is little to no relationship that can be drawn about the behavior of the local estimate. This makes sense because the local area estimate contains no consideration of time in the computation of this estimate. So, whether measurements are taken at one collection or over the extent of a long-term study, will make no difference in the computed values of the Local Area Neighborhood estimate.

The above summaries were describing the relationships of the local variance one-at-a-time to the linear model variance components of site, time and residual. It was also of interest to explore a multiple relationship such that we could perhaps predict the local area estimate as the sum of all three components of variance: Site, Time and Residual. In this analysis, it was determined that there is no difference between the regression that resulted from using the 35 variables of the entire coastal region than that of the region partitioned into the five Monitoring Areas as defined by the Oregon Department of Fisheries and wildlife within the coastal region. So, there were no gains by pooling the information and the arbitrary assignment of regions was insignificant.

Conclusions

The ODFW data was a test data set to answer the following question: when is it appropriate to use the linear model estimates over the local area estimate of variance and vice versa?

To answer this, consider that it is known that the local area neighborhood estimate of variance will contain all information about the residual variance and time. We know that there is no variance due to time. Thus, we need to consider to what extent the local estimate accounts for the variance that is due to site. The local estimate was on average 52%, or roughly half, of the simple random sample estimate of variance. And the results of this study show that the local estimate accounts for roughly 60-70% of the variance due to the site where the measurements are taken.

For a region that will naturally produce higher variance one should consider the effects of using the NBH. By this we expect there to be higher than normally observed site variance. As such, the local area estimate of variance may be misleadingly low. For example, in the ODFW data set, the Umpqua Monitoring Area covers a wide geographic region. One side of the region is to the west of the Oregon Coastal Range and consists of Douglas Fir forests. The larger, eastern portion of the region is the Umpqua River drainage basin, to the east of the coastal range, includes lowland meandering streams making suitable farming land. By these differing landscapes of the streams in this region, it is expected that there would naturally be wide ranges of the variables of interest. In this situation, it was the case that the relationships described in the results section were observed, however this relationship was very weak.

Overall, site accounted for approximately two-thirds of the variance whereas residual accounted for the remaining one-third of the variance. The local estimate was on average one-half the size of the linear or simple random sample variance. We can deduce that since all the residual variance is contained in the local estimate that approximately half of the variance due to site will be accounted for by using the local area estimate.

References

Hocking, R.R. **Methods and Applications of Linear Models**. Wiley, 2002

Horvitz, D.G. and D.J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**: 663–685.

Sarndal, C., Swensson, B., and Wretman, J. **Model Assisted Survey Sampling**. Springer-Verlag, 1992.

Stevens, D.L., and Olsen A.R. 1999. Spatially Restricted Surveys Over Time for Aquatic Resources. *Journal of Agricultural, Biological and Environmental Statistics*, Vol.4, No.4: 415-428.

Stevens, D.L. 2003. Sampling Design and Statistical Analysis Methods for the Integrated Biological and Physical Monitoring of Oregon Streams. The Oregon Plan for Salmon and Watersheds.

Stevens, D.L., and Olsen, A.R. 2003. Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* **14**:593-610.

Strahler A.N. 1957. Quantitative analysis of watershed geomorphology. *Transactions of the American Geophysical Union*, 21, 913-920.

Urquhart, N.S., Overton W.S., and Birkes D.S. 1993. Comparing Sampling Designs for Monitoring Ecological Status and Trends: Impact of Temporal Patterns**. Statistics for the Environment**, chapter 3:71-85.

Urquhart, N.S., Paulsen, S.G. and Larsen, D.P. 1998. Monitoring for policy-relevant regional trends over time. *Ecological Applications*, 8: 246-257.

Urquhart, N.S., and Kincaid, T.M. 1999. Designs for detecting trend from repeated surveys of ecological resources. *Journal of Agricultural, Biological and Environmental Statistics*, 4: 404-414.

The Oregon Plan for Watersheds and Salmon, Biennial Report, 2005.

www.epa.gov/emap

www.dfw.state.or.us
Map obtained from http://nrimp.dfw.state.or.us/crl/default.aspx?PN=GCAs

Appendices

A: Tables of obtained variance estimates

Response	Region	Count	Slope	Site	Year	Residual	SiteRes	Local	SRS
WATERSHED	**1**-NC	329	0.06 **	709.83	0.00	0.25	710.08	422.34	708.60
	2-MC	307	0.02	1568.99	0.00	0.04	1569.03	826.03	1569.67
	3-MS	293	-0.03	2094.53	0.00	0.39	2094.92	697.63	2076.68
	4-UMP	298	0.10	942.05	-0.06	9.93	951.98	501.35	943.78
	5-SC	320	0.12 **	2136.14	-0.09	2.88	2139.02	615.53	2135.44
	ALL	1547	0.06	1479.50	0.00	2.65	1482.15	606.40	1483.56
PRICHNLL	**1**-NC	329	23.29 **	46628.15	527.11	19170.72	65798.87	35658.53	67248.77
	2-MC	308	5.05	58913.80	-193.61	8905.11	67818.91	36870.36	67245.98
	3-MS	294	13.31 *	47293.51	498.81	18755.71	66049.22	34011.36	65242.61
	4-UMP	299	2.84	50155.14	673.44	12868.42	63023.56	31284.48	62590.17
	5-SC	323	0.27	44830.65	-151.79	5073.72	49904.37	24665.76	52419.54
	ALL	1553	8.20	53408.27	393.31	12984.90	66393.17	32324.29	66803.03
SECCHNLL	**1**-NC	314	0.62	8996.91	323.71	4631.94	13628.85	8865.62	13944.09
	2-MC	304	3.65 **	10509.10	-215.27	5403.48	15912.58	6048.40	15815.55
	3-MS	286	0.81	8635.38	234.64	1226.14	9861.53	6718.65	9954.72
	4-UMP	272	3.20 *	1181.89	41.48	1194.75	2376.64	1608.33	2522.09
	5-SC	304	1.24	6366.80	484.89	3243.74	9610.54	5110.84	9765.78
	ALL	1480	1.70	7536.16	42.98	3306.78	10842.94	5823.28	10905.75
PRICHNAREA	**1**-NC	329	88.38 *	14629967.42	-12698.67	1232099.61	15862067.03	10297290.00	15816634.88
	2-MC	308	-26.79	15736641.98	-36845.04	1290427.27	17027069.26	8982976.00	16990754.79
	3-MS	294	29.47	22405369.54	150572.36	2630135.49	25035505.03	10078740.00	24548387.92
	4-UMP	299	-26.04	8699169.45	5758.61	576325.79	9275495.23	4316992.00	9226905.21
	5-SC	323	-10.32	7860774.38	-49363.29	1191647.25	9052421.63	3949771.00	9163129.82
	ALL	1553	9.84 **	14042231.09	3183.29	1382718.01	15424949.10	7460706.00	15427343.90
SECCHNAREA	**1**-NC	314	5.30	247805.84	-9.22	92810.91	340616.75	225939.60	342010.30
	2-MC	304	14.87 **	191217.93	-3020.04	73495.43	264713.36	77529.41	262645.77
	3-MS	286	9.16 *	48221.19	-154.31	21056.20	69277.39	44262.48	69658.16
	4-UMP	272	1.09	48744.10	143.67	7334.76	56078.86	37507.50	57134.55
	5-SC	304	6.20	257152.82	9281.23	154926.41	412079.24	164045.10	405046.59
	ALL	1480	7.70	163931.28	871.82	74237.91	238169.19	113843.40	238242.17
GRADIENT	**1**-NC	329	-0.01	16.87	0.17	1.54	18.41	8.78	18.31
	2-MC	308	0.02	19.77	0.00	0.30	20.07	11.74	20.11
	3-MS	294	-0.08	35.25	0.35	3.33	38.58	19.54	38.19
	4-UMP	299	-0.07	33.54	1.20	1.43	34.97	13.91	35.78
	5-SC	323	-0.20 *	48.41	0.16	1.29	49.69	26.64	49.41
	ALL	1553	-0.07	32.78	0.08	1.84	34.62	16.11	34.66
VWIRCH	**1**-NC	329	-0.50	59.61	2.92	41.20	100.81	57.93	106.15
	2-MC	308	0.79 **	2.19	-1.24	40.03	42.21	27.83	43.76
	3-MS	294	0.23 *	29.51	0.15	2.35	31.85	15.65	31.96
	4-UMP	299	1.05 **	93.66	-0.54	45.46	139.11	70.59	144.61
	5-SC	323	0.47 **	38.22	-1.34	82.43	120.65	55.12	120.38
	ALL	1553	0.91 **	43.60	-0.50	45.71	89.30	45.84	90.34
WIDTH	**1**-NC	329	-0.06 **	10.73	-0.08	1.42	12.15	7.61	12.27
	2-MC	308	-0.02	12.01	-0.20	4.25	16.26	8.78	16.31
	3-MS	294	-0.01	10.61	0.01	0.87	11.48	5.02	11.63
	4-UMP	299	-0.03	5.15	0.03	0.36	5.51	2.58	5.52
	5-SC	323	-0.02	5.60	-0.02	0.93	6.53	3.07	6.52
	ALL	1553	-0.04	9.21	-0.01	1.50	10.71	5.40	10.73
ACW	**1**-NC	329	0.04 *	39.25	1.19	5.22	44.47	28.99	46.58
	2-MC	308	-0.14	32.45	0.19	2.74	35.18	18.32	35.75
	3-MS	294	-0.43	49.78	-0.12	54.48	104.27	40.23	103.11
	4-UMP	299	-0.01	17.17	-0.03	1.32	18.49	8.87	18.35
	5-SC	323	-0.14	30.07	0.43	4.66	34.73	14.57	34.71
	ALL	1553	-0.18 *	34.50	0.06	13.69	48.19	22.01	48.44

** Indicates significance of slope estimate at the 0.01 level
* Indicates significance of slope estimate at the 0.05 level

Variable descriptions:
WATERSHED – Watershed area from 7.5 min topo (blanks acceptable)
PRICHNLL – Length of primary channel (m)
SECCHNLL – Length of secondary channel (m)
PRICHNAREA – Area of primary channel (m^2)
SECCHNAREA – Area of secondary channels (m^2)
GRADIENT – Average of unit gradients (percent slope) for reach, weighted by unit length
VWIRCH – Valley width index. The ratio of active channel to valley floor
WIDTH – Average channel width (m) of the wetted portion of the channel
ACW – Active channel width (m). Horizontal distance across the channel at annual high flow line.

Response	Region	Count	Slope	Site	Year	Residual	SiteRes	Local	SRS
NOUNITS	1-NC	329	0.68	88.10	17.82	129.72	217.82	143.57	231.48
	2-MC	308	1.10	235.37	7.06	130.90	366.27	242.42	394.18
	3-MS	294	0.33	131.25	7.34	107.78	239.03	130.60	248.61
	4-UMP	299	1.77	148.00	43.24	106.39	254.39	189.68	291.14
	5-SC	323	0.44	129.21	9.17	110.03	239.24	120.99	244.53
	ALL	1553	0.85	151.62	1.18	129.46	281.08	165.17	288.30
NOPOOLS	1-NC	329	-0.19 **	38.34	15.43	44.17	82.51	56.26	89.26
	2-MC	308	-0.04	94.39	-0.73	43.54	137.94	77.72	141.34
	3-MS	294	-0.09	45.49	4.70	34.66	80.15	45.30	83.66
	4-UMP	299	0.55	52.93	10.32	40.01	92.94	59.60	99.74
	5-SC	323	0.07	32.59	5.08	26.72	59.32	33.87	60.92
	ALL	1553	0.12	58.32	2.16	41.14	99.46	54.69	100.91
PCTPOOLS	1-NC	329	-0.07	501.29	49.07	219.94	721.23	358.03	742.12
	2-MC	308	1.08 **	487.72	-3.65	141.12	628.83	311.60	643.19
	3-MS	294	-0.03	501.81	26.09	188.99	690.81	348.69	713.25
	4-UMP	299	-0.09	392.08	19.70	92.03	484.11	213.08	479.56
	5-SC	323	-0.03	178.92	3.18	84.26	263.17	124.88	261.18
	ALL	1553	0.11	468.87	9.89	150.41	619.28	270.98	625.36
PCTSCPOOL	1-NC	329	-0.34	303.53	33.82	121.31	424.84	263.79	429.37
	2-MC	308	-0.21	305.32	9.56	115.43	420.75	215.96	424.17
	3-MS	294	0.10	382.27	39.47	180.31	562.58	302.26	594.12
	4-UMP	299	0.40	304.35	24.58	96.52	400.88	205.74	401.33
	5-SC	323	-0.13	153.67	1.94	68.91	222.58	105.89	220.49
	ALL	1553	-0.05	305.07	10.26	123.29	428.36	219.93	435.72
PCTSWPOOL	1-NC	329	0.28	363.66	13.83	181.84	545.50	257.14	552.27
	2-MC	308	1.37 **	187.60	-2.61	169.09	356.69	196.62	366.39
	3-MS	294	-0.58 **	221.30	-3.26	81.98	303.28	132.96	300.52
	4-UMP	299	-0.62 **	36.15	-1.72	35.95	72.10	30.29	71.94
	5-SC	323	0.09	5.77	-0.07	22.70	28.47	16.15	28.26
	ALL	1553	0.27 *	178.26	-0.29	98.86	277.12	126.56	277.42
SCRPOOLD	1-NC	314	-0.01078	0.07635	0.00346	0.01566	0.09201	0.05685	0.09676
	2-MC	305	-0.00025	0.05103	0.00180	0.04333	0.09436	0.06208	0.09678
	3-MS	283	0.00807	0.10928	0.03494	0.03694	0.14622	0.08138	0.16232
	4-UMP	287	-0.01955	0.07210	0.00246	0.03143	0.10353	0.04804	0.10778
	5-SC	304	-0.01507	0.11028	0.00288	0.01403	0.12431	0.05259	0.13015
	ALL	1493	-0.00959	0.08241	0.00270	0.03353	0.11594	0.06006	0.11875
RESIDPD	1-NC	319	-0.00017	0.04746	0.00073	0.01336	0.06082	0.03668	0.06174
	2-MC	306	0.00802	0.03053	0.00049	0.00818	0.03870	0.02473	0.04109
	3-MS	289	0.00873	0.08588	0.00409	0.04997	0.13585	0.06325	0.13539
	4-UMP	284	-0.01582 *	0.03089	0.00083	0.02913	0.06002	0.03134	0.06221
	5-SC	305	-0.01170	0.07292	0.00232	0.01001	0.08293	0.03562	0.08601
	ALL	1503	-0.00352	0.05369	0.00125	0.02271	0.07640	0.03814	0.07753
RIFFLEDEP	1-NC	287	-0.01042	0.00429	0.00189	0.00339	0.00768	0.00650	0.00940
	2-MC	291	0.00410 **	0.00471	-0.00005	0.00240	0.00711	0.00472	0.00707
	3-MS	267	-0.00125	0.00350	0.00226	0.00186	0.00536	0.00429	0.00702
	4-UMP	260	-0.00807	0.00444	0.00087	0.00268	0.00712	0.00548	0.00809
	5-SC	262	-0.00179	0.00748	0.00046	0.00262	0.01010	0.00555	0.01078
	ALL	1367	-0.00379	0.00517	0.00069	0.00310	0.00827	0.00531	0.00883
LRGBLDR	1-NC	327	2.10	71237.86	2876.89	30846.69	102084.54	48591.47	103771.74
	2-MC	308	16.36 **	144888.82	-681.37	12570.25	157459.07	90491.26	157120.43
	3-MS	292	20.67 **	80054.39	299.55	36184.82	116239.21	58339.18	116839.33
	4-UMP	296	1.23	59995.89	-265.02	20587.88	80583.77	49356.86	82981.78
	5-SC	323	5.56	194304.17	9367.47	46174.98	240479.15	117830.90	245942.33
	ALL	1546	6.69	114417.59	1030.24	32022.40	146439.99	73698.63	147960.51

** Indicates significance of slope estimate at the 0.01 level
* Indicates significance of slope estimate at the 0.05 level

Variable Descriptions:
NOUNITS – Number of habitat units in the reach (total of pools, riffles, rapids, etc)
NOPOOLS – Combined count of scour and dammed pools in reach
PCTPOOLS – Combined percentage (by area) of scour and dammed pools in reach
PCTSCPOOL – Percent of scour pools (lateral, straight, plunge, trench)
PCTSWPOOL – Percent of slow water pools (dam, beaver, backwater, alcove, isolated)
SCRPOOLD – Average depth of scour pools
RESIDPD – Average residual depth of pools
RIFFLEDEP – Average depth of riffles
LRGBLDR – Number of boulders = 0.5m diameter in reach

Response	Region	Count	Slope	Site	Year	Residual	SiteRes	Local	SRS
PCTSNDOR	**1**-NC	329	-1.81	596.63	40.64	63.81	660.43	235.68	719.04
	2-MC	308	-0.23	361.75	11.28	92.93	454.68	187.30	454.47
	3-MS	294	1.83 **	711.92	4.66	87.40	799.33	245.75	796.95
	4-UMP	299	0.26	438.74	11.87	64.61	503.35	202.62	500.41
	5-SC	323	0.04	282.05	13.28	63.14	345.18	144.39	355.57
	ALL	1553	-0.11	484.01	1.84	95.20	579.20	204.15	579.20
PCTGRAVEL	**1**-NC	329	1.78 **	127.38	12.86	76.63	204.01	113.62	226.39
	2-MC	308	0.02	146.40	6.08	110.79	257.19	150.16	259.32
	3-MS	294	0.05	183.72	22.37	98.30	282.02	125.36	287.79
	4-UMP	299	0.18	127.49	26.48	55.29	182.77	98.77	192.64
	5-SC	323	-0.19	87.63	1.33	50.54	138.17	68.75	136.90
	ALL	1553	0.38	131.45	1.18	89.50	220.94	112.66	223.12
PCTBEDROCK	**1**-NC	329	-0.25 **	87.59	-0.69	15.84	103.42	60.88	103.62
	2-MC	308	0.04	192.39	0.69	17.56	209.94	80.02	206.04
	3-MS	294	-0.48 **	136.53	-1.09	24.84	161.38	66.70	160.35
	4-UMP	299	0.15	216.29	0.42	19.37	235.66	109.91	233.08
	5-SC	323	0.49 *	125.45	1.22	13.07	138.51	55.74	137.42
	ALL	1553	0.17 *	153.10	0.02	18.08	171.18	73.56	170.67
POOLS_KMTL	**1**-NC	314	-0.72	61.74	13.65	41.51	103.25	67.13	112.34
	2-MC	304	-0.23	61.80	-0.01	77.51	139.31	80.22	145.20
	3-MS	286	-0.62	93.75	2.29	74.52	168.26	92.63	170.86
	4-UMP	272	0.51	90.19	25.49	89.41	179.61	133.62	199.14
	5-SC	304	0.16	61.25	8.90	61.67	122.92	76.90	131.52
	ALL	1480	-0.08	76.36	6.17	71.40	147.76	89.82	152.67
POOLS_KMPL	**1**-NC	329	-0.88	59.35	39.13	67.69	127.04	91.36	144.48
	2-MC	308	-0.18	70.45	1.11	85.77	156.22	95.06	162.19
	3-MS	294	-0.61	103.59	4.39	77.13	180.72	99.54	184.83
	4-UMP	299	0.51	88.65	32.19	100.72	189.38	141.09	212.03
	5-SC	323	0.16	86.18	9.15	66.51	152.69	87.18	161.43
	ALL	1553	-0.10	81.67	7.42	86.75	168.43	103.15	174.52
POOL1P_KM	**1**-NC	329	-0.24 **	6.78	0.10	4.27	11.04	6.79	11.48
	2-MC	308	0.01	4.49	0.01	1.83	6.33	3.02	6.29
	3-MS	294	0.09 **	9.49	-0.13	3.20	12.69	8.13	12.73
	4-UMP	299	-0.02	5.39	0.09	0.73	6.12	2.74	6.14
	5-SC	323	-0.07 **	7.90	-0.05	1.68	9.59	5.18	9.56
	ALL	1553	-0.09 **	6.92	0.00	2.36	9.28	5.20	9.35
PCTSHADE	**1**-NC	329	-0.65	188.82	23.79	107.07	295.88	173.97	307.21
	2-MC	308	0.29	96.13	7.81	69.40	165.53	103.49	170.84
	3-MS	294	-0.43	294.44	6.64	75.03	369.48	160.68	368.07
	4-UMP	299	0.28	96.67	23.23	137.12	233.80	167.25	264.77
	5-SC	323	-0.48	248.11	3.43	73.12	321.23	112.02	318.73
	ALL	1553	-0.11	183.80	5.44	100.47	284.27	143.26	286.97
PCTEROSION	**1**-NC	329	-6.10 **	-52.13	72.03	321.47	269.34	281.58	400.18
	2-MC	308	0.49	73.48	30.42	82.01	155.49	114.44	171.48
	3-MS	290	0.78 **	62.46	1.00	63.41	125.86	87.98	129.19
	4-UMP	295	0.18	99.52	0.41	82.76	182.28	101.12	189.18
	5-SC	321	-0.11	50.48	26.25	99.87	150.35	118.26	169.17
	ALL	1543	-0.67	40.60	8.89	177.20	217.80	142.99	224.42
PCTUNDERC	**1**-NC	324	-1.15 **	7.21	1.51	39.19	46.40	32.07	53.06
	2-MC	307	0.51	27.38	7.83	26.85	54.22	37.84	59.19
	3-MS	293	1.03 *	16.89	6.80	39.77	56.66	32.47	63.13
	4-UMP	292	-0.24	34.17	0.25	23.31	57.48	34.98	58.03
	5-SC	320	-0.27	-2.20	0.42	35.79	33.59	22.18	35.00
	ALL	1536	-0.02	16.47	1.03	37.43	53.90	31.85	54.34

** Indicates significance of slope estimate at the 0.01 level
* Indicates significance of slope estimate at the 0.05 level

PCTSNDOR – Average percent of sand, silt and organics in surface substrate of all units
PCTGRAVEL – Average percent of gravel in surface substrate of all units
PCTBEDROCK – Average percent of bedrock in surface substrate of all units
POOLS_KMTL – Number pools per kilometer of total stream length
POOLS_KMPL – Number pools per kilometer of primary channel length
POOL1P_KM – Number pools deeper than 1 meter per kilometer of total stream length
PCTSHADE – Percentage of 180° amount of shade provided to stream by riparian vegetation and topography
PCTEROSION – percent reach length of channel units with banks classified as actively eroding
PCTUNDERC – Undercut bank unit average as percent of unit length

Response	Region	Count	Slope	Site	Year	Residual	SiteRes	Local	SRS
PIECESLWD	1-NC	329	-2.49	4784.57	411.56	6831.30	11615.87	7136.24	11670.93
	2-MC	308	5.40 **	3611.16	-25.46	1730.48	5341.63	2973.51	5449.41
	3-MS	294	2.10	4706.21	486.43	1157.37	5863.58	3142.27	6084.12
	4-UMP	299	0.22	2050.46	29.35	925.02	2975.47	1616.90	2989.08
	5-SC	321	-2.09	2355.42	64.16	849.33	3204.75	1255.90	3248.62
	ALL	1551	0.30	4123.98	127.89	2376.13	6500.11	3283.84	6544.56
VOLUMELWD	1-NC	329	-16.20	18521.70	3960.53	25540.38	44062.08	29292.42	45627.16
	2-MC	308	8.97 **	9864.01	36.53	14922.26	24786.27	13582.28	25614.56
	3-MS	294	-10.36 *	16930.80	416.43	8838.14	25768.94	12100.35	26178.47
	4-UMP	299	-1.01	12771.11	946.96	4198.87	16969.98	9364.76	17099.72
	5-SC	321	-3.61	8969.30	125.28	4517.96	13487.26	6641.18	13668.51
	ALL	1551	-4.38 *	14065.13	20.80	12881.04	26946.17	14288.96	27066.81
LWDPIECE1	1-NC	329	-1.23	77.73	13.54	149.59	227.32	138.47	234.91
	2-MC	308	0.21 *	81.11	-0.43	14.43	95.54	52.19	96.13
	3-MS	294	-0.15	178.85	9.30	21.48	200.34	113.16	200.97
	4-UMP	299	-0.21	76.14	0.95	13.96	90.10	40.00	89.75
	5-SC	321	-0.43	63.51	2.55	21.93	85.44	33.58	86.88
	ALL	1551	-0.37	99.54	2.47	48.12	147.65	75.46	148.72
LWDVOL1	1-NC	329	-3.12	596.51	84.91	481.12	1077.63	729.76	1131.97
	2-MC	308	0.62	281.65	8.76	183.23	464.88	253.95	473.98
	3-MS	294	-2.66 **	846.00	2.05	387.60	1233.60	693.72	1255.59
	4-UMP	299	-0.47	373.06	33.96	115.62	488.68	247.65	498.37
	5-SC	321	-0.72	295.69	6.50	202.37	498.06	249.17	504.80
	ALL	1551	-1.25 **	469.05	0.78	307.02	776.07	433.50	785.12
KEYLWD	1-NC	329	-0.91	81.67	12.24	81.97	163.65	123.57	170.36
	2-MC	308	0.39	33.98	2.50	29.02	63.01	36.61	65.80
	3-MS	294	-0.67 *	45.97	1.71	40.32	86.29	42.47	87.42
	4-UMP	299	0.07	34.08	4.84	18.22	52.29	29.80	54.74
	5-SC	321	-0.15	28.21	0.41	12.23	40.44	22.50	40.82
	ALL	1551	-0.27 *	44.82	0.15	40.62	85.44	51.62	86.20
KEYLWD1	1-NC	329	-0.16	1.89	0.26	1.46	3.34	2.47	3.55
	2-MC	308	0.02	0.63	0.05	0.46	1.09	0.63	1.12
	3-MS	294	-0.12 **	0.86	0.02	1.52	2.38	1.27	2.41
	4-UMP	299	0.00	1.05	0.15	0.54	1.59	0.86	1.68
	5-SC	321	-0.04	1.03	0.01	0.31	1.34	0.77	1.36
	ALL	1551	-0.06	1.07	0.00	0.95	2.02	1.21	2.05
WOODJAMS	1-NC	291	-0.53 **	9.12	-0.25	9.64	18.75	12.56	20.02
	2-MC	261	-0.14	13.39	-0.08	5.68	19.07	10.29	19.41
	3-MS	249	0.15	12.41	0.57	8.14	20.56	12.51	21.58
	4-UMP	262	0.01	5.21	0.18	4.04	9.25	5.61	9.34
	5-SC	274	0.00	5.89	0.18	4.82	10.71	4.78	10.65
	ALL	1337	-0.09 **	10.34	-0.03	6.51	16.85	9.22	16.99
JAMS_KM	1-NC	291	-0.62 **	22.50	-0.78	20.60	43.09	28.71	45.37
	2-MC	261	-0.35 **	29.49	-0.32	8.69	38.18	22.75	39.27
	3-MS	249	0.26	35.78	1.28	22.20	57.98	35.19	58.99
	4-UMP	262	-0.03	17.68	1.19	10.75	28.43	16.54	29.08
	5-SC	274	0.05	16.92	0.33	13.98	30.91	13.17	30.56
	ALL	1337	-0.36 *	25.86	-0.12	15.62	41.48	23.31	41.62

** Indicates significance of slope estimate at the 0.01 level
* Indicates significance of slope estimate at the 0.05 level

PIECESLWD – Pieces of large woody debris (=0.15m diameter and =3m long) in reach
VOLUMELWD – Volume of large woody debris (m^3) in reach
LWDPIECE1 – Pieces of large woody debris per 100 meters of primary stream length
LWDVOL1 – Volume of large woody debris per 100 meters of primary stream length
KEYLWD – Key pieces of large woody debris (=0.60m diameter and =12m long) in reach
KEYLWD1 – Key pieces of large woody debris per 100 meters of primary stream length
WOODJAMS – Number of occurrences of debris jams
JAMS_KM – Number of occurrences of debris jams per kilometer

B: SAS Program

```
/********************************************************************/
/*                   7-16-06                                        */
/*                                                                  */
/*  N.S. Urquhart                                                   */
/*  S.J. Williams                                                   */
/*                                                                  */
/*  Use the program to run the mixed linear model regression and    */
/*obtaining estimates of components of variance. Output is tables   */
/*consisting of these estimates as well as other data parameters of */
/*interest.                                                         */
/*                                                                  */
/********************************************************************/


LIBNAME ODFW2005 'D:\SAS\ODFW2005';
      RUN;

PROC IMPORT OUT= ODFW2005.DATA3
            FILE= "D:\Documents\STARMAP\Students\Sarah\R\AIP Reach DATA
1998-2005.xls"
            DBMS=EXCEL REPLACE;
     SHEET="AIP_Reach_DATA_1998_2005";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

PROC sort DATA=ODFW2005.DATA3 out=ODFW2005.DATA3;
      by GCG; RUN;

DATA ODFW2005.DATA3;
      set ODFW2005.DATA3;
      if GCG in('1-NC','2-MC','3-MS','5-SC','4-UMP');
      TIME = YEAR - 1997;
RUN;

DATA ODFW2005.DATA3;
      ATTRIB GCG LENGTH=$5 FORMAT=$char5.;
      Set ODFW2005.DATA3;
      RUN;

PROC contents DATA=ODFW2005.DATA3;
RUN;
```

```
/*
  Create a data set that contains each of the variables as an entry.
  For use in macro
*/

DATA ODFW2005.dependent_vars;
 input index varnames $16.;
 cards;
1 WATERSHED
2 PRICHNLL
3 SECCHNLL
4 PRICHNAREA
5 SECCHNAREA
6 GRADIENT
7 VWIRCH
8 WIDTH
9 ACW
10 NOUNITS
11 NOPOOLS
12 PCTPOOLS
13 PCTSCPOOL
14 PCTSWPOOL
15 SCRPOOLD
16 RESIDPD
17 RIFFLEDEP
18 LRGBLDR
19 PCTSNDOR
20 PCTGRAVEL
21 PCTBEDROCK
22 POOLS_KMTL
23 POOLS_KMPL
24 POOL1P_KM
25 PCTSHADE
26 PCTEROSION
27 PCTUNDERC
28 PIECESLWD
29 VOLUMELWD
30 LWDPIECE1
31 LWDVOL1
32 KEYLWD
33 KEYLWD1
34 WOODJAMS
35 JAMS_KM
;
RUN;


%macro odfw;
  %do i = 1 %to 35;

            *brings in each of the 35 variables, one at a time;

  DATA _null_;
  set ODFW2005.dependent_vars;
  if index = &i then
    call symput("curr_var",varnames);
  RUN;
```

```
            *the next two mixed procedures are identical in the ;
            *analysis. The first calls the 5 separate MA regions;
            *by the line "BY GCG";

  TITLE "The current variable is &curr_var";
  PROC mixed DATA=ODFW2005.DATA3 method=type3 noclprint;
        class ID_NUM YEAR;
      model &curr_var = TIME/cl covb s;
      random ID_NUM YEAR;
        BY GCG;
        ODS OUTPUT covparms=ODFW2005.CP SolutionF=ODFW2005.Slope
NObs=ODFW2005.NObs1;
    RUN;

  PROC mixed DATA=ODFW2005.DATA3 method=type3 noclprint;
        CLASS ID_NUM YEAR;
        MODEL &curr_var = TIME/SOLUTION;
        RANDOM ID_NUM YEAR;
        ods output covparms=ODFW2005.CP2 SolutionF=ODFW2005.Slope2
NObs=ODFW2005.NObs2;
    RUN;

            *Places the label 'all' in the region column;

  DATA ODFW2005.CP2;
      GCG="ALL";
      SET ODFW2005.CP2;
      RUN;

            *merging the parameter estimates into one temporary data
set;

  DATA ODFW2005.VarComps_temp;
      Variable="&curr_var";
      set ODFW2005.CP ODFW2005.CP2;
      RUN;

  %If &i=1 %then %Do;
      DATA ODFW2005.VarComps;
            Set ODFW2005.VarComps_temp;
            %End;
  %Else %Do;
  DATA ODFW2005.VarComps;
      set ODFW2005.VarComps ODFW2005.VarComps_temp;
      %End; RUN;

  DATA ODFW2005.Slope2;
      GCG="ALL";
      SET ODFW2005.Slope2;
      RUN;

            *merging the slope estimates into one temporary data set;
```

```
  DATA ODFW2005.Slopes_temp;
      Variable="&curr_var";
      set ODFW2005.Slope ODFW2005.Slope2;
      RUN;


  %If &i=1 %then %Do;
      DATA ODFW2005.SlopesOut;
            Set ODFW2005.Slopes_temp;
            %End;
  %Else %Do;
  DATA ODFW2005.SlopesOut;
      set ODFW2005.SlopesOut ODFW2005.Slopes_temp;
      %End; RUN;

  DATA ODFW2005.NObs2;
      GCG="ALL";
      SET ODFW2005.NOBS2;
      RUN;

  DATA ODFW2005.NObs_temp;
      Variable="&curr_var";
      set ODFW2005.NObs1 ODFW2005.NObs2;
      RUN;

  %If &i=1 %then %Do;
      DATA ODFW2005.NObsOut;
            Set ODFW2005.NObs_temp;
            %End;
  %Else %Do;
  DATA ODFW2005.NObsOut;
      set ODFW2005.NObsOut ODFW2005.NObs_temp;
      %End; RUN;

  %end;
%mend odfw;
RUN;

%odfw; RUN;

      *end of macro;


/*
 mean and variance of full DATA;
*/

PROC means DATA=ODFW2005.DATA3 var;
TITLE "OVERALL SUMMARY";
var WATERSHED PRICHNLL SECCHNLL PRICHNAREA SECCHNAREA GRADIENT VWIRCH
      WIDTH ACW NOUNITS NOPOOLS PCTPOOLS PCTSCPOOL PCTSWPOOL SCRPOOLD
      RESIDPD RIFFLEDEP LRGBLDR PCTSNDOR PCTGRAVEL PCTBEDROCK
      POOLS_KMTL POOLS_KMPL POOL1P_KM PCTSHADE PCTEROSION PCTUNDERC
      PIECESLWD VOLUMELWD LWDPIECE1 LWDVOL1 KEYLWD KEYLWD1 WOODJAMS
      JAMS_KM;
RUN;
```

```
/*
 mean and variance by region;
*/

PROC means DATA=ODFW2005.DATA3 var;
TITLE "REGIONAL SUMMARIES";
var WATERSHED PRICHNLL SECCHNLL PRICHNAREA SECCHNAREA GRADIENT VWIRCH
      WIDTH ACW NOUNITS NOPOOLS PCTPOOLS PCTSCPOOL PCTSWPOOL SCRPOOLD
      RESIDPD RIFFLEDEP LRGBLDR PCTSNDOR PCTGRAVEL PCTBEDROCK
      POOLS_KMTL POOLS_KMPL POOL1P_KM PCTSHADE PCTEROSION PCTUNDERC
      PIECESLWD VOLUMELWD LWDPIECE1 LWDVOL1 KEYLWD KEYLWD1 WOODJAMS
      JAMS_KM;
by GCG;
RUN;

PROC print DATA=ODFW2005.varcomps;
      TITLE "ESTIMATES OF VARIANCE COMPONENTS";
      RUN;

PROC print DATA=ODFW2005.SlopesOut;
      TITLE "ESTIMATES OF SLOPES AND ALLIED QUANTITIES";
      RUN;

/*
 The following four data steps extract only the desired information
 from the macro data sets and places them into one final data set
 in table 4.
*/

DATA ODFW2005.Table1;
      SET ODFW2005.Varcomps;
      IF CovParm="ID_NUM" THEN DO;
            Site=Estimate; RETAIN Site;
      END;
      ELSE IF CovParm="YEAR" THEN DO;
            Year=Estimate; RETAIN Year;
      END;
      ELSE IF CovParm="Residual" THEN DO;
            Residual=Estimate; SiteRes=Site+Residual;
            OUTPUT;
            End;
      Drop CovParm Estimate;
      RUN;
```

```
DATA ODFW2005.Table2;
      SET ODFW2005.SlopesOut;
    IF Effect="TIME" THEN DO;
       Slope=Estimate; KEEP Variable GCG Slope;
         OUTPUT;
    END;
    ELSE PUT;
    RUN;


DATA ODFW2005.Table3;
      SET ODFW2005.NObsOut;
       IF Label="Number of Observations Read" THEN DO;
       KEEP Variable GCG NObsRead NObsUsed;
         OUTPUT;
    END;
    ELSE PUT;
    RUN;

DATA ODFW2005.Table4;
      MERGE ODFW2005.Table3 ODFW2005.Table2 ODFW2005.Table1;
    RUN;
```

C: Schematic of ODFW panel revisit plan