# Comparison of the statistical power of BACI and hierarchical-staircase experimental designs to detect changes in fish abundance and habitat change within the framework of the Asotin Creek Intensively Monitored Watershed Project:

# Progress Report

# September 30, 2010

# Prepared by:

Thomas M. Loughin

Statistics and Actuarial Science

Simon Fraser University,

Surrey, British Columbia

[www.stat.sfu.ca/~tloughin/STATPAGE.html](http://www.stat.sfu.ca/%7Etloughin/STATPAGE.html)

# Prepared for:

Stephen Bennett

Eco Logical Research Inc.

456 South 100 West,

Providence, Utah

# Introduction

## Description of analysis model

The proposed design currently being considered by Eco Logical Research Inc. contains the following features:

The study area consists of three **creeks**: Charley, North Fork, and South Fork.

It is proposed that Charley Creek be a treated stream, while North Fork and South Fork remain as untreated controls. Charley would receive restoration treatments as prescribed later.

A 12-km stretch of each creek is divided into three 4-km **sections**.

Within each Charley Creek section, there are two **reaches**. Within the sections of the other two creeks there will be one reach. The reaches represent sampling locations.

The monitoring of the watershed is scheduled to run for 12 **years**, 2008-2019.

Within each year there is will be two sampling times corresponding to spring and fall **seasons**.

The long-term time-associated effects of the restorations will be measured in terms of any changes associated with Years After Restoration (**YAR**). For example, the 2011 measurements on Section 1 of Charley will be considered as comparable to the 2014 measurements on Section 2 and the 2017 measurements on Section 3 of Charley. All measurements prior to treatment in Charley and all measurements in N. and S. Forks are considered to have YAR=0. The staggering of the restoration treatments applied to Charley Creek allow better separation of the effects due to restoration from the effects due to natural annual variations inherent in the ecosystem: this creates a sort of replication of the restoration process that would not be available if all treatments were applied in the same year.

Figure 1 shows a schematic of the design proposed for the experiment.

Based on these features, the experiment contains *factors*: Creek, Section within Creek (written “Section(Creek)”), Reach within Section of Creek (written “Reach(Section\*Creek)”), Year, Season, and YAR. Among these, the factors Creek, Season, and YAR are considered as *fixed effects* factors. This means that we restrict attention only to those levels actually present on the experiment and do not infer to other creeks, seasons, or years after restoration. The factors Year, Section(Creek) and Reach(Section\*Creek) are considered as *random effects* factors. This means that we consider these years to be representative of years to which the results may apply; they are a sample of possible years in which the experiment could have been run and are used to quantify the variability that might be seen across years in the future. Similarly, the sections and reaches actually used in the study are used to represent the entire creeks and to allow us to quantify the variability that is seen spatially within a stream.

Figure 1: Schematic of Asotin Restoration Experiment



The model that has been developed for all of the response measurements is based on experimental design principles. We identified the experimental units to which each factor or interaction of factors was assigned or observed. These are determined from the rectangles of various sizes and shapes that represent each factor in Figure 1. Then a model was derived containing terms corresponding to each different size of experimental unit. Fixed and random effects were identified as above, and also using the convention that interactions involving random effects are also random.

Not all combinations of YAR and Creek are present in the experiment; indeed, these two factors are highly unbalanced, because the levels of YAR are only ever 0 in the two forks of Asotin. Main effects of YAR and Creek can therefore not be estimated separately without interference from each other’s effects. Instead, the effects of YAR must be estimated using contrasts within the context of the YAR\*Creek interaction. This puts restrictions on the terms that can be included in the model.

The resulting model can be expressed as an ANOVA table or as an equation. Both are shown in the table below.

Model ANOVA Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | DF | Fixed or Random | Symbol | Subscript |
| Year | 11 | Random | q | h |
| Season | 1 | Fixed | α | i |
| Year\*Season | 11 | Random | (qα) | hi |
| Creek | 2 | Fixed | β | j |
| Year\*Creek | 22 | Random | (qβ) | hj |
| Season\*Creek | 2 | Fixed | (αβ) | ij |
| Year\*Season\*Creek | 22 | Random | (qαβ) | hij |
| YAR\*Creek | 9 | Fixed | (βτ) | jk |
| Section(Creek) | 6 | Random | s | jl |
| Year\*YAR\*Section(Creek) | 57 | Random | (qτs) | hjkl |
| YAR\*Season\*Creek | 9 | Fixed | (αβτ) | ijk |
| Year\*YAR\*Season\*Section(Creek) | 63 | Random | (qατs) | hijkl |
| Reach(Section\*Creek) | 3 | Random | r | jlm |
| Year\*YAR\*Reach(Section\*Creek) | 33 | Random | (qτr) | hkjlm |
| Residual Error | 36 | Random | e | hijklm |

The statistical research plan is to investigate several things in the context of the Asotin IMW experiment: **Updated April 29, 2010**

1. Relating to this model
   1. Investigate relationships between this model and the standard BACI-type analyses that one might use on an experiment like this.
   2. Demonstrate strengths and weaknesses of this model and the BACI-type analyses.
   3. Use simulations to assess the power of this model to identify interesting effects, primarily contained within the YAR\*CREEK effect.
2. Consider alternative experimental designs
   1. Develop alternative designs that could conceivably be used for this experiment
   2. Develop analyses models for these designs
   3. Assess the power of these design/analysis combinations under the same circumstances as used in simulations of the current design

At present, literature review has largely been completed for (1a), leading to the conclusion that a direct BACI-type analysis cannot be conducted on an experiment of this complexity. Toward (1b), we have shown that there exists a mixed model that can perform the equivalent analysis as a BACI-type analysis, and further that it can actually improve on the power of the BACI analysis. We therefore are comfortable with moving forward using mixed models for all analyses of data from this experiment. Steps (2a) and (2b) are also finished. Simulations will be conducted on the combined goals (1c) and (2c). A SAS program has been written that can be used as a template for simulations. Inputs required for the program are being determined using historical data as well as data collected from the initial stages of this study. Details of some of this work follow.

## Investigation of Mixed Model vs. BACI analysis in simplified version of problem

To better understand the analyses recommended for BACI-type studies, I first used computer-generated data to perform some simulation comparisons between the analysis recommended by Stewart-Oaten and Bence (2001) and a mixed-model analysis. To avoid confounding factors that might cloud the comparisons, I used a simplified version of the experimental design that we are considering, consisting of 3 creeks and a 20-year monitoring period, with a single intervention taking place on one creek after year 10. I generated data based on a model that is appropriate for both analysis types, so that neither would have an inherent advantage. The data generation model included factors year, creek, and YAR and certain interactions among them. Year and year\*creek were taken as random effects along with a random error term, while creek, YAR and creek\*YAR were fixed. The variance of the error term was held fixed at 1 and all other factors’ effects allowed to vary in relation to this. I initially considered year variance components of between 25 to 625 times the error variance and year\*creek variance components of between 4 to 100 times the error variance, based on experience in other types of studies that suggest that year-to-year variation can dominate all other forms of variance and expecting a potentially intermediate amount of variability in creek effects among years. All error terms were considered independent, ignoring the potential for serial correlation. I considered a simple model for the intervention response in which all years following intervention have a constant intervention effect. This is the easiest model to assess as it allows all comparisons to be based on before- and after- intervention means.

I performed the BACI analysis as per Stewart-Oaten and Bence (2001) as follows: First, the two control creeks’ responses were averaged. Next, yearly differences between the treated creek and the control creek were computed, creating 20 annual differences. Finally, a pooled-variance t-test was performed comparing the average annual difference after intervention to that before intervention. The p-value for the test was recorded.

The mixed-model analysis was performed separately on the original responses (not on annual differences) according to the model consisting of fixed effects for Creek and YAR\*Creek and with year as a random effect. A contrast was computed to compare the mean response before and after intervention. The p-value testing the significance of this contrast was recorded.

Data were generated for various combinations of year, creek, year\*creek, YAR, and YAR\*creek effects. For each combination, both analyses were performed on each of 10,000 sets of randomly-generated data. Then the proportion of significant tests at α=.05 was computed for each type of analysis, estimating the type I error rate or power of the tests depending on whether or not a YAR effect was included in the simulations. Plotting these proportions against the size of the YAR effect estimates the statistical power curve for each test.

All computations were done using SAS Version 9.2. The mixed-model analysis was performed using PROC MIXED, while the BACI used customized programming.

# RESULTS

The BACI and mixed model analyses are very similar with respect to their type I error rates and their power (Figure 2). Neither is affected by increasing the Year variance component, as both methods explicitly remove variability due to years from their respective analyses (BACI through taking annual differences and mixed model through the random Year effect). Both analyses treat added year\*creek variability as if it were added error variability, which is appropriate in this design because there it contains no means by which to separate these two sources of error (they are both measured on the same experimental units). Hence, the simulations are presented with respect to the ratio of the size of the YAR effect to the “total error”, defined as the square root of the sum of the year\*creek and error variances.

Both analyses maintained their type I error rates to within simulation error of the nominal level in each of four separate sets of 10,000 data sets that were generated with no YAR effect (but various combinations of year and year\*creek variance components). The estimated power curves for the two procedures are very close, with the mixed model analysis maintaining a very slight advantage in all settings tested. Finally, the mixed model analysis was quite easy to perform in SAS Version 9.2, requiring only 5 lines of required PROC MIXED code.

Figure 2: estimated power curves for mixed model and BACI analyses.

It is reasonable to conclude based on this simulation that in a situation with one intervention and two controls and 10 years of data collection before and after intervention, there is no particular advantage to BACI analysis over a mixed model analysis. Having observed the consistency with which this conclusion holds over the range of observed YAR effect sizes, I conclude that it suffices to limit YAR effects for future comparisons in this vein to two values: 0 (to establish that type I error rate is maintained) and a moderate effect size such as 1.0 times the total error.

**Feb, 2010**

## Follow-Up: Mixed Model that duplicates BACI analysis

Upon further investigation of the two analyses, I have identified exactly what differentiates the BACI analysis from the mixed model analysis that was performed above, and can now duplicate the BACI *exactly* using a mixed model. The mixed model analysis described above can be written alternatively in terms of different factors that are more natural to the BACI approach:

|  |  |  |
| --- | --- | --- |
| **Factor** | **Label** | **Meaning** |
| Treatment | TRT | Distinguishes whether creek is treated or control |
| Creek within treatment | Creek(TRT) | Distinguishes the two control creeks |
| Period | Period | Before vs. after intervention |
| Year within period | Year(Period) | Unites measurements taken in the same year in a given period |

Then the possible terms in an ANOVA can be written using this factorization as

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **DF** | **Fixed or Random** | **Symbol** | **Subscript** |
| TRT | 1 | Fixed | τ | i |
| Creek(TRT) | 1 | Fixed | γ | ij |
| Period | 1 | Fixed | π | k |
| Year(Period) | 18 | Random | r | kl |
| TRT\*Period | 1 | Fixed | (τπ) | ik |
| Period\*Creek(TRT) | 1 | Fixed | (γπ) | ijk |
| TRT\*Year(Period) | 18 | Random | q | ikl |
| Residual Error | 18 | Random | e | ijkl |

As other authors have noted, the term that represents the hypothesis of interest is the highlighted TRT\*Period effect: it measures whether the difference between the two periods is different in the treated creek vs. the control creeks. The ANOVA table that I used in the analyses in the previous section recognized that the last two random terms in the model —TRT\*Year(Period) and Residual Error — are measured on the same experimental unit and are often assumed to be measuring the same variation. In the analyses in the section above, I pooled these two terms into one term with 36 df, which served as the error term for the TRT\*Period tests. The BACI analysis uses just the TRT\*Year(Period) as the error term.

Which is “correct”? Pooling these two terms is justified by the following argument. The former term measures how variable the annual differences are between the treated creek and the average of the control creek, after adjusting for any possible treatment effect. The latter measures how variable the annual differences are between the control creeks, after adjusting for mean differences between the creeks. After adjusting for effects of treatment and creek on the means, then, the three creeks should not be different from one another in any meaningful way, so that differences between any pair of creeks should be indistinguishable from one another. If that is so, then these two terms in the ANOVA table are measuring exactly the same variability and should be pooled to give a more powerful test. This explains the slight power advantage enjoyed by the mixed model analysis.

It is conceivable that these terms are measuring different types of variability. These two terms may differ if the treatment has an effect on the *variability* of the responses, which may be conceivable in applications like ours where the overall habitat improvement may conceivably lessen the effects of annual environmental variations or might drastically increase output in very good years. Then these two terms *might* be measuring something different.

It should be noted, however, that if this is the case, then neither of these analyses is correct. The TRT\*Period(Year) effect is then a combination of two variances: the variability of annual differences before treatment and the variability of annual differences after treatment. A proper analysis should explicitly identify and model the different variance in the treated creek after treatment using a heterogeneous-variance model. This extension is not difficult in a mixed-midel setting.

**UPDATE April 29, 2010**

## Extensions of Comparison of Mixed Model and BACI Analysis to Other Settings for Same Design

Comparing mixed models to BACI analysis in other experimental designs presumes that a BACI-type analysis can be developed for these experiments. However, in a staggered-start (staircase) design like ours, it becomes difficult to identify exactly what differences should be computed at different times. Specifically, the BACI approach requires that differences be computed between the treated units and the control units both before and after treatment is applied. But in our setting the treated unit is split into sub-units and these sub-units are treated at different times. When the first sub-unit is treated, the other two sub-units in the treated unit are still controls, so it is inappropriate to include them with the treated sub-unit. But they will eventually be treated, so it is inappropriate to include them with the control units. If instead they are omitted until they are treated, then the omission biases the initial pre-treatment comparisons between treated and control units.

The only reasonable way out of this quandary that I can see is to perform a separate BACI analysis for each treated subunit, using the same non-treated units as controls in each case. That is, for each intervention time, compare only that treated sub-unit to the untreated units, ignoring the other sub-units that have been or will eventually be treated. This is obviously an awkward and inefficient thing to do as it (a) fails to use relevant treatment and control information from other treated units in each analysis, and (b) provides three separate, inter-dependent analyses to try to answer one question.

I therefore do not consider BACI to be a viable analysis approach for experiments using a staircase design and will no longer consider it in comparative simulations.

## Simplification of the main model to single-season data

Initial tests of the model derived in the first section ran into computational difficulties associated with the complexity of the model. This is probably the main drawback of using mixed models: the methods can sometimes be computationally sensitive and fail to produce answers. It is not always clear why this happens, but sometimes simplifying the model can lead to better results. In trial runs using simulated data, it was found that reducing the data to analyze one season at a time resulted in much more stable computational performance. Following discussions with Eco Logical Research Inc., we determined that interest in formal comparisons of seasonal effects and interactions was minimal, and that comparing “comparable” numbers from the same season across years was of much greater importance. We therefore decided to run separate analyses of Fall and Spring data. This results in the simplified model given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | DF | Fixed or Random | Symbol | Subscript |
| Year | 11 | Random | q | h |
| Creek | 2 | Fixed | β | i |
| Year\*Creek | 22 | Random | (qβ) | hi |
| YAR\*Creek | 9 | Fixed | (βτ) | ij |
| Section(Creek) | 6 | Random | s | ik |
| Year\*YAR\*Section(Creek) | 57 | Random | (qτs) | hijk |
| Reach(Section\*Creek) | 3 | Random | r | ikl |
| Residual Error | 33 | Random | e | hijkl |
| TOTAL | 143 |  |  |  |

The mathematical model and definitions of effects are:

yhijkl = qh+βi+(qβ)hi+(βτ)ij+sik+(qτs)hijk+rikl+ehijkl

yhijkl = Response in year h, creek i YAR j section k reach l

All Random effects are assumed to have Normal distributions with mean 0 and different variance components (one for Year, one for Year\*Creek, one for Section(Creek), one for Year\*YAR\*Section(Creek), one for Reach(Section\*Creek) and one for Error)

Year effect accounts for variability from year to year that simultaneously affects all creeks

Creek effect accounts for fixed differences among creeks

Year\*Creek effect accounts for different effects of years on different creeks

YAR\*Creek effect contains differences among the Years after Restoration, as well as interaction effects between YARs and Creeks

Section(Creek) accounts for variability among sections within a creek

Year\*YAR\*Section(Creek) measures variability of differences among YARs and Sections across years. This term contributes to the error term for treatment comparisons, along with Residual Error.

Reach(Section\*Creek) measures variability of sampling sites within a section

Error measures variability over time of response measurements made on the same sampling site.

## Alternative design

The current plan for the Asotin IWM design is to treat three sections of Charley Creek in a staggered fashion, leaving the North Fork and South Fork of the Asotin as control creeks. An alternative design that could have better statistical properties would be to treat one section of each creek in the same staggered fashion. I outline potential advantages and disadvantages below.

The original plan for treating only Charley was based on the concern that the measurable effect of the interventions might be small relative to variability, and on the belief that the continued improvements applied to a single stream could eventually synergize to create larger and clearer effects. As I understand it, the synergism argument would seem more likely to hold for fish measurements than for habitat measurements, because the latter are expected to be fairly localized relative to the treatments.

This design would also leave two completely untouched streams for control comparisons, assuming that the spatial correlations do not allow effects from one altered stream to carry over into another.

A potential disadvantage of this design is that it confounds effects of treatments with random year-by-stream interactions. It is expected that the staggering of treatment applications will reduce this confounding considerably. However, if the differences between measurements made on different streams vary widely across years, I would expect this design’s power to suffer, meaning that even large-ish effects might be difficult to identify.

An alternative design would be to treat one section of each stream, maintaining the staggering among the three treatment times. An example of this is depicted below. (Note that the picture assumes two sampling reaches in each treated section. This is not how the additional sampling reaches are currently arranged, but it is also not a central requirement of the design.)

Figure 3: Schematic of an Alternative Experimental Design



The main advantage of this design comes if the sections within a stream are much more similar to one another than sections in different streams, and if differences among sections vary much less across years than do differences between streams. If these two conditions hold, then I would expect that this design would have much more power than the first, analogous to the gains experienced by blocked experimental designs. Here, the stream acts like a block effect, with the within-block variability (section-to-section) being much less than the between-block variability (stream-to-stream).

Furthermore, it appears to me that the three sections within a stream have certain properties that might carry over from stream to stream. Each stream has a “top” section, a “middle” section, and a “bottom” section, and it is conceivable (to me) that there are some consistent properties among tops, among middles, and among bottoms. If this is the case, then we can use section position as a second block factor and run the experiment in a Latin Square design, simultaneously blocking on stream and section position. This could conceivably reduce the variability of treatment vs. control comparisons further, enabling detection of smaller effects than in the first design.

What remains to be determined is critical:

(1) Are the assumptions that favor each design valid? If so, to what degree? (I don’t know how we will test some of the structural assumptions such as synergism. It would seem to me that only past experience will tell us this.)

(2) What sorts of gains/losses are possible with each design, in terms of power comparisons? This will be one of the main goals of the simulations.

(3) Is the alternative design even logistically feasible? I certainly can’t answer that, but if it can’t be done, then power comparisons are moot.

## Simulation Planning

I am currently working on developing the simulation model and estimating some of the parameters that need to be set in the model. The simulation model will look much like the analysis model in the sense that I will simulate measurements from normal distributions (alternatives are possible) with variability being added in from a number of sources: year, stream, year\*stream, section, year\*section, and others. I will also consider the addition of temporal correlation (years close together tend to have more similar measurements than years far apart) and spatial correlation (sampling sites close together tend to have more similar measurements than sampling sites far apart). I will check in with you when a proposed simulation model has been worked out, to confirm whether it all makes sense to you.

The sizes of effects that are introduced into the simulated data need to be determined. This will be done, to the extent that it is possible, using estimates derived from historical data and preliminary measurements. I will analyze the data sets that Eco Logical Research Inc. has sent me and determine realistic values for the needed variance components and correlations. We are focusing on two measurements that are of primary interest and may have some different properties: Fish Abundance (count per 100 m2) and Pool count. Pool counts may exhibit more spatial independence due to dependence on purely local features, but may show temporal correlation as alterations in stream structure may be slow to occur but lasting when they do. On the other hand, Abundance may have little temporal correlation as it will be affected greatly by variable facets of the environment that are independent of the Asotin watershed (e.g. marine survival, conditions during the run, and so forth). Abundance may exhibit some spatial correlation, however, if there is any “herd mentality” among fish, or if their selection of a spawning site can be affected by conditions near, but not on, the actual site.

Given the limitations of the preliminary and historical data relative to the extent of sampling over time and space that will take place, there will undoubtedly be some quantities that cannot be estimated and that will have to be filled in by guessing.

## Analysis of Preliminary and Historical Data Updated May 5 2010

Eco Logical Research Inc. sent me 5 data sets to work with: two based on 2 years of data collection (2008, 2009) within the current study (one for pools and one for abundance); two based on historical data reaching back to the 1980’s (one for redds and one for abundance), and one that summarizes some estimated from current and past observation. The data from 2008-9 is useful for estimating certain spatial trends, as the sampling is done under the current Asotin IMW plan, but the fact that there are only two years limits its utility for estimating year-related variance components (estimates would have very high uncertainty). The historical data serves the opposite purpose: it can be used for year and year\*stream effects, but cannot help at all with spatial effects on a scale finer than a stream (redds) or stream section (abundance).

I have been working with log(abundance), because the variance of abundance tends to grow as abundance measurements grow, in violation of assumptions of the mixed models. The log transformation seems to alleviate this problem. For the same reason I have been working with the square root of pool frequency. I could not assess the variance-mean relationship for redds, but I took a square root for the redd count on the principal that the variances of counts usually grow as the means grow, and square roots are often appropriate to stabilize these variances.

### Temporal Analysis of Abundance

The first data set is on yearly juvenile abundance from the Washington Department of Fish and Wildlife. It contains population estimates for age 0 and age 1+ fish from 1983-2006 from 4 streams (Charley, North Fork, South Fork, and Mainstem), three of which were measured in two sections (all but Charley).

I fit models assuming that Stream is a fixed effect, as we will do in our analyses, and that Year and Section(Stream) are random effects. The Year\*Stream interaction is also random. There is potential for serial correlation in two different ways. First, the Year effects that impact all streams simultaneously (we can call these “broad” Year effects) may be correlated. The factors that increase or decrease fish counts from year to year may linger and continue to have residual effects in later years. I assume an autoregressive correlation process with a 1-year lag (AR-1), which says that correlation between years decreases as a power of the distance between years (if ρ is the correlation of abundances in consecutive years, then ρ|i-j| is the correlation between year i and year j. This form of correlation means that even measurements taken of different streams are correlated with each other. Second, the unique features in each stream section may result in annual measurements that are temporally correlated within the stream, but not correlated to measurements taken on other streams. These effects can be called “local” year effects.

I fit models consisting of different combinations of random effects and correlations to the available data on log(abundance). The models were chosen to contain as many variance components as possible that are needed in the simulations. Correlation was optional. Adding Section(Stream) explicitly as a random effect allows Year\*Section(Stream) to be estimated as a separate variance component in the Residual Error. However, it was found that this was incompatible with estimating local correlations, so only one of these features could be used at a time. I used the corrected version of Akaike’s Information Criterion (AICc) and Akaike Weights (Burnham and Anderson 2002) to help select the model that seemed to fit the data best. The results are below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Random Effects | Correlations | AICc | Weight |
| 1 | Year, Year\*Stream | Broad, Local | -12.4 | 9 |
| 2 | Year, Year\*Stream | Local | -14.6 | 85 |
| 3 | Year, Year\*Stream | Broad | -9.4 | 1 |
| 4 | Year, Year\*Stream |  | -11.4 | 4 |
| 5 | Year, Year\*Stream, Section(Stream) | Broad | -8.0 | 0 |
| 6 | Year, Year\*Stream, Section(Stream) |  | -10.1 | 1 |

Considering the Akaike weights, Model 2 appears to be a fairly strongly favored model. This model states that there no broad serial correlation, but that there *is* local serial correlation. Estimates are given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Random Effects | | |  | Fixed Effects | | |
| Effect | Estimate | 95% Conf Int |  | Stream | Mean log(abun) | 95% Conf Int |
| Var(Year) | 0.195 | 0.009, 0.632 |  | Asotin | 1.18 | 1.04, 1.31 |
| Var(Year\*Stream) | 0.0063 | 0.0015,0.55 |  | Charley | 1.46 | 1.30, 1.61 |
| AR-1 Corr Local | 0.42 | 0.10,0.74 |  | NFork | 1.31 | 1.20, 1.42 |
| Var(Section (Stream))\* | 0.026 | 0.015,0.056 |  | SFork | 1.35 | 1.24, 1.47 |

\* Residual error from model

The random effects are interpreted as follows. Var(Year) measures the variability of annual mean log(abundance) measurements. Var(Year\*Stream) measures the variability of stream-to-stream differences in log(abundance) across years. AR-1Corr is the correlation between log(abundance) measurements made on the same stream section 1 year apart. Taking this number to the kth power gives the correlation of log(abundance) measurements taken on the same stream k years apart. Section (Stream) measures the variability of log(abundance) measurements taken the same year in different sections of the same stream.

The total variance of log(abundance) measurements is found from the sum of the variance components. Here, the sum is .195+.0063+.026 = .227. The average mean is 1.33. Therefore, the estimated CV for the log(abundance) measurements across years is 36%.

***Spatial Analysis of Abundance***

This analysis is conducted on data from the first two years of sampling from the Asotin IWM project. In 2008, one summer measurement was taken in each of three sections in Charley, North Fork, and South Fork. In 2009, summer and fall measurements were taken and in Charley measurements were taken at two reaches (habitat sites) per section in each season. These three sets are analyzed together as one set.

Model selection is not carried out, as only one model is considered. This model contained Stream as a fixed effect and Section(Stream) as a random effect. Further random effects were added to account for differences among the three year\*season data sets: Year\*Season, Year\*Season\*Stream, and Year\*Season\*Section(Stream). The residual error measures the FishSite(Section,Stream) variance component. The results are below (year-based variance components are omitted, as they are not of primary interest here and are not reliable, being based on only two years).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Random Effects | | |  | Fixed Effects | | |
| Effect | Estimate | 95% Conf Int |  | Stream | Mean log(abun) | 95% Conf Int |
| Var(Section(Stream)) | 0.0057 | 0.009, 0.632 |  | Charley | 1.34 | 1.10, 1.59 |
| Var(FishSite(Sectn\*Stream)) | 0.0027 | 0.0015,0.55 |  | NFork | 1.13 | 1.06, 1.30 |
|  |  |  |  | SFork | 1.30 | 1.30, 1.54 |

Here, Var(Section(Stream)) measures the variance of log(abundance) across sections of the same stream in the same year, while Var(FishSite(Sectn\*Stream)) is the variance of measurements from different fish sites within the same section and year. The total within-year variance is .0084, and the average mean is 1.26, so that the estimated within-year CV for log(abundance) is 7.3%.

### Spatial Analysis of Pool Frequency

The structure of this data set is similar to the one used for the spatial analysis of abundance. There are two key differences: (1) there is only one measurement time per year, and (2) in 2009 there was an added measurement taken at each reach/fish site. These are called “panel” in the data, and can also be called “habitat sites”.

The initial strategy was much like the one used for the spatial analysis of abundance, but this became untracked early on. It turns out that this data set contains a pair of pool frequency measurements from different habitat sites from the same fish site in the same year that are much more different from each other than would be expected based on all of the other such pairs of pool frequencies. This pair creates such a large estimate for the variability between habitat sites within a fish site that models do not recognize the possibility of *further* variability coming from any source. We therefore get no estimates of variance components associated with spatial features Section(Stream) and FishSite(Section,Stream). As it is believed that *some* variability among these features exists, an alternative plan had to be undertaken in order to estimate these corresponding variance components.

For this data set, therefore, three separate subsets of data were analyzed. One was from 2008, which yielded only an estimate of Section(Stream) variability. One was from 2009 the annual panels only. This eliminated the question of habitat site variability and enabled estimation of Section(Stream) and FishSite(Section,Stream) variance components. Finally, all data from 2009 was used in a third analysis, which resulted in only the HabSite(FishSite,Section,Stream) variance component being estimated. The results are below. In all cases, the response variable was square root of pool frequency/km

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Random Effects | | |  | Fixed Effects | | |
| Data | Quantity | Section | FishSite | HabSite |  | Charley | North Fork | South Fork |
| 2008 | Est | 2.03 | - | - |  | 6.1 | 4.5 | 6.2 |
|  | Conf Int | 0.84, 9.9 | - | - |  | 4.1, 8.1 | 2.5, 6.5 | 4.2, 8.2 |
| 2009 Annual | Est | 1.79 | 0.78 | - |  | 6.6 | 4.3 | 4.8 |
|  | Conf Int | 0.55, 29.5 | 0.25, 11.6 | - |  | 5.4, 7.7 | 2.7, 6.0 | 3.2, 6.5 |
| 2009 All | Est | - | - | 1.64 |  | 6.5 | 4.3 | 5.6 |
|  | Conf Int | - | - | 0.97, 3.35 |  | 4.4, 8.6 | 2.0, 6.6 | 3.3, 7.9 |

In 2008, the total variance was 2.03 and the average mean was 5.8, leading to an estimated CV for square-root pool frequency of 25%. In 2009 using the annual habitat sites, total variance was 2.57 and the average mean was 5.5, so that the estimated CV was 29%. Finally, using both habitat sites in 2009, total variance was 1.64, the average mean was 5.2, and so the estimated CV was 25%.

### Temporal Analysis of Redds

The redds data set contains annual red counts from 1986-2008 (with some missing values) for Charley, North Fork, South Fork, and Mainstream Asotin. There are no spatial features besides stream.

The analysis is similar to the temporal analysis of abundance and consists of models based on annual variability. Stream is fixed, Year is random, and both broad and local autocorrelations. The residual error measures the year\*stream variability. The response is square root of red count. Model fitting results are below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Random Effects | Correlations | AICc | Weight |
| 1 | Year | Broad, Local | 310.1 | 80 |
| 2 | Year | Local | 311.5 | 20 |
| 3 | Year | Broad | \* | - |
| 4 | Year |  | 317.5 | 0 |

\* Did not converge

Considering the Akaike weights, Model 1 appears to be a fairly strongly favored model. This model states that there is both broad and local serial correlation. Estimates are given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Random Effects | | |  | Fixed Effects | | |
| Effect | Estimate | 95% Conf Int |  | Stream | Mean sqrt(redds) | 95% Conf Int |
| Var(Year) | 6 | 1, 150000 |  | Asotin | 14.1 | 11.3, 17.1 |
| AR-1 Corr Broad | 0.78 | -0.13, 1.69 |  | Charley | 6.1 | 3.3, 9.0 |
| AR-1 Corr Local | 0.40 | 0.09,0.72 |  | NFork | 9.8 | 7.0, 12.6 |
| Var(Year\*Stream) | 3.1 | 2.0, 5.3 |  | SFork | 6.7 | 3.9, 9.6 |

\* Residual error from model

I am suspicious of this model. Given that a simpler model containing only the broad correlations could not be estimated, I expected this one to fail to converge as well. The same model run on the untransformed data failed to converge. Also, the confidence intervals for Year and for the broad correlation are useless. Analysis of simulated data shows me that this model actually is not fitting this data structure properly, so I will dismiss it. Here are the estimates from the second-best fitting model:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Random Effects | | |  | Fixed Effects | | |
| Effect | Estimate | 95% Conf Int |  | Stream | Mean sqrt(redds) | 95% Conf Int |
| Var(Year) | 2.8 | 1.4, 8.5 |  | Asotin | 13.9 | 12.3, 15.5 |
|  |  |  |  | Charley | 5.9 | 4.4, 7.4 |
| AR-1 Corr Local | 0.46 | 0.17,0.75 |  | NFork | 9.5 | 8.1, 10.9 |
| Var(Year\*Stream) | 3.3 | 2.1, 6.1 |  | SFork | 6.4 | 5.0, 7.8 |

I find these to be sensible estimates and will use these instead.

## SIMULATIONS OF PROPOSED DESIGN AND ALTERNATIVE Updated Aug 18, 2010

An initial, fairly extensive round of simulations is now complete. I have not investigated every possible variable: to do so would take prohibitive amounts of time. But I have the effects investigated many variables that I guessed might be important or interesting (with some help from you). I am quite open to exploring specific questions that might arise, but will not have the huge amounts of time again that I spent on this initial study.

### Simulation Study description

The simulation study was performed by the following sequence of steps:

1. building computer models of watersheds according to a specified structure of certain factors and responses,
2. generating many “pseudo-watersheds” (PWs) from the model, including responses of abundance or pool counts,
3. applying one of several candidate sampling plans to each PW,
4. applying one of several experimental designs to the data from each sampling plan and imposing a treatment effect on responses that come from units subject to restoration treatment
5. analyzing the responses according to model appropriate for the combination of design and sampling plan
6. recording whether or not a treatment effect was detected in each analysis

Details on these steps are given below:

1. ***Computer model for watershed***: The computer model was based on the Asotin watershed to describe the spatial and temporal layout of the study. There were three streams, which were treated as independent of one another (responses on one stream were not affected by responses on another). Within each stream there were 3 sections; within each section there were 2 fish sites (“f-sites”). Within each fish site there were three habitat sites (“h-sites”). Thus there were a total of 3x3x2x3=54 locations within the watershed at which measurements *could* be taken. This spatial structure is observed for 12 years, so that a total of 648 potential observations could be created.

The model used to generate data for this spatio-temporal structure is an additive normal linear model using terms based upon the analysis model described previously. Stream and treatment effects were taken to be fixed; year, section, f-site, and h-site effects were taken to be random, and any interactions with random effects were taken to be random. Specifically, year\*section, year\*f-site, and year\*h-site random effects were added.

Random effects: The variance components for random effects were based upon the estimates from the pilot study above. Where a variance component could not be directly estimated from one of the existing data sets, educated guesses were used to fill in these missing values, according to the principle that smaller-size spatial units generally have smaller variance components (more similar responses) than larger size units. Because the variance components were estimated with (sometimes substantial) uncertainty, three different variability scenarios were considered. The first was used the estimated variance components, which represents our “best guess” as to the actual variability present. The second used variance components set to the lower limits of their respective confidence intervals, representing a “best case” for variability. The third used variance components set to the upper limits of their respective confidence intervals, representing a “worst case” for variability. The complete listing of variance components is given in the tables below.

Because annual correlations were estimated to exist in both the abundance and pool pilot data, the watershed models contained the potential for serial autocorrelation. Annual means (year effects) were subject to autocorrelation of either 0 or 0.5. This correlation was considered in combination with each of the three levels of variance components, creating 6 different spatio-temporal structures that could define a watershed for each of the two response measures.

Variance components for log(Abundance)

|  |  |  |  |
| --- | --- | --- | --- |
| Factor | Estimated | Best Case | Worst Case |
| Year | 0.195 | 0.009 | 0.632 |
| Section(Stream) | 0.026 | 0.015 | 0.056 |
| FSite(Section) | 0.0027 | 0.0011 | 0.013 |
| HSite(FSite) | 0 | 0 | 0 |
| Year\*Stream | 0.0063 | 0.0015 | 0.56 |
| Year\*Section1 | 0.00321 | 0.00081 | 0.0282 |
| Year\*FSite3 | 0.00163 | 0.00043 | 0.0064 |
| Year\*HSite5 | 0 | 0 | 0 |
| Random Error6 | 0.037 | 0.024 | 0.078 |

1Assumed to be ½ x Year\*Stream because section is subunit of stream

2 Assumed to be ½ x Section because changes in differences between sections across years are likely to be smaller than the original differences

3 Assumed to be ½ x Year\*Section because f-site is subunit of section

4Assumed to be ½ x Fsite because *changes* in differences between f-sites across years are likely to be smaller than the original differences.

5 Assumed to be ½ x Year\*FSite because h-site is subunit of f-site

6Set to 0 because this term is redundant with Year\*Hsite

Variance components for

|  |  |  |  |
| --- | --- | --- | --- |
| Factor | Estimated | Best Case | Worst Case |
| Year1 | 2.8 | 1.5 | 8.5 |
| Section(Stream) | 2.0 | 0.84 | 9.9 |
| FSite(Section) | 0.78 | 0.25 | 11.6 |
| HSite(FSite) | 1.64 | 0.97 | 3.4 |
| Year\*Stream | 3.3 | 2.1 | 6.1 |
| Year\*Section2 | 1.6 | 1.0 | 3.0 |
| Year\*FSite3 | 0.8 | 0.5 | 1.5 |
| Year\*HSite4 | 0.4 | 0.3 | 0.8 |
| Random Error5 | 0 | 0 | 0 |

1Estimated from Redds analysis

2Assumed to be ½ x Year\*Stream because section is subunit of stream

3 Assumed to be ½ x Year\*Section because f-site is subunit of section

4 Assumed to be ½ x Year\*Hsite because h-site is subunit of f-site

5Set to 0 because this term is redundant with Year\*Hsite

Fixed Effects: The means for each stream were also estimated from the means in the pilot analyses. For log(abundance), the means were 1.46 for Charley, 1.31 for North Fork, and 1.35 for South Fork. For the means were 6.3, 4.4, and 5.5, respectively.

Treatment effects were added in to the response values for all observations taken on units after restoration was applied in accordance with the design being considered. Because there are many possible patterns in which the treatment effects could appear, we cannot investigate the ability of each design to detect all possible manifestations. Instead, two possible scenarios for treatment effects were considered: “sudden” and “ramped”. For sudden effect, it was assumed that the full treatment effect was observed in the first measurement subsequent to treatment and that this effect persisted through the remainder of the study. For the ramped effect, the treatment effect reached its full level 4 years after treatment, and rose by increments of ¼ of the full level in each if the intervening years. *Added 6 Sept 2010: Following consultation with Eco Logical Research Inc., a second round of simulations was performed assuming one further “likely” treatment-effect pattern for each response measurement. For log(abundance) we created a different “ramped” effect wherein 60% of the full effect was realized in the first year, 90% in the second year, and 100% from the third year onwards. For we assumed that 100% of pool creation would take place within one year, but that after 3 years this number would degrade at a rate on 10% per year.*

The ultimate full levels were determined as follows. For abundance, we wish to be able to detect a 25% increase in mean abundance due to restoration. Thus, log(1.25) was added to all treated units after treatment, incremented as necessary for the ramped effect. For pools, the use of a square root transformation makes assigning treatment effects slightly more complicated. The desired change that we wish to detect is a doubling in the number of pools. This was achieved by multiplying each stream’s mean by after treatment, again ramped as needed.

For simplicity, no specific stream\*treatment interaction was created, because two of the experimental designs considered place all treatments in one stream (see below, this should be noted as a *weakness* of these designs).

Experimental designs: Three different experimental designs were compared for assigning restoration treatments to units in the study. The first was the design that is currently planned to be used in the Asotin study (referred to as the “current”). In this design the three sections of one stream, Charley, are to be restored one-at-a-time in three-year intervals. The current design is depicted in Figure 1 above.

The second design is the alternative design that I proposed and that is depicted in Figure 3. In this design, the staggering of treatment applications in three-year intervals continues, but the sections treated at different times are in different streams. This is called the “alt” design.

The third design is the simple design that might be used by many researchers. It consists of a single treated section, restored at the midpoint of the experiment (i.e. after 6 years). Without loss of generality, the middle section of Charley was used as the treated section. This is called the “1-time” design.

1. ***Generating pseudo-watersheds***: Once the model was defined, PWs were generated by simulating pseudo-random data to represent the potential measurement at each of 54 h-sites across 12 years. Random effects were generated independently according to their respective variance components and stream means were added in. One PW consisted of 648 potential measurements. For each of the 12 combinations of response variable, variability scenario, and autocorrelation, 1000 PWs were simulated. This number allows Type I error rates of analyses conducted at the 5% level to be estimated to less than ±1.5% error with 95% certainty. Power estimates similarly can be estimated to at worst ±3.2% error with 95% certainty.
2. ***Sampling from a PW***: Abundance measurements are taken on f-sites, while pool counts are taken on h-sites. Therefore different sampling plans were considered for the two response variables. In each case, the goal was to investigate what relative differences in power could be obtained by different levels of sampling intensity.

For abundance, 5 different sampling plans on f-sites were considered:

1. “1-per-stream”, in which one f-site is chosen at random from the middle section of each stream and measured in each year. This represents the barest minimum measurement that could take place in a BACI-type study, and is used only with the 1-time experimental design.
2. “1-per-section”, in which one f-site is randomly chosen from each section of each stream and measured in each year. This represents a minimum sampling plan design in which all three designs can be run and compared.
3. “Planned”, which consists of the same measurements as in 1-per-section, plus a second f-site in each section in Charley, the treated stream.
4. “Alternative”, which follows the same spirit Planned, but matches the extra measured f-site with the treated sections from the Alt design.
5. “Full”, in which 2 f-sites are measured in each section (twice as much measurement as in 1-per-section, 50% more than Planned).

For pool counts, 7 different sampling plans on h-sites were considered:

1. “1-per-stream”, in which one h-site is randomly selected from one randomly selected f-site in the middle section of each stream and measured in each year. This represents the barest minimum measurement that could take place in a BACI-type study, and is used only with the 1-time experimental design.
2. “3 per stream”, in which all three h-sites are measured from a random f-site in the middle section of each stream. This, too, is combined only with the 1-time design.
3. “1-per-section”, in which one f-site is selected from each section and one h-site is selected from each f-site.
4. “Planned-Lite”, augmentation of 1-per-section, in which one h-site is selected from both f-sites per section in Charley.
5. “Planned-Panel” is the current sampling plan. It augments Planned-lite with a second h-site sampled from each sampled f-site in each section, on a rotating panel basis. That is, one h-site per f-site is always measured, and a second is measured by sampling each of the two remaining h-sites in every f-site every other year on a rotating basis.
6. “Alternative”, which mimics Planned-Panel, but assigns the extra sampled f-site to the treated sections in the Als experimental design (and is used only with this design).
7. “Full”, in which all three h-sites are used in each f-site in each section of each stream in every year.

Each sampling plan is applied to each PW. That is, a PW represents a set of known measurements, some of which get to be observed or are left unobserved by a given sampling plan, and 12 parallel samples are drawn from each PW according to these 12 sampling plans. In this way, each sampling plan is examined under exactly the same sets of potential measurements, so any differences in performance are due entirely to the effects of the plans and not due to differing random effects. Through this arrangement of different sampling plans, we will be able to see the impact of additional sampling on different levels and identify which plan(s) might be most appropriate for studies of this type.

1. ***Apply a design to the measurements***: Once the responses have been sampled from a PW, the treatment effects are assigns as described in (1) to the measurements corresponding to all treated units in accordance with each designs. That is, three parallel data sets are created for each PW: one assuming the 1-time design would be applied in this PW, one assuming the Current design, and one assuming the ALT design. Because the only differences between these parallel data sets is in the locations of the treatment effects, and all other random effects are the same for all three, differences in performance between the designs is due entirely to the designs and not any differences in random effects.
2. ***Analysis of data***: There are 11 combinations of design and sampling plan for abundance, and 15 for pools. Steps 1-4 create 6x1000x11 + 6x1000x15 = 156,000 data sets to be analyzed. Each is analyzed using a mixed model that is appropriate for that combination of design and analysis. All analyses were carried out in SAS version 9.2 using PROC MIXED, with models appropriate to the combination of design and sampling plan. These models are available upon request.

For each analysis model, a linear contrast was written to estimate a possible treatment effect. Generally, contrasts are designed to look for specific patterns in the treatment means. While they may have some power for detecting effects with other patterns, they are best at detecting their intended pattern. Contrasts were therefore designed separately to estimate either a sudden treatment effect or a ramped treatment effect. This represents the (optimistic) case where we know what sort of pattern to look for in advance. However, we can use each contrast to look for both effects, and hence have an idea how similar contrasts might perform in less serendipitous circumstances.

1. ***Recording of results***: For each contrast estimated on each data set a confidence interval was constructed. From each interval, two important features were recorded: (1) the length of the interval, and (2) where the interval was with respect to its null hypothesis value of 0, representing no treatment effect. A good contrast should (a) cover the value 0 roughly 95% of the time when there is no treatment effect, (b) lie above the value 0 as often as possible when there *is* a treatment effect, and (c) be as short as possible, so that the interval can be useful as an estimate of the treatment effect. Thus, across all data sets from a given scenario, the average length and the proportion of intervals covering vis. above 0 for null vis. non-null cases were the recorded responses.

*Added 6Sept 2010*: Two separate rounds of simulations were performed. The first used the first definitions of “sudden” and “ramped” treatment effects and tested the performance of the contrasts that were written to perform the analyses. Results from the first round led to modifications used in the second round. There were no substantive differences in the relative performance of the designs or sampling plans between the two rounds. The results from the second round are those reported, except as noted.

An important observation from the first round of simulations was the extreme slow completion times under some sampling plans. This was coupled with problems with the convergence (successful completion) of the computational algorithms. The difficulties were traced to the effort to account for the correlation among Year effects within the analysis. As noted below, the presence of correlation played no role in affecting the results. As a result, this stumbling block was removed in the second round and all analyses were done without estimating the autocorrelation among year effects. The execution times for the simulation runs were reduced by at least an order of magnitude.

# RESULTS ADDED 6SEPT10

Certain observed features were common to both response variables:

1. In all designs and sampling plans, the analyses maintained their specified coverage levels (Type I error rates if interpreted as a test). That is, all coverages were within simulation error of 95%.
2. The presence of correlation among year effects had no impact on the coverages or average lengths.
3. As expected, the coverages dropped and lengths increased as the variability profiles were changed from best case to worst case.

### Abundance

Under the best case for variability, all designs and sampling plans have 100% detection of the 25% increase. Even under the estimated variability, all designs and sampling plans have at least 95% power to detect the treatment effect except the BACI combination, 1-per-stream sampling with a 1-time design, which has just over 70% power (Figures below). Once measurements are made in each section, confidence interval lengths do not change much with additional subsampling within the sections. The alt design has the shortest intervals, while the 1-time design has the longest.

Under the worst-case variability, greater differences among the methods begin to emerge (Figures below). The 1-time and current designs have very similar powers and lengths regardless of the subsampling intensity. However, the alt design distinguishes itself in terms of both power and length of confidence interval. Powers range between 60-70%, compared to 25-35% for the other designs. Confidence interval lengths are roughly 2/3 those of the other designs.

### Pools

All powers for detecting a doubling in pools are lower than those for detecting a 25% increase in abundance (figures below). Power under estimated variability is clearly better for both the current and the alt design than for the 1-time design. Under planned sampling, both designs have more than 80% power. The benefits of added subsampling are slightly greater here than for abundance, but there are still no great differences among sampling plans that sample all sections employing varying levels of subsampling. Improvements due to subsampling are clearest with the 1-time design. The lengths of confidence intervals again tend to be smaller with the alt design than with the others.

# Discussion

First and foremost, these results must be viewed as somewhat speculative. The process of estimating all of the needed variance components for the watershed was not straightforward. The historical data were sparse, particularly at the subsampling levels (F-site and H-site). Data sets tended to have either spatial or temporal components to them, so combined spatio-temporal random effects (year\*unit interactions) may not be well estimated at all. Despite the use of upper endpoints of confidence intervals in formulating a “worst case”, some variance components are pure guesses and therefore may be subject to far greater variability than presumed. Hence, great caution should be applied in interpreting the values of the powers.

On the other hand, the principles that drive the *comparisons* among powers do not depend on the actual values of the variance components, but rather on their relative sizes. Fundamentally, treatments are applied to sections and subsequently measured in different years. The analyses include terms that account for any variability that occurs on a larger scale, and hence this variability does not affect the designs or sampling plans’ relative powers of confidence interval lengths. (As a check on this, I inflated individual variance components and reran a couple of simulations. There was zero change in power when variance components were inflated that correspond to units larger than Year\*Section.)

Similarly, subsampling of F-sites and H-sites is predictably less effective than measuring more sections. The difference between the 1-per-stream and 1-per-section sampling plans for the 1-time design was very large. The difference between taking 1 measurement per site and full subsampling was not generally very large. However, this is largely due to the relative sizes of the variance components for the subsampling effects. It is conceivable that lower-level variability is much greater than assumed, in which case subsampling becomes an effective and relatively inexpensive means of improving precision. The only way to k now this is to collect data on a finer scale than what is presently available.

The simulation model and analysis models were based on the same distributional assumptions, specifically normality of all random effects. This would help to explain why coverages all maintained the specified 95% level. Nonetheless, it is a comfort that, despite the complexity of the watershed system model, it is possible to construct an analysis that achieves a certain sense of legitimacy.

The fact that correlation plays no role in the analysis results undoubtedly stems from the fact that the random effect of Year does not enter into the error term for the treatment effects of each design. The treatments are applied to sections of a stream and hence their effects are tested against variability among section measurements in different years. When all measurements in a given year increase or decrease together, as is common with abundance measurements, these effects cancel in differences taken at the section level within a year. Thus, correlation among these larger-scale effects is not noticeable in any analysis of treatment effects.

The high power for detecting treatment effects for log-abundance is based on the estimated variability from past data and assumptions that were made pertaining to variability at spatial scales finer than those at which data were collected. If these variance components turn out to be much larger than estimated, or in particular if the variability among fish sites within a section is much larger than assumed, the actual power to detect a 25% increase could look more like the worst case estimates. As data are collected, even before treatments are applied, we may be able to update the estimates.

The lower power for detecting effects in pools vs. abundance is because the size of the effect we are trying to detect for pools is larger than the one we seek for abundance, relative to the variability estimated to be present in the two measures. It does not mean that pools are harder or less likely to change. It does mean that, if the variance components estimated from past data are anything like those for future data, then we will less be able to detect a doubling in number of pools, should it occur, than a 25% increase in abundance.

Additional runs were performed under planned sampling, varying the ultimate treatment effect from a 5% increase to a 40% increase. This was intended to allow more detailed comparison of the current and alternative designs, specifically addressing the concern that a multiple treatments applied in different sections of the same stream may synergize to generate a larger treatment effect in each treated section than would be observed by treating only one section of a stream. By looking at the power curves for the two designs, we can see how much synergy would need to take place in order to make the current design favored over the alternative.

Under estimated variability the curves are separated only for detecting changes of 20% or less. If synergy of multiple treatments in one stream accounts for the horizontal difference between the curves, then the designs are equivalent. Here we see that the horizontal difference is never more than 5%, so that the synergism does not have to be large for the current design to have power that is favored over the alt design. However, the difference in confidence interval lengths is independent of the treatment effect, so the current design really needs to be *more* powerful than the alt design in order to make up for the fact that it produces longer, less useful intervals.

In the worst case variability, the horizontal difference can be as much as 15%. There would need to be considerable synergy, making up a sizable portion of the total treatment effect, before the current design would be favored over the alt design.

Whether such synergism exists is unknown. If it exists, its exact nature is also unknown. Given published reports that suggest that 90% of fish never leave their native stream as juveniles, then improvements in habitat will have a very limited effect toward attracting fish from other streams. Therefore we can expect that there is little negative dependence from stream-to-stream. We may speculate that a large percentage of fish do not roam far from their natal habitat within the same stream, although \*I\* don’t know whether this is proven.

Furthermore, pool creation is a local phenomenon. It is unlikely that a treatment applied to one section of a stream will produce pools several km away. Therefore, we can expect that there is no synergism in the pool effects, and comparative power curves like those above are unnecessary. If section-to-section wandering of fish is low, then we can expect that the treatment effects on abundance will be mostly independent from section to section, and any synergism is quite limited. We will be able to test this hypothesis more stringently when data arrive.

The relatively similar performance of the 1-time and current designs for detecting differences in abundance is a bit disheartening, considering the extra effort that the current design requires. The current design suffers because comparison between treated and untreated sections cannot be made within the same stream. It is the variability of such sections that is the most important component of the error term for testing and forming confidence intervals for treatment effects. This increases the variability of treatment effects estimated later in the design, in particular those associated with times more than 6 years after treatment. So the advantages of multiple treated sections are diminished by the disadvantages of increased difficulty in separating treatment effects from inherent variability. The relative magnitudes of these two facets, plus and minus, vary with the sizes of the variance components. Thus, we see a large advantage for the current design over the 1-time in the analysis of pools, and little advantage in the worst-case analysis of abundance.

It should be noted that the alt design overcomes this issue by having treated sections spread among three streams, with untreated sections in the same streams. This creates a situation akin to blocking in that treatment comparisons against controls are made within stream rather than between streams, and therefore incur less variability in estimating effects. This explains the improvement enjoyed by the alt design, both in terms of power and, crucially, confidence interval length.

An important feature of the staggered-start designs is that they are more robust against outliers caused by unusual circumstances in a given year, which can ruin single-treatment experiments completely. An ill-times flood that shifts a stream dramatically or drought that causes increased mortality may leave lasting effects on a section of a stream differently from other sections or streams. This can have devastating effects on classic BACI studies, as has been extensively noted. It is primarily as a defense against such problems that Walters et al (1988) and Loughin et al (2007) recommend staggering the start of a long-term experiment. The simulation study did not incorporate random floods to try to quantify the gains achieved by staggering treatments on time.

# Recommendations

**IF** it is believed that treatments applied to different sections of the same stream synergize to create a broader, more favorable environment for fish, then the application of multiple treatments to sections of the same stream has the potential to create an environment that is overwhelmingly favorable in a single stream. This can lead to a larger overall treatment effect which, if sufficiently larger, would be easier to detect than effects caused by other designs.

If, on the other hand, the potential for synergy is viewed as minimal or nonexistent, then the clear favorite is the alt design, in which a section of each stream is eventually treated over time.