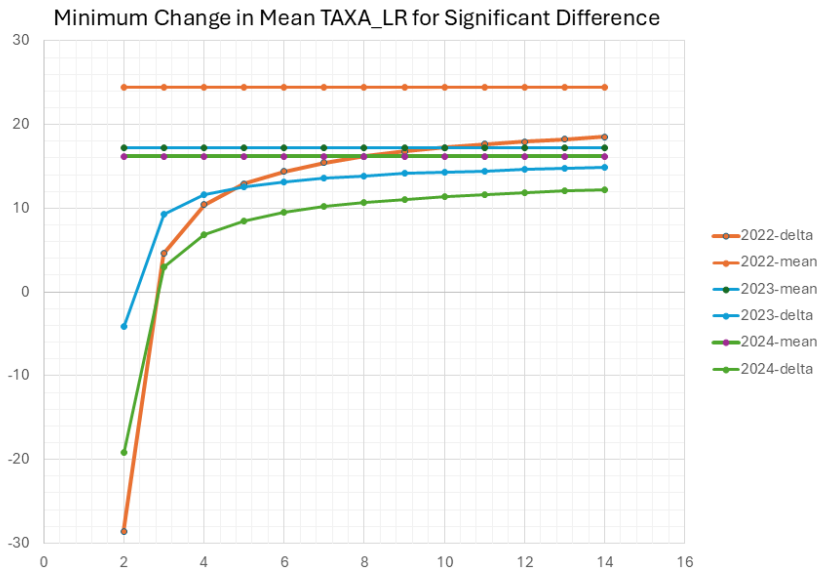
LIMITED POWER ANALYSIS TO OPTIMIZE SAMPLING EFFORT FOR MACROINVERTEBRATE COMMUNITIES OF THE WOOD RIVER BASIN

River Continuum Concepts Technical Memo RCC2025-02

B. Marshall, River Continuum Concepts, Inc. Manhattan MT, 59741.



Technical Memo: RCC-2025-02

TO: Cory McCaffrey, WRLT
By: Brett Marshall, RCC
RE: Design optimization / Statistical Power Analysis
CC: Jackson Birrel, SFP

Introduction

During a recent meeting with Wood River Land Trust (WRLT) and the Salmon Fly Project (SFP) on 4/17/25 we discussed the allocation of effort in the field, lab, and design for optimization. When I worked on some benthic samples from Maine in the early 2000s, I remember encountering a composite sample that contained an estimated 1.5 million macroinvertebrates. For perspective, the group I was working with had 55 staff members processing macroinvertebrate samples who collectively (and coincidentally) happened to identify a total of approximately 1.5 million macroinvertebrates over the course of that entire year. This underscores the kinds of costs that can occur if samples are processed in their entirety—especially composite samples. If the cost of paying 55 people to work for an entire year on a single sample seems prohibitive imagine the cost of a small study involving 5 sites with 5 statistical replicates. This "small study" would take a team of 55 people 25 years just to complete the laboratory work—or it could be completed in about 1 year by employing 1375 biologists. Fortunately, not all samples are this large, but this example underscores the importance of standardized laboratory subsampling procedures in providing timely and cost-effective macroinvertebrate data. Thus, it is important to carefully consider the assumptions associated with different field, laboratory and design.

Budget is always a factor influencing the design of ecological studies. When developing an ecological monitoring program, one of the first decisions facing investigators is how to best apply a fixed sampling budget to meet the primary goals of the monitoring program; should the program focus on extensive sampling (broadly spatial in scope, ideal to describe longitudinal or regional trends) or intensive sampling (focusing on fewer sites with more rigorous investigation, ideal for assessing site change over time (inc. BACI and BACI+ designs)). Statistical Power Analysis is an ideal way to assess the optimization of sample allocation for benthic studies.

Methods

Statistical Power is a function of α (type-1 error rate), β (type 2 error rate), σ (standard deviation of the sampled population, δ (effect size), and *n* (replication). If you know these variables, you can solve for the others. Fortunately, we have three years of data to allow us to estimate and solve for any of the values needed. I decided to solve for δ because it reflects the amount of change in a metric that is required for statistical significance (the amount of change for rejection of the no-difference null

hypothesis). I selected taxa richness because it represents diversity in a way that is easily digestible for the general public. For example, most people would agree that a 100% loss of species is undesirable. I used the large-rare-corrected taxa richness value because it was more accurate and tended to have lower variation than the purely quantitative estimate of richness.

After this decision, I set the type-1 and type-2 error rates constant and equal; both at $\alpha=\beta=0.10$. Although this is a slightly relaxed critical P (as determined by α), the goal of most applied ecological monitoring is not to describe world-changing phenomena (such as cold fusion), but to detect ecological changes, often anthropogenic ecological impacts, before they become detrimental. Furthermore, the costs of failing to detect an ecological impact are usually greater than reporting a false positive because the corrective action for restoration, reintroduction, or extirpation may require active management. One could probably argue that even looser restrictions (e.g., $\alpha=\beta=0.15$) could be used; this is about mathematically equivalent to the 25% quartiles that were used in the development in most indices of biotic integrity (e.g. Karr 1991).

Each year we observed different mean richness, and different levels of within site variance. Although we discussed[1] potentially high within-site variance in the WRLT data, a cursory review did not find the observations to be excessively high—they are typical for within site (within riffle) variances for the region (Western WY, Henry's Fork, WRLT, Big Hole River, Gallatin River, Streams in Gallatin National Forest, and Yellowstone National Park). Nonetheless, with three years of data archived, the act of choosing a level of variance required a decision: use all variances, use one variance (high, pessimistic), use one variance (low, optimistic), or use average variance. Because of the time involved, I decided to use one estimate of variance and means from each site. I specifically opted for the mean variance and average richness. The analysis was set as a two-tailed test for each site.

[1] During the 4/17/25 meeting.

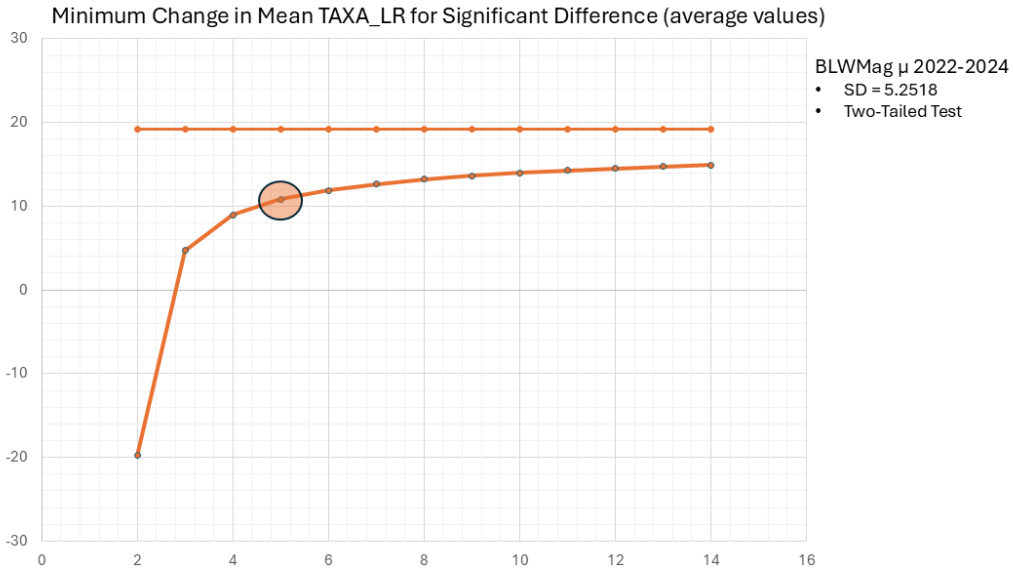
Results

BLWMag

Although I decided to use average standard deviations and average means, I began the analysis by generating power curves for each separate year at the site BLWMag (see cover illustration). The horizontal lines along the top are the average for each of the three years. The curves of the corresponding color below the horizontal line show the decline in richness to reject the no-effect/null hypothesis. At BLWMag the greatest mean was observed in 2022 (blue line), and the lowest mean was observed in 2024 (green). The greatest standard deviation was observed in 2022 (blue) whereas the lowest standard deviation was observed in 2023 (orange). The averages were not significantly different at the level of replication used for the WRLT sites (n=5). The x-axis is the number of collected samples collected. Thus, in 2022, the mean taxa richness (TAXA_LR) was 24.4. With 5 replicates richness had to decrease to 12.9 taxa (11.2 Δ[2] taxa, 42%Δ) for statistical significance given the model parameters discussed in the methods section. This was the worst-case scenario for detecting change. In 2023, the mean of 17.2 taxa had to become reduced to 10.8 taxa (a reduction of 6.4 taxa, 37%). Finally, in 2024, the mean of 16.2 taxa had to be reduced to 8.53 taxa (a reduction of 7.67 taxa, 44.6%). Using the mean of annual means and annual variation estimates (μ=19.3 taxa, σ = 5.25) we found that the current design could detect an average reduction in richness of 8.5 taxa (43.9%). Thus, the current design, using two-tailed tests can detect about 40-45% change in richness, regardless of the year used. Analysis of additional sites will focus on the mean of annual means and standard deviations.

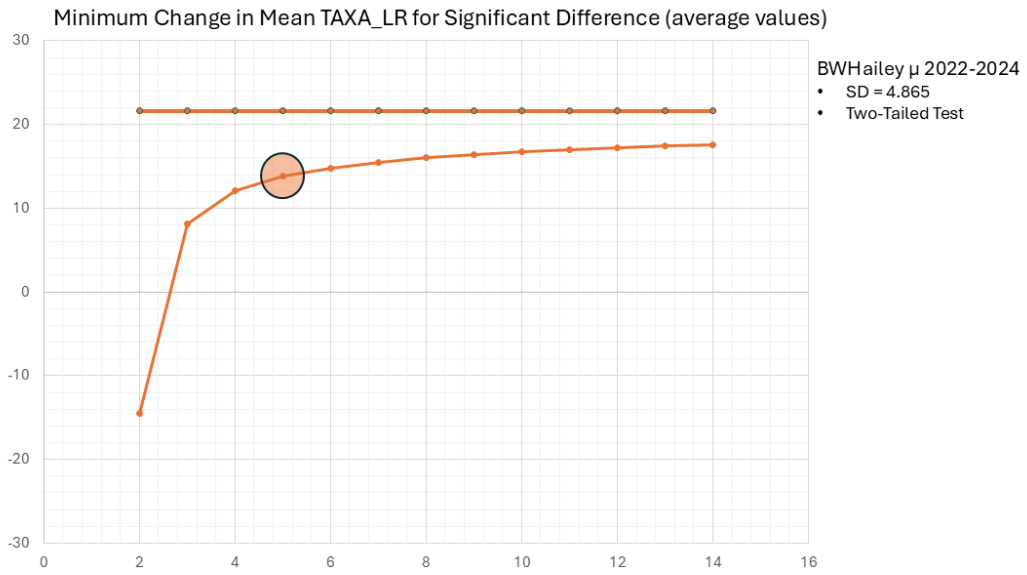
This form of power analysis is best applied to study optimization. Each single sample added to the program adds a fixed cost. However, the benefits of each successive sample have diminishing returns. For example, if we collect only two replicates at BLWMag, the total taxa richness would need fall all the way down to -19.2 taxa (a change of 35.4 taxa, more taxa than occurred there originally). Adding one more replicate only requires taxa richness to drop to 2.9 taxa; a net change (loss) of 13 taxa. Thus, the addition of one-sample improved the amount of change that can be statistically detected by 22.4 taxa. That is a huge improvement in sensitivity by adding one sample. In contrast, the change from 14 to 15 replicates only improves the sensitivity by a fraction of taxon (0.15 taxa). The optimized sample effort occurs where the power curve (δ) begins to flatten out, begins to reach an asymptote. Decreasing the sampling effort from 5 replicates to 4 replicates per site reduces the minimal detectable difference by 1.7 taxa, whereas increasing the sampling effort from five to six replicates increases the minimum detectable difference by 1 taxon (0.995 taxa). I usually conclude the design is optimized where an additional replicate improves the design by less than 1 taxon. The current design is very close to optimization for BLWMag, if the replication is changed, we should consider 1 additional replicate for a 5% improvement of sensitivity (q.v., cover figure, and BLWMag fig, below).

[2] The use of δ and Δ are similar. Lower case δ is the effect size, whereas upper case Δ is change.


Minimum Change in Mean TAXA_LR for Significant Difference (average values)
30
20
10
0
-10
-20
-30
0
2
4
6
8
10
12
14
16
BLWMag μ 2022-2024
• SD = 5.2518
• Two-Tailed Test

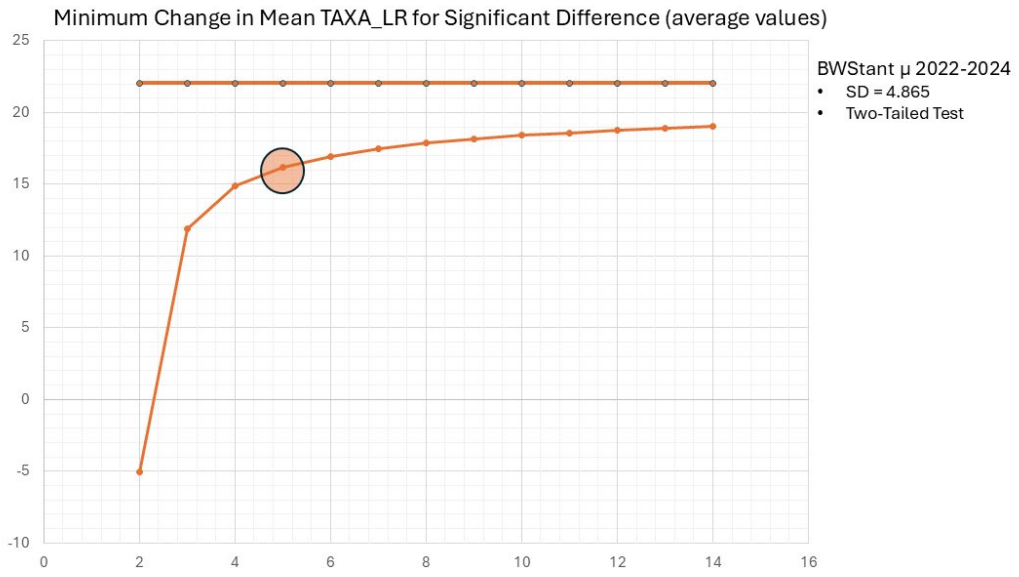
BWHailey

The current design allows us to statistically detect a reduction from the mean annual richness of 21.6 taxa to 13.8 species (Δ=6.8 taxa, 31%). The addition of one and two more replicates (n=6, n=7) adds 1.01 and 0.7 taxa, respectively, incrementally. The loss of replicate (n=4) reduces the detectable difference by nearly 2 whole taxa (1.76 taxa). Thus, the design is nearly optimized, but could potentially benefit from one more replicate.



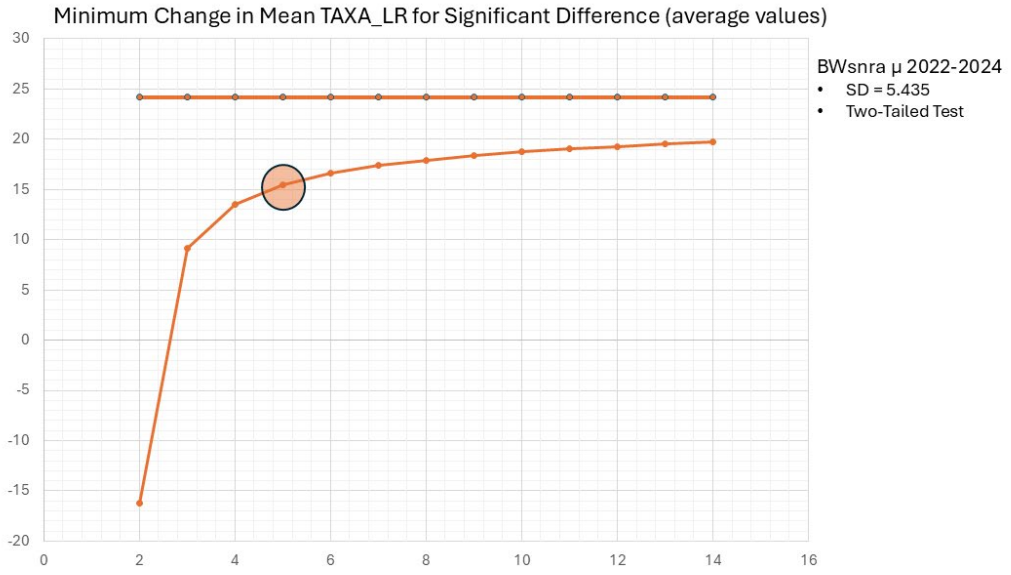
STANT

The average of three year's richness estimates at STANT was 22.07 and the current study design requires a reduction in average richness to 16.19 taxa—a loss of 5.8 taxa (26.6% change). A reduction in replication (from n=5, to n=4) an increase in the number of lost taxa required for significant 0differences to 7.2 taxa (32%Δ)—a loss of sensitivity by 1.4 taxa. The addition of one more replicate (n=6) reduces the number of lost taxa required for significant differences to 5.11 taxa (23.2% Δ)—only an improvement of 0.76 taxa. Thus, for average data, the design is near-optimal cost-benefit. However, the greatest variation and average richness occurred in 2022 (vs. 2023-2024) and in this instance, the increase from n=5 to n=6 improved the sensitivity of the design by 1.2 taxa—a marked improvement. Regardless, the site is near-optimal at n=5, but could sometimes be improved by adding one more replicate.



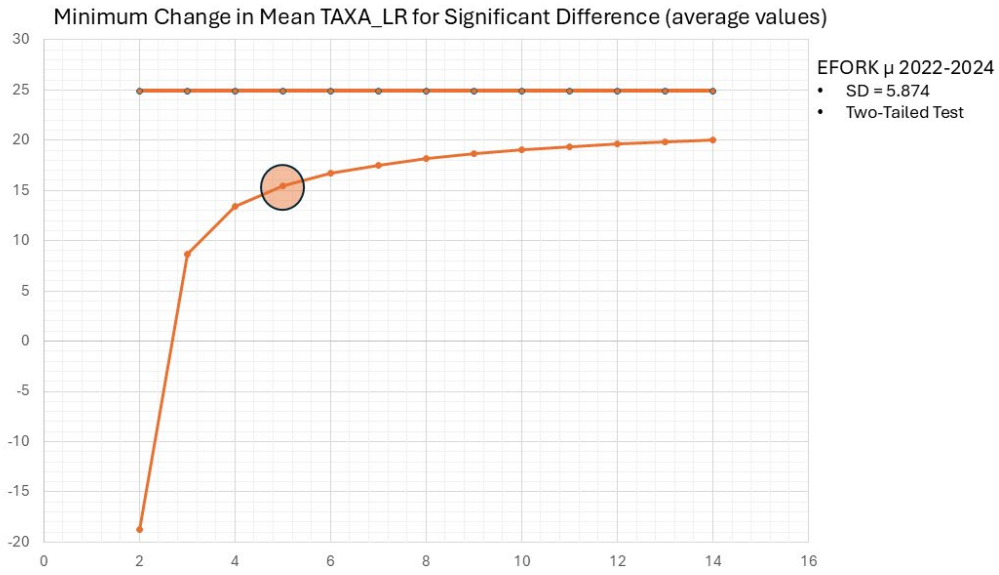
BWSNRA

This site had high richness and high variance, typical for less disturbed ecosystems. The current design requires a loss of 8.75 taxa from the average of 24.2 taxa (36.2%Δ) for statistical significance. Because of the higher variances observed (relative to the sites discussed earlier) each sample added to the sampling regime increases the sensitivity (decreases the amount of change required for statistical significance) by >1 taxon (n=6 improved sensitivity by an additional 1.4 taxa Δ, n=7 improved by 1.26 taxa Δ, n=8 improved by 1.15 taxa Δ, n=9 improved by 1.07, n=10 improved by 1.00 taxa Δ, and n=11 improved the design by 0.95 taxa Δ). The fifth replicate improved the design by 1.6 taxa relative to n=4. Although the sample plan is not quite optimized for BWSNRA, the amount of percent change required for statistical significance is about the same as for the sites discussed earlier. Using averages, a loss of about 8.7 taxa is required for statistical significance at n=5. Since each sample beyond n=5 increases the sensitivity by >1 taxon, by the time 10 replicates are collected a change of only 5.5 taxa would be needed for statistical significance.



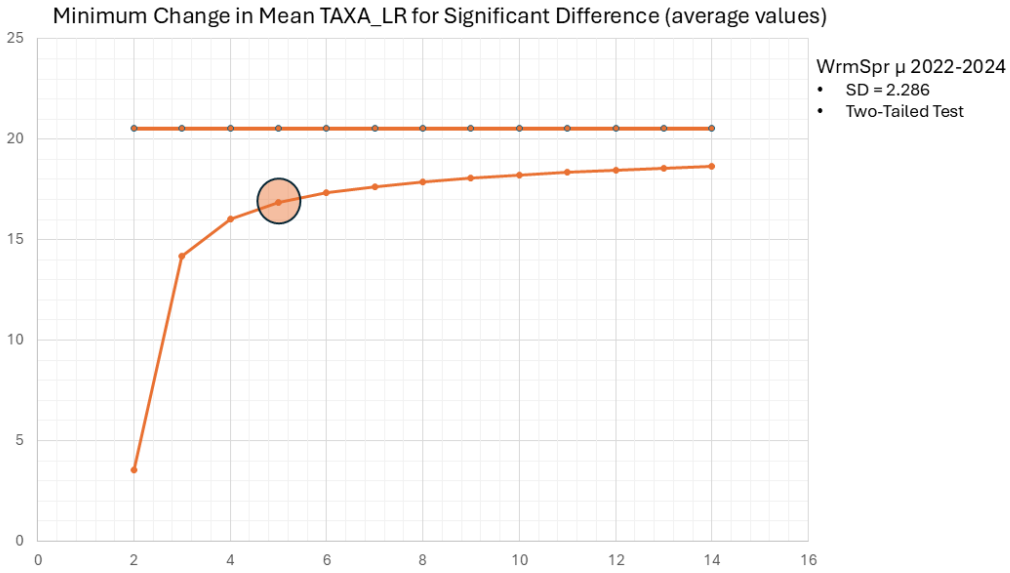
EFORK

The EFORK site was similar to BWSNRA in that it generally produced greater richness estimates with greater variances, as is typical of sites with less anthropogenic disturbance. At the current level of replication, the design required a change of 9.75 taxa (39% Δ) for statistical significance given the model parameters discussed in the method section. Successive samples added >1 taxon of sensitivity incrementally in a pattern similar to the observations for BWSNRA (i.e., improved by 1.61Δ @ n =5, by 1.4 Δ more @ n=6, by 1.25 Δ more @ n=7, by 1.15 Δ more @ n=8, by 1.07 Δ more @ n=9, by 1.00 Δ more @ n= 10, and by 0.95 Δ more @ n=11). Although each successive replicate has diminishing returns in improving the sensitivity (as usual), by the time 10 replicates are used at this site, a markedly smaller change (5.9 taxa Δ) is needed for statistical significance. This is an improvement from 39% Δ to change needed to 23.6 % Δ.



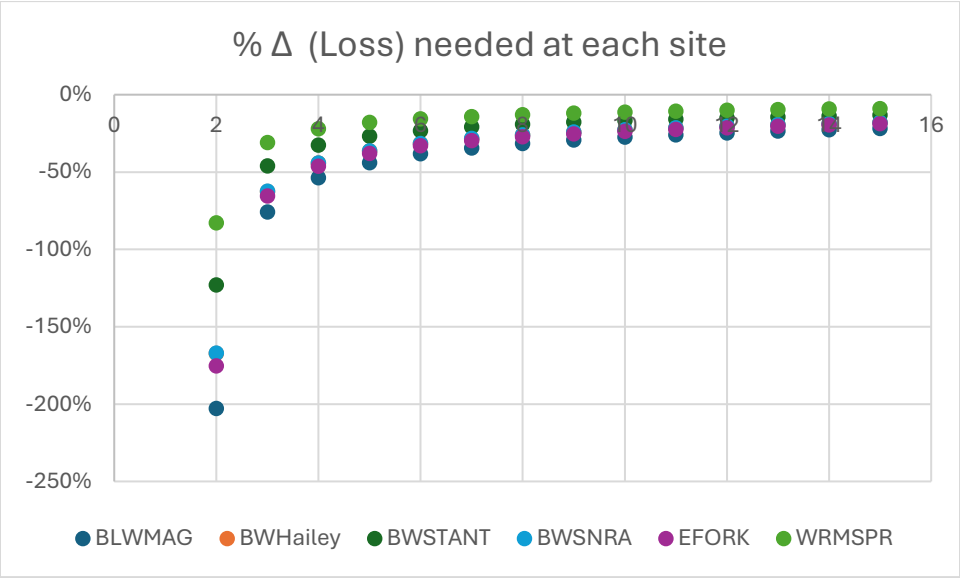
WRMSPR

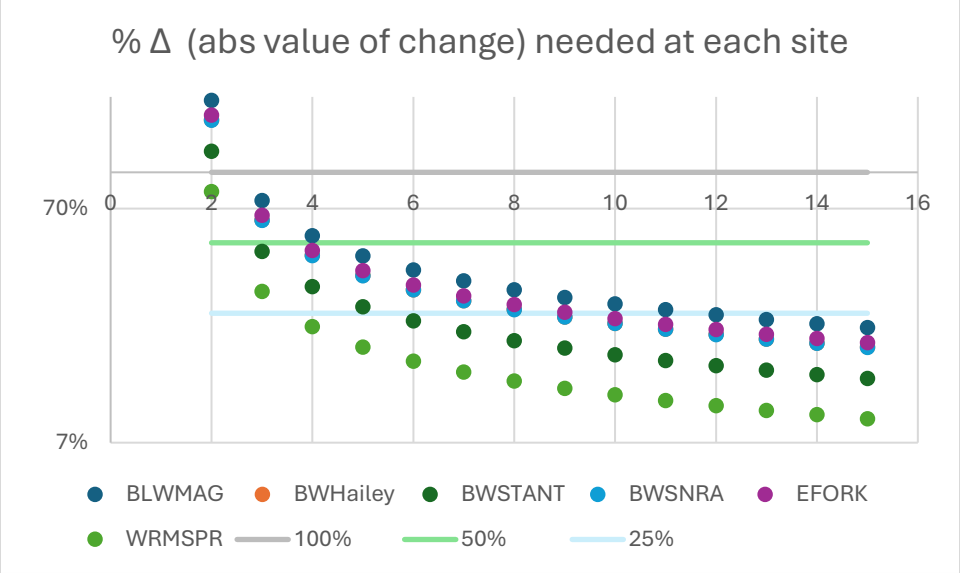
As is typical for streams with anthropogenic or more homogeneous systems, the variances at WRMSPR were lower and more consistent among years than the other sites. Average annual variances fluctuated similarly in range (uncorrelated by date) with BLWMag. Using average statistics and the current sampling plan a change of 4.5 (21.9% Δ) taxa can be statistically detected for this site—an improvement of about 0.8 taxa from using reduced sampling (n=4). According to Cohen 1988, this is the only site where a “small change” (~20%) could be routinely detected. Additional samples did not improve the design markedly at this site.



Summary

1. All the sites were approximating optimization in terms of the cost-benefits of replicates.
2. More data will allow stronger estimation of site richness estimates.
3. Variances in richness were not excessively high and were about what was observed for other locations in the region.
4. The amount of change required for statistical significance is not ideal at most sites and could be improved by attempts to homogenize the variance.
5. The two most variable (EFORK and BWSNRA) sites could benefit measurably from several additional replicates.
6. These analyses used a T-test design to compare the population at each site for the amount of change required for statistical significance. In application, as years are added to the design, other model parameters (e.g. "k," number of treatments in GLM etc.) would be added, which should improve the power beyond the thresholds defined here. However, this complicates the analysis and would require more assumptions and more iterations (e.g., multiple variances for each year, with multiple years). As such it was beyond the scope of this memo. Nonetheless these analyses do provide a standardized approach to the response of each site to detect changes in species richness from individual year to individual year on average.
7. The power—expressed as minimal detectable change here—should increase when stream communities become more influenced by disturbance regime (dams, groundwater, urbanization, and other anthropogenic disturbances) because patch dynamics become less influential. Consider this when planning the design. For example is a site expected to improve in diversity over time—then you may want to anticipate this and be prepared to add samples in subsequent years. Conversely if a site is declining in condition, you may be able to reduce replication somewhat once the community is degraded. Hopefully this is not the case.
8. I believe this gives us enough context to discuss the current levels of replication and the effects of changing the level of replication. It seems that all sites except WRMSPR would benefit from an additional sample or two—especially EFORK and BWSNRA. Given the marked cost increase associated with increasing the subsample target from 200 to 500, I would not recommend changing the subsample target count. This would reduce the level of replicates to 2-3 per site to keep the cost equivalent—resulting in a significant loss of power.
9. In the final figures (below) I expressed the amount of change in richness required for statistical significance as percent. The first figure shows the amount of change in richness as % loss from the mean (negative). However, I wanted to overcome the obfuscation caused by scale by expressing the results logarithmically. Since there are no logs of negative numbers, the absolute values of %Δ were used. The grey, green and blue reference lines indicate 100%, 50% and 25% change, respectively. Notice that reducing the level of replication will mean that most sites must loose half of their species before the change is significantly different—this makes a very weak monitoring program, and I do not recommend under replicating.





Recommended Citation for Discussion:

Marshall, B.D. 2025. Limited Power Analysis to Optimize Sampling effort for Macroinvertebrate Communities of the Wood River Basin. River Continuum Concepts Technical Memo RCC2025-02. Prepared for Wood River Land Trust, Hailey, ID 83333.