Ferry County VSP Power Analysis Summary
May 7th, 2020

Levi Keesecker and Brian Cochrane, Washington State Conservation Commission

Summary

- We used freely available water quality data to explore how well it could detect changes between baseline (2011) and 5-years post-baseline (2016) for Ferry County.
- Data were inadequate for detecting changes due to low sample sizes and high variability in the reported data; this is likely in part because data were collected for other purposes not directly related to the Ferry County VSP.
- Reported water quality data were used to estimate sample sizes necessary to detect changes in water quality in the future. The goal of this approach was to provide Ferry County actionable information (e.g., how many samples are needed to demonstrate changing conditions in a statistically robust way?) as they continue/develop their monitoring activities.
- We expect that future on-the-ground sampling by Ferry County VSP will be aided if accompanied by clear questions, logical and thoughtful sample design (e.g., utilizing power analysis), and tight linkages between the questions asked and the data collected.

Introduction

The Ferry County VSP Work Plan defines the following as indicators of water quality (i.e., Critical Area function; pg. 62 Table 5-8):

- Track turbidity relative to baseline 2011 levels
- Track agricultural toxins or nutrients relative to baseline 2011 levels
- Track dissolved oxygen/temperature relative to baseline 2011 levels
- Track agriculture related contaminants relative to baseline 2011 levels

The plan specifies the use of freely available public data for monitoring of some Critical Area indicators. The purpose of this analysis is to determine the applicability of using freely available water quality data from the USGS/EPA Water Quality Data Portal (WQP)[1] for VSP monitoring, to support VSP implementation, and to satisfy the monitoring requirements of the Ferry County VSP plan. Of primary interest is whether or not the freely available data contains enough samples required to demonstrate whether or not changes in water quality are occurring, given the variability of the sample data and effect size desired by the VSP program. In general, larger numbers of samples are required to show smaller

[1]Water Quality Portal (WQP) - https://www.waterqualitydata.us/portal/. "The Water Quality Portal (WQP) is a cooperative service sponsored by the United States Geological Survey (USGS), the Environmental Protection Agency (EPA) and the National Water Quality Monitoring Council (NWQMC) that integrates publicly available water quality data from the USGS National Water Information System (NWIS) the EPA STOrage and RETrieval (STORET) Data Warehouse, and the USDA ARS Sustaining The Earth's Watersheds - Agricultural Research Database System (STEWARDS)….Organizations, including states, tribes, watershed groups, other federal agencies, volunteer groups and universities, submit data to the STORET Warehouse in order to make their data publicly accessible."

differences in water quality, and *vice versa*. Water quality sampling protocols are often created based upon the number of samples that are thought to be needed to demonstrate a change, or not, from baseline conditions.

One approach to this is "power analysis", which essentially estimates the "power" of a data set to confirm or reject a hypothesis. For example, if the hypothesis is "there is no difference in stream turbidity between today and the baseline conditions", then a power test can illustrate how many samples you will need (approximately) to be able to show this in a statistically significant way. Significance, statistically speaking, means that the observed differences are not likely to be caused by chance alone due to the amount of noise in the data and the size of the difference the hypothesis seeks to 'see'.

Using this data and approach, we can compare sets of water quality samples between two conditions (baseline and 5-years post baseline), estimate the number of samples required to determine statistically different differences based on the observed measurements (i.e., observed means), and then estimate the number of samples required to determine statistical differences for a range of water quality changes. We conducted this analysis on data from two sample locations: (i) KNRD_WQX-SFT1 (KNRD), located on the south fork of Tacoma Creek and sampled by the Kalispel Indian Community of the Kalispel Reservation, and (ii) WA_ECOLOGY-60A070 (ECOLOGY), located on the Kettle River near Barstow and sampled by the Washington State Department of Ecology (Fig. 1).

The results for Ferry County indicate that the existing free data from the WQP are insufficient to see a -42% change (from 2.9 NTU pre-2011 to 1.67 NTU post-2011) and a -22% change (from 3.03 NTU pre-2011 to 2.36 NTU post-2011) in turbidity for the KNDR and ECOLOGY stations, respectively (Figs. 1 and 2). Using turbidity as an example, KNRD would need 30 samples (each period) to show a -97% change (from 2.9 NTU pre-2011 to 0.1 NTU post-2011; difference in turbidity = 2.8 NTU, Fig. 4b); ECOLOGY would need 30 samples (each period) to show a -86% change (from 3.03 NTU pre-2011 to 0.43 NTU post-2011; difference in turbidity = 2.6 NTU, Fig. 5b). We found similar results for nearly all water quality characteristics (Figs. 6a, 7a, and 9a), with the exception of dissolved oxygen at the KNRD site (8a), which appears to have not changed across time periods. Ultimately the free WQP data did not contain sufficient sample resolution to detect changes in water quality for most parameters.

While not the desired outcome, options for moving forward include:

- Supplementing the existing data and/or narrowing the sample protocol to obtain less variable data
- Refining the goals/objectives of the monitoring (i.e., what's really of interest?)
- Explore additional tools, methods, and approaches that can measure Critical Area functions and values
- Look for water quality characteristics to measure that are less variable
- Lastly, compare and contrast methods based on cost efficiency and accuracy/precision (e.g. photos, settleable solids [e.g., use of an Imhoff cone], turbidity meters, Total Suspended Solids [TSS; lab method]), but only after clarifying the question(s) the monitoring is attempting to answer

Approach

We selected five water quality parameters that fall within the categories outlined in the Ferry County VSP Work Plan: turbidity, inorganic nitrogen (agricultural nutrient), orthophosphate (agricultural nutrient), dissolved oxygen, and fecal coliforms (agricultural contaminant, if originating from livestock). We selected these parameters because they had among the largest sample sizes of the parameters available in the WQP database.

We first selected all water quality data from the period 7/22/2006 to 7/22/2016. The database is large and expansive, thus we selected only sample locations within streams and rivers. From this subset of water samples we focused on the characteristics above. We pooled data into two time periods: pre-2011 (7/22/2006 to 7/22/2011; baseline) and post-2011 (7/23/2011 to 7/22/2016; 5-years post baseline).

During these time periods, 382 locations were sampled throughout Ferry County. Of these, two locations (KNRD and ECOLOGY) were sampled frequently enough to provide sufficient sample resolution (station/parameter with >10 samples) to the power analysis based on the characteristics outlined above. Stations with fewer than 10 total samples were excluded because sample sizes were deemed too low to provide useful information with respect to power analysis. The stations are highlighted in Figure 1, with KNRD and ECOLOGY differentiated as squares.

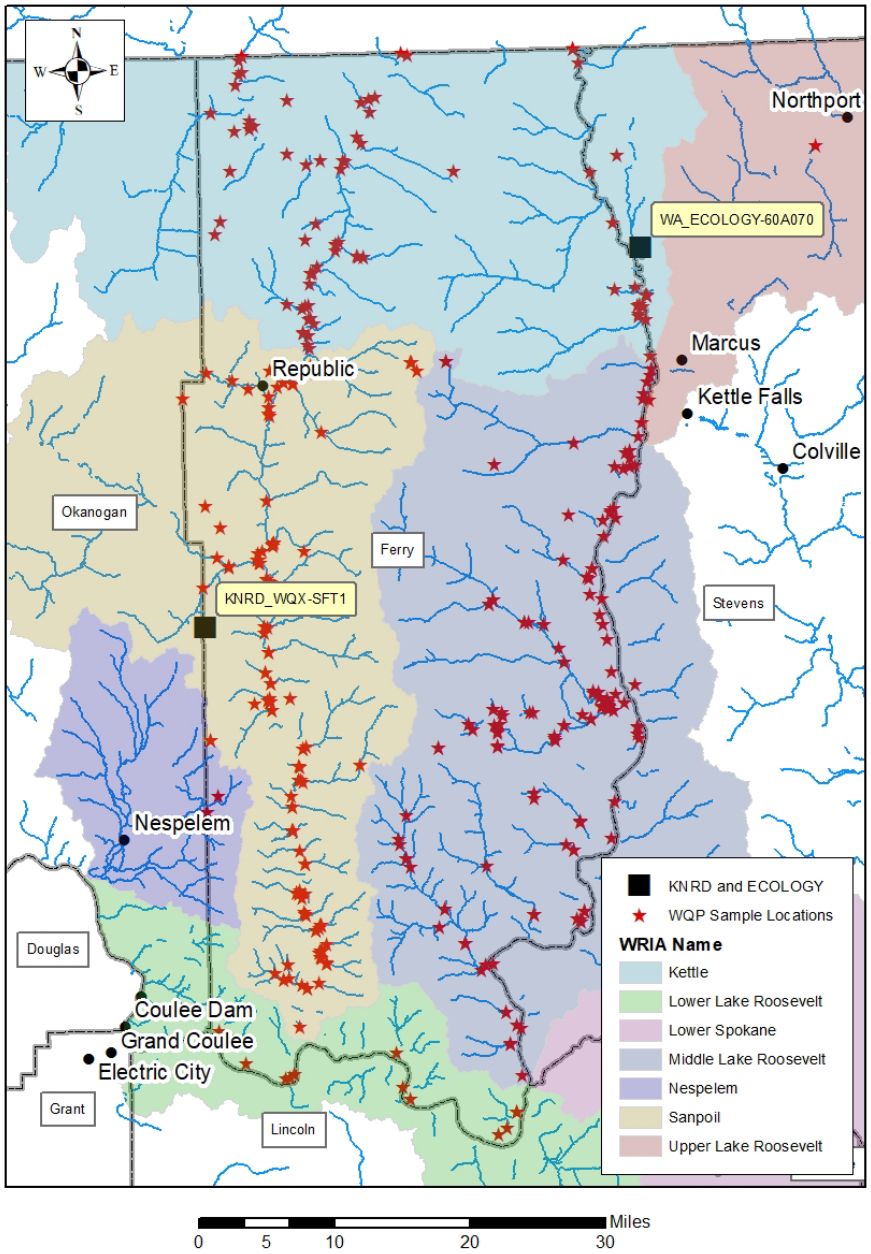


Figure 1. Map of Water Quality Portal (WQP) sample locations with KNRD and ECOLOGY locations differentiated as squares.

Example and Explanation: Case study with turbidity

The first chart (Fig. 2) shows mean levels of turbidity (bars) and with the standard deviation (lines). Standard deviation is a measure of how variable the data is. When the standard deviation is high, relative to the mean, then it is difficult to statistically distinguish patterns. Standard deviation is a measure of how variable (i.e., noisy) the data are. When the variability of the data is lower, the lines become narrower, and it is generally easier to identify differences between groups. If the standard deviation lines overlap, as is the case for turbidity at both stations in both time periods, then no statistical difference between groups can be inferred. More samples usually results in narrower standard deviation lines, thus making it easier to test for true differences.

The numbers above the bars represent the number of samples collected. In this case there were no statistically significant differences in turbidity between sample periods (pre- and post-2011) at either site. You can see this visually where neither bar is outside of its comparison's standard deviation lines. More samples were collected from the KNRD site than the ECOLOGY site; the standard deviation lines are a little narrower in the KNRD site.

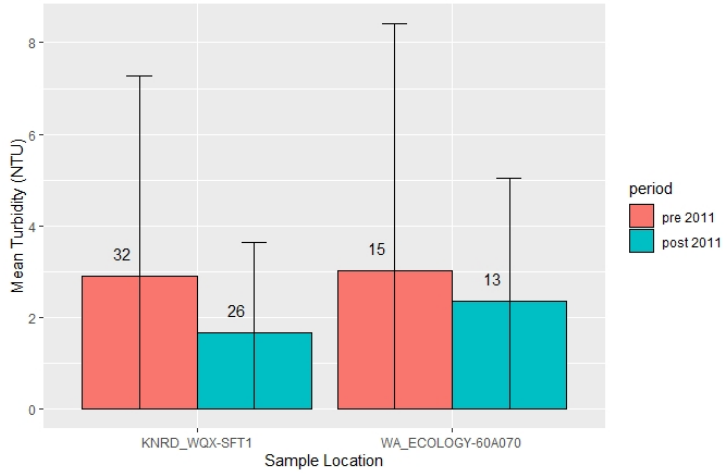


Figure 2. Mean turbidity for the KNRD and ECOLOGY stations comparing pre- and post-2011 sample events. The bars represent mean turbidity; lines represent one standard deviation.

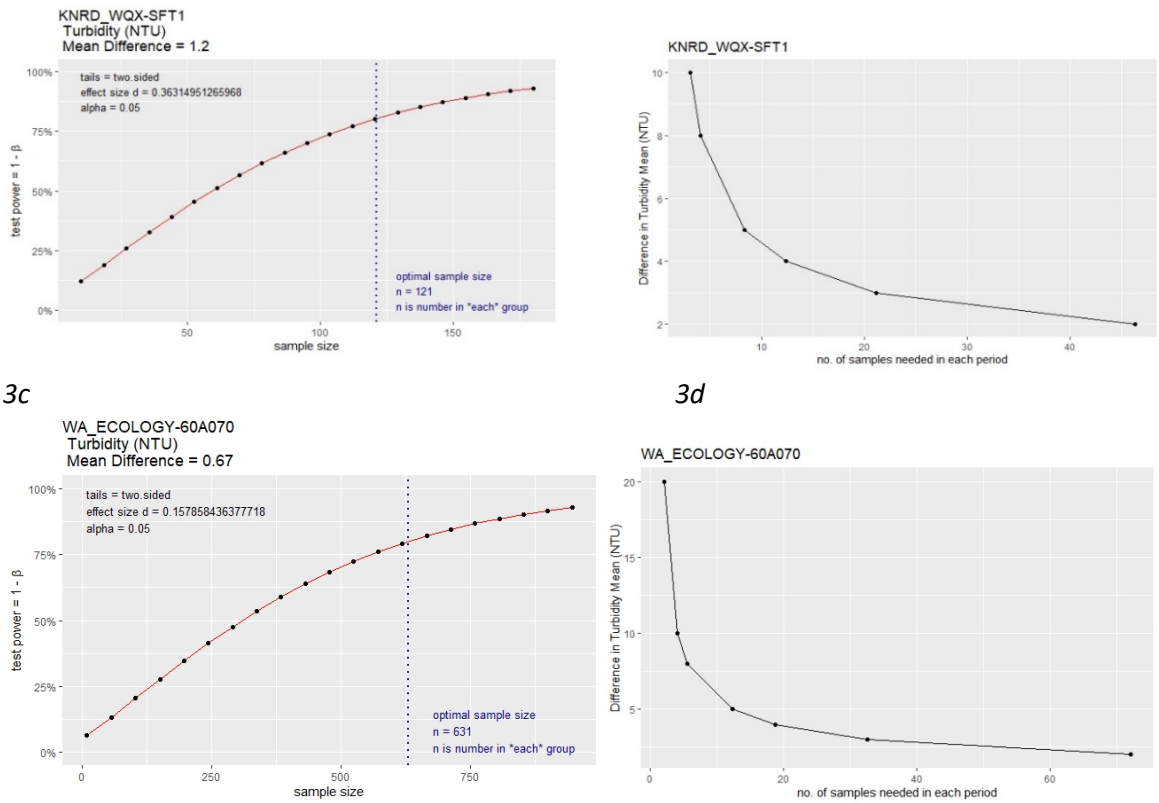
The difference in mean turbidity is also modest at both sites (1.2 and 0.67 NTU). Figures 3a and 3c show how many samples would be needed to determine if the difference in means is significant. About 120 samples during each period (240 samples total) would be required to determine whether pre-2011 turbidity was in fact 1.2 NTUs higher than post-2011 at station KNRD. Only 58 samples were actually collected for turbidity at KNRD, suggesting that nearly 180 more samples would have been needed to assess a change of 1.2 NTU.

Over 600 samples during each period (1,200 samples total) would be required to demonstrate whether pre-2011 turbidity was in fact 0.67 NTUs greater than post-2011. Only 28 samples were actually collected for turbidity at ECOLOGY, suggesting that nearly 1,170 more samples would have been needed to assess a change of 0.67 NTU.

This approach assumes an 80% chance of correctly identifying whether a true difference is occurring (i.e., a power of 0.8) if there indeed is a difference. If a greater power to detect differences is desired, more samples will be required.

Figures 3b and 3d are based on the means and standard deviations of samples for each period at each station. These charts illustrate an estimate of how many samples would be required during each period (baseline and post-baseline) to reject or accept a hypothesis that there was "no change" in turbidity between the two time periods. Note that detecting smaller changes requires more sampling.

3a *3b*



3c *3d*

Figure 3. Results of the power analysis applied to the turbidity data. Figures 3a and 3c estimate the sample size (x-axis) required each period to assess the observed differences in turbidity from the WQP data with a power of 80% (y-axis). The power to detect differences decreases with sample size (dotted black line). Figures 3b and 3d illustrate estimates for how many samples will be needed (x-axis) in each period (pre- and post-2011) to detect a given change in turbidity (y-axis). More samples are required to detect smaller differences (dotted black line).

The magnitude of difference that is important to your county's VSP may vary by resource concern and water quality characteristic. This approach is meant to provide you with some basic guidelines as you establish/continue monitoring efforts. A general rule of thumb is that a *minimum* of 30 samples are desirable as a starting point in water quality sampling – though we do not suggest that 30 samples be the default sample frequency for any given water quality parameter. We recommend that sample frequency be tailored to the question(s) of interest as well as specific indicators chosen (e.g., fecal coliforms may have more variability than dissolved oxygen, thus requiring more samples); the nature of

these decisions should be based on data. Moreover, depending on the water quality characteristic, other factors may influence the ability to assess changes, such as consistency of sampling before/after rainfall and in different seasons, among others. Data here were not thoroughly explored for these factors, but we encourage you to consider these details as you monitor.

Another consideration is the distribution of the underlying data used for power analysis. The underlying mechanics of a power analysis assume the data is "normally distributed" – forming more or less a bell-shaped curve (upper right panel). Figures 4a and 4b represent turbidity and dissolved oxygen across stations and time periods. On the x-axis are the reported water sample values (range: 0-30 NTUs for turbidity, 6-15 mg/L for dissolved oxygen). The y-axis represents the frequency of samples collected at a specific value. The turbidity data does not look normally distributed because of high frequencies of low turbidity levels below the mean, relative to low frequencies of values greater than the mean (i.e., doesn't fit the bell-shaped curve). In contrast, the dissolved oxygen data conform more to a normal distribution. While the dissolved oxygen does not fit the normal distribution exactly, the pattern looks more or less okay; most of the reported dissolved oxygen values are within 1 standard deviation of the mean and there are more or less equal frequencies of samples above and below the mean. Before conducting a power analysis, it is important to first explore the data to ensure it meets the assumption of normality.

The same approach to power analysis was conducted for the remaining water quality characteristics. Summary charts are provided in the Appendix (Figs. 6-9).

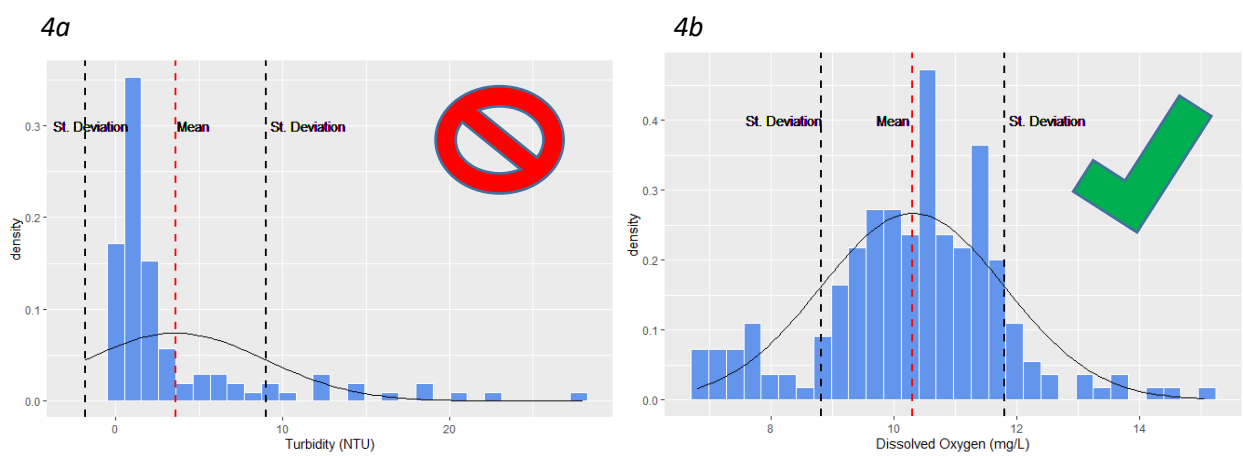


Figure 4. Histograms illustrating the frequency of parameter values reported for turbidity (a) and dissolved oxygen (b).

Conclusion

While we hoped that the freely available data from the Water Quality Portal could aid the Ferry County VSP with respect to monitoring, we conclude that the data is not sufficient to indicate clear patterns (i.e., whether or not there are differences in water quality between pre- and post-2011). Several factors likely contributed to our inability to detect differences, including: (i) the data were collected for different purposes than VSP, (ii) sample sizes were relatively low, and (iii) variability within the data was high relative to the difference in means being tested (e.g., small differences between mean values between periods). We utilized the data to inform future field sampling by conducting a power analysis. Results of

the power analysis indicate that a much larger number of samples will be required to determine whether or not water quality is improving, declining, or not changing. We encourage you to consider these factors as you continue to develop your monitoring approaches. One option for your consideration is the development of an Excel template with the power analysis calculations embedded within it. Such a template could be populated with data collected by your VSP so that you can gauge how sufficient your sampling approach will be for detecting changes. Additionally, SCC staff, and potentially sister agency partners, can provide further input and work with your VSP to develop more targeted monitoring approaches, including sample frequency, locating of sample sites, and sampling protocols.

Appendix

Figure 6a - Inorganic Nitrogen

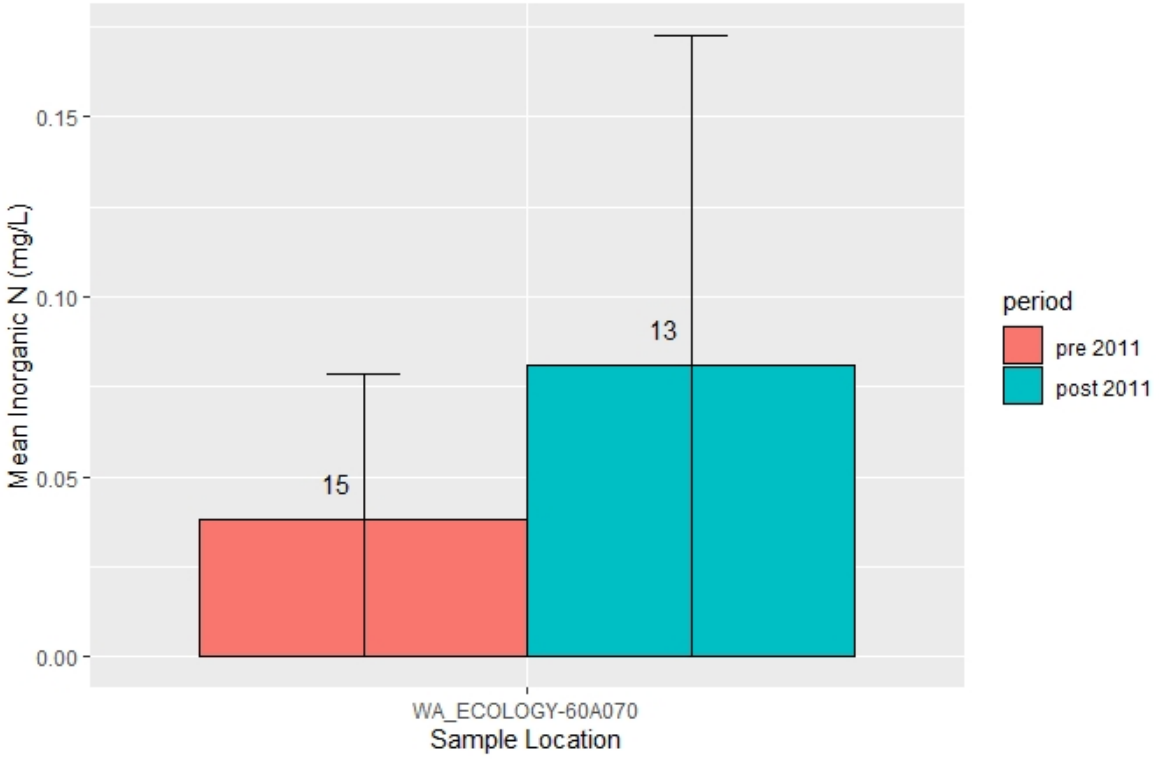


Figure 6b - Inorganic Nitrogen

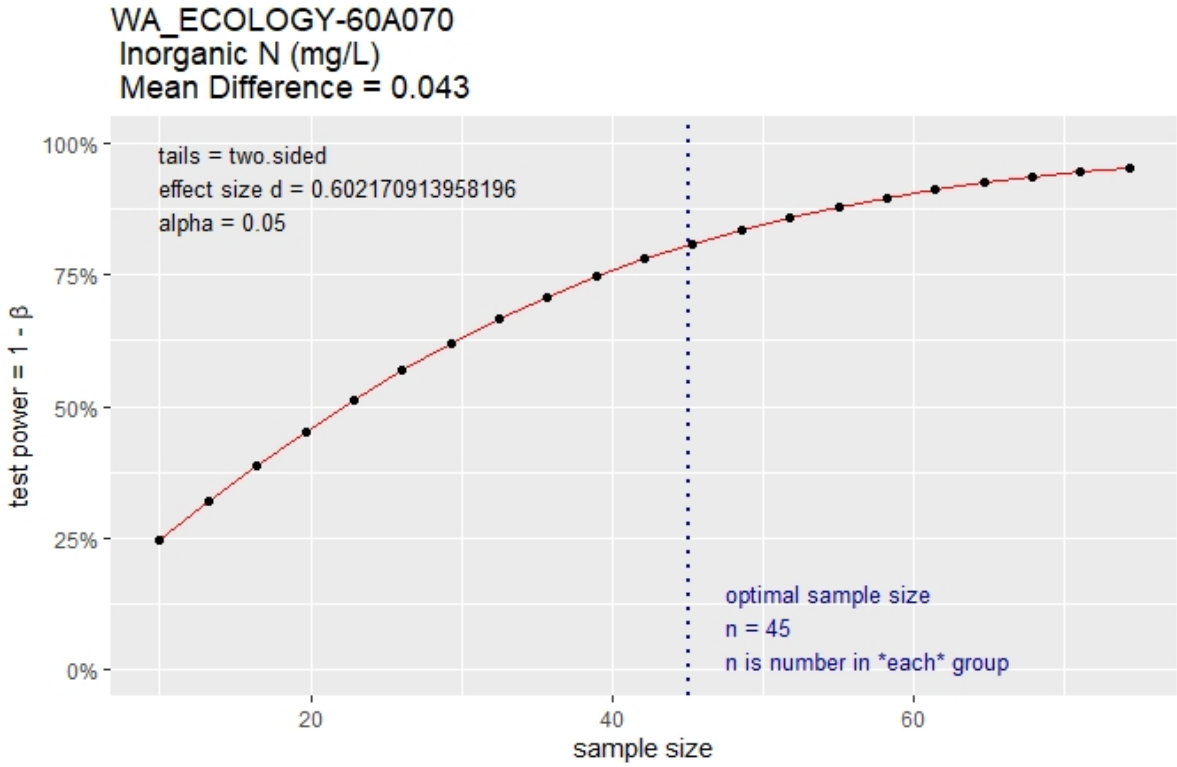


Figure 6c - Inorganic Nitrogen

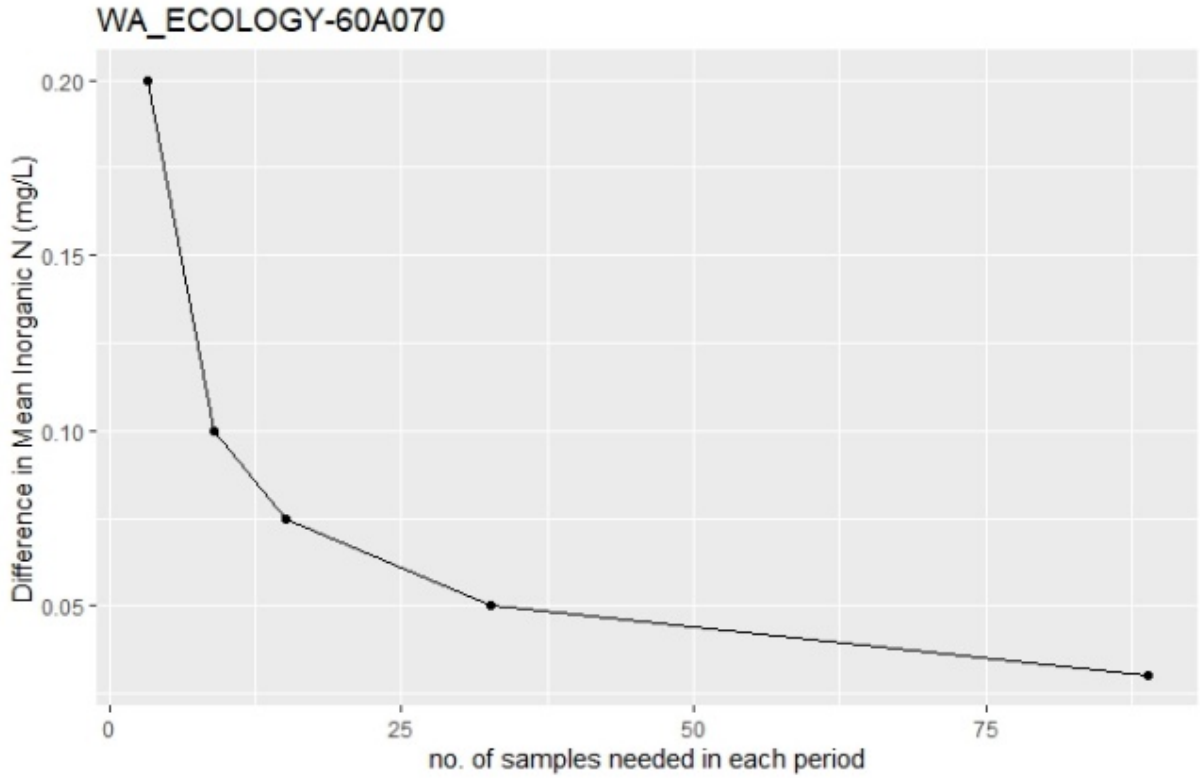


Figure 7a - Orthophosphate

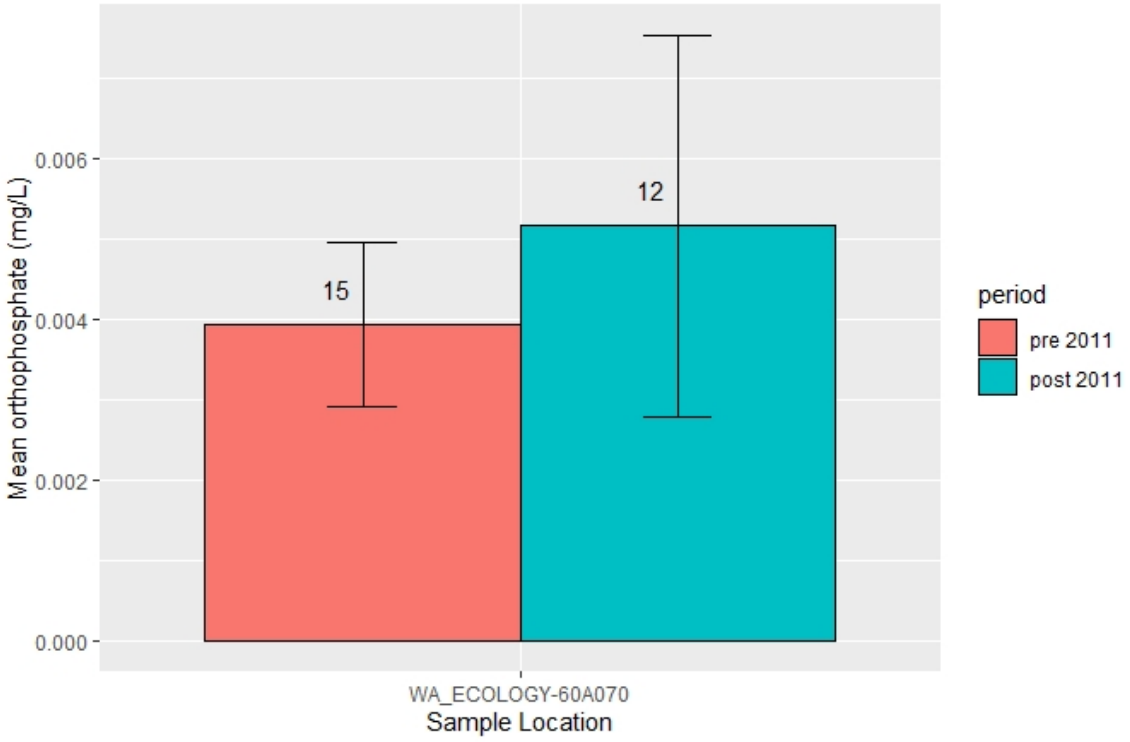


Figure 7b - Orthophosphate

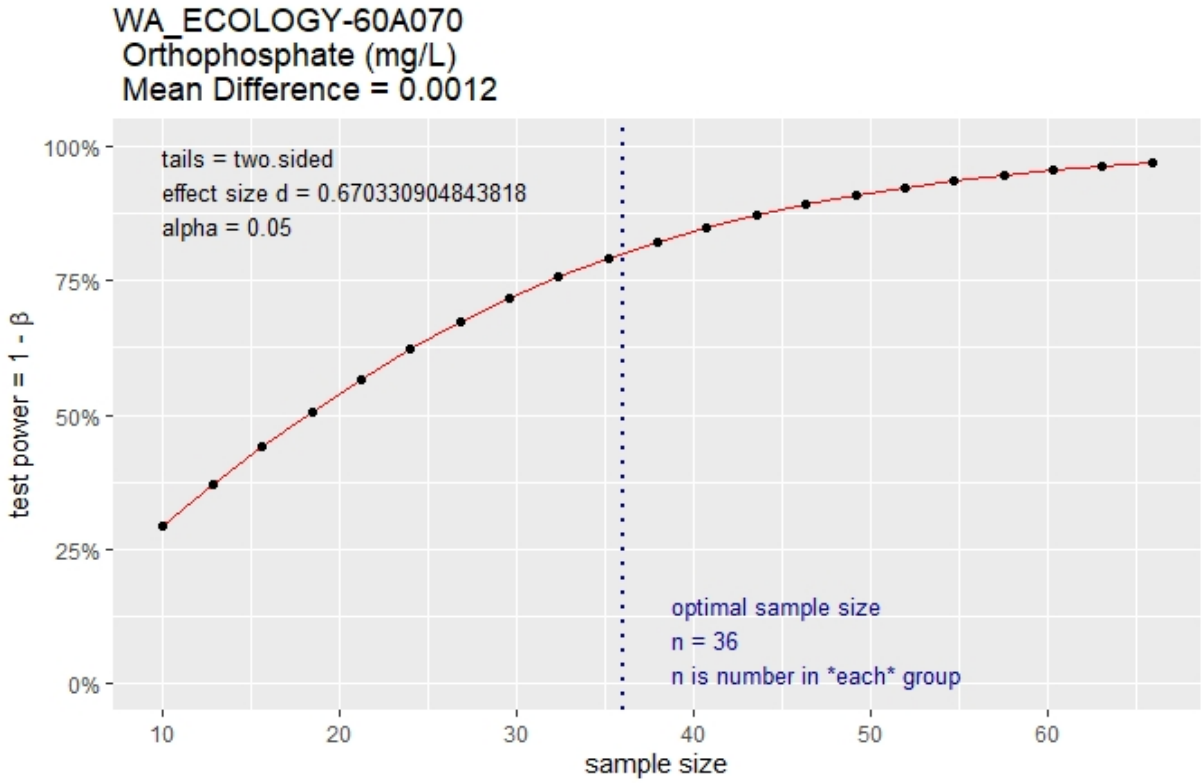


Figure 7c - Orthophosphate

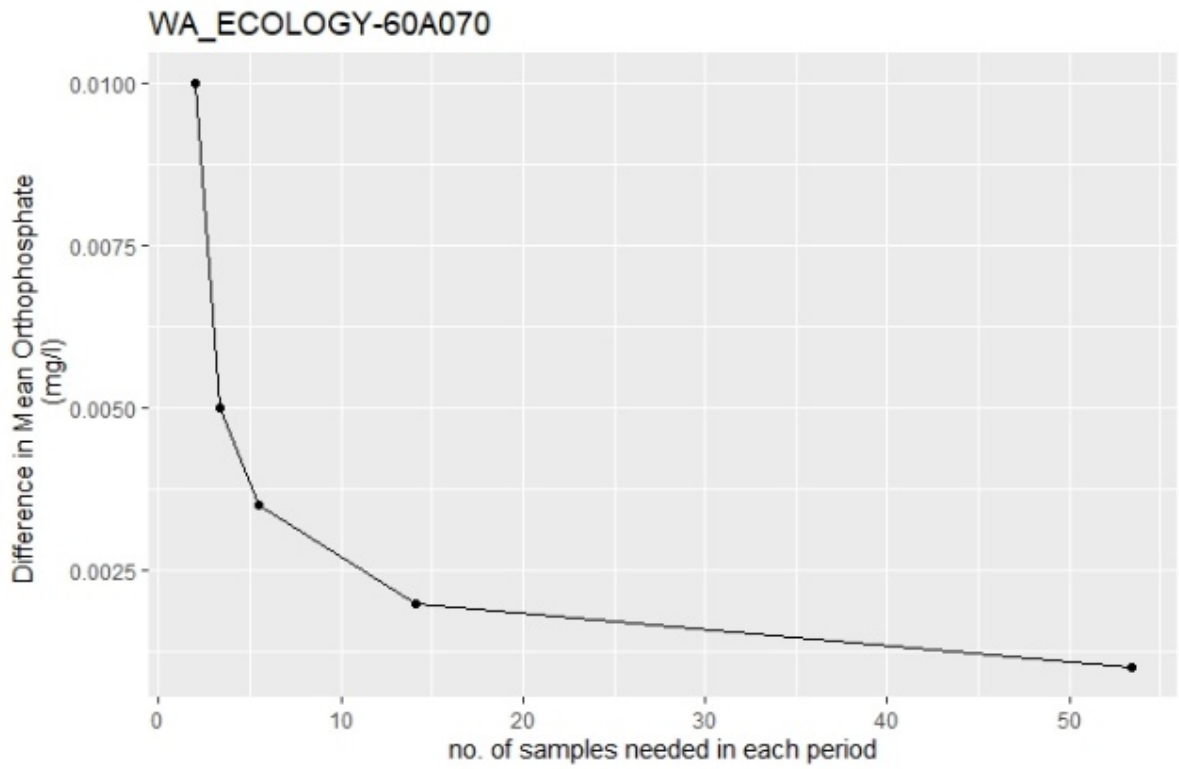


Figure 8a - Dissolved Oxygen

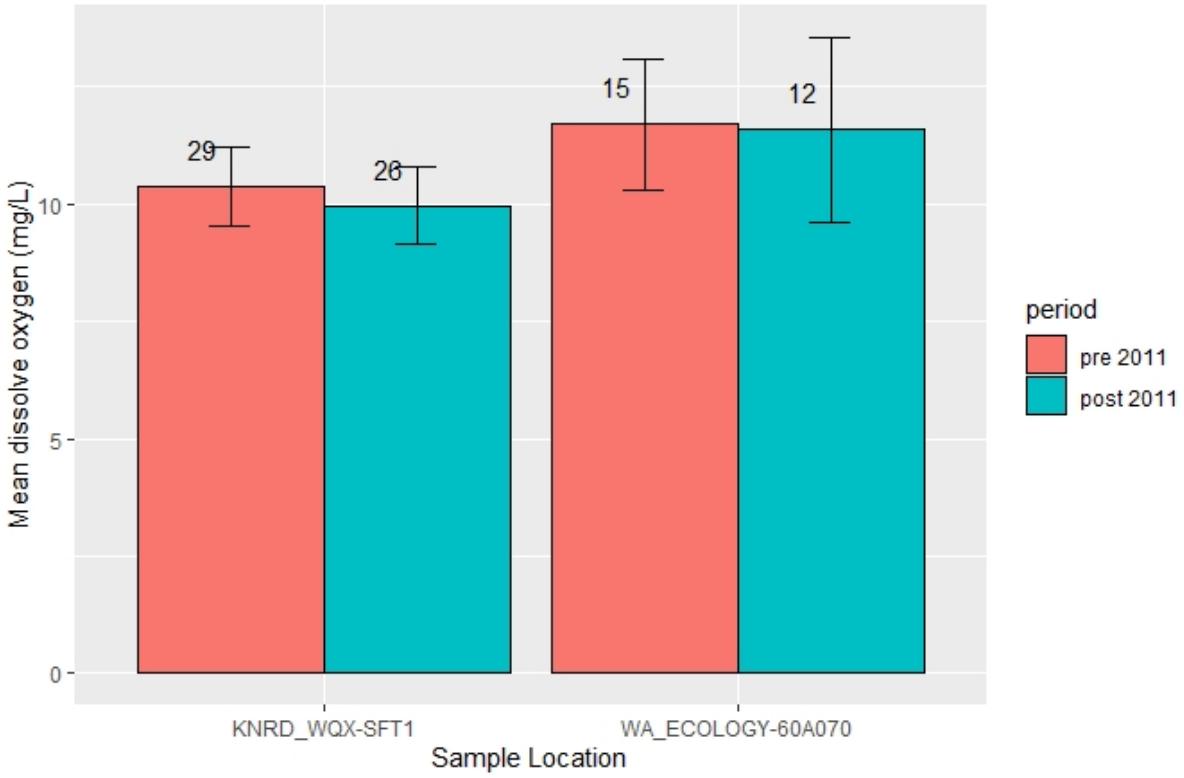


Figure 8b - Dissolved Oxygen

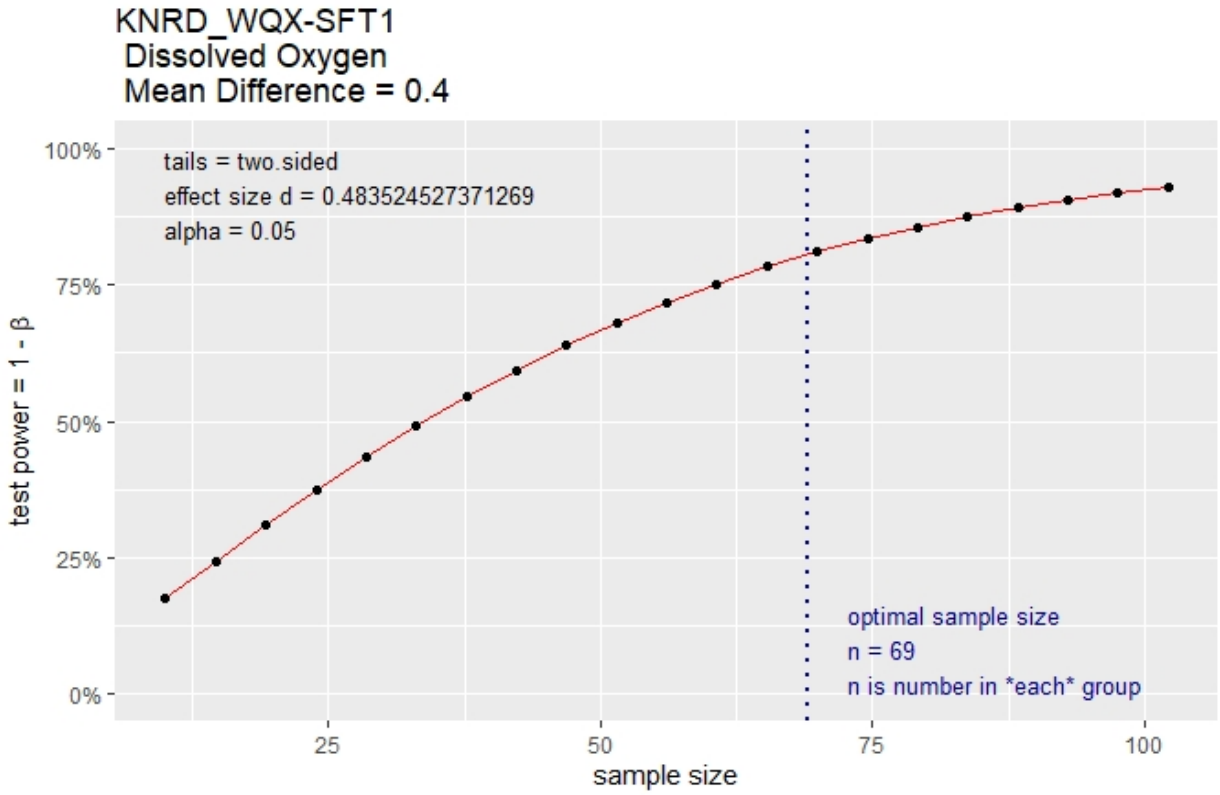


Figure 8c - Dissolved Oxygen

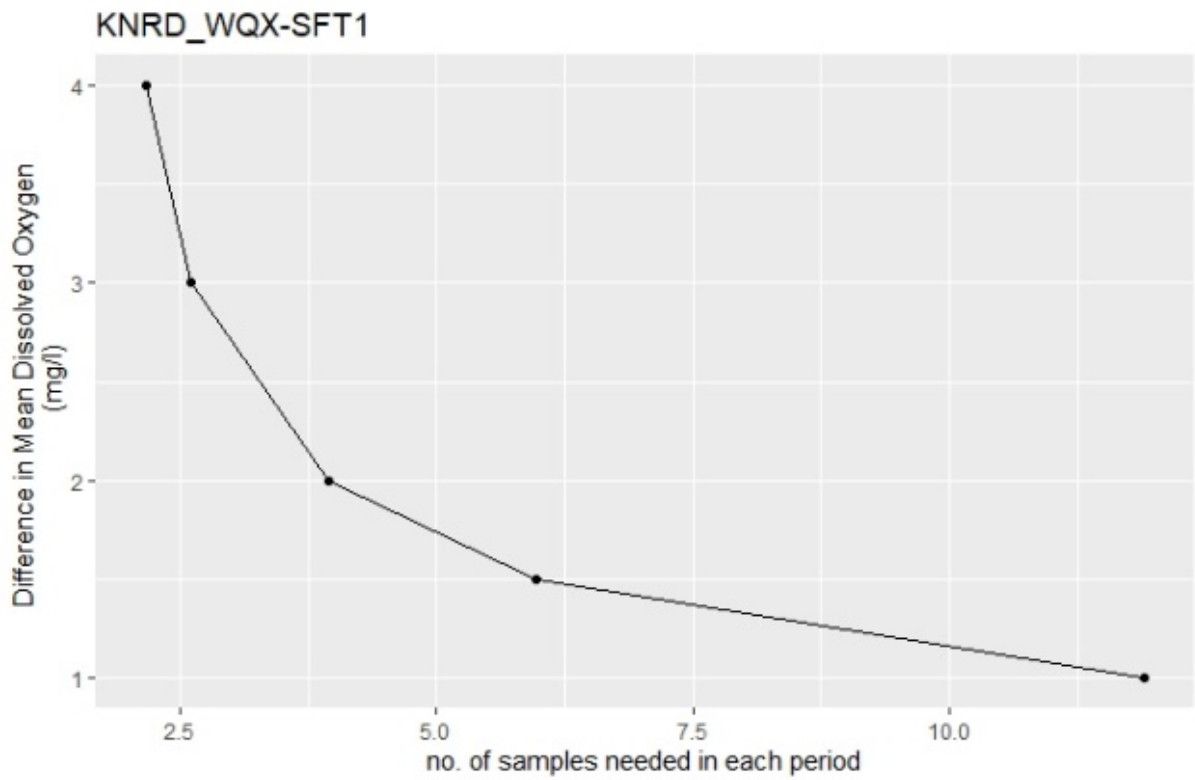


Figure 8d - Dissolved Oxygen

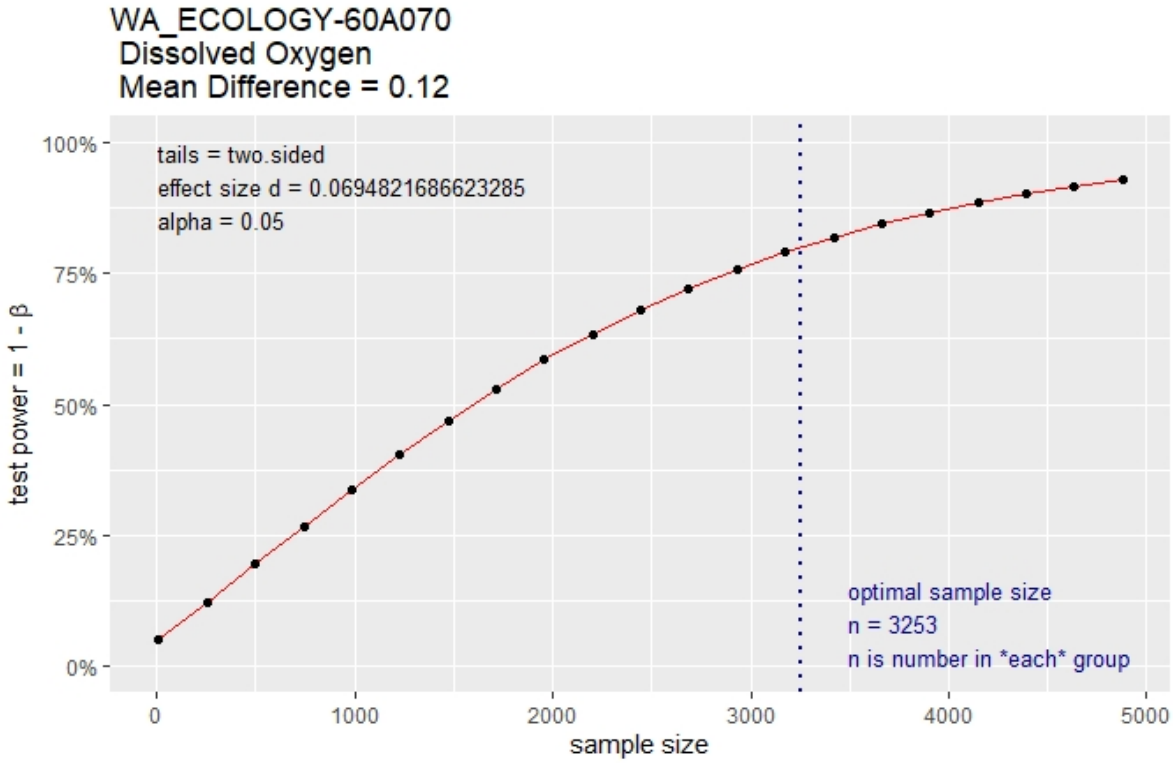


Figure 8e - Dissolved Oxygen

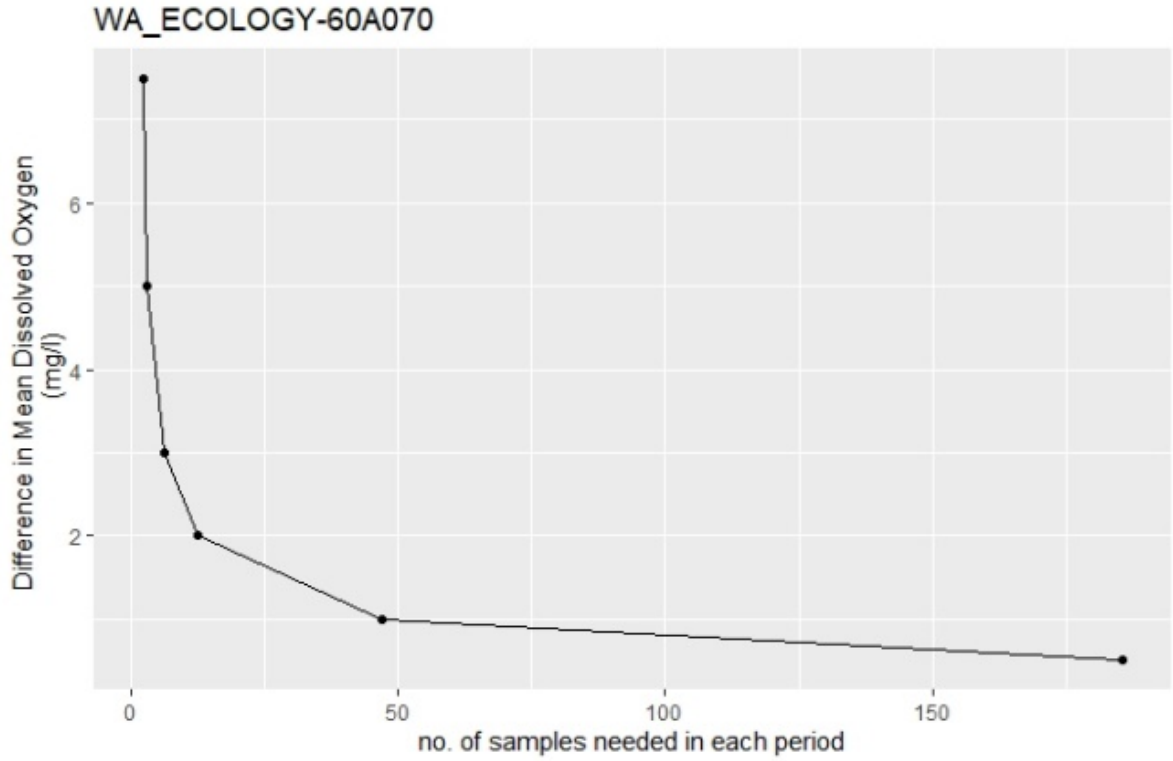


Figure 9a - Fecal coliforms

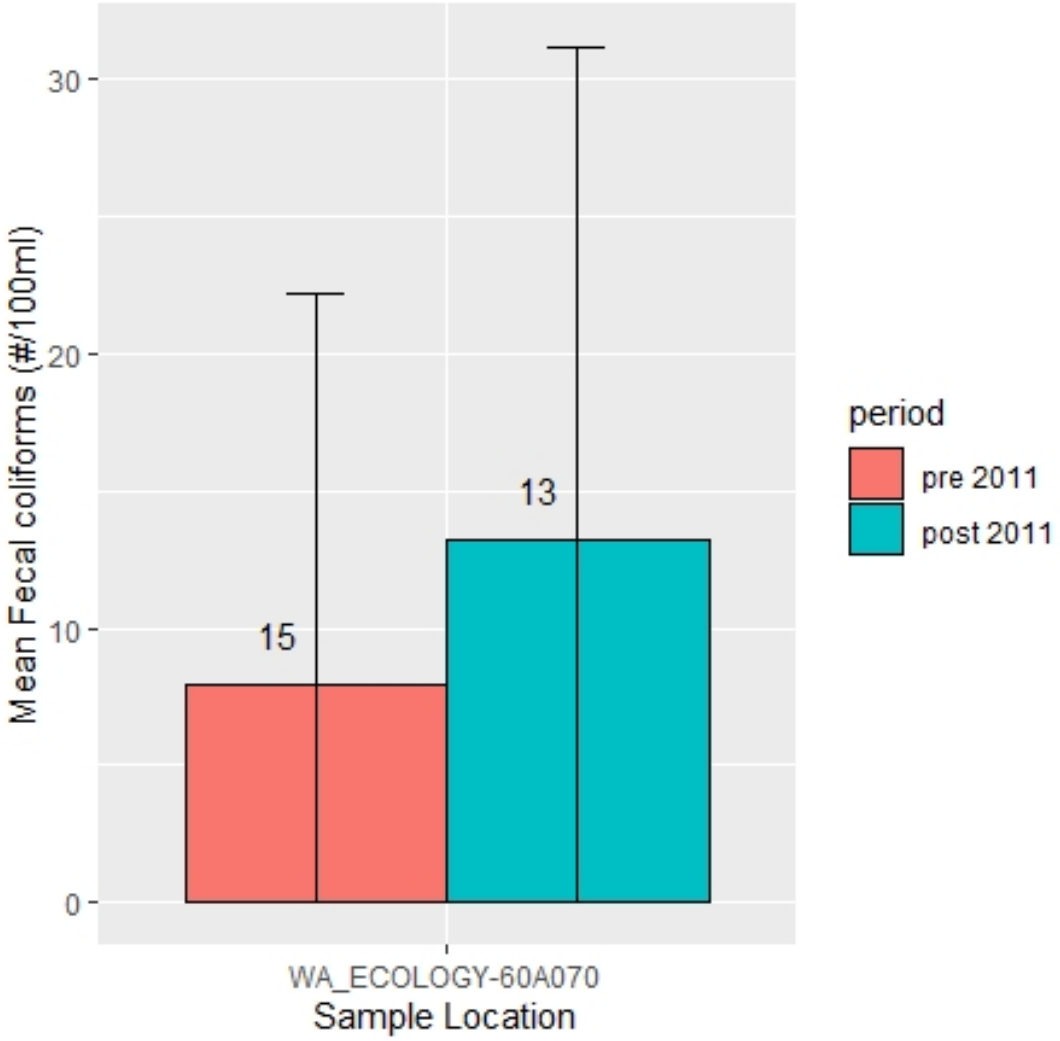


Figure 9b - Fecal coliforms

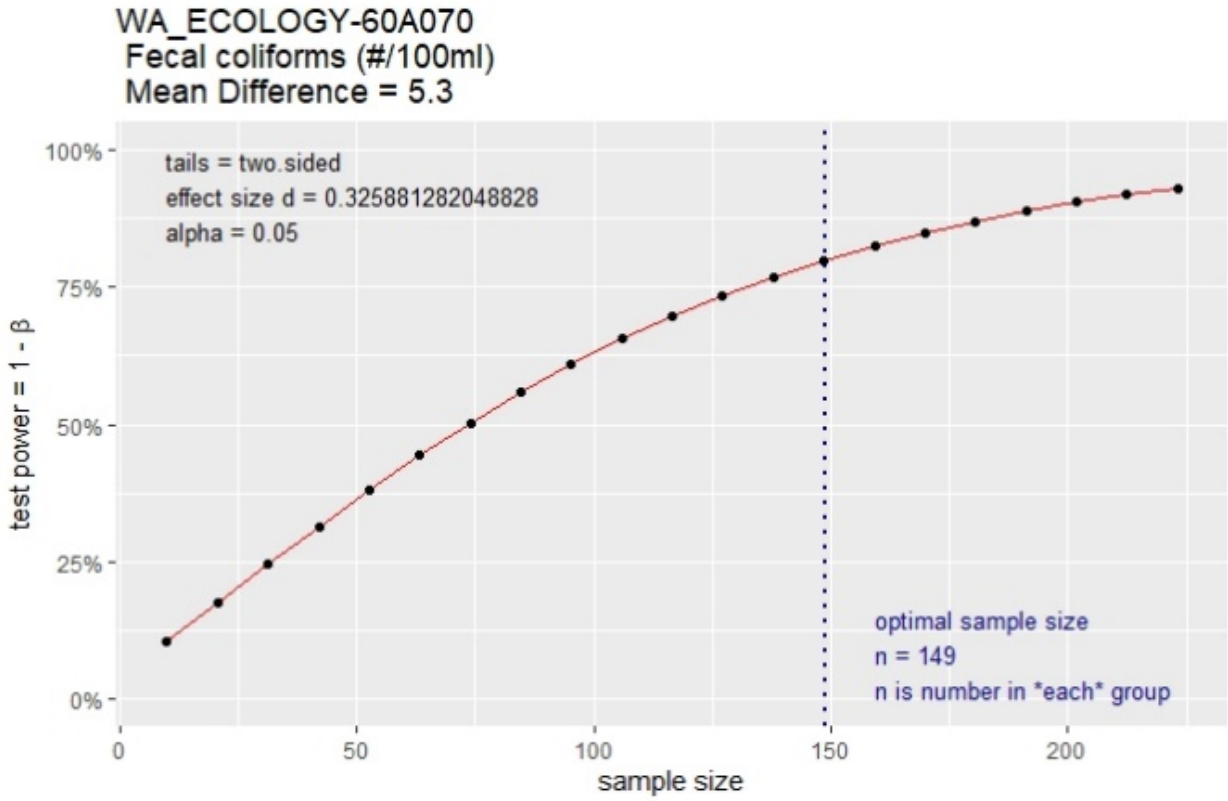
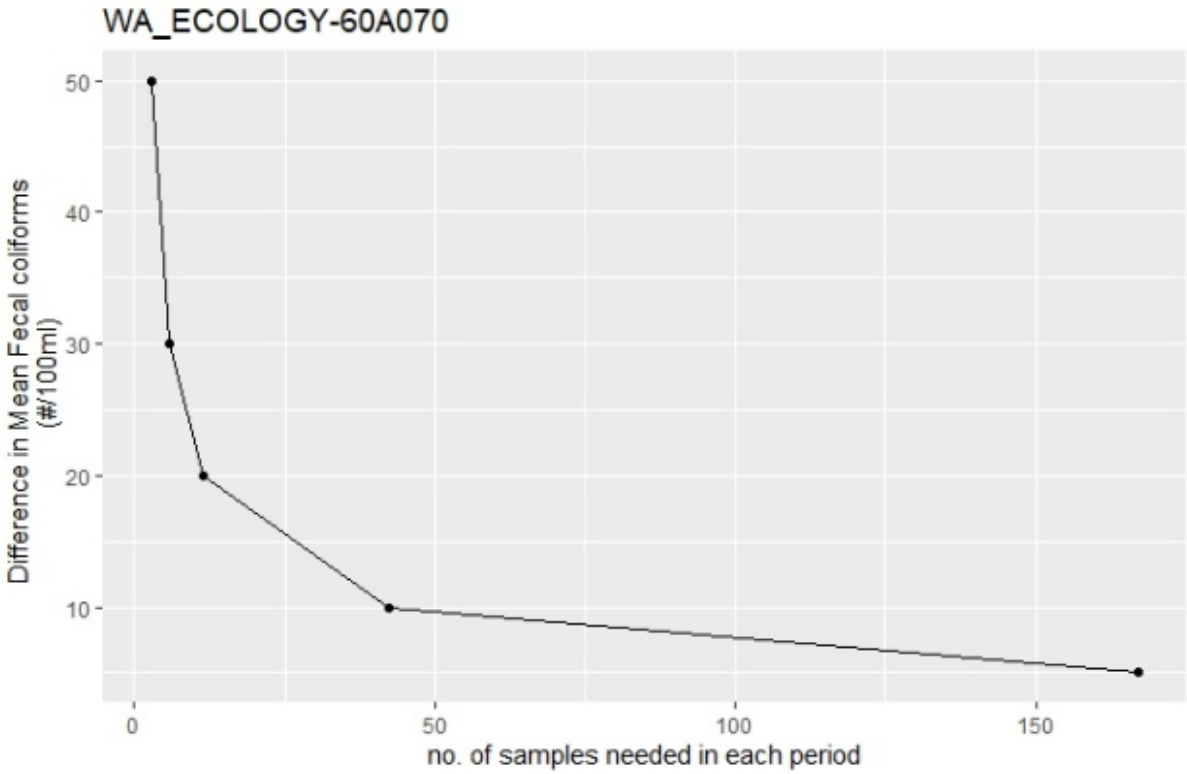



Figure 9c - Fecal coliforms