Data Analysis Review of the “Lower Deschutes River Macroinvertebrate & Periphyton Study”
(by R2 Resource Consultants, Inc., for Portland General Electric)

Reviewed by:
John Van Sickle, Environmental Statistics (May 16, 2016)
For the Deschutes River Alliance.

This review examines the data analytic and statistical methods employed in the macroinvertebrate and periphyton study. The review’s goal is to evaluate whether appropriate methods were used, and whether the report’s conclusions are supported by the results of the statistical analyses.

The study employed 4 types of statistical analyses to support its conclusions: 1) ANOVA and t-tests, 2) Linear regression to test longitudinal trends, 3) Visual inspection of graphical and tabular data summaries, and 4) Ordination of sampled species assemblages. These methods are generally appropriate for the study’s objectives, sampling design, and assemblage data.

However, I found that the methods were misapplied in several cases, leading to questionable results. Most of my comments address two general issues: a) the study did not use widely-accepted statistical methods of community ecology, methods that have been developed expressly to deal with the nonstandard “shape” of macroinvertebrate and periphyton species assemblage data sets (e.g., McCune and Grace 2002, Legendre and Legendre 2012), and b) the study did not recognize and treat lack-of-independence issues, in its applications of ANOVA and regression.

The review is organized as a sequence of distinct topics, followed by a final summary. My comments focus on the analyses of macroinvertebrate assemblage data. They apply equally well to the periphyton assemblage data, which was generally analyzed in the same ways.

Study Objectives and Sampling Design

The study objectives and hypotheses were clearly stated. The sampling design was generally well-conceived and appropriate for those objectives. In particular, by sampling the same river sites Pre-and Post –SWW, the study was able to make a statistically-sound estimate of the change before and after the SWW were begun.

Data Transformations

ANOVA and regression were applied to various abundance-derived metrics, such as invertebrate densities (e.g, Figure 1, Appendix 4). My concern is the great variability, across kick-net replicates, in abundance-based metrics, as indicated by some wide confidence intervals (CIs) in Figure 1, Appendix 4. To satisfy ANOVA assumptions, the study employed the Shapiro-Wilks test of Normality, followed by the awkwardness of outlier removal, in an effort to deal with this variability (p. 42). However, this approach fails to address the more critical ANOVA assumption of variance homogeneity (Miller 1986), which can also affect raw abundance/density data.

To deal with this issue more effectively, it is common to log-transform abundance and density variables prior to ANOVA or regression (McCune and Grace 2002). Likewise, relative abundances of taxonomic groups would benefit more from a log-transform than the arcsine-square-root (p. 51). Log-transformations may not be necessary when analyzing abundance metrics that have been averaged or composited across replicate kick nets (e.g., Figure 6), because those actions reduce variability and trim extreme values.

***** *Two-way ANOVAs of sites, versus seasons or sampling events, for post-SWW sampling.***

The independent-observations assumption of two-way ANOVA requires that each observation be statistically independent from the other observations. However, this study repeatedly sampled the same sites for the different seasons and sampling events, and the repeated observations from each site cannot be expected to be independent.

In addition, the 4 kick-net samples at each of 4 sites are not statistically independent because the 4 kick nets at any one site were collected closely in space and time (within tens of meters, and probably within minutes), relative to the spatial and temporal separation of sampling efforts at different sites. Thus, these kick-net samples might be more accurately viewed as “pseudoreplicates” (Hurlbert 1984) , for the purpose of inferring seasonal differences over the full longitudinal scope of the study. To understand this better, imagine if 4x4=16 kick-net samples had instead been collected at 16 sites along the river, one sample per site. This alternative sampling design would have more strongly supported the study’s desire to draw conclusions that are valid for the entire length of river.

For these reasons, report’s ANOVA p-values are likely to be too small, overstating the statistical significance of seasonal differences.

At a minimum, I suggest treating the seasonal or sampling-event observations as “repeated measures” at each site (e.g., Ramsey and Schafer 1997). There are several ways to analyze repeated-measures data (e.g., mixed models), some of which require larger sample sizes than are available here. For this sampling design, the simplest approach might be to conduct a paired t-test that compares one season (e.g., Fall) to one other season (e.g., Spring), by pairing each site's mean or composited values (to eliminate pseudoreplication). (This is the same approach that the study appropriately used to test Pre-SWW versus Post-SWW metric values.) Then, repeat these paired t-tests for other combinations of two seasons. If multiple pairs of seasons are compared by paired-t testing, then a general-purpose multiple-comparisons correction, such as the Bonferroni, is needed (Ramsey and Schafer 1997).

This same repeated-measures problem confronts the two-way ANOVAs (site by season) done to assess the gravel augmentation effects.

Reporting of ANOVA, Paired Test, and Regression Results

The study reports its statistical testing results solely in terms of p-values. This approach is increasingly unpopular in the applied statistics world, because "naked" p-values convey little information and are hard to interpret (Wasserstein and Lazar 2016).   For example, when presenting results from Table 9, this report makes the very common mistake of interpreting nonsignificant p-values as indicating that "no difference" or "no change" has occurred. However, a nonsignificant p-value actually implies a more vague result, namely, that a difference or change has not been detected in the data (see Nuzzo 2014, or any statistics textbook). Given the small sample sizes and high variability of invertebrate metrics in this report, equating nonsignificance with "no difference" is a potentially serious mistake.

In addition, readers need to assess the biological significance as well as the statistical significance of the study's tests. For example, a downstream trend in macroinvertebrate density may be statistically significant, and yet be too small in magnitude to indicate any meaningful biological change. However, trend direction and magnitude is difficult to estimate from the report's plots. Although the sites on Figure 6 and similar figures are consistently ordered, left to right, in the downstream direction (except, inexplicably, the above-dam reference sites), the figure does not incorprate any between-site distances. This makes it difficult to visualize any trend, or to estimate its magnitude, from the plot. Likewise, trend slopes (e.g., density change per kilometer) are not reported.

Effect magnitudes that were tested by ANOVA are equally difficult to visualize in the report's plots and tables (except for Table 9). To improve this, one could plot the mean and 95%CI of density for each of the site-by-season cases on the same plot, using different colors and symbols, thus displaying the estimated effects that correspond directly to the p-values in site-by-season ANOVAs. Plots that correspond directly to statistical tests should be up front in the report text, with summary data plots such as Figures 6-9 relegated to appendices. This would also help avoid overinterpretations of patterns that are observed in plots but have no tests or CIs, such as the report's statements about observed seasonal differences in functional feeding groups (p. 52, p.97).

Testing Selected Metrics

The study calculated 37 macroinvertebrate metrics (Appendix III). However, some statistical analyses were run on only a small subset of metrics. The report did not explain how this subset was selected, nor did it state that they were selected prior to viewing test results.

At the other extreme, a large number (30) of the 37 metrics were each tested for Pre versus Post-SWW differences (Table 9). This strategy is weak because, after conducting so many nonindependent statistical tests, one would expect at least some of them to be significant ($P < 0.05$) due to chance alone. (The tests are not independent because many of the metrics are not mutually independent, all having been calculated from the same site-by-species abundance data set). Thus, the significances of some test results (in bold) in Table 9 would likely evaporate if the study could be replicated.

The Bonferroni procedure can be used to adjust declarations of significance when performing many simultaneous, nonindependent tests (Miller 1981). A more powerful and reasonable approach is to control the false discovery rate, rather than the Type 1 error rate, of a large number (i.e., a "family") of nonindependent tests (Waite and Campbell 2006).

Having said this, the most effective way to avoid chance significances in metric testing, and to clarify the interpretation of multiple tests, would be to calculate, and to test, a much smaller number of nonredundant metrics. For example, one could adopt the strategy behind the development of multimetric indices, such as the ODEQ index. That strategy assesses a macroinvertebrate assemblage using a small set of minimally-redundant metrics (about 4 to 8), each reflecting a different, biologically-meaningful attribute of the assemblage. By testing far fewer metrics having minimal redundancy, one could argue that there is no need for multiple-testing adjustments.

Paired tests for the mean difference between Pre-SWW versus Post-SWW metrics values

These tests are generally sound, and Table 9 usefully reports the magnitude of the mean difference (which is equal to the difference of the averages). This is fortunate, because the Pre- versus Post-SWW comparison was the study's primary objective. It is too bad that a nonparametric paired-sample test was used, because then one cannot construct corresponding CI's on the mean difference. Constructing such CI's is a big advantage of using the paired t-test instead, which is fairly robust to nonNormality. Also, log-transforming the metric values prior to testing might have enhanced Normality.

Ordination of assemblages

The report uses Principal Component Analysis (PCA) to ordinate the macroinvertebrate assemblage samples from all sites and sampling events. Results are plotted as biplots with sites plotted as points and vectors representing taxa (e.g., Figure 10).

PCA is not recommended for the ordination of species assemblage data, because of the many zeros occurring in a sample-by-species, macroinvertebrate abundance data matrix (McCune and Grace 2002, Legendre and Legendre 2012). In addition, the Euclidean distances used by PCA are highly sensitive to the great variability in species abundances across taxa and samples (often 2-3 orders of magnitude), which creates a highly skewed species abundance distribution. McCune and Grace (2002) discuss these issues in depth.

The report's authors are aware of the many-zeros problem, and they tried to solve it by including only the 35 most abundant taxa in the PCA. This strategy probably reduced the many-zeros problem, but it is unclear that it would ameliorate the skewness problem. The authors did not report any evaluation of how well their strategy worked. In addition, some macroinvertebrate ecologists would argue that, by eliminating less-than-common taxa, as well as truly rare taxa, one is no longer adequately representing how the full community is responding to its environment.

McCune and Grace (2002) present alternative ordination methods designed to tackle the challenges of modeling assemblage abundance data, while including all but the truly rare taxa. The methods are based on nonmetric dissimilarity measures, such as the Bray-Curtis , Sorenson, and chi-squared measures for abundance data, and the Jaccard measure for presence/absence data, that can represent high-richness assemblages more robustly than Euclidean distance. These nonmetric dissimilarities can then be used as input to various ordination methods, of which nonmetric multidimensional scaling (NMS; McCune and Grace 2002) is probably the most reliable and popular among community ecologists. Unlike PCA, NMS makes no assumptions of linear relations between samples and/or species. Its algorithm simply represents samples on a 1-D, 2-D, or 3-D plot, so that between-sample distances on the plot best represent the matrix of between-sample dissimilarities. One can also generate NMS biplots having the same format as Figure 10 (McCune and Grace 2002).

Software note: The book by McCune and Grace(2002) was written as a companion to the commercial “PC-ORD” program that implements its methods. However, nearly all of the book’s methods can also be implemented using the “vegan” package of the free statistical software program, “R” , available at *https://**cran**.r-project.org/*.

Summary

The report’s most important analysis (Box 1, p. 20) is to test for changes in macroinvetebrate and periphyton metric values, Pre- versus Post-SWW. I found this analysis and its results (Tables 9 and 10) to be statistically sound, apart from the problem of performing many dependent statistical tests. All other ANOVA and regression methods in the report would be made more rigorous and accurate by implementing data transformations, by using repeated-measures analyses, by avoiding potential pseudoreplication, by adjusting for multiple testing, and by accurate interpretation of nonsignificant p-values. Finally, the ordination results would be more credible to macroinvertebrate and periphyton ecologists if the report were to use widely-accepted methods tailored especially for the configuration of taxa-rich assemblage data.

References

Legendre, l. and P. Legendre. 2012. Numerical Ecology (3rd ed.) Elsevier, Amsterdam.

McCune, B. and J.B. Grace. 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, OR.

Miller, R.G., Jr. 1981. Simultaneous Statistical Inference (2nd ed.) Springer-Verlag, New York.

Miller, R.G., Jr. 1986. Beyond ANOVA: Basics of Applied Statistics. John Wiley & Sons, New York.

Nuzzo, R. 2014. Statistical errors. Nature 506, 150-152.

Ramsey, F.L. and D.W. Schafer (1997). The Statistical Sleuth: A Course in Methods of Data Analysis. Duxbury Press, Belmont, CA.

Waite, T.A. and L.G. Campbell. 2006. Controlling the false discovery rate and increasing statistical power in ecological studies. Ecoscience 13, 439-442.

Wasserstein, R.L. and N.A. Lazar (2016) The ASA's statement on p-values: context, process, and purpose, The American Statistician, DOI:10.1080/00031305.2016.1154108, available at http://dx.doi.org/10.1080/00031305.2016.1154108