Learned Discourses: Timely Scientific Opinions

Timely Scientific Opinions

Intent. The intent of Learned Discourses is to provide a forum for open discussion. These articles reflect the professional opinions of the authors regarding scientific issues. They do not represent SETAC positions or policies. And, although they are subject to editorial review for clarity, consistency, and brevity, these articles are not peer reviewed. The Learned Discourses date from 1996 in the North America *SETAC News* and, when that publication was replaced by the *SETAC Globe*, continued there through 2005. The continued success of Learned Discourses depends on our contributors. We encourage timely submissions that will inform and stimulate discussion. We expect that many of the articles will address controversial topics, and promise to give dissenting opinions a chance to be heard.

Rules. All submissions must be succinct: no longer than 1000 words, no more than 6 references, and at most one table or figure. Reference format must follow the journal requirement found on the Internet at http://www.setacjournals.org. Topics must fall within IEAM's sphere of interest.

Submissions. All manuscripts should be sent via email as Word attachments to Peter M Chapman (peter_chapman@golder.com).

Learned Discourses Editor

Peter M. Chapman
Golder Associates Ltd.
200-420 West Hastings Street
Vancouver, BC V6B 1L1
peter_chapman@golder.com

SETAC's Learned Discourses appearing in the first 7 volumes of the SETAC Globe Newsletter (1999–2005) are available to members online at http://communities.setac.net. Members can log in with last name and SETAC member number to access the Learned Discourse Archive.

In a Nutshell. . .

Ecotoxicology

Pseudoreplication in ecotoxicology, by Marcos Krull, Francisco Barros, and Michael Newman

Examples of pseudoreplication in ecotoxicology, and recommendations to address this issue, are provided.

Sampling and Modeling Issues

Influences of subsampling and modeling assumptions on the USEPA field-based benchmark for conductivity, by Shaun A Roark, Craig F Wolf, Grant D De Jong, Robert W Gensemer, and Steven P Canton

The subsampling and modeling assumptions made by Cormier et al. (2013) lead to biases that challenge the accuracy of the final conductivity benchmark of 300 μS/cm.

When is enough sampling enough?, by Mark F Shibata and Natasha T Hasumann

The importance of determining sampling sufficiency a priori is illustrated by a case study.

How do you deal with an organic sediment sample for benthic macroinvertebrate community analyses?, by Fernanda Lage and Ana Lúcia Brandimarte

Benthic surveys in water bodies dominated by aquatic plants need to carefully examine what may be living associated with aquatic plants but not directly associated with the sediments.

Environmental Risk Assessment

Environmental risk assessment of shipwrecks: A fault-tree model for assessing the probability of contaminant release, by Hanna Landquist

Shipwrecks can have short- and long-term environmental, economic, and social issues whose risks need to be assessed for appropriate management decision-making.

Human Health

Fish and shellfish consumption: Traditions, regulations, and a cleaner environment in the US Pacific Northwest, by Jennifer Stiles and Ruth Sofield

Uncertainties in risk assessments including balancing risks and benefits, and biological organisms that don't always follow the rules, are discussed.

DOI: 10.1002/ieam.1455

PSEUDOREPLICATION IN ECOTOXICOLOGY

Marcos Krull*† Francisco Barros† and Michael Newman‡
†*Universidade Federal da Bahia, Salvador, Brazil*
‡*Virginia Institute of Marine Science, Gloucester Point, Virginia, USA*
*marcoskpl@gmail.com

DOI: 10.1002/ieam.1440

As defined by Hurlbert (1984), pseudoreplication occurs during ''the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (although samples may be) or replicates are not statistically independent.'' Inferential statistics have been used for decades in ecotoxicology with different approaches such as analysis of variance and linear regression to calculate the values of no observed effect concentrations (NOEC)/lowest observed effect concentrations (LOEC) and 50% effect concentrations (EC50), respectively, and their confidence intervals. Routine use of standard protocols and guidance in ecotoxicology can unintentionally foster the propensity of researchers to take less time and effort to properly design experiments that produce the right data to answer the question. It is important to understand that conventional protocols are intended to provide general guidance, not a complete recipe for a successful experiment. This Learned Discourse aims to draw attention to one of the most common design errors in ecotoxicology.

In ecotoxicology, there are 2 different types of replicates that often are confounded: (i) true or experimental replicates, which are sampled at the experimental unit level and used in inferential statistics (e.g., aquariums, beakers, or even sampling sites); and, (ii) sampling replicates for which repeated measurements are taken (e.g., fish, algae, or samples of water/sediment). Thus, choosing the appropriate experimental and sampling units will depend on what question you intend to answer. For instance, if you want to infer something about sediment quality of an estuary, your experimental unit will be sampling sites within that estuary. However, if you want to evaluate the quality of a specific site, then your experimental unit will be independent samples of sediment within that site (see Hurlbert 1984, for further examples).

Sampling replicates could also be provided from field samples that are split into subsamples, perhaps as part of an analytical procedure and quality assessment. As such they would be considered sampling replicates, not experimental replicates. They are not statistically independent and their utility is primarily for estimating method precision. When using conventional inferential statistics, replication is required at least at 1 level (i.e., experimental unit). Replication at other levels may be unnecessary (Hurlbert 1984). Although sampling replicates might be optional, they can increase precision within each treatment and could be important in some designs. They have been used, in most cases, during ECx evaluation in which the sum value of each treatment is used in hypothesis testing. Therefore, true and sampling replicates are intended for different reasons.

Sampling replicates are not independent, so they are not appropriate for analysis of variance (ANOVA) analyses such as those for the dubious NOEC estimation. The current use of NOEC, if true replication is used, does not provide direct information on the precision of the method. If pseudoreplicates have been used to estimate the NOEC values for decision makers, we might have been making misleading inferences based on inadequate sample designs and improper use of inferential statistics. For instance, the mean value of laboratory replicates may be used in EC50 evaluation only if 1 concentration is independent from another although, in reality, most of them are not. For example, if we are testing the toxicity of an effluent, we sample a gallon of raw effluent from an equalization storage tank and from this gallon we make serial dilutions. Because none of these samples were independent, the only answer we could address is about that particular gallon of water which is related to a specific and very limited space and time.

To make sure that samples are independent, each replicate should be randomly sampled and stored in different containers so that you have independent samples from the population you are trying to understand, that is, the effluent. Collecting a single composite sample would not completely avoid the problem.

Pseudoreplication might also be common in sediment testing. If replication is made from only 1 sample of reference and 1 of contaminated sediments, then the null hypothesis being tested is ''there is no difference between these 2 samples'' instead of ''there is no contamination effect.'' Logistical issues may compromise collection of independent sediment samples; however, researchers should be aware of the compromise and temper their conclusions accordingly.

There are other potential problems related to pseudoreplication, such as taking multiple measures from the same experimental replicate. Among the most common examples are: 1) taking several subsamples from the same micro/mesocosms and treating them as true replicates (Landis et al. 1997), 2) evaluating different groups of organisms from the same aquarium in multispecies tests as if they were independent (e.g., algae, bacteria and protozoa), 3) measuring multiple endpoints from the same test (e.g., survival, growth and reproduction), and 4) measuring multiple biomarkers from the same set of organisms.

Taking the latter as an example, different tissues are frequently sampled from the same group of individuals to evaluate different biomarkers. However, they should not be considered as independent measures as they can be highly correlated (Quinn and Keough 2002). Using ANOVA analysis for each biomarker as if they were independent measures is not appropriate if we are making inferences about possible effects of contamination in this specific environment. Unless experiment-wise error rate adjustments are made, this approach would increase the chance of Type I error. Once again, the only response we could have is about the sampled individuals or which is the best biomarker to be tested in this specific site, but not much about the system from which they are taken. Nevertheless, it would be infeasible in most cases to collect such large numbers of organisms to meet the independence assumption of ANOVA. Furthermore, a high number of true replicates are usually prohibitively costly, such as in multispecies tests.

As suggested by Landis et al. (1997) multivariate methods are more suitable for data analyses of multispecies experiments. In these cases, multiple measures from the same experimental replicate could be used for risk evaluation using multivariate analysis of these measures (e.g., ANOSIM). Multiple samples from the same replicate could also be used to increase precision or to evaluate internal system variability; however, only the sum or the mean values of these subsamples should be used in hypothesis testing. Moreover, continuous experimental designs in which replication at the same treatment are generally not mandatory have been an option for ecotoxicologists being criticized for pseudoreplication (Belanger 1997). However, as discussed above, continuous designs are not necessarily free of pseudoreplication and should be carefully applied.

We believe that pseudoreplication is common in the field of ecotoxicology. To avoid this issue, we make some recommendations for both scientists and editors, some of which are very similar to Hurlbert (1984): 1) sample and dilution procedures should be clearly described in papers so editors and readers can evaluate potential problems related to pseudoreplication, 2) better training in experimental design and statistics is needed, 3) clear statements of what question(s) you intend to answer are necessary, because small changes in sampling design influence the validity of testing the hypothesis of interest, 4) make sure your design involves true replication and interspersion of treatments, and 5) protocols and guidance should provide detailed information on true and laboratory replication, if they are to be used (e.g., OECD 2006).

REFERENCES

Belanger SE. 1997. Literature review and analysis of biological complexity in model stream ecosystems: Influence of size and experimental design. *Ecotox Environ Safe* 36:1–16.

Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. *Ecol Monogr* 54:187–211.

Landis WG, Matthews RA, Matthews GB. 1997. Design and analysis of multispecies toxicity tests for pesticide registration. *Ecol Appl* 7:1111–1116.

OECD (Organization for Economic Cooperation and Development). 2006. Current approaches in the statistical analysis of ecotoxicity data: A guidance to application. OECD Environmental Health and Safety Publications, Series on Testing and Assessment No. 54. Paris (FR). ENV/JM/MONO(2006)18.

Quinn G, Keough M. 2002. Experimental design and data analysis for biologists. New York (NY): Cambridge University Press. 556 p.

INFLUENCES OF SUBSAMPLING AND MODELING ASSUMPTIONS ON THE US ENVIRONMENTAL PROTECTION AGENCY FIELD-BASED BENCHMARK FOR CONDUCTIVITY

Shaun A Roark,* Craig F Wolf, Grant D De Jong,
Robert W Gensemer, and Steven P Canton
GEI Consultants, Inc., Denver, Colorado, USA
*sroark@geiconsultants.com

DOI: 10.1002/ieam.1438

Cormier and Suter (2013) proposed a field-based method to derive protective aquatic life thresholds (i.e., ‘‘benchmarks’’) as an alternative to traditional laboratory-based methods (i.e., Stephen et al. 1985). The method was applied using paired stream benthic macroinvertebrate and water quality surveys to derive a benchmark for conductivity (300 μS/cm), applicable to Appalachian Region waters dominated by salts of sulfate and bicarbonate at neutral to mildly alkaline pH (USEPA 2011; Cormier et al. 2013). We have significant concerns with this benchmark regarding causation, confounding factors, and mechanisms, but we focus here on inherent characteristics of the macroinvertebrate data and arbitrarily selected model parameters that substantively affect the resulting benchmark.

The data used by Cormier et al. (2013) were collected by the West Virginia Department of Environmental Protection (WVDEP) to develop the West Virginia Stream Condition Index (WVSCI). Cormier et al. (2013) narrowed the data set from 2668 to 2210 samples based on ecoregion, watershed size, ionic mixture, and pH. The total number of genera included was also narrowed from approximately 500 to 163 by requiring that any genus included occur in ≥25 samples, including 1 reference station. To adjust for uneven sampling density (more samples at mid-range conductivity), the conductivity range was divided into 60 even (log10) bins, and samples in each bin were weighted in inverse proportion to the number of samples in the bin. For each genus, the 95th centile of the weighted cumulative distribution function (CDF) of conductivity was determined for samples where the genus was present; this was termed the ‘‘extirpation concentration’’ (XC95). The 5th centile of the ranked XC95 values for 163 genera (the HC05) was set as the aquatic life benchmark for conductivity.

We obtained the same WVDEP database and replicated as closely as possible the data set and calculations of Cormier et al. (2013) using R 2.15.2 (R Development Core Team 2013). Our XC95 estimates for each genus matched those of the authors, and our HC05 of 297 μS/cm was similar to the result of Cormier et al. (2013) of 295 μS/cm. Nonetheless, based on our review, we conclude that 300 μS/cm has little relevance as an effect threshold for macroinvertebrates in these streams.

First, the macroinvertebrates in the WVDEP database were enumerated in a manner inconsistent with the implicit assumption of Cormier et al. (2013) that absence, the foundation of the extirpation coefficient, from any location is known. Specifically, invertebrates were enumerated on fixed-count subsamples of 200 ± 40 individuals from each sample. Absence from a station was inferred based on absence from this small subsample; with a median sample size of 1500 in the database, there is a significant likelihood of missing taxa in a subsample of only 200. Whereas fixed-count subsampling is appropriate for developing regional multimetric index scores, it underestimates true taxa richness because less abundant taxa are missed in the subsampling procedure (Courtemanch 1996). Therefore, an undeterminable number of locations had taxa present that were not captured in the WVDEP subsample and, as a consequence, the inference of absence by Cormier et al. (2013) is biased.

To evaluate the subsampling effect on the XC95 and HC05, we used a resampling approach based on the WVDEP data set. Using 250 iterations each, we subsampled the macroinvertebrate data set (all genera in 2213 samples) with fixed-count subsamples ranging from 100 to 200. For each genus, smaller subsamples resulted in absence at more stations and, in turn, smaller subsamples resulted in a lower and less certain (i.e., more variable) HC05 (Figure 1A). We conclude that, if the original WVDEP method had used a larger subsample, the HC05 would have been greater than 300 μS/cm.

Second, the number of bins used for weighting across the conductivity range affects each XC95 and the final benchmark value. Cormier et al. (2013) used a weighted CDF for each XC95 to account for uneven sampling density across the range of observed conductivity. While weighting can be an appropriate method to account for uneven sampling, no sensitivity analysis of the weighting method was presented, and no explanation was provided for the selection of 60 bins. We found that the HC05 varied with the number of bins selected (Figure 1B), and that using an unweighted CDF the conductivity benchmark was 347 μS/cm. It is not clear from our analysis what weighting approach, if any, or number of bins would be appropriate, but it is clear that the weighting and the number of bins selected affects the HC05.

Finally, Cormier et al. (2013) also made the decision to require that any genus used in the model occur in ≥25 samples and in at least 1 reference location. These rules reduced the number of genera included in derivation of the benchmark from approximately 500 to 163. The authors stated this was to ensure ‘‘reasonable confidence in the evaluation of the relationship between conductivity and the presence or absence of a genus.’’ They also stated, ‘‘the benchmark varied within <5% when SSD models were constructed from 20 or more occurrences of each genus, whereas the benchmark steadily became lower when XC95 values were derived from <15 occurrences.’’ However, the