The Multiple-Comparison Trap and the Raven's Paradox – Perils of Using Null Hypothesis Testing in Environmental Assessment

Song S. Qian* Thomas F. Cuffney†

June 1, 2018

Abstract

Detecting and quantifying environmental thresholds is frequently an important step in understanding ecological responses to environmental stressors. We discuss two statistical issues often encountered in threshold detection and quantification when statistical null hypothesis testing is used as a main analytic tool. The hidden multiple-comparison trap (leading to a much higher risk of a false detection) and Raven's paradox (rendering a "detection" meaningless) are often obscured when hypothesis testing is used as part of a more elaborate model, especially models based on computer-intensive methods. Using two examples, we show that the hidden multiple-comparison trap can be exposed using computer simulation to estimate the probability of making a false detection and the Ravens paradox can be avoided by clearly stating the null and alternative hypotheses using scientific terms to substantiate that the rejection of the null is equivalent to proving that the alternative of interest is true. The hidden multiple comparison trap implies that a null hypothesis testing based on a computer-intensive method should be used with caution. The implications of Raven's paradox requires that we focus on providing evidence supporting the proposed hypothesis or model, rather than seeking evidence against the frequently irrelevant null hypothesis. These two problems, and many others related to null hypothesis testing, suggest that statistical hypothesis testing should be used only as a component of the body of evidence, perhaps, as the devil's advocate.

Keywords: deviance reduction methods, null hypothesis testing, statistical inference, threshold

*Department of Environmental Sciences, The University of Toledo, Toledo, OH, USA, Email: song.qian@utoledo.edu

†USGS South Atlantic Water Science Center, Raleigh, NC, USA

1 Introduction

Null hypothesis testing (NHT) is widely used in environmental assessment and management. Open any issue of this journal and we can find examples where NHT is used as the main analytic tool for inference. Anytime we see a p-value, an NHT is implied. The use of NHT is, however, controversial. The controversy is often centered around the misuse and misinterpretation of the key statistic, the p-value. In 2014, the journal *Ecology* published a number of papers in a forum on the use of p-value and model selection (Ellison et al. 2014). The American Statistical Association issued a statement on p-values (Wasserstein & Lazar 2016) arguing, through six principles, that the proper use of null hypothesis testing requires a careful consideration of the context of the problem, the data collection process, and the purpose of the study to avoid misuse and misinterpretation of the p-value. With increased awareness, we are now better at discovering and avoiding misuses of p-values. However, many statistical issues of NHT are not quite obvious. In this paper, we discuss two well-known issues that are rarely mentioned in environmental and ecological studies – the hidden multiple-comparison trap and Raven's paradox. These issues are of particular concern in methods used for detecting and quantifying ecological thresholds because of the increased use of NHT in conjunction with computer-intensive methods in this area. We introduce these issues with two examples from published papers and discuss how to detect and avoid them. Our discussion goes beyond the strict definitions of these two issues to represent similar misuses and misinterpretations of NHT.

1.1 The Multiple-Comparison Trap

The issue of multiple comparisons concerns the increased likelihood of incorrectly rejecting the null hypothesis (a statistically significant result) when multiple tests are carried out. In a one-way ANOVA problem, the simplest multiple comparison scenario is to carry out all possible pairwise t-tests. The problem arises when the ANOVA null hypothesis (i.e., all means are equal) is not rejected and the pairwise t-tests use the same significance level of $\alpha = 0.05$. For each t-test, we have a 5% chance of erroneously rejecting the null hypothesis ($H_0 : \mu_i = \mu_j$). The chance of at least one erroneous rejection (among the multiple t-tests) is, however, larger than 0.05. In other words, we are more likely to find a statistically significant difference when we make multiple comparisons even when all the means are the same. This problem is often masked when only the largest difference among all pairwise comparisons is tested. Although recognized in the 1940s (Tukey 1949), the multiple comparison problems did not receive wide attention until the 1970s, marked by the publication of the Holm-Bonferroni method (Holm 1979). The issue came back in the 1980s and 1990s as part of the study on false discovery rate (e.g., Benjamini & Hochberg (1995)). In detecting an environmental threshold, the issue of multiple comparisons can often be hidden because the threshold is frequently "detected" through computer-intensive methods, for example, an automatic search for a maximum (or minimum) "difference." Such a search can be hidden in the computer program that exhausts all possible comparisons. But the consequence is the same – the stated type I error probability of $\alpha = 0.05$ is an underestimate

of the actual type I error probability (probability of obtaining a false detection). In practice, an NHT is not always used when employing a computer-intensive method for threshold detection. Specifically, when using algorithms similar to the classification and regression tree (CART) approach, we often do not apply an NHT on the resulting change point. As a result, the chance of a false detection of a threshold is even higher than when an inappropriate NHT was used because a CART-based algorithm will always "detect" a change point, regardless of the existence of a threshold response.

1.2 Raven's Paradox

Raven's paradox, also known as the Hampel's paradox of confirmation, is a paradox about what constitutes evidence for confirming a hypothesis. The logical contrapositive of the statement "all ravens are black" is "everything that is not black is not a raven." In mathematics, proof by contrapositive is an oft-used inference method. That is, we can show "If not B then not A" to support "If A then B." The significance of Raven's paradox is that it shows that the proof by contrapositive method can be paradoxical because it suggests that observing, for example, a white shoe is an evidence supporting the claim that all ravens are black (a white shoe is a non-white and non-raven object). Good (1983) proposed a Bayesian explanation of the paradox by stating that the weight of evidence provided by observing a white shoe (or any non-raven and non-black object) is positive, but small if the number of ravens is small compared to the number of all non-black objects. But how is the paradox relevant to statistical hypothesis testing? The misuse of NHT in many fields is very similar to proof by contrapositive. For example, in a one-sample t-test problem, we are interested in showing that the alternative $\mu \neq 0$ is true, but the test provides evidence against the null hypothesis of $\mu = 0$. We often want to know, for example, whether μ is equal to a specific value μ_a. The Bayesian explanation establishes that $\mu = 0$ is not true provides some, but not much, evidence to support $\mu = \mu_a$. Consequently, using the rejection of $\mu = 0$ to test $\mu = \mu_a$ is a misuse of the t-test. A statistical paradox is often a manifestation of a misuse of a method or misunderstanding/misinterpretation of the result. In most CART-based threshold detection methods with an NHT, the null hypothesis is invariably a hypothesis of no change (the response does not change along an environmental gradient). Using a significant result (rejecting such a null hypothesis) as a proof of the existence of a threshold is akin to showing a white shoe as proof of the blackness of a raven. We use Raven's paradox to dramatize such a misuse of hypothesis testing in detecting environmental thresholds.

The practical ramifications of these two issues are the increased likelihood of false detections of a threshold. A false detection may be a result of at least two scenarios. First, the target stressor is not the primary cause of ecological degradation (hence the null hypothesis is true) and a false detection is a result of a type I error. Second, an ecological response to the stressor does not have a threshold pattern (e.g., increasing linearly) and a detected "threshold" is irrelevant. In both scenarios, a false or irrelevant threshold will result in ineffective management practices.

2 Methods

Many statistical results are counterintuitive because we are not accustomed to thinking in terms of probability or average behavior (Kahneman & Tversky 1972). Because the concept of probability is easily explained based on a long-run frequency, Monte Carlo simulation is often the best way to expose hidden problems of a statistical procedure. A Monte Carlo simulation can be designed to replicate the theoretical characteristics of a method or model, such that we can examine the outcome under known scenarios. For an NHT, we often use simulations to verify the test's prescribed behavior. Specifically, a test's type I error probability should be at the commonly accepted level of 0.05. In fact, testing an NHT's type I error probability is a standard method of verifying a test's robustness against departures from various assumptions, a practice started well before the advent of high-performing personal computers (see, for example, Box (1976)). Using a simulation, we repeatedly generate data from the distribution defined by the null hypothesis and carry out the test. The test should reject the null hypothesis about 5% of the time, the predetermined acceptable type I error probability. When a test is subject to the multiple-comparison trap, we will reject the null hypothesis more frequently. Using Monte Carlo simulation, we can approximate the type I error probability of a test.

Raven's paradox is about the use of irrelevant evidence to support a claim. In the context of threshold detection and quantification, we use the term Raven's paradox to highlight the fact that the null hypothesis of no change is mostly irrelevant to the goal of detecting a threshold. The best practice to avoid the paradox is to clearly state the null and alternative hypotheses of a test in scientific terms. The question we must ask is whether the rejection of the null hypothesis leads to the validation of the alternative hypothesis in which we are interested. Without clearly stating the null and alternative hypotheses, we cannot readily answer this question. This is why Qian (2014*b*) recommended verbally describing a statistical model, even for a simple t-test.

2.1 Examples

We use two examples to illustrate the use of simulation for assessing the false detection probability, as well as to explain why the paradoxical nature of the results from such applications can be exposed when the null and alternative hypotheses are verbally described. The CART-based threshold method in (Qian et al. 2003) uses the CART model as the basis to select a split point along a gradient. The CART-selected split point or change point is the point that results in the largest changes in the response variable deviance, a statistical measure of variation. The selected split point signifies the likelihood of a change in the response variable.

Instead of using the statistical deviance as a criterion for selecting the change point, the algorithm of Baker & King (2010) uses an indicator species approach (Dufrêne & Legendre 1997) for determining a potential change point along a gradient. A key parameter of the approach is an indicator value (IV) of a taxon, which measures the association of the taxon with a group of sites. It is the product of the taxon's specificity (ratio of the abundance of

the taxon in one group over the abundance in all groups) and fidelity (frequency of occurrence of the taxon in the group). In this case, sites are diveded into two groups separated by the change point along a gradient. The ecological relevence (or the lack thereof) of the IV is discussed by Cuffney, Kashuba, Qian, Alameddine, Cha, Lee, Coles & McMahon (2011).

Both Qian et al. (2003) and Baker & King (2010) have been cited more than 160 times according to Google Scholar. Before 2012, most papers citing Qian et al. (2003) used the method in some way, many of them for supporting the development of environmental standards. After 2012, Qian et al. (2003) was mostly cited as a reference for using the methods of Baker & King (2010). Many applications of these methods used environmental monitoring data across multiple spatiotemporal scales with a goal of setting environmental standards.

2.1.1 Using CART-based Method for Detecting and Quantifying Ecological Threshold

One of the two methods for detecting an environmental threshold in Qian et al. (2003), the deviance reduction method, uses a classification and regression tree (CART) model. The basic idea is to find a breakpoint in a measure of ecological response along an environmental gradient to reduce the total deviance of the response variable. Specifically, the deviance reduction method is a CART model with one predictor variable (the gradient) and the estimated threshold is the first split point. Qian et al. (2003) included a generic χ^2-test for testing the statistical significance of the estimated threshold. The null hypothesis of this test is that the response variable distribution is the same across the entire gradient and the alternative hypothesis is that the response variable has two distributions, separated by the split point. The -2 log-likelihood-ratio (null over alternative) has a χ^2 distribution if the null model is true. The degrees of freedom (df) of the χ^2 distribution is the difference between the df of the null model and the df of the alternative model. For example, when the response variable is a count variable approximated by the Poisson distribution, the null hypothesis describes the response variable using a single Poisson distribution with one parameter and the alternative hypothesis describes the response variable using two Poisson distributions (two parameters). The χ^2 test df is then 1. If the response variable distribution is normal or log-normal, χ^2 test df can be either 2 (assuming both mean and variance are different before and after crossing the split point) or 1 (assuming a constant variance). When assuming constant variance, the test is reduced to a two-sample t-test for normal response variables.

The test can, however, be deceiving. The CART model selected split point is the point that results in the largest reduction in deviance, among all possible split points. This process is analogous to selecting and testing the largest difference among all pairwise comparisons in an ANOVA problem. Because the CART model does the selection "behind the scene," the fact that the resulting split point is selected among all possible split points (i.e., multiple comparisons) is "hidden." Although we set the test significance level at 0.05, the actual probability of making a type I error (erroneously rejecting the null hypothesis and claiming a threshold response) can be much higher, resulting in a test that rejects the null hypothesis

far more frequent than we expected. Many authors citing Qian et al. (2003) used the R (S-Plus at the time) code distributed by Qian, but most did not use the χ^2-test. The CART-detected change point is directly used as the threshold in many cases when the method is used for establishing an environmental standard. Miltner (2010) correctly argued that the χ^2-test should not be used because the threshold was not pre-selected. However, graphical depictions of the data and results (Figures 2-5 of Miltner (2010)) showed no clear evidence of a threshold response. Because the CART-based method will always detect a change point regardless whether a threshold response is appropriate (Qian 2014*a*), our simulations represent an optimistic scenario, where the NHT provides an inadequate regulation of the outcome. Without the test, the likelihood of a false discovery of a threshold is higher.

We use two Monte Carlo simulation approaches for this example, assuming that the response variable has a normal distribution. For this example, the null hypothesis assumes the response variable has the same normal distribution along the environmental gradient. We can draw random numbers from this normal distribution and pair them to randomly drawn gradient values. This step produces one realization of the data under the null hypothesis. We then apply the CART model to find the split point and conduct a two-sample t-test to compare the means of the two groups created by the split point. The process is repeated many times, each time we record whether the null hypothesis is rejected (or the p-value is below the predetermined significance level of 0.05). The proportion of times we reject the null hypothesis is an estimate of the actual probability of making a type I error. Alternatively, we can use an existing data set to carry out the simulation by randomly permuting the observed response variable data along the gradient, thereby removing the gradient effect. In the online supporting materials, we provided a simple R function to carry out both simulations. In the same simulation programs, we also estimate the probability of detecting a change point, regardless of the statistical significance.

2.1.2 Permutation Test for Species Aggregation

The second example is a component of a computer program using an indicator of species aggregation to detect a threshold along an environmental gradient (Baker & King 2010). In this program, occurrences of a group of ecologically relevant species are used as a measure of a community level response to environmental changes. At the species (or taxon) level, the program measures species aggregation using IV of Dufrêne & Legendre (1997). Specifically, the computer program finds all possible split points along the gradient to divide the species count data into two groups. For each possible split, an IV value is calculated. The split point that resulted in the largest difference in indicator values between the two groups is selected as a potential threshold. (These taxon-specific thresholds are then combined to develop a community-level threshold.) A permutation test is used to determine whether the taxon-specific potential threshold is "statistically significant." The underlying mechanism of leading to the multiple-comparison trap is the same as the first example – a statistical test is applied to the two subsets of the data selected to have the maximum difference in IV, among all possible groupings. In this analysis, we focus on the permutation test at the taxon level.

Baker & King (2010) is cited more than 160 times according to Google Scholar. The majority of these papers implemented the method using the R code accompanying Baker & King (2010) with the permutation test built-in for each taxon. In at least one case (Cao et al. 2016) the CART-based method was also applied on a community level indicator to serve as a comparison.

Although we did not find a clear description of the null and the alternative hypotheses of the test in the original paper and in all papers using the method, the use of a permutation test implies a null hypothesis that the distribution of the response variable (abundance) is not a function of the gradient. For a specific split point, the sample sizes of the two subsets (n_1 and n_2) are known. The permutation test tabulates (more specifically, samples from) all possible permutations of splitting the data into two subsets of sample sizes n_1 and n_2. For each permutation, an IV is calculated. The collection of these IVs is used to form an empirical distribution, and the p-value is calculated as the fraction of these permutation-simulated IVs exceeding the observed IV. In other words, the distribution of these permutation-derived IVs is taken as the null hypothesis distribution. Consequently, the null hypothesis of the test is that taxon abundance is not a function of the gradient, which is equivalent to a constant expected value of taxon abundance along the gradient.

To simulate data from the null hypothesis model, we can, as in the previous example, either draw random samples from a specific null hypothesis distribution or randomly permute the existing abundance observations along the gradient. In the online supporting materials, we provide the computer code for the Monte Carlo simulations. When using simulated or synthetic data, we explored the relationship between the type I error probability and sample size. Using the Everglades data accompanying Baker & King (2010), we illustrate the consistency between the results of the synthetic data simulation and of the real data based simulation.

For both examples, we explicitly present the null and alternative hypotheses to show that rejecting the null in both cases does not imply a threshold response. Online supplement, including R code and example data, is available at `https://github.com/songsqian/perils`.

2.2 Bootstrapping May Not Work for CART-Based Algorithms

In many applications of the two methods, bootstrapping is used to develop a confidence interval of the derived threshold. It is often used as an alternative to a formal test of whether the resulting threshold is statistically significant. This is because the other method of detecting a change point in Qian et al. (2003), the Bayesian change point model, tests the existence of a change point using the estimated probability of either end of the gradient being the change point. When using the CART-based method, the equivalent of a test of whether either end of the gradient is the change point does not exist. Instead, many applications define a statistically significant threshold as a 95% confidence interval that does not include either end of the gradient data range. Numerically, however, this approach is impossible because the calculation of a deviance (or IV) change requires two groups each with a minimum number (e.g., five or more) of data points. That is, a bootstrapping-derived

confidence interval will never include one of the ends of the gradient.

Furthermore, the bootstrapping approach may be misleading for this particular situation because the bootstrapping estimated confidence interval of a change point is likely too narrow. Bühlmann & Yu (2002) detailed the statistical background of bootstrapping and provided a mathematical proof that the bootstrap-estimated confidence interval is always too narrow for a split point problem. Banerjee & McKeague (2007) developed a statistical method for calculating the confidence interval of a split point, which is wider than the bootstrap-estimated confidence interval of the same split point. Qian (2016) used a Monte Carlo simulation algorithm to study the behavior of a bootstrap estimated confidence interval and he showed that a bootstrap estimated 95% confidence interval often contains the true split point less than 70% of the time (should be 95% of the time). With a narrower than expected confidence interval, we are more likely to be overly confident about the resulting threshold.

3 Results

3.1 Multiple Comparison Trap

3.1.1 The CART-based Method

We performed two sets of simulations. First, we simulate the distribution defined by the null hypothesis, no change of the response variable along the gradient, by drawing random numbers from a normal distribution paired with a set of randomly drawn gradient values. The deviance reduction method is applied to the resulting data and a two-sample t-test is used and the p-value is recorded. This process is repeated many times (10,000) and the fraction of the p-values smaller than 0.05 is an estimate of the test's actual type I error probability. These simulations show that the type I error probability of the test is a function of the sample size (Figure 1) and will exceed 0.2 when the sample size is larger than 40.

Using the data from Qian et al. (2003), we perform the simulation by randomly permuting the gradient (TP) data. For each random permutation, we apply the deviance reduction method to divide the data into two subsets and carry out a two-sample t-test to compare the means. After repeating the process many times (10,000), the fraction of tests with p-values below 0.05 is an estimate of the type I error probability. We estimated the type I error probability using data for Bray-Cutis dissimilarity index (BCD) and percent of tolerant macroinvertebrate taxa for each sampling date. The estimated probabilities are similar to those obtained from the synthetic data simulation (Figure 1).

In the same simulations, we estimated the probability that the CART-based model returns a threshold that is not at either end of the gradient (suggesting the existence of a change point). The estimated probability ranges from 0.9 to 0.97. In other words, the CART-based method will report a change point more than 90% of the time when the response variable does not change along the gradient.

3.1.2 Permutation Test on *IV*

We simulate the permutation test null hypothesis by drawing abundance data from a Poisson distribution (with a mean of 5) and randomly assign each abundance value to a gradient value. The resulting abundance distribution is not a function of the gradient. Under this null hypothesis, the expected abundance is a constant along the gradient. The resulting type I error probability is a function of sample size and consistently larger than 0.2 (Figure 1). Using the Everglades macroinvertebrate taxa abundance data from Baker & King (2010), the estimated actual type I error probabilities for individual taxa are consistently larger than 0.25 and follow the pattern of the synthetic data simulation but with a large variation. The variation of the false detection probability is inversely proportional to the number of zeroes in the data (Figure 2). This pattern suggests that taxa that are more tolerant to the disturbance (i.e., present along the entire gradient) are more likely to result in a false positive, that is, finding a threshold where a threshold response doesn't exist.

3.2 Raven's Paradox

The null hypotheses for both examples describe a constant expected value of the response variable along the gradient of interest. For the CART-based method, the intended alternative hypothesis is clearly stated as a step-function model (the mean response does not change along the gradient before crossing the threshold and the mean jumps to a different stable value after crossing the threshold) (Qian et al. 2003). This model is consistent with the CART model. Clearly, when the null hypothesis of a constant mean response is rejected, we cannot readily claim that the response pattern is a step function; the response pattern can be anything other than a constant one. Likewise, the null hypothesis of the IV permutation test is a constant mean taxon abundance along the gradient. Rejecting the null hypothesis implies that the taxon abundance may vary along the gradient. But the significant test result yields no specific information on how the abundance will change.

4 Discussion

The hidden multiple-comparison trap and Raven's paradox have direct management and policy implications. The CART-based method and the species aggregation index (IV) method will always return a change point whether the ecological response has a threshold pattern or not. When a threshold response is invalid, the multiple-comparison trap will increase the likelihood that a detected threshold is false. Consequently, environmental criteria derived based on these methods are likely irrelevant.

With the advent of inexpensive computing power, we are increasingly relying on computer-intensive methods. As a result, the multiple-comparison trap is increasingly relevant. Fortunately, with affordable computing power, we can easily implement Monte Carlo simulations to characterize the behavior of new methods. As a rule, we should always evaluate an NHT-based method's type I error probability before it can be presented for wide application.

Statistical null hypothesis testing (NHT) represented an intellectual leap in the development of modern statistics (Stigler 2016). Using NHT allows us to formally summarize information in noisy data to support the hypothesis of interest. However, the NHT formulation of Neyman-Pearson cannot "provide any valuable evidence of the truth or falsehood of (a specific) hypothesis" (Neyman & Pearson 1933). Threshold detection and quantification are part of the process of understanding the ecological response to environmental changes. That is, we are interested in the validity of the hypothesis of a threshold response. As such, emphasis should be placed on developing evidence directly supporting this hypothesis, rather than rejecting an irrelevant null hypothesis. An NHT is conceptually appropriate in environmental standard compliance assessment after a criterion is established, perhaps, based on an established threshold response pattern. Once the pattern is established, the "rule of behaviour" set up by an NHT will ensure that in a long run the compliance assessment will be correct most of the time, that is, in Neyman and Pearson's words, "we shall reject the hypothesis H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false" (Neyman & Pearson 1933).

We believe that the definition of an ecological threshold is vague. By vague we mean that the definition does not allow the formulation of a specific quantitative model. Without a specific model, the meaning of a threshold is unclear (Cuffney, Qian, Brightbill, May & Waite 2011, Cuffney & Qian 2013). As in any statistical modeling problem, establishing the likely quantitative model forms should be the first step of a threshold detection process (Qian 2014*a*, Qian & Cuffney 2012). Using a formal model fitting and assessment process, we can then establish the validity of the proposed model as well as the model's inferential limit (McElreath 2016). The formulaic process of NHT is inadequate for establishing the statistical validity of a model. The difficulty of a modeling approach is the selection and justification of a proper model. New computer-intensive nonparametric methods are now available for us to carry out exploratory analysis to aid the development of a proper model. For example, Wagenhoff et al. (2017) used boosted regression tree models to better understand ecosystem responses to agricultural stressors.

Statistics should be used to make a principled argument to support the hypothesis of interest. A principled argument starts with a clear statement of the goal. In our examples, we should start with a definition of a threshold response and derive a method to demonstrate its validity. As "a null hypothesis test is a ritualized exercise of devil's advocacy," it should not be used as the only piece of evidence; a persuasive case should be presented using the MAGIC (magnitude, articulation, generality, interestingness, and credibility) criteria of Abelson (1995).

5 Acknowledgment

We thank Ian Waite, Chad Wagner, and two anonymous reviewers for reviewing an early version of the paper. Any use of trade, product, or firm names in this publication is for descriptive purposes only and does not imply endorsement by the U.S. Government. The

authors declare no conflict of interest.

References

Abelson, R. (1995), *Statistics as Principled Argument*, Psychology Press, New York.

Baker, M. & King, R. (2010), 'A new method for detecting and interpreting biodiversity and ecological community thresholds', *Methods in Ecology and Evolution* **1**(1), 25–37.

Banerjee, M. & McKeague, I. W. (2007), 'Confidence sets for split points in decision trees', *The Annals of Statistics* **35**(2), 543–574.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289–300.

Box, G. (1976), 'Science and statistics', *Journal of the American Statistical Association* **71(356)**, 791–799.

Bühlmann, P. & Yu, B. (2002), 'Analyzing bagging', *The Annals of Statistics* **30**(4), 927–961.

Cao, X., Wang, J., Liao, J., Sun, J. & Huang, Y. (2016), 'The threshold responses of phytoplankton community to nutrient gradient in a shallow eutrophic Chinese lake', *Ecological Indicators* **61**, 258–267.

Cuffney, T., Kashuba, R., Qian, S., Alameddine, I., Cha, Y., Lee, B., Coles, J. & McMahon, G. (2011), 'Multilevel regression models describing regional patterns of invertebrate and algal responses to urbanization across the USA', *Journal of the North American Benthological Society* **30**(3), 797–819.

Cuffney, T. & Qian, S. (2013), 'A critique of the use of indicator species scores for identifying thresholds in species responses', *Freshwater Science* **32(2)**, 471–488.

Cuffney, T., Qian, S., Brightbill, R., May, J. & Waite, I. (2011), 'Response to King and Baker: limitation on threshold detection and characterization of community thresholds', *Ecological Applications* **21**(7), 2840–2845.

Dufrêne, M. & Legendre, P. (1997), 'Species assemblages and indicator species: The need for a flexible asymmetrical approach', *Ecological Monographs* **67**(3).

Ellison, A., Gotelli, N., Inouye, B. & Strong, D. (2014), 'P values, hypothesis testing, and model selection: it's déjà vu all over again', *Ecology* **95**(3), 609–610.

Good, I. (1983), *Good Thinking: The Foundations of Probability and Its Applications*, University of Minnesota Press, Minneapolis, MN.

Holm, S. (1979), 'Simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**(2), 65–70.

Kahneman, D. & Tversky, A. (1972), 'Subjective probability: A judgment of representativeness', *Cognitive Psychology* **3**, 430–454.

McElreath, R. (2016), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Chapman & Hall/CRC Press.

Miltner, R. (2010), 'A method and rationale for deriving nutrient criteria for small rivers and streams in Ohio', *Environmental Management* **45**, 842–855.

Neyman, J. & Pearson, E. (1933), 'On the problem of the most efficient tests of statistical hypotheses', *Philosophical Transactions of the Royal Society A* **CCXXXL**(702), 289–337.

Qian, S. (2014*a*), 'Ecological threshold and environmental management: A note on statistical methods for detecting thresholds', *Ecological Indicators* **38**, 192–197.

Qian, S. (2014*b*), 'Statistics in ecology is for making a "principled" argument', *Landscape Ecology* **29**(6), 937–939.

Qian, S. (2016), *Environmental and Ecological Statistics with R*, 2nd edn, Chapman and Hall/CRC Press.

Qian, S. & Cuffney, T. F. (2012), 'To threshold or not to threshold? That's the question', *Ecological Indicators* **15**(1), 1–9.

Qian, S., King, R. & Richardson, C. (2003), 'Two statistical methods for the detection of environmental thresholds', *Ecological Modelling* **166**, 87–97.

Stigler, S. (2016), *The Seven Pillars of Statistical Wisdom*, Harvard University Press, Cambridge, MA.

Tukey, J. (1949), 'Comparing individual means in the analysis of variance', *Biometrics* **5**(2), 99–114.

Wagenhoff, A., Liess, A., Pastor, A., Clapcott, J., Goodwin, E. & Young, R. (2017), 'Thresholds in ecosystem structural and functional responses to agricultural stressors can inform limit setting in streams', *Freshwater Science* **36**(1), 178–194.

Wasserstein, R. & Lazar, N. (2016), 'The ASA's statement on p-values: context, process, and purpose', *American Statisticians* **70**(2), 129–133.

6 Figures

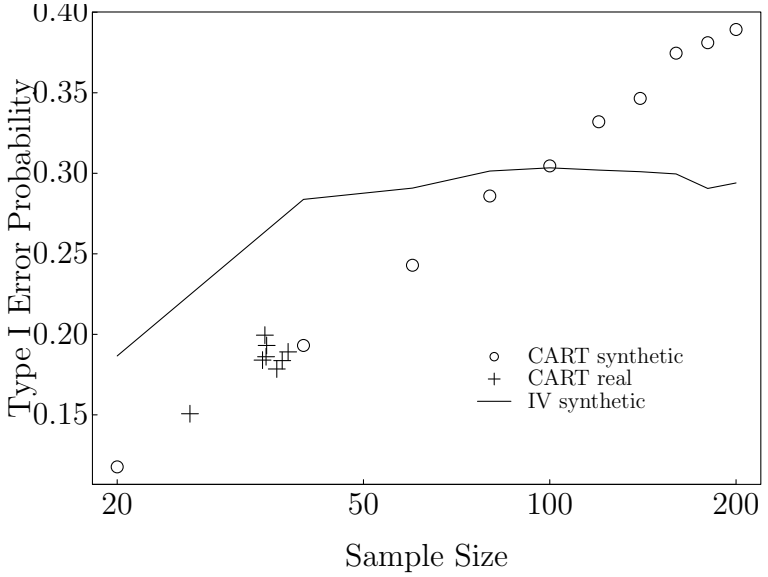

Figure 1: Simulation estimated actual type I error probabilities of the CART-based model and the *IV* permutation test increase as sample size increases. CART and *IV* synthetic are based on simulated data and CART real are based on data from Qian et al. (2003)

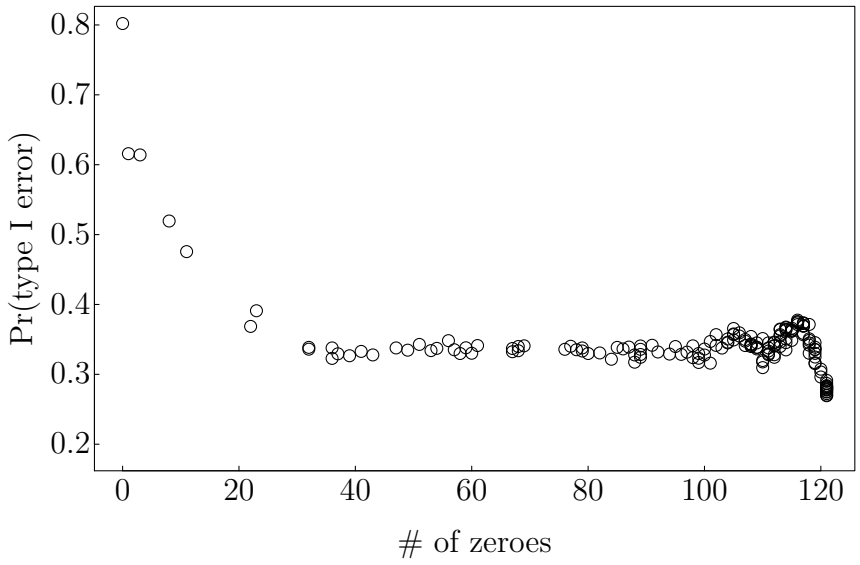

Figure 2: Estimated actual type I error probabilities using the Everglades data from Baker & King (2010) vary as a function of the number of zeroes in the abundance data. All taxa have the same sample size of 126