Journal of Environmental Management 192 (2017) 89–93

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

The Sequential Probability Ratio Test: An efficient alternative to exact binomial testing for Clean Water Act 303(d) evaluation

Connie Chen [a, 1], Matthew O. Gribble [b, *, 1], Jay Bartroff [c], Steven M. Bay [d], Larry Goldstein [c]

[a] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, United States
[b] Department of Environmental Health, Emory University Rollins School of Public Health, Atlanta, GA, United States
[c] Department of Mathematics, University of Southern California, Los Angeles, CA, United States
[d] Southern California Coastal Water Research Project, Costa Mesa, CA, United States

ARTICLE INFO

Article history:
Received 12 September 2016
Received in revised form
14 January 2017
Accepted 18 January 2017

Keywords:
Sequential probability ratio test
Water quality
Clean water act
Study design
Sediment quality objectives

ABSTRACT

The United States's Clean Water Act stipulates in section 303(d) that states must identify impaired water bodies for which total maximum daily loads (TMDLs) of pollution inputs into water bodies are developed. Decision-making procedures about how to list, or delist, water bodies as impaired, or not, per Clean Water Act 303(d) differ across states. In states such as California, whether or not a particular monitoring sample suggests that water quality is impaired can be regarded as a binary outcome variable, and California's current regulatory framework invokes a version of the exact binomial test to consolidate evidence across samples and assess whether the overall water body complies with the Clean Water Act. Here, we contrast the performance of California's exact binomial test with one potential alternative, the Sequential Probability Ratio Test (SPRT). The SPRT uses a sequential testing framework, testing samples as they become available and evaluating evidence as it emerges, rather than measuring all the samples and calculating a test statistic at the end of the data collection process. Through simulations and theoretical derivations, we demonstrate that the SPRT on average requires fewer samples to be measured to have comparable Type I and Type II error rates as the current fixed-sample binomial test. Policymakers might consider efficient alternatives such as SPRT to current procedure.

1. Introduction

In the United States, the Clean Water Act (CWA) Section 303(d) requires states to identify impaired water bodies and to recommend total maximum daily loads (TMDLs) for contaminants affecting impaired waters, such that water bodies adhering to those TMDLs will eventually comply with water quality standards (United States Environmental Protection Agency, 2012). After the USEPA Administrator's approval of a state's recommended list of impaired water bodies and implementation of TMDLs, impaired water bodies are then monitored to determine whether they have attained or not yet attained the water quality standards. If a water body meets water quality standards it may be removed from the list of impaired waters (e.g., delisted). There are major regional differences within the United States in how the 303(d) listing criteria are implemented (Keller and Cavallaro, 2008), so we focus on the regulatory framework within California, although our findings from this example may be informative for other settings considering efficient alternative study designs for 303(d) evaluation. In particular, we focus on the decision rule for sediment quality as an indicator of whether a water body is impaired under the Clean Water Act.

The California Water Code section 13191.3(a) requires the state to develop standards for listing and delisting water bodies per the CWA (California Water Code, 2014). Listing decisions are based on the frequency of exceedance of water quality standards (binary decision variable), which, for constituents such as bacteria, dissolved oxygen, contaminants, or nutrients, are numeric criteria or objectives (Gibbons, 2003). For listing evaluations based on sediment quality in bays and estuaries, California has adopted a sediment quality objective based on a “multiple lines of evidence” approach that considers contaminant levels, sediment toxicity and sediment macrofaunal community condition (Bay and Weisberg, 2012; Beegan and Bay, 2012; Bay et al., 2012). These multiple lines of evidence are integrated and assessed to determine whether

* Corresponding author. Department of Environmental Health, Emory University Rollins School of Public Health, 1518 Clifton Rd NE, Mailstop 1518-002-2BB, Atlanta, GA 30322, United States.
E-mail address: matt.gribble@emory.edu (M.O. Gribble).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jenvman.2017.01.039
0301-4797/

the sediment quality objective has been attained at a given station (State Water Resources Control Board (SWRCB), 2008) which reduces the multiple possible considerations for sediment quality into a binary decision variable suitable for evaluating exceedance frequency and responding to the 303(d) listing and delisting requirements of the CWA.

There have been several methods proposed for the analysis of binary water or sediment quality data. In 2003, Shabman and Smith recommended striking a balance between the desired Type I and Type II error rates for any 303(d) regulatory test (Shabman and Smith, 2003). The California EPA in 2004 considered several fixed sample size methods for Section 303(d) analyses (State Water Resources Control Board (SWRCB), 2004) and opted for a variation on the exact binomial test with upper limits of 0.2 for both the Type I and Type II error rates as the basis for its listing decisions; delisting decisions are based on maximum error rates of 0.1. California has specified the number of maximum number of exceedances (failures) for a specified number of total samples, leading to a range of Type I and II error rates allowed for different sample sizes (Fig. 1, Fig. 2, Supplementary Material).

Application of the exact binomial test in California's 303(d) decisions requires a substantial number of samples to attain the specified error rates. For example, a minimum of 28 samples, with no more than 2 exceedances, is required to remove a site (water or sediment segment) from the 303(d) list (State Water Resources Control Board, 2015). For evaluating sediment quality, monitoring costs to obtain the minimum sample size to evaluate delisting could easily exceed $200,000 (State Water Resources Control Board (SWRCB), 2004). Use of an alternative method with similar performance (i.e., Type I and II error rates), but reduced sample size requirements, would reduce the cost of compliance monitoring.

An alternative approach that can make the same assumptions as the exact binomial (i.e., independent and identically-distributed Bernoulli observations), called the *sequential probability ratio test* (SPRT), uses the data obtained from previous testing to evaluate whether adequate evidence exists at that time to favor a null or alternative hypothesis (Bartroff et al., 2013; Wald, 1947). This is conceptually similar to the sequential Bayesian updating proposed by Qian and Reckhow for longitudinal environmental monitoring data to determine if a water body following a TMDL has attained water quality standards (Qian and Reckhow, 2007), but here, in addition to being focused on a binary variable, our inferential goal includes making the decision of whether a water body should be listed as impaired under 303(d) and subjected to TMDL requirements. California's current regulatory testing paradigm has a parallel structure for the listing and the delisting decisions, and in this analysis we are comparing against an alternative test that also facilitates parallelism between the listing and delisting procedures.

The objective of this study is to contrast the performance of the sequential probability ratio test with California's current procedure, a fixed-sample exact binomial test, through theoretical derivations and simulation studies. Our comparison metrics are the expected number and standard deviation of the number of required samples to obtain the same Type I and Type II error rates as for the corresponding fixed-sample binomial tests. Our comparison here focuses on one alternative approach making similar assumptions to current regulatory practice, as this apples-to-apples comparison can best highlight the potential gains from more efficient methods. However, it should be noted that other efficient designs and analysis approaches, making alternative assumptions, might be even more useful for informing 303(d) listing and delisting decisions.

2. Experimental (materials and methods)

2.1. Historical sediment quality data

The Southern California Bight Monitoring Program has, since 1994, coordinated a regional sediment quality monitoring survey approximately every 5 years from up to 400 stations around Southern California (Schiff et al., 2016). This large longitudinal sediment quality monitoring dataset provides a useful resource for evaluating the performance of methods to assess 303(d) compliance on realistic datasets. We recoded the five sediment quality assessment categories in the public database into the following binary categories, consistent with how these data would be used for regulatory decision-making under current practice: *meets the*

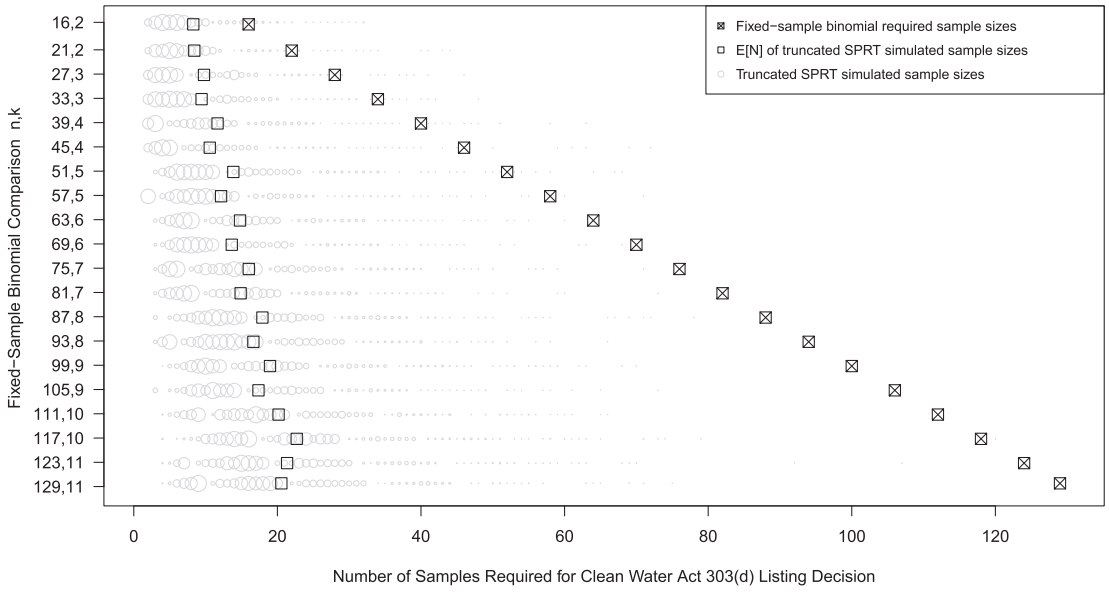

Fig. 1. Simulations Comparing Sample Sizes for 303(d) Listing: Sample Sizes under Truncated SPRT vs. Simultaneous Testing. The grey circles have proportionate area to frequency of each sample size observed across simulations (e.g., these are top-down views of histograms). The black squares represent the mean expected sample sizes, per row. The black squares with X through them represent the required sample size from the corresponding fixed-sample test per the state of California's current requirements. The required number of samples under exact binomial is “n” and the minimum number of exceedances for a listing decision is “k”.

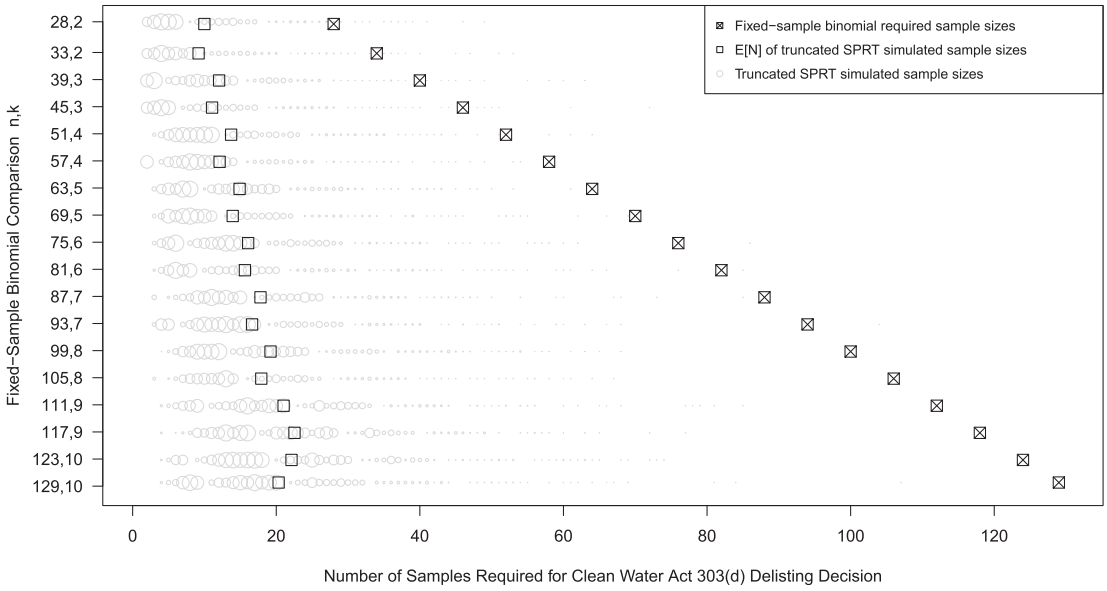

Fig. 2. Simulations Comparing Sample Sizes for 303(d) Delisting: Sample Sizes under Truncated SPRT vs. Simultaneous Testing. The grey circles have proportionate area to frequency of each sample size observed across simulations (e.g., these are top-down views of histograms). The black squares represent the mean expected sample sizes, per row. The black squares with X through them represent the required sample size from the corresponding fixed-sample test per the state of California's current requirements. The required number of samples under exact binomial is "n" and the maximum number of exceedances for a delisting decision is "k".

sediment quality objective ("unimpacted", "likely unimpacted") or *fails to meet the sediment quality objective* ("possibly impacted", "likely impacted", "clearly impacted"). Data assessments in the database classified as "inconclusive" were excluded from analysis, just as they would also have been excluded from informing regulatory decisions. For this analysis, we focused on regional monitoring data (N = 46) from San Pedro Bay in southern California, which includes Los Angeles and Long Beach Harbors. To facilitate replication of our simulations, these data are available in the Supplementary Material.

2.2. Simulated data

We simulated Bernoulli data with the same success probability (i.e., "exceedance rate") as observed in the historical regional monitoring data (code available in Supplementary Material).

2.3. Statistical approach

2.3.1. Idealized comparisons of exact binomial test vs. SPRT

Theoretical sample sizes for the exact binomial test and for the sequential probability ratio test for specified Type I and Type II error rates are derived and presented in Tables 1 and 2. Details on the theory for sequential probability ratio test are provided in the Supplementary Material. Because the sample size of the sequential probability ratio test is a random variable, we also assessed the performance of this procedure using simulated datasets. We first evaluated the empirical sample size requirements of the sequential probability ratio test to obtain the previously specified Type I and Type II error rates (Table 3). The code for all simulations is provided in the Supplementary Material.

Table 1
Fixed-sample exact binomial test: sample sizes needed to achieve type I error and power.

Type I error	Power = 0.8	Power = 0.9	Power = 0.95
$\alpha = 0.05$	78	109	135
$\alpha = 0.1$	61	86	112
$\alpha = 0.2$	39	63	82

2.3.2. Simulations comparing current California regulatory test procedure vs. truncated SPRT

The simultaneous test employed by California for regulatory purposes is a decision-rule defined by the observed numbers of failures and total numbers of trials, with different combinations detailed in tables in the *Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List* (State Water Resources Control Board, 2015). We converted these decision-rules into corresponding Type I and Type II error rates (Supplementary Material) for comparisons with SPRT. To summarize, California's decision rules currently allow Type I error rates ranging from 0.0009 to 0.16 and Type II error rates ranging from 0.0008 to 0.1885 (Supplementary Material), consistent with never allowing Type I nor Type II error rate to exceed 0.20 for listing decisions, or 0.10 for delisting decisions.

The fact that there is no fixed, *a priori* limit to the number of samples required by the SPRT to arrive at a decision could be an obstacle for environmental decision-makers. Therefore, as a proof-of-feasibility, we modified the SPRT by adding a truncation rule that would declare a water body "impaired" if no decision had yet been reached by the truncation point. We used a conservative truncation cutoff of twice the required number of observations from the corresponding current regulatory test under comparison. For example, in comparison with the state's decision-rule for 50 samples, the truncated SPRT was forced to make a decision by the 100th sample, although it typically terminated in less than 50 samples (see Figs. 1 and 2). In general, the lower a truncation threshold for declaring impairment, the fewer extremes are included in the sample mean (and standard deviation) number of samples for making a decision. If we apply the Precautionary Principle (Jordan and O'Riordan, 2004) and interpret ambiguous truncated outcomes as "impaired" when truncated, then for 303(d) listing decisions the Type I error rate increases and Type II error rate decreases, while for 303(d) delisting decisions, the Type I error rate decreases and Type II error rate increases. Therefore, the comparisons provided

Table 2
Sequential probability ratio test: theoretical sample size needed for given type I error and power.

Type I error	Under the Null Hypothesis			Under the Alternative Hypothesis		
	Power = 0.8	Power = 0.9	Power = 0.95	Power = 0.8	Power = 0.9	Power = 0.95
$\alpha = 0.05$	36.6	54.4	72.2	52.0	64.8	72.2
$\alpha = 0.1$	31.2	47.9	64.8	37.1	47.9	54.4
$\alpha = 0.2$	22.7	37.1	52.0	22.7	31.2	36.6

Table 3
Sequential probability ratio test: expected sample size needed for given type I error and power, and standard errors (in parentheses) from 1000 simulated trials.

Type I error	Under the Null Hypothesis			Under the Alternative Hypothesis		
	Power = 0.8	Power = 0.9	Power = 0.95	Power = 0.8	Power = 0.9	Power = 0.95
$\alpha = 0.05$	39.3 (31.6)	58.2 (43.9)	72.5 (47.1)	49.0 (36.9)	57.8 (40.4)	65.0 (48.4)
$\alpha = 0.1$	34.9 (26.4)	50.4 (33.4)	67.6 (43.8)	36.3 (26.9)	45.3 (34.2)	50.6 (42.0)
$\alpha = 0.2$	27.4 (19.5)	42.6 (27.7)	60.0 (39.3)	22.6 (18.7)	31.3 (27.5)	34.8 (32.9)

between the truncated SPRT and the state's current test are approximate matches, as the two tests will have slightly different empirical error rates. How different the error rates are between these two methods depends on how often the stochastic process underlying SPRT invokes truncation.

3. Results and discussion

Theoretical and simulation-estimated sample sizes for the sequential probability ratio test without truncation were lower than the exact binomial test, for the same power and type I error rate (Tables 1–3). For a fixed Type I and Type II error rate, the required sample sizes for non-truncated SPRT in simulations were slightly higher than the theoretical sample sizes, but had standard deviations almost as large as the expected values (Tables 2 and 3). Thus, the performance of the non-truncated SPRT varies according to the data on which it is used, and truncation might protect against excessive sampling.

In the applied comparison against the current listing tests used by California, the truncated SPRT also required on average fewer samples both for listing (illustrated in Fig. 1) and delisting (illustrated in Fig. 2) decisions.

To make our simulations as comparable the California current practice as possible, in this analysis we applied the special case of SPRT where the observations in sequence are Bernoulli random variables, but the idea of SPRT is more general and can be used for other sets of independent and identically distributed objects, for example potentially modeling independent and identically distributed, vector-valued *batches of data* as sequential objects, rather than single Bernoulli variables. The alternative flavor of SPRT for batches of data is called group sequential testing (Dennison and Turnbull, 1999). It could be an interesting extension of this work on 303(d) evaluation methods to assess how varying the number of samples per stage in a sequential collection of batches could help optimize the procedure for real-world practicality and statistical efficiency.

Although it appears from this analysis that there could be major efficiency gains in shifting from a simultaneous to a sequential testing framework, in particular with a sequential sampling design and truncated SPRT for delisting decisions, we do not recommend that the SPRT (or its truncated version) be adopted in its current form for regulatory purposes, because there are still major limitations (e.g., failure to account for dependence between samples) to both SPRT and conventional approaches. Rather, we hope our analysis will advance a conversation leading to development of even more efficient and appropriate methods. Our analysis has focused on more efficient designs for future assessments (e.g., for future regulatory testing programs assessing regulatory compliance using data collected over a fixed period such as new data collected within a 2-year regulatory window, comparable to how current Clean Water Act determinations are made using data from the most recent assessments). Neither the SPRT nor the current approach explicitly provides for how to make use of previous observations collected prior to the current regulatory observation window. Several methodological papers focused on improving efficiency of CWA regulatory testing have noted there is often prior information available on the historical water quality and information on data from neighboring sites. The Bayesian power prior method advocated by Duan, Ye and Smith incorporates historical and adjacent site data into a binomial-model water quality assessment via a power prior, but treats “current” data as a batch to be tested simultaneously to provide a likelihood of the parameter value of interest (Duan et al., 2006). Similarly, the Bayesian approach encouraged by McBride and Ellis uses a simultaneous sample binomial likelihood but with beta priors (McBride and Ellis, 2001). The TMDL compliance method recommended by Qian and Reckhow allows for sequential updating of the likelihood, but was only developed for continuous variables (in the context of attainment of TMDL goals) (Qian and Reckhow, 2007), whereas California determines CWA compliance via a binary decision variable per sample. Combining these ideas, for example by basing a prior for a “failure” parameter on historical information and updating sequentially with each sample using a group sequential test, could make better use of the complete monitoring record available. Another extension that could further strengthen the real-world appropriateness of this method is to formulate a model that would explicitly account for the spatial and temporal autocorrelation between regulatory samples collected across different stations and years; none of the regulatory decision rules we have encountered yet take this spatio-temporal autocorrelation into account.

One concern that might be raised for sequential testing in an environmental monitoring context is that the data points could be “cherry-picked” by an unscrupulous assessor to prioritize the cleanest sampling sites within a water body to be the first sampling stations evaluated, biasing CWA 303(d) decisions toward false “attainment”. However, similar concerns about using unreasonable environmental sampling sites also apply to simultaneous study designs and the current “exact binomial test” as well. Any study used for regulatory decisions must include strict requirements for informative sampling.

In conclusion, the SPRT offers an efficient alternative to the current regulatory framework for CWA 303(d) listing and delisting

decisions in California. In particular, for CWA 303(d) delisting decisions, adoption of a truncated SPRT that parses any inconclusive (i.e., truncated) results as "impaired" could reduce both the Type I error rate, thus better protecting public health and the environment, and the average required sample sizes, thus reducing the cost of monitoring for adherence, relative to the status quo. Further work is needed to develop and make accessible to stakeholders (e.g., through easy to use software) new methods for CWA 303(d) evaluation that incorporate historical data, account for the autocorrelation of samples, and update evidence sequentially in order to efficiently make appropriate decisions.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgements

M. Gribble was supported during this project by T32 training grant support from the National Institute for Environmental Health Sciences (T32 ES013678) and the HERCULES Exposome Research Center funded by the National Institute for Environmental Health Sciences (P30 ES019776). J. Bartroff was supported in part by grant DMS-1310127 from the National Science Foundation and grant R01 GM068968 from the National Institutes of Health.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jenvman.2017.01.039.

References

Bartroff, J., Lai, T.L., Shih, M.C., 2013. Sequential Experimentation in Clinical Trials: Design and Analysis. Springer Series in Statistics. Springer, New York (NY), USA, p. 240.

Bay, S.M., Weisberg, S.B., 2012. Framework for interpreting sediment quality triad data. Integr. Environ. Assess. Manag. 8 (4), 589–596.

Bay, S.M., Ritter, K.J., Vidal-Dorsch, D.E., Field, L.J., 2012. Comparison of national and regional sediment quality guidelines for classifying sediment toxicity in California. Integr. Environ. Assess. Manag. 8 (4), 597–609.

Beegan, C., Bay, S.M., 2012. Transitioning sediment quality assessment into regulations: challenges and solutions in implementing California's sediment quality objectives. Integr. Environ. Assess. Manag. 8 (4), 586–588.

California Water Code, 2014. Section 13191.3. Updated March 17. http://law.onecle.com/california/water/13191.3.html (Accessed 7 June 2014).

Dennison, C., Turnbull, B.W., 1999. Group Sequential Methods with Applications to Clinical Trials (Chapman & Hall/CRC Interdisciplinary Statistics), first ed. Chapman & Hall, p. 416. ISBN: 978-0-8493-0316-6. eBook ISBN: 978-1-58488-858-1.

Duan, Y., Ye, K., Smith, E.P., 2006. Evaluating water quality using power priors to incorporate historical information. Environmetrics 17 (1), 95–106.

Gibbons, R.D., 2003. A statistical approach for performing water quality impairment assessments. J. Am. Water Resour. Assoc. 39 (4), 841–849.

Jordan, Andrew, O'Riordan, Timothy, 2004. The precautionary principle: a legal and policy history. In: Martuzzi, Marco, Tickner, Joel A. (Eds.), The Precautionary Principle: Protecting Public Health, the Environment and the Future of Our Children. World Health Organization, pp. 31–48. Chapter 3. http://www.kinderumweltgesundheit.de/index2/pdf/dokumente/50023_1.pdf.

Keller, A.A., Cavallaro, L., 2008. Assessing the US Clean Water Act 303(d) listing process for determining impairment of a waterbody. J. Environ. Manag. 86, 699–711.

McBride, G.B., Ellis, J.C., 2001. Confidence of compliance: a Bayesian approach for percentile standards. Water Res. 35 (5), 1117–1124.

Qian, S.S., Reckhow, K.H., 2007. Combining model results and monitoring data for water quality assessment. Environ. Sci. Technol. 41 (14), 5008–5013.

Schiff, K., Greenstein, D., Dodder, N., Gillett, D.J., 2016. Southern California Bight regional monitoring. Reg. Stud. Mar. Sci. 4, 34–46.

Shabman, L., Smith, E., 2003. Implications of applying statistically based procedures for water quality assessment. J. Water Resour. Plan. Manag. 129 (4), 330–336.

State Water Resources Control Board, 2015. Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List. Division of Water Quality. State Water Resources Control Board, Sacramento, CA.

State Water Resources Control Board (SWRCB), 2004. Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List. Final Functional Equivalent Document. Division of Water Quality. State Water Resources Control Board, Sacramento, CA.

State Water Resources Control Board (SWRCB), 2008. Water Quality Control Plan for Enclosed Bays and Estuaries; Part I: Sediment Quality. Division of Water Quality. State Water Resources Control Board, Sacramento, CA.

United States Environmental Protection Agency, 2012. Clean Water Act Section 303. Updated March 6. http://water.epa.gov/lawsregs/guidance/303.cfm (Accessed 26 May 2014).

Wald, A., 1947. Sequential Analysis. John Wiley and Sons, Inc., New York (NY), USA, p. 212.