Model-based Tests for Standards Evaluation and Biological Assessments

Zhengrong Li

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Eric P. Smith, Chair
Samantha C. Bates Prins
John P. Morgan
Eugene R. Yagow
Keying Ye

August 15, 2006
Blacksburg, VA

Keywords: Model-based tests, Fixed effects models, Random effects models, Regression-based tests, Redundancy analysis, Reduced-rank regression, Water quality assessment

This work received support from the U.S. Environmental Protection Agency's Science to Achieve Results (STAR), Grant No. RD-83136801-0.

Model-based Tests for Standards Evaluation and Biological Assessments

Zhengrong Li

ABSTRACT

Implementation of the Clean Water Act requires agencies to monitor aquatic sites on a regular basis and evaluate the quality of these sites. Sites are evaluated individually even though there may be numerous sites within a watershed. In some cases, sampling frequency is inadequate and the evaluation of site quality may have low reliability. This dissertation evaluates testing procedures for determination of site quality based on model-based procedures that allow for other sites to contribute information to the data from the test site. Test procedures are described for situations that involve multiple measurements from sites within a region and single measurements when stressor information is available or when covariates are used to account for individual site differences.

Tests based on analysis of variance methods are described for fixed effects and random effects models. The proposed model-based tests compare limits (tolerance limits or prediction limits) for the data with the known standard. When the sample size for the test site is small, using model-based tests improves the detection of impaired sites. The effects of sample size, heterogeneity of variance, and similarity between sites are discussed. Reference-based standards and corresponding evaluation of site quality are also considered.

Regression-based tests provide methods for incorporating information from other sites when there is information on stressors or covariates.

Extension of some of the methods to multivariate biological observations and stressors is also discussed. Redundancy analysis is used as a graphical method for describing the relationship between biological metrics and stressors. A clustering method for finding stressor-response relationships is presented and illustrated using data from the Mid-Atlantic Highlands. Multivariate elliptical and univariate regions for assessment of site quality are discussed.

Dedicated

To my parents.

Acknowledgements

Completion of this degree would not have been possible without the continuous guidance and support from my advisor Dr. Eric P. Smith. I am extremely thankful for his patience and thoughtfulness throughout this research. I would also like to express my sincere appreciation to Dr. Keying Ye for his help as a mentor and a friend during these tough graduate school years. I really want to give my special thanks to Dr. John P. Morgan for his suggestions and guidance in finalizing my dissertation.

Also, many thanks go to the other committee members, Dr. Samantha Bates Prins and Dr. Eugene R. Yagow for their valuable comments.

I am grateful to the faculty of the Department of Statistics for building the bridge from statistics to the real world. My four years of learning, researching, consulting and teaching have all been rewarding and enjoyable. Thanks also go to all of the staff and graduate students for their warm-hearted assistance and concerns.

I would want to express my deepest gratitude to my parents, my husband and my brother for their love. They are always there and do their best to encourage and support me.

The last but not the least acknowledgement is left to my son. He changes my life……

—— Zhengrong Li

The apparition of these faces in the crowd
Petals on a wet, black bough.
— Ezra Pound

Contents

List of Figures

List of Tables

1. Introduction

Water plays an important role in the proper functioning of the ecosystems and in human health. Appropriate monitoring and assessment of water quality help prevent water pollution and provide efficient management. In environmental statistics, water quality evaluation is a vital component. There are a number of statistical issues in this area, such as setting standards (Barnett and O'Hagan, 1997), estimating trends (Helsel and Hirsch, 2000), and testing compliance. The general assessment procedure involves the development of standards, the determination of associated numerical criteria, the collection of information related to the standards, and the evaluation of that information to assess compliance. Approaches that have been discussed in the literature include the US Environmental Protection Agency's (USEPA) raw score method, the binomial approach (Smith et al., 2001; Lin et al., 2000), a permit limit approach (USEPA, 1991), acceptance sample by variables (Smith et al., 2003), tolerance intervals (Gibbons, 2003), prediction intervals (Kilgour et al., 1998), and Bayesian tests on a percentile (Ye and Smith, 2002).

One of the statistical issues in the application of statistical tests is power. In some cases sample sizes are small and often inadequate to provide parameter estimates with good properties and high power for evaluating a specific site. In this situation, regional information, i.e., information from a neighborhood of the test site, can be very helpful in evaluating current water quality. Borrowing information from a neighborhood needs model-based techniques. To address this issue, Robert W. Smith (2002) proposed a model for a two-way crossed random design to estimate tolerance intervals and set numerical criteria for environmental regulations.

However, the region of interest and hence the data for neighborhoods is complicated in practice. The quality of a neighborhood influences site-oriented tests. The method by which the numerical criterion for a standard is determined affects any testing procedure. Sampling strategies also impact the selection of an appropriate test. To deal with these issues and other practical concerns, my research proposes model-based tests in the univariate case using analysis of variance and evaluates the improvement of model-based tests compared to the single site test. A general regression-based test is also provided. To

deal with multiple biological responses in complicated and large environmental data (the multivariate case), my research proposes the integration of model-based clustering and assessment. Specifically, in this study, I probe practical situations with different models, such as fixed effects models and random effects models with corresponding estimation approaches. A method with relatively good power is suggested for application. Limits based on tolerance intervals and prediction intervals are investigated for their effectiveness in judging impairment. Further, the causes of water quality impairment are explored by using regression-based tests and some multivariate techniques. The effect of covariate adjustment is also addressed with the regression model. Clustering based on random tessellations and redundancy analysis is proposed for an exploratory examination of the response-stressor relationship. Reference sites (sites which meet water quality standards) are used as an alternative to the use of preset criteria. In addition, reduced-rank regression is carried out for impairment detection. The graphical display of assessment results is presented.

In my study, the site of interest is called the test site which is evaluated for its water quality. The non-test sites are other sites where information is borrowed for estimation. These sites are in the neighborhood of the test site. The non-test sites can be either reference or non-reference sites (See Section 2.2.8 for definitions). The test site and non-test sites form a region of interest.

The rest of this dissertation is organized as follows. Chapter 2 reviews the literature on water quality assessment, briefly summarizes current model-based estimation for assessment, and introduces the framework of my research. Chapter 3 investigates compliance tests using tolerance intervals when the numerical criterion is a fixed value. Theoretical analyses and simulations are used to evaluate different methods for making decisions. Chapter 4 addresses prediction issues in assessment when the numerical criterion comes from reference conditions. In Chapter 5 water quality assessment is carried out by regression-based tests. The effect of covariate adjustment on tests is discussed by the comparison with the ordinary test(s). Chapter 6 applies random tessellations and redundancy analysis to the Mid-Atlantic Highlands benthic data to find the optimal cluster with the strongest response-stressor relationship. Prediction intervals based on the reduced-rank regression and full rank multivariate regression are compared

for biological assessment. The graphical display of acceptance region is also discussed. Finally, in Chapter 7, the broad model-based tests are summarized and future research is sketched.

2. Water Quality Assessment Methods: A Review

Water quality is constantly threatened by many different sources and types of pollution. The two general classifications of pollution sources are point and non-point sources. Point source pollution comes from a concentrated originating point (such as industrial and sewage treatment plants) while non-point source pollution comes from diffuse (or non-point) sources and is caused by rainfall or snowmelt moving over and through the ground. In the past decades, the US has made tremendous advances to clean up the aquatic environment by controlling pollution from industries and sewage treatment plants (i.e., point source pollution). But pollution from non-point sources has not been fully monitored. Non-point source pollution has harmful effects on drinking water supplies, recreation, fisheries, and wildlife. This type of pollution (ranging from chemical contamination to habitat alteration) has been the leading cause of water quality problems (Barbour et al., 1996). The related water quality assessment (monitoring programs) has thus become the focus, recently. There is a variety of legislations driving the monitoring of water quality, but the focus here is on the Clean Water Act (USEPA, 1987) and assessment of standards under this act.

Usually three aspects of water quality are measured under the Clean Water Act: biology, chemistry, and nutrient quality. Biological quality indicates the overall ‘health’ of water segments. Chemical quality is often an indicator of organic pollution in water. Nutrient status typically refers to phosphate and nitrate in water. With the increasing emphasis on non-point source pollution and biological communities, biological water quality is attracting more attention. Biological quality concerns structural and functional aquatic community characteristics and concerns the ability of water to support aquatic life uses. Therefore, even though the monitoring of chemicals is still a priority, emphasis on biological monitoring has become more important since biological data may provide valuable insights into the nature and causes of impairment.

Under the Clean Water Act, every state must adopt water quality standards to protect, maintain, and improve the quality of the nation’s surface waters. These standards describe a level of water quality that will support designated uses of water. From a statistical perspective, there are two main themes in the area of water quality research: assessing the water quality and exploring the cause of a perceived problem. Assessment of water quality typically involves

setting standards and associated numerical criteria, collecting information related to the standard, and evaluating that information to assess compliance (Barnett and O'Hagan, 1997). A recent focus has been on the use of biological responses in the decision process and the interpretation of these responses (Smith et al., 2001). Evaluation of causes of biological impairment often involves the development of stressor-response relationships. This evaluation usually fits a regression model relating the available stressor and response variables. Due to the complexity of an ecosystem, multiple stressor and/or response variables may be available. Multivariate methods are thus applied to detect relationships. This relationship exploration provides information about the stressors and how these stressors affect the biological community. Such knowledge is especially useful in understanding how to manage impaired systems.

This chapter begins with the first theme (assessment), introducing the need for statistical compliance assessment and discussing prior researches in this area. Then a brief summary is addressed for current study concerns and my research focus related to assessment and cause exploration.

2.1 Standards development

Standards have various names in the programs of the US Environmental Protection Agency (EPA), such as ARARs (Applicable or Relevant and Appropriate Requirements), concentration limits, limitations, regulatory thresholds, action levels, and criteria (Gilbert et al., 1996). This dissertation uses criteria and standards somewhat synonymously although a standard may be narrative while a criterion typically refers to a fixed numerical value. Water quality criteria are viewed here as numerical values setting limits on a particular component. Categories of water quality criteria include aquatic criteria, biological criteria, nutrient criteria, sediment criteria, and so on (USEPA, 2000). Water quality criteria help managers in the protection and restoration of the quality of surface waters consistent with the requirements of the Clean Water Act (CWA). Before the 1980's, water quality assessment under the CWA only used indirect measures of the aquatic community. These indirect measures include quantities, such as dissolved oxygen and biological oxygen demand, which are related to ecological response. Other measures, such as bacteria counts, are related to human health and focus on the use of waters for everyday (swimming-related) activities. Because of concerns that chemical protection does not imply ecological protection, biological criteria were developed to protect water resources in an

ecologically and environmentally efficient manner (Yuan and Norton, 2003). They can provide regulatory mechanisms to assess and help protect biological resources at risk from chemical, physical, or biological impacts (e.g., the invertebrate fauna, dominated by sewage-tolerant worms, midges, and snails).

2.1.1 Standards and numerical criteria development in some EPA programs

There are different ways to classify standards. In terms of regulatory scales, water quality criteria can be developed at the national level or the state level (USEPA, CWA, 1987). Under Section 304(a) of the CWA, national water quality criteria consist of "scientific information regarding concentrations of specific chemicals or levels of parameters in water that protect aquatic life and human health". At the state level, water quality criteria are elements of state water quality standards typically adopted under Section 303(c) of the CWA. Recently, water quality standards have been developed at a regional level. In related EPA programs (e.g., Mid-Atlantic Integrated Assessment), geographical patterns of similarity among ecosystems are used to define ecoregions and develop standards (National Research Council, 1992). Environmental regulations are enforced by partnerships among different levels of governments. This development approach integrates regional inputs and provides a comprehensive picture of water quality.

Water quality criteria in EPA programs can be established based on technologies, risk assessments, or site-specific data (Gilbert et al., 1996). This categorization scheme has more environmental emphasis and helps in interpreting the assessment procedure. From a technology or engineering perspective, criteria are developed to define the effectiveness of pollution abatement technology. They are based on the performance of treatment and control technologies. Variability in the operation of a technology system is considered in this type of standard. Once the technology has been installed and is operating properly, the likelihood of exceeding the standards is rare.

A risk-based standard is a specified concentration value that is assumed to be known with certainty. This type of standard is usually determined by ecological risk assessment or laboratory assessment. Risk-based standards have been the main type for decades and have included standards for bacterial contamination.

Roughly after 1994, the EPA switched focus (especially for biologically-based criteria) from risk-based standards to background-based or reference-based standards. The new focus considers

the presence of any existing contamination and uses the available data to set the standard. Given a region of interest, sites are selected that are considered minimally impaired and treated as 'healthy' or as sites with desirable attributes (reference conditions). The 'healthy' sites (reference sites) are then used to represent the expected biological integrity of other sites in the same waterbody or nearby waterbodies. Nowadays, reference sites/regions are attracting more interest for developing standards (USEPA, EMAP research strategy, 1997).

2.1.2 Development tools for biological criteria

Among the three general categories of water quality (biology, chemistry, and nutrient quality) chemistry and nutrient are easily measured and interpreted in aquatic systems, while biological information is more difficult to interpret. Biological data typically consist of counts (benthic or fish counts) or a health indicator. Examples include the number of tolerant taxa or the EPT index (EPT refers to three taxa indicative of healthy streams: Ephemeroptera, Plecoptera, and Trichoptera).

As an overall indicator, a biological criterion describes the minimal desired condition for aquatic life inhabiting waters with a designated aquatic life use. There are four approaches to developing biological criteria: (1) to define biological integrity in terms of environmental functions; (2) to interpret the natural geographic and temporal variability of data by their integration within regions of ecological similarity; (3) to use reference sites to obtain reference conditions for specific areas; and (4) to combine several metrics to produce a single numeric index (Davis and Simon, 1995). These tools basically date back to the use of indicator organisms, which originated in the Saprobien system (Bartsch and Ingram, 1966; Mackenthum and Ingram, 1967). Based on these indicator organisms, numeric biological indices were developed. Soon after, the aggregation of several numeric biological attributes into multiple metric indices was used to measure biological integrity. Nowadays composite indices are widely used for biological criteria, especially in the United States. They are developed by (1) carrying out beneficial use assessments for aquatic life support based on measures of biological integrity, then (2) using multiple reference sites to define attainable conditions within ecoregions. In this study, I will use composite biological criteria to illustrate univariate methods proposed for water quality assessment.

2.1.3 Measurement forms of standards

Standards from different development procedures lead to numerical criteria with different statistical behaviors. Statisticians put more effort into risk-based standards and background-based standards than technology-based standards. A risk-based standard leads to a criterion that is a single numerical value specified by regulators or stakeholders. This is a stable guideline and usually doesn't change over long periods. For example, a common toxicological index is the non-observed effect level (NOEL), a threshold above which observable effects in test water regions are believed to occur and below which no toxicity is observed (Gilbert et al., 1996). Background-based standards for biological metrics are derived from reference data and used only for sites/regions with ecological similarity. They can be the mean, the median, or some percentile of the distribution of a biological metric for the reference data. Even though background-based criteria take numeric values, they are random variables and have statistical properties that should be evaluated, thereby incorporating further sources of uncertainty into the assessment procedure.

While risk-based criteria are often viewed as fixed, some treat them as random quantities. For example, in toxicological risk assessment, ED50s (doses resulting in an effect in 50% of the test animals) might be calculated on different species. Because different species have different sensitivities, ED50s are expected to vary and hence might be used to describe a "species-sensitivity distribution" (Posthuma et al., 2001). The uncertainty in the ED50s is incorporated into the criterion through the use of a confidence limit or tolerance limit.

Given a numerical criterion, assessment at a particular site (referred to as a test site) requires a decision process. For environmental assessment the decision process can be based on an acceptable frequency, magnitude, and/or duration of the standard that will allow the designated use for the particular site (USEPA, Chesapeake Bay Program, 2000). 'Frequency' is the term for how often a criterion is exceeded. 'Magnitude' indicates how much of a pollutant, expressed as a concentration, is allowable. Duration refers to the period of time (averaging period) over which average concentration for a water segment is in violation. Among these three components, water quality standards assessment involves mainly the frequency of standards violation. The acceptable frequency of standards violation refers to the proportion of measurements in violation under 'safe' conditions. That is, the proportion of samples that exceed the criterion at a specific site is less than or equal to the regulatory-specified proportion and the site still supports its designated use. In terms of a statistic, the acceptable frequency is called the acceptable

exceedance probability or the percentile of a measurement distribution. The acceptable frequency is the focus for evaluating water quality (USEPA Statistical Guidance Documents, 1989, 1992, 2000). Compared with an assessment based on the direct comparison between measurements and the standard, the decision process using a statistical testing method increases the degree of certainty that there is a persistent pollution problem in water bodies before the water bodies are listed.

2.2 Methods for standards assessment

Data for standards assessment typically consist of measurements over time at a particular test site. In the state of Virginia, the Department of Environmental Quality (DEQ) routinely monitors many stations (sites) and considers many sources of data in the assessment of water quality. Data may be collected by colleges and citizen monitoring groups or by the DEQ staff. The collected data are multivariate in nature. For example, water column data might contain measurements on dissolved oxygen (DO), biological oxygen demand, pH, and fecal coliform bacteria. Because DEQ reports are required on a regular basis, typically two-to-five years of data are available for assessment.

Figure 2.1 displays an example dataset that is similar to many datasets for standards assessment. In this figure, the horizontal line at 5 indicates the criterion for DO level and each of the other three lines indicates one site. The three sites in this figure come from Gaston Lake (on the boundary of North Carolina and Virginia). At one or more time points in the years 1991, 1996, and 2000, dissolved oxygen levels were measured at three sites. Any of these sites can be the test site.

Although the data that arise are often multivariate, assessment is typically done using simple techniques and multivariate methods are seldom considered for standards assessment (Multivariate methods are sometimes used in some assessment programs with biological data. See Wright et al., 2000, or Reynoldson et al., 2000 for examples and details). The method used for assessment varies considerably with types of data. For example, with a toxic chemical, some consider any violation of the criterion as an indication of a problem. With other measurements such as the water column measurements listed above, commonly used methods involve simple decision procedures as well as univariate parametric or nonparametric hypothesis tests for assessing compliance (Gilbert, 1987; Lin et al., 2000).

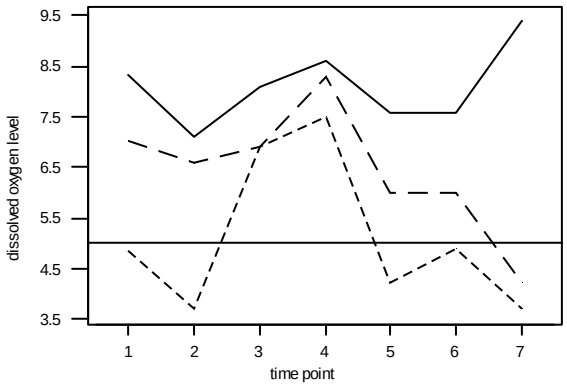

Figure 2.1 Dissolved oxygen levels at three sites from Gaston Lake in 1991, 1996, and 2000.

The simplest approach considers only the proportion of samples (typically the proportion of times) that exceed the criterion. In standards assessment, the word ‘exceed’ indicates that a measured value falls in the violation region, regardless of whether that region is expressed as “greater than criterion” or “less than criterion”. For instance, with dissolved oxygen (DO), a site is declared healthy if its DO level is 5.0 mg/L or above. The criterion is thus set at 5.0 mg/L. Any measurement below this criterion is called a measurement exceeding the criterion. If the percentage of samples with dissolved oxygen below 5.0 mg/L is more than 10%, it is considered that the site does not meet the standard. This procedure will be referred to as the “raw score” approach.

Different from the raw score approach using the direct comparison between the measurements and the criterion (standard), the hypothesis test approach assumes a probability distribution for the data and uses a statistic from data at the test site for the comparison with the standard to make an impairment decision. Furthermore, the hypothesis test approach allows considerations of error rates (probability of false declaration of impairment and unimpairment). The statistic from the sampled dataset can be the mean, the median, or a specified percentile. For example, a specified percentile from the sample is evaluated to determine if it exceeds a fixed criterion, either from a guideline (i.e., risk-based standard) or from reference data.

When a water segment is claimed to be impaired (i.e., listed), a Total Maximum Daily Load process may be initiated (USEPA, CWA, 1987). This process is typically complicated and

potentially expensive. On the other hand, when a truly impaired site is not listed, the potential harm falls on the public. The influence is long-term and the cost is immeasurable. An efficient evaluation method is needed in assessing water quality to avoid incorrectly listing or delisting sites. Widely discussed methods include the raw score method, the binomial test, and acceptance sampling by variables. Each method only uses one sample from the test site, i.e., the dataset is based on a single site.

2.2.1 Raw score method

The policy document on water quality assessment (USEPA, 1987) suggests that a stream segment be listed as impaired when more than 10% of the measurements of water quality conditions exceed a numeric criterion. This is often referred to as a "raw score" assessment method. It checks compliance simply by evaluating the proportion of observations (measurements) that are in violation of the criterion.

The 1987 policy explained the application of this method using a chemical criterion. For water measurements of dissolved oxygen, the site is declared impaired when a minimum of one measurement is beyond the State's numeric surface water quality criterion within the most recent five-year period that the data have been collected (typically one observation from each year). Actually when the number of observations at one site is lower than 10, no observation is allowed to exceed the standard if the site is to be judged as an unimpaired site. When the number of observations is between 11 and 19, the site is treated as an unimpaired site if there is one or no observation exceeding the standard (Smith et al., 2001).

The raw score method is a non-statistical method in the sense that it makes no attempt to control error rates as a function of sample size. However, its error rates can be calculated exactly by an analytical approach with the null hypothesis that the site of interest is unimpaired (Table 2.1). Let n be the sample size and p_0 (the acceptable exceedance proportion) be 0.1. Consider the 'event' characterized by an observation exceeding the criterion. If the number of exceedances, X, is equal to or less than integer(np_0)[1] (the maximum acceptable exceedance number), the site is not listed. Let k be the maximum acceptable exceedance number. Assuming independence and that X has a binomial $B(n, p_0)$ distribution, the Type I error rate is the sum of

[1] Integer(np_0) stands for the integer part of np_0. For example, when n=18, p_0=0.1, integer (np_0)=integer (1.8)=1.

the binomial probabilities over the value exceeding k in a Binomial distribution with parameters n and p_0, i.e., Type I error rate=P($X>k \mid n,\ p_0$). The Type II error rate is the probability of not listing an impaired site. Given the true exceedance proportion p_1, the Type II error rate is the cumulative probability of no more than k acceptable exceedances with a Binomial distribution, $B(n, p_1)$.

Table 2.1 Analytical error rates for raw score method

		Decision based on sample	
		List	Don't list
Truth about population	Unimpaired	$\alpha = P(X>k \mid n,\ p_0)$ $=P(X<n-k \mid n,\ 1-p_0)$	$1 - \alpha = P(X \le k \mid n,\ p_0)$
	Impaired	$1 - \beta = P(X>k \mid n,\ p_1)$	$\beta = P(X \le k \mid n,\ p_1)$

When the sample size is up to 50, the error rate of the raw score method is plotted in Figure 2.2. In this example, the raw score method has a high Type I error rate. The Type I error is calculated for an assumed acceptable exceedance proportion equal to 0.1. The true exceedance proportion is set to 0.2 in Figure 2.2. The raw score method leads to frequent decisions that sites are in violation when in fact they are not. The significance of falsely listing a site as impaired leads to increased regulatory burdens and costs for permit compliance. As the true exceedance proportion becomes larger, the Type II error rate decreases (Figure 2.3), i.e., the raw score approach is more powerful when the true exceedance proportion is larger.

For a good assessment of water quality one must be cognizant of Type I (a false declaration of standards violation) and Type II (a false declaration of no violation) errors. The raw score method achieves good rates of Type II error while losing control of the Type I error rate. Moreover, this method doesn't set a minimum sample size, which is an important practical concern in environmental studies. A statistical procedure is thus needed. Smith et al. (2001) and McBride and Ellis (2001) pointed out that the evaluation of a numerical criterion can be viewed as a statistical decision process so to improve its performance. Considering sampling resources, naturally varying measurements, possible measurement errors, and tolerable standards violations, it is reasonable to view the assessment process as a statistical decision problem.

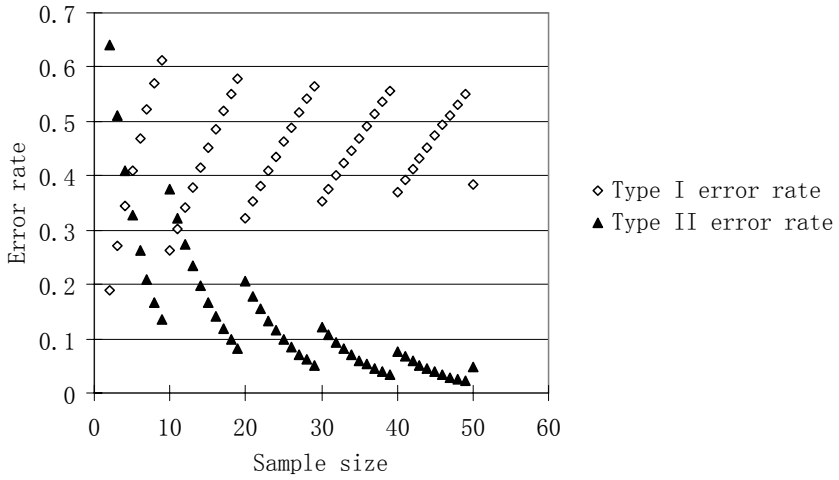

Figure 2.2 Error rates for different sample sizes (up to 50) for the raw score method. Here $p_0 = 0.1$.

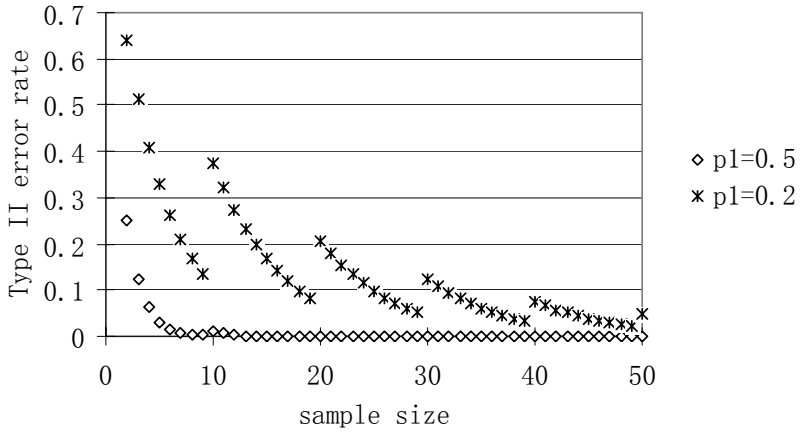

Figure 2.3 Type II error rates for different sample sizes at a fixed true exceedance proportion.

2.2.2 Frequentist binomial test

When a statistical decision process is implemented to evaluate compliance, the hypothesis test for standards assessment is generally set up as H_0: $p \leq p_0$ (unimpaired, don't list) versus H_1: $p > p_0$ (impaired, list), where p is the true probability of exceeding the standard and p_0 is a constant between 0 and 1 (Smith et al., 2001; Lin et al., 2000). In my study, p_0 is always 0.1, corresponding to an acceptable violation rate less than or equal to 10%. With this hypothesis

framework, under the null hypothesis each measurement in the dataset is viewed as a Bernoulli trial with the success defined as the measurement exceeding the standard. The total number of successes, under independence, thus has a Binomial distribution. The test is based on *x*, the observed number of exceedances, and $\hat{p}$, the sample proportion of exceedances. This method uses the same information as the raw score method to make decisions but allows for the control of the error rates. Therefore, sampling plans may be developed to set error rates at satisfactory levels for sufficient sample sizes.

In the frequentist binomial test, the Type I error rate is still calculated using $P(X > k \mid n, p_0)$, where *k* is the largest *x* that satisfies $P(X \leq x \mid n, p_0) \leq \alpha$. The maximum acceptable exceedance is achieved with the bounded error rate α. With this condition, the Type I error rate is always less than or equal to α. The Type II error rate is $P(X \leq k \mid n, p_1)$.

Type I and Type II error rates are displayed in Figure 2.4 when the sample size varies from 2 to 50 and the true exceedance proportion is 0.2. The Type II error rate is quite high for small sample sizes but is bounded by the Type I error rate. When a conventional (small) α value is used, the binomial test tends to not list sites that should be listed, especially when sample size is very small. Compared to the raw score method, the binomial test has lower Type I error and higher Type II error rates, i.e., a tendency to not list. Neither the raw score method nor the binomial test considers the actual numerical value of the standard. Moreover, the cutoff in the raw score method is determined by selecting an acceptable number of exceedances without acknowledging the presence of error rates, while the cutoff in the frequentist binomial test is determined by bounding the Type I error rate based on a selected acceptable error rate. Therefore, the raw score method can be viewed as a binomial method with a changing Type I error rate while the binomial method is a raw score method with a varying cutoff point.

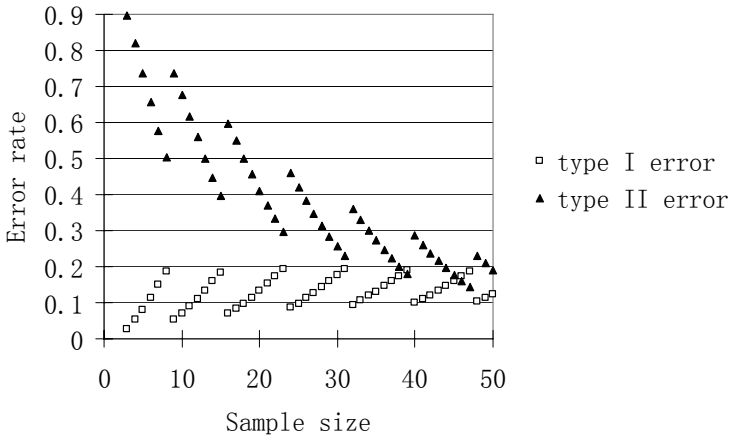

Figure 2.4 Error rates for different sample sizes (up to 50) with binomial test. Here p_0=0.1, p_1=0.2.

2.2.3 Bayesian approaches

Different from the raw score method and the frequentist binomial test, the Bayesian approaches consider prior information about violation. In the available literature, three types of Bayesian approaches are described to evaluate site impairment (Table 2.2). They are the general Bayesian approach (Ye and Smith, 2002), the Bayesian binomial approach (Smith et al., 2001; McBride and Ellis, 2001), and power prior approach (Duan, Ye and Smith, 2006). These three approaches all treat the parameter of interest (e.g., the exceedance probability, the percentile corresponding to the exceedance probability under normality, the mean and standard deviation of a normal distribution for measurements) as a random variable with an associated distribution. The prior distribution comes from previous information or expert opinion. Once data are collected, the information from the prior and from the data is combined to form the posterior distribution of impairment probability by Bayes rules. Based on the posterior distribution, a decision can be made using a posterior cutoff approach or Bayes factor approach.

Ye and Smith (2002) proposed a general Bayesian approach to use the raw numerical data (measurements) instead of the exceedance data (i.e., binomial data) for impairment evaluation. A reference prior (which can maintain certain frequentist properties while incorporating Bayesian flexibility) is applied to measurements from a normal population. The comparison of error rates shows that the Type I error rate of the general Bayesian approach is approximately equal to that of the binomial test. When sample size becomes large, the general Bayesian approach produces

similar Type II error rates as the raw score method. The main advantage of this approach is its flexibility to incorporate historic information into the current data.

Table 2.2 Prior distributions for different Bayesian frameworks

Method	Parameter of interest	Prior distribution
General Bayesian method	Parameters in underlying distribution (e.g., μ, σ in a normal distribution)	Reference prior, $\pi(\mu,\sigma) \propto \frac{1}{\sigma}$
Bayesian binomial approach	the exceedance proportion , *p*	1)Uniform prior, $\pi(p) \propto 1$ 2)Jeffrey's prior $\pi(p) \propto Beta(\alpha,\beta)$
Power prior method	the exceedance proportion, *p*, and power parameter, δ	$\pi(p) \propto 1$ $\pi(\delta) \propto Beta(\alpha,\beta)$ or $\propto 1$

Smith et al. (2003) considered a Bayesian binomial approach with a uniform prior on *p*. The Type II error rates are reasonable when sample sizes are large (e.g. >20) but the power is low for smaller sample sizes, which leads one to claim that a site is not impaired when in fact it is. McBride and Ellis (2001) put a prior distribution on the exceedance rate and compared performances of two reference prior (uniform and Jeffrey's), an optimistic user-defined prior and a pessimistic user-defined prior in the form of a beta-distributed prior. Jeffrey's reference prior is suggested for application to make compliance assessment less onerous especially for small sample sizes. When sample size is large, the information in the data dominates and all priors produce similar results.

Uncertainty due to insufficient data suggests incorporating information from adjacent sites or previous reports (i.e., historical data) and considering the weight or importance of this information. Duan et al. (2006) used a modified power prior to assess water quality. A power parameter is introduced to determine how much historical data is used in the current study. They use a uniform prior as the initial prior for *p* and adopt the posterior cutoff method as the decision rule. Relative to the Bayesian binomial approach, the power prior method has lower Type II error rates. This method improves power with small sample size and is potentially quite useful for water quality assessment.

Bayesian approaches seem to give more direct answers to balancing (McBride and Ellis, 2001) compliance probability (the probability that the test site has exceedance proportion less than or equal to the maximum acceptable exccedance proportion) and breach probability (the probability that the test site has exceedance proportion greater than the maximum acceptable exceedance proportion), since Type I and Type II errors are less relevant in a Bayesian framework and switching null and alternative hypotheses will not lead to different decisions.

2.2.4 Acceptance sampling by variables

Acceptance sampling is an important field in statistical quality control where "specific implemented sampling plans indicate the conditions for acceptance or rejection of the immediate lot that is being inspected" (Taylor, 1995). One categorization of this field consists of two types of sampling plans, acceptance sampling by attribute and by variables. The former deals with binary evaluation (the item is either defective or non-defective) and the number of defectives is used as the test statistic. This is ultimately a binomial test. The latter category takes the measured value of the item into account. Smith et al. (2003) adopted this technique to assess compliance with information on the frequency and magnitude of violations. The general procedure compares the allowable impairment probability with the estimated proportion of samples exceeding the standard, assuming a particular distribution for observations. An alternative approach is to compare the numerical value of a parameter (or a function of parameters) from the distribution with the estimate of this parameter from the sample. The distribution the parameter of interest comes from is a distribution that meets the standard and is related to the associated probability of the impairment.

Smith et al. (2003) discussed the acceptance sampling by variables technique for normally distributed data in the context of water quality assessment. Suppose a random sample with n measurements is taken from a water segment and the measurements have a normal distribution with unknown mean μ and standard deviation σ. Assume that the standard of interest is a specified lower limit, L. When the measurement is less than this lower limit, an 'exceedance' is recorded and thus the exceedance proportion is obtained from the n observations. Further, suppose that p is the true proportion exceeding the criterion and p_0 is the boundary value of p assuming compliance, i.e. the largest exceedance proportion under compliance. Under the

hypotheses, H_0: $p \le p_0$ (unimpaired, don't list) versus H_1: $p > p_0$ (impaired, list), the test statistic is

$$t = \frac{\bar{y} - L}{s/\sqrt{n}}$$

which has a non-central t distribution with (n-1) degrees of freedom and non-centrality $\lambda = -z_{p_0}\sqrt{n}$ under the null hypothesis. z_{p_0} is the $(100\,p_0)^{th}$ percentile of the standard normal distribution. $\bar{y}$ and s are the sample mean and sample standard deviation, respectively.

An alternative, equivalent version of the test is based on the statistic $t = \frac{\bar{y} - L}{s}$ with a rejection decision made when $t < m$, where $m = t(n-1, \alpha, \lambda)/\sqrt{n}$ and $t(n-1, \alpha, \lambda)$ is the critical coefficient from a non-central t distribution with (n-1) degrees of freedom and noncentrality parameter λ at the significance level of α. In terms of the relationship between mean and exceedance proportion, the above hypotheses for p can be equivalently set in terms of $\tau(p)$, the $(100\,p)^{th}$ percentile of the distribution. That is $H_0 : \tau(p_0) \ge L$ versus $H_1 : \tau(p_0) < L$, where $\tau(p_0)$ is the $(100\,p_0)^{th}$ percentile of the test site distribution. When the true exceedance proportion is p_1, the power for the null hypothesis boundary ($H_0 : \tau(p_0) = L$) is obtained by the following calculation.

$$\begin{aligned}
power &= P(\bar{y} - ms < L) \\
&= P\left(\frac{\bar{y} - L}{s/\sqrt{n}} < m\sqrt{n}\right) \\
&= P\left(\frac{\bar{y} - L}{s/\sqrt{n}} < \frac{t(n-1, \alpha, z_{1-p_0}\sqrt{n})\sqrt{n}}{\sqrt{n}}\right) \\
&= P\left(\frac{\frac{\bar{y} - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - L}{\sigma/\sqrt{n}}}{s/\sigma} < t(n-1, \alpha, z_{1-p_0}\sqrt{n})\right) \\
&= P(T(n-1, \lambda) < t(n-1, \alpha, z_{1-p_0}\sqrt{n}))
\end{aligned}$$

where $\lambda = \frac{\mu_1 - L}{\sigma/\sqrt{n}} = z_{p_1}\sqrt{n}$, μ_1 is the true site mean, and $T(n-1, \lambda)$ represents a random variable from a $t(n-1, \lambda)$ distribution.

Figure 2.5 compares the Type II error rate for the raw score approach, binomial test, and the acceptance sampling by variable method. All three methods have p_0=0.1 and p_1=0.2. In the binomial test, the Type I error rate varies from 0.028 to 0.184 depending on the sample size; in the non-central *t* test, $\alpha = 0.05$. Even though the Type I error rates are not the same for all the methods, the information from Figure 2.5 helps in evaluating the tradeoff of error rates for various methods. The Type II error rate for the raw score and binomial test has a saw shape (due to the discrete property of the binominal distribution) but is displayed as a decreasing curve for acceptance sampling by variables (due to the continuousness of the underlying distribution). The raw score approach has the smallest Type II error rate although with an increased Type I error rate. The approach of acceptance sampling by variables has similar error rates as the binomial test.

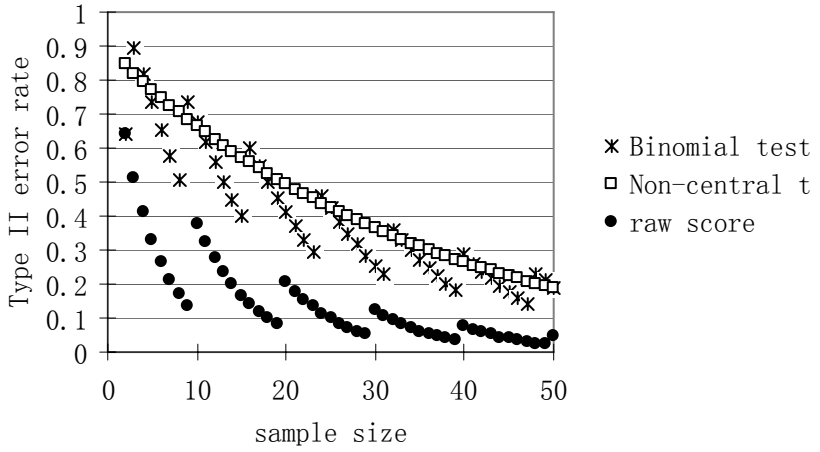

Figure 2.5 Type II error rate for raw score, binomial and acceptable sampling approaches.

2.2.5 Tolerance intervals and confidence limits for percentiles

A tolerance interval is an interval intended to cover a fixed proportion, *p,* of the population with specified degree of confidence. Applied to standards evaluation, the one-sided tolerance interval approach detects impairment by comparing a 100(1-α)% lower (or upper) tolerance limit calculated from the test site data for the $(100p)^{th}$ percentile with the criterion. This method ensures that no more than a specified percentage of water quality samples will exceed a standard with a high level of confidence. The limits in this method can be computed under the assumption

of a distribution (e.g., normal distribution). This approach is equivalent to acceptance sampling by variables. A variation on this idea is the use of a one-sided confidence limit for the desired percentile. Thus, given data, an upper or lower value is computed that gives a 100(1-α)% interval for the $(100\,p_0)^{th}$ percentile.

Similar to the acceptance sampling by variables approach, the tolerance interval approach also uses a non-central *t* distribution. An example of this approach is discussed by Gibbons (2003), who sought the 100(1-α)% lower bound on the$[100(1-p_0)]^{th}$ percentile (the upper percentile) of the distribution of sampled water quality measurements, assuming normally distributed data. The lower bound is computed as

$$LCL_{1-\alpha,p_0} = \bar{y} + K_{\alpha,p_0} s$$

where K_{α,p_0} is the one-sided normal tolerance limit factor for $100(1-\alpha)$% confidence ($\alpha < 0.5$) and $[100(1-p_0)]$% coverage. Specifically, $K_{\alpha,p_0} = m = t(n-1,\alpha,\lambda)/\sqrt{n}$, where $t(n-1,\alpha,\lambda)$ comes from the non-central *t* distribution with $\lambda = z_{1-p_0}\sqrt{n}$. When this estimated lower bound is greater than the criterion, the test site is claimed as impaired. This test is interested in the lower bound for $\tau(1-p_0)$ (the$[100(1-p_0)]^{th}$ percentile). When the percentile of interest is the lower percentile, the upper bound for $\tau(p_0)$ will be the focus. TL[11] in Section 3.2.5 is a general form of this type of test. Gibbons' test is a special case of a single sample.

Gibbons (2003) also examined lognormally distributed data and nonparametric confidence intervals based on order statistics for water quality assessment.

2.2.6 Prediction intervals

Prediction intervals are statistical intervals that are intended to describe possible future values with a specified degree of confidence. Prediction intervals or related tests have been suggested for two possible applications to water quality. In the context of groundwater monitoring, a fundamental problem is the "prediction of future measurements based on a background sample of historical measurements" (Gibbons, 1994, page 8). In environmental studies, sometimes the impairment is signaled by two-sided exceedance, i.e., low and high concentrations both indicate impairment. With this type of exceedance and assuming that there are n_r observations associated with baseline or reference conditions and interest is in a single future or test observation, the

corresponding interval at a specified confidence level α is $\bar{y}_r \pm t(n_r - 1, \alpha) s_r \sqrt{1 + 1/n_r}$, where $\bar{y}_r$ is the baseline mean and s_r is the baseline standard deviation. The test site is in compliance if a future value, y_{test}, is contained in the interval.

An equivalent view is to use the interval $(\bar{y}_r - y_{test}) \pm t(n_r - 1, \alpha) s_r \sqrt{1 + 1/n_r}$. Coverage of zero indicates the test site is not declared as different from the reference. Hence, a prediction interval may be used to test if a site is consistent with reference conditions.

A similar approach was proposed by Kilgour et al. (1998) to test biological metrics using reference criteria. Reference sites are usually employed to define the limits of acceptable and unacceptable conditions. To make a decision with a single observation from an impact site, Kilgour et al. (1998) formulated one possible test

$$F = \frac{(\bar{y}_r - y_{test})^2}{(s_r \sqrt{1 + 1/n_r})^2}.$$

This test statistic has an $F(1, n_r - 1)$ distribution under the null hypothesis, $H_0 : | \mu_r - \mu_{test} | = 0$, where μ_{test} is the mean at the test site. This approach is suitable for biological data since there is often only a single biological sample taken at each test site. This *F*-test is equivalent to the preceding *t*-approach as mentioned in their paper. The multivariate case is also discussed in their paper. A disadvantage of this *F* test is that it cannot give information about the violation frequency (i.e., the impairment proportion).

Gibbons (1994) describes variations of tests based on prediction intervals for groundwater applications that may be useful for the evaluation of biological metrics.

2.2.7 Permit limit approach

This approach implements an estimated permit limit based on a small number of concentration measurements (USEPA, 1991). The EPA proposed this approach for effluent data. Since effluent values are non-negative and positively skewed, a log-normal distribution is recommended for this type of data. A permit limit is defined as the $(100p)^{th}$ percentile of the concentration distribution. The calculation of the permit limit follows the equation,

$$C_p = Y_{(n)} \exp(\beta \hat{\sigma})$$

where C_p is the estimated $(100p)^{th}$ percentile, $Y_{(n)}$ is the maximum observation, $\hat{\sigma}$ is the estimated standard deviation and β is a coefficient. The value of β depends on the $(100p)^{th}$ percentile of interest and the significance level, α. It's obtained by the formula

$$\beta = \Phi^{-1}(p) - \Phi^{-1}(\alpha^{1/n}),$$

where $\Phi^{-1}(\bullet)$ is the inverse cumulative probability for a standard normal distribution. The variance is approximated from the empirical coefficient of variation by

$$\hat{\sigma}^2 \approx \ln(1 + C\hat{V}^2) = \ln(1 + s^2 / \bar{Y}^2)$$

where $s^2 = \dfrac{\sum (Y_i - \bar{Y})^2}{n-1}$ and $\bar{Y} = \dfrac{1}{n}\sum Y_i$.

The permit limit estimate depends on two features of the whole dataset, namely, the maximum observed concentration and the empirical coefficient of variation. When the calculated permit limit is greater than the standard, impairment is claimed. This approach tends to overestimate the true permit limit. It leads to conservative upper bounds that exceed actual concentrations by an order of magnitude (USEPA, 1991).

2.2.8 Concerns in standards assessment

Figure 2.6 categorizes some of the current methods used for standards assessment. The EPA raw score approach is a simple tally of exceedances. This non-statistical method has high power with the consequence of high Type I error rate. To make an impairment decision based on more rigorous methods, a number of statistical techniques may be implemented, including a binomial test, acceptance sampling, tolerance interval approaches, and Bayesian implementations using prior distributions. These techniques try to balance Type I and Type II error rates. In practice, the small sample size and reference conditions have been the recent focuses in environment studies. These concerns are not incorporated in current standards assessment methods.

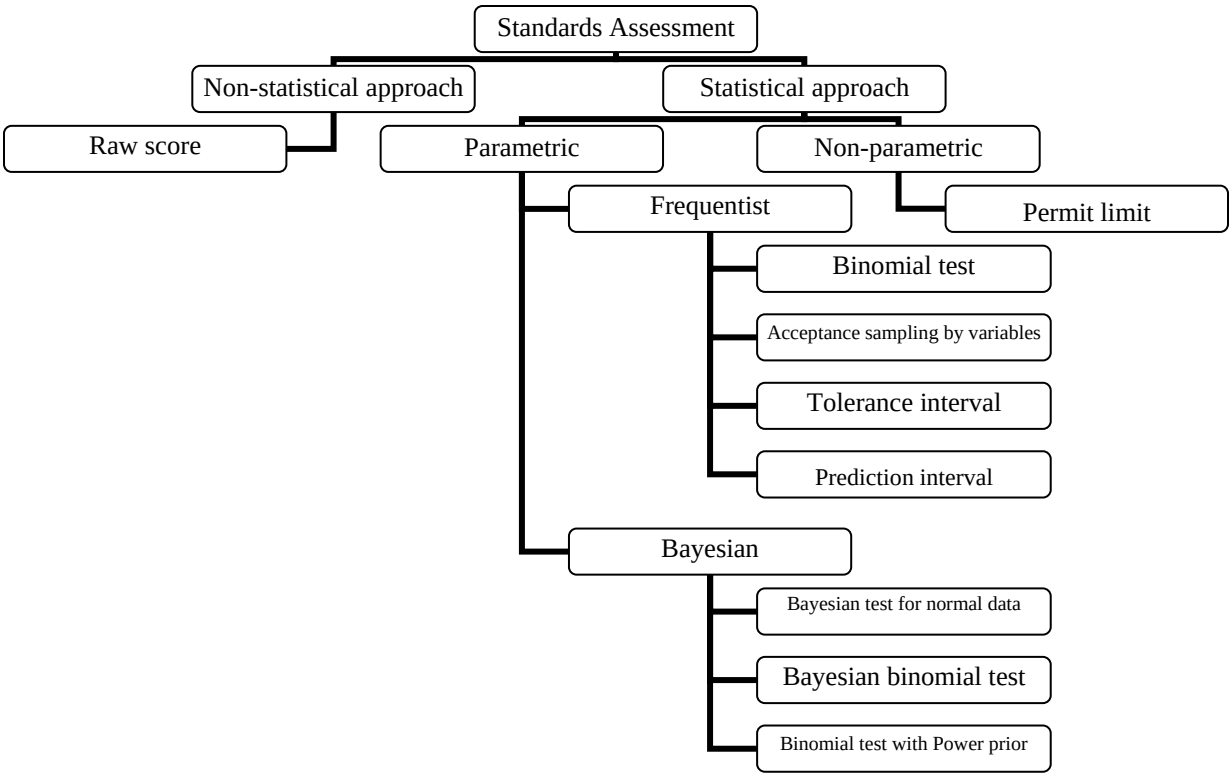

Figure 2.6 Diagram of current assessment approaches.

Power for small sample sizes and regional concern

The lack of power is a big concern in standards assessment since sample sizes for assessment monitoring are frequently small. For example, in Virginia, assessment decisions are made every two years using the past five years of data. When data are sampled quarterly, there will only be 20 samples if there are no missing values. Biological data tend to have even fewer samples. Given the increase in the number of constituents that are measured, it is unlikely that sample sizes will increase in routine monitoring. It's important to find appropriate statistical techniques to improve assessment accuracy for small samples. An additional concern is the change in sampling programs from a fixed set of stations to randomly selected stations. The use of probabilistic sampling approaches may exacerbate the problem since these programs often use only a single sample per site (i.e., a sample at one time) or a rotating scheme for sampling that might include only a few sampled times per site.

Although the sampling design tends to focus on regions, decisions are typically made on a site-by-site basis. A suggestion was recently made to use the probabilistic sampling strategy

based on a five-year period for sampling (Urquhart and Kincaid, 1999). Each site may be sampled only once in each period. This scheme is useful for regional inference although lacking power for site-specific inference. Also this strategy has some potential problems, such as the homogeneity of sites. The possibility of auto-correlated sampling error across time should also be considered before carrying out tests.

Reference sites

With the emergence and popularity of the River Invertebrate Prediction and Classification System (RIVPACS) (Wright et al., 2000), approaches based on reference conditions have attracted attention in water quality assessment. The sites that meet reference conditions are called reference sites. The deviation of a test site from reference conditions is used as a measure of the effect of stressors on the ecosystem (Bailey et al., 2005).

With biological data, a criterion is based on sampled data from sites that are considered to be reference sites. The value of the criterion is either a value that represents a specific percentile of the estimated distribution of reference values or is an adjusted value based on expert opinion or studies comparing reference and non-reference sites. In my study, the 10^{th} percentile of the distribution of reference values will be assumed to be the basis for the biological criterion. Since stressors typically harm biological communities, larger values of biological metrics will be viewed as healthy.

When there are multiple samples taken at a given test site, the means of the test samples are compared by a two-sample test. If tests are non-significant, multiple samples are treated as one sample. Otherwise, a multiplicity adjustment will be carried out to protect the site-wide (or region-wide) Type I error rate when evaluating the individual samples. The multiplicity adjustment is widely used in many areas but has not been considered in environmental studies. An example of a conservative approach to multiplicity adjustment is the Bonferroni approach. Instead of α, one would use α/k in the decision rule, where k is the number of samples taken at the site. The test site is claimed to be impaired if any of the samples is not in the adjusted interval.

To tackle these problems and concerns, model-based tests have come to the fore. A brief summary of this type of tests will be given in Section 2.3. My proposed model-based tests will be depicted in Section 2.4.

2.3 Model-based estimation for assessment

When sample size is very small, direct estimators are likely to have unacceptably large standard error. To get smaller error rates for small sample sizes, techniques of small area estimation are adapted to standards evaluation (Bell and Carolan, 1998). Information from related areas is borrowed to get indirect estimators and predict impairment probability with associated uncertainties. One of the common small area estimation methods is to use the mixed model. The use of model-based estimation may increase the power of the assessment test and allow for a regional view for site assessment. This method can increase the power of the test. Also, it incorporates concerns discussed in the preceding section.

The general form of the linear mixed model is

$$\underline{y} = X\underline{\beta} + Z\underline{u} + \underline{\varepsilon},$$

where $\underline{y}$ denotes the vector of observed values across all sites and times, X and Z are known matrices describing fixed and random effects, $\underline{\beta}$ is the vector of unknown fixed-effects parameters, $\underline{u}$ is a vector of unknown normal random-effects parameters, and $\underline{\varepsilon}$ is a vector of normal random errors independent of $\underline{u}$ such that $\underline{\varepsilon} \sim \boldsymbol{MVN}(\underline{0}, \Gamma)$ and $\underline{u} \sim \boldsymbol{MVN}(\underline{0}, G)$. The matrix Γ describes the variance properties of errors and the matrix G describes the random effects. These quantities ($\underline{\beta}$, $\underline{u}$, Γ, and G) are estimated by estimated generalized least squares (EGLS). The fixed effects models, random effects models, and regression models discussed later are all special cases of the linear mixed model.

2.3.1 Smith's random-model tolerance interval

An example of using the linear mixed model is given by Smith (2002) who used a two-way random effects model to construct tolerance intervals in environmental regulations. His proposed model consists of four variances components: a temporal random effect (δ_i), a spatial random effect (β_j), spatial-temporal interaction (γ_{ij}), and measurement error (e_{ij}). This leads to the model as below.

$$y_{ijl} = \mu + \delta_i + \beta_j + \gamma_{ij} + e_{ijl}$$

Here, y_{ijl} is the l^{th} observation at time point *i* and site *j*. The general formula for a parametric one-sided lower tolerance interval bound for the overall mean using the reference sites is

$$b_{p,\alpha} = \bar{y} - K_{p,\alpha} s$$

where $\bar{y}$ is the estimate of an overall mean μ (computed as the mean of all observations in the sample) and *s* is the estimated standard deviation which depends on the computation method. The constant $K_{p,\alpha}$ is computed so that the estimated interval bound will "fail to cover the underlying $(100p)^{th}$ population percentile (100α)% of the time" (Smith, 2002), i.e., $P(b_{p,\alpha} \geq \tau(p)) = \alpha$. The tolerance-interval bound is calculated by a computational or bootstrap method. When a single observation is less than the computed tolerance-interval bound, $b_{p,\alpha}$ (also called the tolerance limit), impairment is declared; otherwise, the test site will be treated as 'good' site.

2.3.2 Regression-based estimation

The random-model-based assessment proposed by Smith (2002) basically implements the methodology of the analysis of variance (ANOVA). The general model discussed before Section 2.3.1 also refers to regression situations. These occur, for example, when there are covariates (such as elevation) or stressors that might be included in the modeling process.

Regression methods are commonly applied to study how biology varies with covariates and/or stressors. One currently popular application of simple regression in environmental assessment is the Benthic Assessment of Sediment (BEAST) method used in Canada (Barbour et al., 1999). BEAST uses regression analysis to adjust a response by the value of the covariate. For instance, McCormick et al. (2001) used regression to adjust variables by watershed area so that all response variables are nonnegative. In their work, response variables were regressed on watershed area using reference data. Then all sites were fit with the reference regression equation to get residuals. After this, residuals were adjusted by a criterion value from the regression equation. No formal assessment tests were simultaneously carried out with regression.

Other regression methods such as generalized additive models (GAM) may be used to estimate the response-stressor relationships at test sites (Heegaard et al., 2001; Yuan and Norton, 2003). This type of model relaxes the assumption of response function and visualizes the relationship.

Among the available literature, there is no research using regression-based tests for impairment assessment in univariate case.

In the multivariate case, ordination methods are widely-used to "reveal the relationships between ecological communities graphically" (Legendre et al., 1998). However, no formal impairment test has been proposed to depict the complex ecosystem by using multivariate approaches.

2.4 Proposed model-based tests for assessment improvement

The region of interest and hence the data for neighborhoods is complicated in practice. The region is a group of sites which have some common attributes. The sites in the region are either test site or non-test sites. My proposed model-based tests deal with these practical concerns, such as the setting of a numerical criterion for a standard, the sample strategy, the use of neighborhoods, and the causes of water quality impairment. Analysis of variance models (specifically fixed and random effects models) and simple regression are proposed and evaluated in the univariate case. The integration of random tessellations and redundancy analysis is proposed to explore the response-stressor relationship and investigate potential impairment in the multivariate case.

2.4.1 Model-based assessment using ANOVA modeling

Tolerance limits provide bounds to distinguish reference from non-reference conditions. When the focus is on a future observation or future site, the prediction interval can also help impairment detection. Moreover, the region of interest is complicated in practice. There is an increasing demand for incorporating practical factors in water quality assessment in model-based tests. Therefore, different from Smith's work (2002), I consider the tolerance limit and prediction limit with model-based estimation to assess impairment. The underlying model comes from analysis of variance (ANOVA). The performance of model-based tests is evaluated according to the sampling strategies and the setting of numerical criterion for a standard.

With the development of probabilistic sampling strategies, measurements are made at k randomly selected sites over either random or systematic time periods. The ANOVA model considered is a fixed or random effects model with one effect, a site effect. It is assumed that there may be multiple observations at the site of interest. For small sample sizes the assumption

of normality of data is important to analyses (Kilgour et al., 1998). In my study, I make use of methods relying on the assumption that data come from normal distributions. Once model-based estimation is obtained, site-oriented comparisons are used to evaluate compliance based on a tolerance limit or prediction limit. Test properties are assessed relative to underlying factors such as sample size, regional impairment proportion, and error structure (heterogeneity of variance and dependence). Chapter 3 and Chapter 4 detail the methods and application.

2.4.2 Regression-based univariate analyses

When the regression method is used in assessment, it may be possible to strength the test of impairment, especially when there is little information for setting a numerical criterion or the criterion varies with a covariate (Urquhart, 1982; McCormick et al., 2001). The generalized regression-based tests in my study consider two scenarios. In the first scenario, the criterion used in impairment assessment is based on the regression model of the stressor-response relationship. The test site is evaluated by comparing its measurements with calculated limits (tolerance limits or prediction limits) as well as the criterion of the stressor for impairment detection. In this scenario, stressor and response variables both contribute directly to defining limits. In the second scenario, regression is used to adjust the estimated limit of the response by covariate. Chapter 5 discusses this topic.

2.4.3 Regression-based multivariate analyses

In the multivariate case, ordination methods are a widely-used family of methods to represent the functioning of ecosystem (Legendre et al., 1998). Canonical (or constrained) ordination is a special type of ordination designed to detect patterns of variation within biological survey data that can be ‘explained’ by observed environmental variables. It uses multivariate techniques to arrange sites along environmental axes which are determined from data collected on species composition (ter Braak, 1986). The technique can be thought of as a type of multivariate regression in which multiple dependent variables (metrics, Y) are regressed on multiple independent variables (environmental attributes, X). Canonical ordination has great appeal as a means of extrapolating biological patterns sampled at a small number of field survey sites across large areas of land. Among the various forms of canonical ordinations, redundancy analysis (RDA) and canonical correspondence analysis (CCA) have become popular choices for ecological research since they recognize different roles for the explanatory and response

variables (Makarenkov and Legendre, 2002). Both are special cases of reduced-rank regression (ter Braak, 1994).

Canonical correspondence analysis (CCA, ter Braak, 1986) is by far the most widely used canonical ordination technique in ecology. This method looks for the 'ideal' linear combination of explanatory variables (i.e., environmental variables) that best explains the count data. The basic procedure of CCA is to linearly regress the transformed Y on the standardized X and then use the singular value decomposition for the matrix of fitted value to compute the site and column scores (Lipkovich, 2002). This analysis allows for simultaneous display of sites, metrics, and environmental variables on the same ordination diagram. Redundancy analysis (RDA) was introduced by Rao (1964). It is an extension of the multiple linear regression. RDA uses a linear model of relationship among the response variables and relationship between the explanatory and response variables. RDA can be looked at as a combination of principal component analysis (PCA) and linear regression.

Environmental data are usually collected over large spatial regions. To distinguish ecosystem patterns in different regions, cluster analysis methods are needed in order to cluster sites in terms of geomorphologic similarity. The proposed model-based clustering combines a special partition method (random tessellations) and a multivariate approach (redundancy analysis) to assess water quality for large ecological datasets. The random tessellation approach is used to generate clusters of sites that may have a similar stressor-response relationship. It can improve assessment by better determining environmental relationships. Chapter 6 addresses this analysis.

The rest of this dissertation is organized as follows. Chapter 3 and Chapter 4 investigate the model-based assessment using the analysis of variance method and some small area estimation techniques. Chapter 5 formulates regression-based tests in impairment assessment. Chapter 6 applies regression-based multivariate tests (based on random tessellations and redundancy analysis) to the Mid-Atlantic Highlands benthic data. Chapter 7 summarizes research findings as well as potential topics for future research.

3. Model-based Assessment Using Tolerance Limits

3.1 Introduction

3.1.1 Water quality assessments

Water quality assessment is concerned with the evaluation of water quality at individual sites or regions. The assessment, as carried out for the Clean Water Act, involves samples collected over time for a given site. From a statistical perspective, evaluation involves a certain exceedance proportion at a specified significance level, which typically leads to a limit "incorporating sampling error and corresponding to a relevant quantile" of the distribution of an environmental variable (Smith, 2002). The limit functions as the numerical representation of a standard, defined in Chapter 2. In terms of setting the numerical criterion for a standard, there are two ways to assess water quality in a statistical context.

When the standard is a fixed value — either defined by regulatory agencies or derived from prior studies, the estimated limits from samples will be compared with the fixed numerical criterion using either a test or tolerance limit (Smith et al., 2003). Error rates are controlled through either the setting of the Type I error rate for testing methods or by specifying a confidence level for interval-based approaches. A common limit is the tolerance limits, defined as "the upper or lower confidence-interval bound of a quantile of the underlying data distribution" (Smith, 2002). The tolerance limit and associated test statistic is discussed in Section 3.2.

When the standard comes from reference or background conditions (this type of standard fully depends on available reference conditions), a single measurement will be compared to the derived numerical criterion (Reynoldson et al., 2000; Smith, 2002; Wright et al., 2000; VDEQ, 2006). The derived criterion can be in the form of a one-sided prediction limit or tolerance limit. When the criterion is derived as a prediction limit, the impairment decision for a single measurement is viewed as a prediction problem (Kilgour et al., 1998) and will be detailed in Chapter 4. When the standard is derived as a tolerance limit (for instance, Smith (2002) compared tolerance limits from

reference data to a single future measurement), the corresponding evaluation follows a procedure similar to the one discussed in this chapter.

In this chapter, the standard is treated as a fixed numerical criterion that corresponds to the maximum acceptable exceedance proportion (set to 0.1). The lower criterion is treated as the quantity of interest, with measurements below the criterion indicating impairment. Thus a one-sided tolerance limit (i.e., the upper $100(1-\alpha)\%$ confidence-interval bound for the 10^{th} percentile) will be calculated from chemical or biological samples (collected from sites) and compared to the fixed value of the standard (criterion). When this estimated one-sided tolerance limit is less than the criterion, the test site is declared to be impaired. When an upper criterion is of interest, similar procedures to the one discussed here can be implemented to make an assessment decision.

When the evaluation of a specific site is of interest, the available measurements at that site may be limited to a small number of samples. To improve estimation and reduce uncertainty, other sites in the same region may be used for estimating parameters common to these sites. Model-based tests are thus proposed to "borrow strength" (Rao, 2003) from related areas and improve the test performance. The underlying models are partly determined by the sampling plan that is used by the researchers and typically included random components as well as fixed ones.

The rest of this chapter is organized as follows. In Section 3.1.2, some of the sampling plans that are used with site evaluation are discussed. These lead to both fixed and random effects models. Section 3.2 will discuss the relationship between the exceedance proportion and one-sided tolerance limit as well as the general model set-up. The test procedure, test statistics, and estimation of unknown parameters will be described. Section 3.3 will compare the model-based tests with a single site test (i.e., acceptance sampling by variables) and investigate the effects of the factors on model-based tests. Section 3.4 will discuss some miscellaneous topics and summarize the key findings.

3.1.2 Sampling plans for environmental assessment

An important component of environmental assessment is the sampling plan used to collect the data. The method for site selection affects the model that is used to evaluate a specific site. Traditionally, sites were selected in a non-random fashion for many years of sampling. This approach leads to good estimates of parameters associated with a particular site but leads to poor estimates of regional parameters. In particular, it is difficult to estimate parameters such as the proportion of sites within a region that are impaired. In this section, some of the strategies that are used to estimate regional effects as well as local effects are described.

The most common sampling plan (design) focuses on fixed sites (stations). Many states implement the "always-revisit" sampling plan (i.e., the fixed sampling plan) in the 303(d) process to evaluate sites under the Clean Water Act. Under this sampling strategy, fixed sites are sampled on a periodic basis whose dates are often the same for a group of sites. The spatial effect (i.e., location of the site) is treated as fixed and a fixed effects model can be used for estimation and tests.

A fixed site approach may lead to non-representative samples relative to a region of interest and may not be useful for estimation of regional status (i.e., estimating the proportion of sites in a region that are impaired). Recently researchers have used the probabilistic sampling for monitoring, especially with biological metrics. The sampling approach divides the region into a hexagonal grid then selects one or more sites within each hexagon. This results in a spatially balanced sample that is quite effective at estimating regional effects. The design is not very useful for local effects since there are often few sites used for collecting data over time.

The need to balance regional estimation with site-specific analysis has lead to the rotating panel sampling plan as a compromise (Urquhart et al., 1993). Generally speaking, a panel is a group of sites which have some common attributes. This group is usually defined spatially. There can be several sites or only one site in a panel. For example, all sites in a watershed can be treated as a panel. A site in a basin can also form a panel. Rotating panel studies vary the panels that are sampled in different time periods. For example, in a 5-year rotating panel design, sites are visited once every five years

(Urquhart and Kincaid, 1999). Site selection and visit schedule (therefore the spatial and temporal effects) are main components of panel studies.

In practice, the 5-year rotating panel design can be implemented in many different ways and two possible implementations are described below. One way is to visit only one panel each year. There is no revisit to a site within the panel during the first 5 year period. After all of the five panels are visited, this procedure is repeated beginning in year 6 (Table 3.1a). In this way, sites in panel 1 are visited in years 1, 6, 11, etc.; sites in panel 2 are visited in years 2, 7, 12, etc. Another way is to visit sites in 5 panels in year 1; in year 2, one of the panels visited in year 1 is dropped and replaced by a new panel. After 5 years, the panels sampled in year 1 are all dropped (Table 3.1b). The revisit proportion is 80% for each year. Random effects models are fit to incorporate random variability in the rotating panel sampling plans (Urquhart and Kincaid, 1999).

Table 3.1a Scheme of five-year rotating panel plans without revisit

	Time periods												
Panel	1	2	3	4	5	6	7	8	9	10	11	12	…
1	X					X					X		
2		X					X					X	
3			X					X					
4				X					X				
5					X					X			

Table 3.1b Scheme of five-year rotating panel plans with a fixed revisit proportion (80%)

	Time periods						
Panel	1	2	3	4	5	6	…
1	X						
2	X	X					
3	X	X	X				
4	X	X	X	X			
5	X	X	X	X	X		
6		X	X	X	X	X	
7			X	X	X	X	
8				X	X	X	
9					X	X	
10						X	

Different sampling plans result in different types of data and require different models to represent spatial and temporal effects (Smith, 2002; Urquhart and Kincaid, 1999). My study will focus on tackling assessment problems when the data are collected within a short period across a relatively large region and will explore the spatial aspect of assessment samples. Thus the spatial effect will be the only effect directly incorporated into the model in my study.

Even though probabilistic sampling plans are becoming more popular, the fixed-sampling plan is still used and applicable at the level of ecoregions. Both the fixed effects model and random effects model will be discussed here for water quality assessment using the one-sided tolerance limit. The standard of interest is assumed to be fixed and a low value (value below a given criterion) indicates impairment. The tolerance limit is calculated based on the one-way fixed or random effects model. The fixed effects model in this chapter corresponds to the "always-revisit" sampling strategy. The random effects model in this chapter corresponds to the sampling strategy that each panel consists of a single site and all panels are randomly selected from a region.

3.2 Model-based tests

If the sample size at the test site is small, then direct estimators (i.e., using data from a single site) are likely to yield large standard errors leading to a high Type II error rate in the decision procedure. This small sample estimation problem can be helped by "borrowing strength" from related areas to get indirect estimators for the test site if the sampled sites in a region are connected in some way. Generally, the variance of an indirect estimator is smaller than that of the corresponding direct estimator since the former uses all available information from the sites in the region. The uncertainty about decisions can be reduced under a reasonable model (Schaible and Casady, 1994). Model-based tests are thus proposed for improving the assessment process.

3.2.1 Exceedance proportion and one-sided tolerance limit

The notation for the general exceedance proportion and lower standard (or limit) are p $(0 \le p \le 1)$ and L, respectively. Evaluation of impairment is based on observations y_{ij}, where i=1,2,…,k+1 represents the site number and j=1,2,…, n_i represents the sample time

points for site *i*. The site corresponding to *i*=*k*+1 is considered the site of interest, i.e., the test site. The measurement, *Y*, at the test site is assumed to come from a normal distribution, $N(\mu_{k+1}, \sigma^2)$. The exceedance proportion for this site is defined along with the standard or numerical criterion, *L*, as follows.

$$p = \Pr(Y < L) = \Pr(\frac{Y - \mu_{k+1}}{\sigma} < \frac{L - \mu_{k+1}}{\sigma}) = \Phi(\frac{L - \mu_{k+1}}{\sigma})$$

where $\Phi(\bullet)$ is the cumulative probability for a standard normal distribution. This definition can be expressed by the equation, $\frac{L - \mu_{k+1}}{\sigma} = z_p$, where z_p is the $(100p)^{th}$ percentile of the standard normal distribution. The exceedance proportion in the samples collected at a site is called the estimated exceedance proportion, denoted by $\hat{p}$.

The maximum acceptable exceedance proportion (p_0) or criterion (*L*) is set by a regulatory agency and used for assessment. In the statistical context of water quality assessment, p_0 and *L* are co-defined when the exceedance proportion of a site is exactly the maximum acceptable exceedance proportion, i.e., $p = p_0$. In that situation, the site is called a baseline site which has the exceedance proportion equal to p_0. With $\tau(p)$ indicating the $100p^{th}$ percentile, this situation implies that $\tau(p_0) = L$. Figure 3.1 displays the relationship between the exceedance proportion (*p*), the percentile ($\tau(p)$), and the site mean (μ) under the normality assumption.

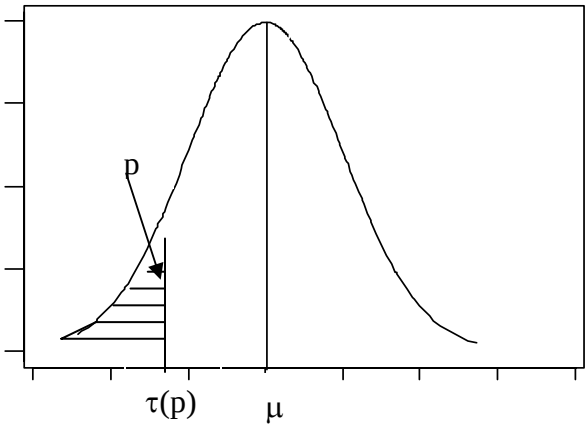

Figure 3.1 Relationship between the $100p^{th}$ percentile ($\tau(p)$) and mean (μ) for a normal distribution.

3.2.2 General model set-up

When setting up the fixed or random effects model for model-based tests, there are several underlying assumptions. First, each site has multiple observations and the site quality (i.e., water quality at this site) is represented by the site mean. Second, normality is assumed for the distribution of measurements. Third, the variance of sites doesn't change with the impairment level (i.e., the same variance holds for all unimpaired or impaired sites in the region).

The basic model is set as follows.

$$\underline{y}_i = \mu_i \underline{1} + \underline{\varepsilon}_i \qquad i = 1,...,k,k+1$$

where $\underline{y}_i$ is a $n_i \times 1$ vector, representing the observed response variable at site *i*, and $\underline{1}$ is a $n_i \times 1$ vector of 1's. The data are collected at n_i time points for site *i*. The sample size may vary across sites. N is the total number of observations i.e. $N = \sum_{i=1}^{k+1} n_i$. The error term $\underline{\varepsilon}_i$ follows a normal distribution, $N(\underline{0}, \sigma_i^2 I_{n_i \times n_i})$. The standard deviation, σ_i, is assumed unknown. With different variance at each site, borrowing information from other sites can bias tests. In practice, there is little information about heterogeneity of variance. In my study, except for Section 3.4.4, the σ_i's are assumed to be the same for the region, equal to σ_e. The $(k+1)^{\text{th}}$ site is viewed as the test site and the other *k* sites are non-test sites. The non-test sites can be either impaired or unimpaired.

When the underlying model is a fixed effects model, the mean of the i^{th} site, μ_i, is an unknown parameter. Comparing parameters from different sites (e.g., the difference between site means, $\mu_i - \mu_j$, and the square root of variance ratio, σ_i / σ_j) provides the means for assessing the similarity among sites. Regional homogeneity occurs when all of the sites in the region come from the same distribution with the population mean equal to μ and the population standard deviation equal to σ_e.

When the underlying model (used for model-based tests) is a random effects model, the mean at the site is viewed as a random variable. Since the randomly sampled sites may be still of interest in the assessment, the site mean μ_i is thus worth investigation.

The site mean μ_i is a random variable, which is assumed to be normally distributed with the mean μ and standard deviation σ_A, i.e., $\mu_i \sim N(\mu, \sigma_A^2)$. This random term is usually expressed as $\mu + A_i$, where μ is the overall mean and A_i is a random effect due to the i^{th} site. The A_i's and ε_{ij}'s are both independent random variables, and their covariance is zero (i.e., $\text{cov}(A_i, \varepsilon_{ij}) = 0$). The random components σ_A and σ_e are unknown. The intraclass correlation $\rho = \dfrac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$ is also unknown but sometimes is treated as a known parameter by replacing it with an estimator. The variance covariance matrix of $\underline{y}_i$ is $\sigma_A^2 J_{n_i} + \sigma_e^2 I_{n_i}$, where J_{n_i} is a matrix of one in all its positions. This implies that observations within sites are correlated over time periods but the correlation is constant. When the random effects model has σ_A equal to 0, the random effects model is reduced to a fixed effects model with the single mean μ.

3.2.3 The null and alternative hypotheses

Fixed standards are usually used for basic water quality parameters (e.g., dissolved oxygen, biological oxygen demand, and coliform bacteria). In practice, exceedance of the standard is an indication that the site may be impaired. An impairment decision is based on a maximum acceptable exceedance proportion, usually 10% (USEPA, 1987). Therefore, the hypotheses to be evaluated are $H_0 : p \le p_0$ (unimpaired, don't list) versus $H_1 : p > p_0$ (impaired, list) (Smith *et al.*, 2003), where p is the population exceedance proportion (i.e., the violation frequency) at the test site, and p_0 is the maximum acceptable exceedance proportion, equal to 0.1. The test statistic employs the boundary of the null hypothesis $H_0 : p = p_0$.

Alternative settings for the hypotheses of impairment can be written in terms of the mean or the associated percentile. Suppose the standard of interest is a lower standard. If the population is normal, the distribution having $p = p_0$ is $N(\mu_L, \sigma_e^2)$ where $\mu_L = L + z_{1-p_0}\sigma_e$. Testing $p \le p_0$ is equivalent to testing $\dfrac{\mu_{k+1} - L}{\sigma_e} \ge \dfrac{\mu_L - L}{\sigma_e}$ under the

normality assumption. The hypothesis for the exceedance proportion can thus be stated in terms of the site mean (i.e., violation magnitude), namely $H_0 : \mu_{k+1} \geq \mu_L$ versus $H_1 : \mu_{k+1} < \mu_L$, where $\mu_{k+1} = L + z_{1-p}\sigma_e$ and $\mu_L = L + z_{1-p_0}\sigma_e$. The null hypothesis can also be written as $\mu_{k+1} - z_{1-p_0}\sigma_e \geq L$, which compares the $(100\,p_0)^{th}$ percentile of the site distribution to the lower standard. In this sense, the original hypotheses of exceedance proportion can be further translated to $H_0 : \tau(p_0) \geq L$ versus $H_1 : \tau(p_0) < L$, where $\tau(p_0)$ is the $(100\,p_0)^{th}$ percentile of the test site distribution.

3.2.4 The test statistic

The water quality assessment test may be based on a tolerance limit associated with an exceedance proportion or a test statistic. The relationship between the one-sided tolerance limit for a normal distribution and the associated non-central *t* distribution can be applied to build the test statistic (Owen, 1968). For a test based on data from a single site, the test statistic is written as $\frac{\bar{y} - L}{s/\sqrt{n}}$ (Smith et al., 2003), where the estimator for the mean of the test site is $\bar{y}$, s is the sample standard deviation of the test site, and *n* is the sample size for the test site (refer to Section 2.2.4 and 2.2.5). This test statistic follows a non-central *t* distribution, $t(df, \lambda)$, whether or not H_0 is true. The term $t(df, \lambda)$ denotes a non-central *t* distribution with degrees of freedom equal to *df* and noncentrality parameter equal to λ.

For model-based tests, I still estimate the site mean by the sample mean, while the standard deviation is a pooled estimator. Under the null hypothesis and normality assumption, the non-central *t* distribution will hold for the test statistic in form of $\frac{\hat{\mu}_{k+1} - L}{se(\hat{\mu}_{k+1})}$. Here $\hat{\mu}_{k+1}$ is the model-based estimator for the mean of the test site and $se(\hat{\mu}_{k+1})$ is the square root of mean squared error (MSE) of the estimator. The test statistic $\frac{\hat{\mu}_{k+1} - L}{se(\hat{\mu}_{k+1})}$ will be referred to as TS[1] in the remainder.

The alternative way to construct the test statistic starts from finding an unbiased estimator of the percentile. Under the normality assumption, an unbiased estimator of the

$(100p)^{th}$ percentile of the test site, $\tau(p) = \mu_{k+1} + z_p\sigma_e = \mu_{k+1} - z_{1-p}\sigma_e$, is obtained by $\hat{\tau}(p) = \bar{y}_{k+1} - z_{1-p}c_n s$, where $\bar{y}_{k+1}$ is the mean of the test site and c_n is the bias correction term for the estimated standard deviation s (Holtzman, 1950; Sobel and Tong, 1976; Bement and Pirkle, 1981) satisfying $E(c_n s) = \sigma_e$ for s^2 being an unbiased estimator of σ_e^2. When the sample standard deviation is estimated by a single site with the sample size equal to *n*, the bias correction term is calculated by $c_n = [(n-1)/2]^{1/2}\Gamma(\frac{n-1}{2})/\Gamma(\frac{n}{2})$ (Equation [3] in Holtzman's 1950 paper), where $\Gamma(\bullet)$ denotes the gamma function, i.e., $\Gamma(y) = \int_0^\infty t^{y-1}e^{-t}dt$. c_n converges to 1 as $n \to \infty$. The variance of this estimator is obtained as follows.

$$
\begin{aligned}
Var(\bar{y}_{k+1} - z_{1-p}c_n s) &= Var(\bar{y}_{k+1}) + (z_{1-p}c_n)^2 Var(s) \\
&= \frac{\sigma_e^2}{n} + (z_{1-p}c_n)^2[E(s^2) - [E(s)]^2] \\
&= \frac{\sigma_e^2}{n} + (z_{1-p}c_n)^2[\sigma_e^2 - (\frac{\sigma_e}{c_n})^2] \\
&= (\frac{1}{n} + (z_{1-p}c_n)^2 - z_{1-p}^2)\sigma_e^2
\end{aligned}
$$

Note that the variance converges to $\frac{\sigma_e^2}{n}$ as $n \to \infty$.

When the sample standard deviation is estimated by all (*k*+1) sites in the region, the bias correction term is derived as follows.

$$c_n = [(k+1)(n-1)/2]^{1/2}\Gamma(\frac{(k+1)(n-1)}{2})/\Gamma(\frac{(k+1)(n-1)+1}{2})$$

The variance of the estimated percentile is $(\frac{1}{n} + (z_{1-p}c_n)^2 - z_{1-p}^2)\sigma_e^2$ with $c_n = [(k+1)(n-1)/2]^{1/2}\Gamma(\frac{(k+1)(n-1)}{2})/\Gamma(\frac{(k+1)(n-1)+1}{2})$. For a fixed value of *k*, c_n again converges to 1 as $n \to \infty$. With this set-up, the test statistic for the compound hypotheses, $H_0 : \mu_{k+1} - z_{1-p_0}\sigma_e \geq L$ versus $H_1 : \mu_{k+1} - z_{1-p_0}\sigma_e < L$ with

$\tau(p_0) = \mu_{k+1} - z_{1-p_0}\sigma_e$, can be easily derived. At the hypothesis boundary, $\tau(p_0)$ takes the value *L*. Based on the pooled estimate of the standard deviation, if the true exceedance proportion is *p*, the test statistic at the hypothesis boundary is as follows.

$$\frac{\hat{\tau}(p_0) - L}{s\hat{e}(\hat{\tau}(p_0))} = \frac{\bar{y}_{k+1} - z_{1-p_0}c_n s - L}{s\sqrt{\frac{1}{n} + (z_{1-p_0}c_n)^2 - z_{1-p_0}^2}} \overset{H_0}{=} \frac{\bar{y}_{k+1} - z_{1-p_0}c_n s - (\mu_{k+1} - z_{1-p}\sigma_e)}{s\sqrt{\frac{1}{n} + (z_{1-p_0}c_n)^2 - z_{1-p_0}^2}}$$

where $s = \hat{\sigma}_e = \sqrt{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 / [(k+1)(n-1)]}$. Denote this statistic as TL[2]. With $a_n = \sqrt{1 + n(z_{1-p_0}c_n)^2 - z_{1-p_0}^2}$, the preceding test statistic can be written as below.

$$\frac{\bar{y}_{k+1} - z_{1-p_0}c_n s - (\mu_{k+1} - z_{1-p}\sigma_e)}{s\sqrt{\frac{1}{n} + (z_{1-p_0}c_n)^2 - z_{1-p_0}^2}} = \frac{\sqrt{n}(\bar{y}_{k+1} - \mu_{k+1})/\sigma_e + \sqrt{n}z_{1-p}}{\frac{s}{\sigma_e}a_n} - \frac{\sqrt{n}z_{1-p_0}c_n}{a_n}$$

$$= \frac{1}{a_n}[TL[1] - b_n]$$

where $b_n = c_n\delta$, $\delta = \sqrt{n}z_{1-p_0}$, and $TL[1] = \frac{\sqrt{n}(\bar{y}_{k+1} - \mu_{k+1})/\sigma_e + \sqrt{n}z_{1-p}}{\frac{s}{\sigma_e}}$. In terms of distributions, the equation for TL[1] has the form of $\frac{N(0,1) + \delta}{\sqrt{\chi^2/df}}$. This reveals that TL[2] is a linear transformation of TL[1]. TL[2] thus has the same probabilistic behavior as TL[1] and so, perhaps surprisingly, use of the bias-corrected estimation will produce exactly the same results as the uncorrected statistic. Therefore, in the remainder only TL[1] (a model-based format of TS[1]) is discussed for the test based on the fixed effects model.

3.2.5 Estimation of the tolerance limit in fixed and random effects models

Given the preceding set-up, this section will describe two forms of test statistics and associated tolerance limits based on the fixed or random effects model. Defined is the standard tolerance limit based on the fixed effects model as well as the tolerance limit

based on the random effects model (by using the concept of the best linear unbiased predictor). In this chapter, interest is on the lower tolerance limit of the test site mean which is given by the upper limit for the lower percentile. When the estimated tolerance limit is less than the lower standard *L*, impairment is declared for the test site.

In the random effects model, the test site mean is a random quantity and is treated as a linear combination of the overall mean and realized value of random effect (i.e. a conditional mean given the random effect). Though the site mean is conditioned on the value of random effect, the variability from the random effect still plays an important role in obtaining mean squared error of the estimated site mean (Fay and Herriot, 1979). This technique is widely used in small area estimation. In the field of small area estimation, sample data from a population (scattered over a large domain) are usually used to make inference about some quantity in sub-domains of that population. In my study, the sub-domain is a specific site of interest and the quantity of interest is the site mean. The data at a specific site come from a small area and strength is borrowed from other 'related' small areas to improve efficiency (Rao, 2003). The primary objective of this technique is to "provide the best possible estimates for areas that contain few sampling units" (Hunting and Harville, 1991).

Below, two different possible tolerance limits are described for evaluation of impairment. First, the basic limit from the previous chapter is extended to the limit based on a fixed effects model. Then the limit for the random effects model is given. This limit is based on an approximate variance estimate and used for site-oriented tests.

Estimated limit in the fixed effects model

Under the basic fixed effects model, the mean of the test site is estimated by the ordinary estimator, i.e. $\bar{y}_{k+1} = \sum_{j=1}^{n} y_{(k+1)j} / n$. When the sample size of each site in the region is equal to *n*, the standard error of this estimator is obtained by $s/\sqrt{n} = \hat{\sigma}_e / \sqrt{n} = \sqrt{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 / [(k+1)(n-1)n]}$, where $\bar{y}_i = \sum_{j=1}^{n} y_{ij} / n$. The estimator of the variance is unbiased since the population distributions of all sites are assumed to have common variance.

As is known, in the fixed effects model, $\frac{(k+1)(n-1)s^2}{\sigma_e^2} \sim \chi^2_{(k+1)(n-1)}$ and $Var(\bar{y}_{k+1}) = \sigma_e^2 / n$. Under the null hypothesis, the test statistic, $\frac{\bar{y}_{k+1} - L}{s/\sqrt{n}}$ (of the form TS[1]) has a noncentral t distribution with $(k+1)(n-1)$ degrees of freedom (i.e. $df = (k+1)(n-1)$) and noncentrality $\lambda = z_{1-p_0}\sqrt{n}$. The lower tolerance limit for the test site mean is $\bar{y}_{k+1} - t((k+1)(n-1), \alpha, z_{1-p_0}\sqrt{n})s/\sqrt{n}$, which will be referred to as TL[11] (i.e. TL[1] in Section 3.2.4). The single site test is a special case of TL[11] with *k*=0 (only one site in the region).

Estimated tolerance limit in the random effects model

Under the random effects model, the site mean is a random quantity. This section will address two approaches to estimate this random quantity based on the balanced random effects model (i.e., all n_i equal to *n*) and focus on the performance of the second approach.

One estimation approach follows the approach in the fixed effects model i.e., estimates the conditional site mean by the ordinary estimator but takes into account the between-site variability in variance estimation. That is, the conditional site mean is estimated by the sample site mean while the variance of the ordinary site sample mean is directly derived based on the random effects model. In this approach, the quantity of interest is the site mean given the realized value of the random effect, $\mu_{k+1} \mid A_{k+1}$. This conditional site mean is estimated by $\bar{y}_i = \sum_j y_{ij} / n$. σ_e^2 / n is the variance of $\bar{y}_i$ conditionally on the realized value of the random effect. Under the small area estimation framework, the random effect variability still contributes to the mean squared error of the conditional site mean. The unconditional variance of $\bar{y}_i$ $(Var(\bar{y}_i) = \sigma_A^2 + \sigma_e^2 / n)$ is proposed to be used in the test statistic so that the random variance has relevance to site-oriented impairment assessment. The boundary of the null hypothesis is expressed as $\mu + a_{k+1} - z_{1-p_0}\sigma_e = L$, where a_{k+1} is the realized value of A_{k+1}. This approach is a pseudo-fixed effects approach and abbreviated as the PFE approach.

The estimated variance of the site mean, $\hat{\sigma}_A^2 + \hat{\sigma}_e^2 / n$, is denoted as s^2. $\hat{\sigma}_A^2$ and $\hat{\sigma}_e^2$ are estimated by $\hat{\sigma}_e^2 = \frac{\sum_{i,j}(y_{ij} - \bar{y}_i)^2}{(k+1)(n-1)}$ and $\hat{\sigma}_A^2 = (\frac{\sum_i (\bar{y}_i - \hat{\mu})^2}{k} - \frac{\hat{\sigma}_e^2}{n})$, respectively. The between-site variability, $SS_A = n\sum_{i=1}^{k+1}(\bar{y}_i - \bar{y}_\cdot)^2 = k(n\hat{\sigma}_A^2 + \hat{\sigma}_e^2) = nks^2$, follows a Chi-square distribution with k *df*, i.e. $k(n\hat{\sigma}_A^2 + \hat{\sigma}_e^2)/(n\sigma_A^2 + \sigma_e^2) \sim \chi_k^2$. Thus, $\sqrt{\frac{\chi^2}{df}} = \sqrt{\frac{SS_A}{k(n\sigma_A^2 + \sigma_e^2)}} = \frac{s}{\sqrt{\sigma_A^2 + \sigma_e^2 / n}}$. The test statistic for this approach (PFE approach) is derived as follows.

$$\frac{\bar{y}_{k+1} - L}{s} = \frac{\frac{(\bar{y}_{k+1} - \mu_{k+1})}{\sqrt{\sigma_A^2 + \sigma_e^2 / n}} + \frac{(\mu_{k+1} - L)}{\sqrt{\sigma_A^2 + \sigma_e^2 / n}}}{s / \sqrt{\sigma_A^2 + \sigma_e^2 / n}}$$

Obviously, this test statistic has two disadvantages compared with TL[11]: 1) it has fewer degrees of freedom (k versus (k+1)(n-1)) and 2) it is divided by a larger standard deviation. The performance of this test statistic is thus expected to be worse than TL[11]. It will not be further discussed in this chapter.

The other estimation method is based on the best linear unbiased predictor (BLUP) method and the restricted maximum likelihood approach (REML). REML is used here for variance estimation so that the possibility of negative estimates for variance of the random effect will be excluded. The potential effect of the ratio of variance components will be studied here. Given data, the estimated overall mean and variance components are used to estimate and adjust the site mean. This type of estimator is an empirical estimator (EBLUP). For the balanced case, the EBLUP of a site mean is $\bar{y}_{i,EBLUP} = \hat{\mu} + \frac{n\hat{\rho}}{(n-1)\hat{\rho}+1}(\bar{y}_i - \hat{\mu}))$, using previously defined notation. This EBLUP can also be written as $\bar{y}_{i,EBLUP} = \frac{(1-\hat{\rho})\hat{\mu} + n\hat{\rho}\,\bar{y}_i}{(n-1)\hat{\rho}+1}$, where $\hat{\rho} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2}$. $\hat{\sigma}_A^2$ and $\hat{\sigma}_e^2$ are REML estimators. In the REML estimation procedure, when the estimated σ_A^2 (i.e., $\hat{\sigma}_A^2$) is negative, $\hat{\sigma}_A^2$ is set as 0 while the estimator, $\hat{\sigma}_e^2$ doesn't change. Denote the variance

ratio σ_A^2/σ_e^2 by θ so that the EBLUP can be written as $\bar{y}_{i,EBLUP} = \frac{\hat{\mu} + n\hat{\theta}\,\bar{y}_i}{n\hat{\theta}+1}$. If the intraclass correlation (defined as $\rho = \frac{\sigma_A^2}{\sigma_A^2+\sigma_e^2}$) is assumed known (i.e., the variance ratio is known), the variance of $\bar{y}_{i,EBLUP}$ is derived as follows. First, note that

$Var(\hat{\mu}) = \frac{n\sigma_A^2+\sigma_e^2}{n(k+1)} = \frac{n\theta+1}{n(k+1)}\sigma_e^2$, $\mathrm{cov}(\hat{\mu},\bar{y}_i) = \frac{n\sigma_A^2+\sigma_e^2}{n(k+1)} = \frac{n\theta+1}{n(k+1)}\sigma_e^2$, and

$Var(\bar{y}_i) = \frac{n\theta+1}{n}\sigma_e^2$. Then,

$$Var(\bar{y}_{i,EBLUP}) = \frac{Var(\hat{\mu}) + (n\theta)^2 Var(\bar{y}_i) + 2n\theta\,\mathrm{cov}(\hat{\mu},\bar{y}_i)}{(n\theta+1)^2}$$

$$= \frac{\frac{n\theta+1}{n(k+1)} + (n\theta)^2\frac{n\theta+1}{n} + 2n\theta\frac{n\theta+1}{n(k+1)}}{(n\theta+1)^2}\sigma_e^2$$

$$= \frac{n^2(k+1)\theta^2 + 2n\theta + 1}{n(k+1)(n\theta+1)}\sigma_e^2$$

$$= [(\theta+\frac{1}{n}) - \frac{k}{n(k+1)} - \frac{k\theta}{(k+1)(n\theta+1)}]\sigma_e^2$$

As the sample size and the number of sites increase, $Var(\bar{y}_{i,EBLUP})$ approaches $(\theta+\frac{1}{n})\sigma_e^2$, i.e. $\sigma_A^2+\sigma_e^2/n$, which also holds when conditioning on the test site (i.e., site *k*+1). Many papers discuss estimation of the mean squared error of the EBLUP (for example, Prasad and Rao, 1990; Das et al., 2004). To simplify the application procedure, I use $\sigma_A^2+\sigma_e^2/n$ for the approximate variance of the EBLUP instead of more complex expressions. The corresponding expression for the test statistic is $\frac{\bar{y}_{k+1,EBLUP} - L}{\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_e^2/n}}$. When the variance ratio is known, the EBLUP is independent of estimated variance components and the approximate non-central *t* distribution is easily derived for this test statistic. When the variance ratio is unknown, the EBLUP and its mean squares error are approximately independent of each other (Bhaumik and Kulkarmi, 1996). The noncentral *t* distribution can still be used for further inference. Therefore, on the boundary of the null

hypothesis that the test site is unimpaired ($H_0 : \mu + a_{k+1} - z_{1-p_0}\sqrt{\sigma_A^2 + \sigma_e^2} = L$), $\frac{\bar{y}_{k+1,EBLUP} - L}{\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_e^2 / n}}$ has an approximate noncentral *t* distribution with noncentrality parameter $\delta_0 = \frac{z_{1-p_0}\sqrt{\sigma_A^2 + \sigma_e^2}}{\sqrt{\sigma_A^2 + \sigma_e^2 / n}}$ and degrees of freedom equal to *k*. The variance components are estimated by REML, and *L* is the lower criterion. The tolerance limit is thus calculated as $\bar{y}_{k+1,EBLUP} - t(k, \alpha, \hat{\delta}_0)\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_e^2 / n}$ with $\hat{\delta}_0 = z_{1-p_0}\sqrt{\frac{n}{(n-1)\hat{\rho} + 1}}$ and $\hat{\rho} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2}$, which will be referred to as TL[EBLUP]. The formula for the interval in terms of $\hat{\rho}$ is given in Table 3.2.

Table 3.2 summarizes two estimated tolerance limits for comparison with the single site test in my study. These data-based limits are compared to the lower standard, *L*. If the limit is less than *L*, impairment is declared for the test site. TL[11] is a widely accepted statistical approach in environmental studies (VDEQ 2006; Yoder and Rankin, 1995; Barbour et al., 1996).

Table 3.2 Estimation of tolerance limit for the test site (site k+1) based on the fixed or random effects models

Model	Approach notation	Tolerance limit*
Fixed effects model	TL[11]	$\bar{y}_{k+1} - t((k+1)(n-1), \alpha, z_{1-p_0}\sqrt{n})s/\sqrt{n}$
	with $\bar{y}_{k+1} = \sum_{j=1}^{n} y_{(k+1)j}/n$, $s = \sqrt{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2/((k+1)(n-1))}$	
Random effects model	TL[EBLUP]	$\bar{y}_{k+1,EBLUP} - t(k, \alpha, z_{1-p_0}\sqrt{\frac{n}{(n-1)\hat{\rho}+1}})\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_e^2/n}$
	with $\hat{\mu} = \sum_{i,j} y_{ij}/[n(k+1)]$, $\bar{y}_{k+1} = \sum_{j=1}^{n} y_{(k+1)j}/n$, $\bar{y}_{i,EBLUP} = \frac{(1-\hat{\rho})\hat{\mu} + n\hat{\rho}\,\bar{y}_i}{(n-1)\hat{\rho}+1}$, $\hat{\rho} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2}$	

* All other notation is defined in the preceding paragraphs.

3.2.6 Evaluation of test performance

Given a specified significance level for the tests (0.05 in this Chapter), the test performance of the limits in Table 3.2 will be evaluated by comparing their corresponding power. For the fixed effects model, analytical expressions are given, while for the random effects model simulations are used for evaluation.

Analytical form of power for test based on the fixed effects model

The decision rule for impairment assessment is based on the comparison between the estimated tolerance limit and the lower standard. It can be written in a general form, $\hat{\mu}_0 - ms < L$, where m is called the critical coefficient, $\hat{\mu}_0$ and s are the estimated site mean and standard deviation, respectively. When this inequality is satisfied, a decision of rejection is made, i.e., the test site is declared as impaired.

Suppose the true exceedance proportion is p_1 and the distribution of the measurements is $N(\mu_1, \sigma^2)$ with $\mu_1 = L + z_{1-p_1}\sigma$. The true exceedance proportion is obtained by $p_1 = \Pr(Y < L)$. The following provides the formula for calculating the power for TL[11]:

$$
\begin{aligned}
power &= \Pr(\bar{y}_{k+1} - ms < L) \\
&= \Pr\left(\frac{\bar{y}_{k+1} - L}{s/\sqrt{n}} < m\sqrt{n}\right) \\
&= \Pr\left(\frac{\bar{y}_{k+1} - L}{s/\sqrt{n}} < \frac{t((k+1)(n-1), \alpha, z_{1-p_0}\sqrt{n})\sqrt{n}}{\sqrt{n}}\right) \\
&= \Pr\left(\frac{\frac{\bar{y}_{k+1} - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - L}{\sigma/\sqrt{n}}}{s/\sigma} < t((k+1)(n-1), \alpha, z_{1-p_0}\sqrt{n})\right) \\
&= \Pr(T((k+1)(n-1), \lambda) < t((k+1)(n-1), \alpha, z_{1-p_0}\sqrt{n}))
\end{aligned}
$$

where $\lambda = \dfrac{\mu_1 - L}{\sigma/\sqrt{n}} = z_{1-p_1}\sqrt{n}$ and μ_1 is the true site mean.

Simulation framework for test performance based on the random effects model

To study the performance of the EBLUP test, a simulation approach is used. SAS (1999) is used for data simulation. In the general simulation approach, I consider a region with one test site and three or nine non-test sites. The sample size is equal for each site. The

significance level is set to 0.05. Observations at the test site are generated using $y_{(k+1)j} = \mu + \varepsilon_j$ with $\varepsilon_j \sim N(0, \sigma_e^2)$. Observations at the non-test sites are generated using $y_{ij} = \mu + A_i + \varepsilon_{ij}$, where i=1, 2, …, k, $A_i \sim N(0, \sigma_A^2)$, $\varepsilon_{ij} \sim N(0, \sigma_e^2)$, and $\text{cov}(A_i, \varepsilon_{ij}) = 0$.

The basis for the parameters of the distribution comes from the reference distribution (the distribution from which reference sites come) of the Index of Biotic Integrity (IBI) with mean 90 and standard deviation 4 (Hughes et al., 2004), i.e., $y_{ij} \sim N(90, 4^2)$. Thus, the population mean (μ) in the simulation is set as $\mu = 90$, the within-site variance σ_e^2 is set as $\sigma_e^2 = 4^2$, and consequently the lower standard (L) is set as $L = 90 - 4z_{0.9} = 84.88$. When the test site is assumed to have an exceedance proportion of 10%, the test site data follow the distribution $y_{(k+1)j} \sim N(90, 4^2)$. When the test site is assumed to have an exceedance proportion of 30%, the data follow the distribution $y_{(k+1)j} \sim N(86.97, 4^2)$. Here, the test site mean μ_{k+1} is obtained by $\mu_{k+1} = L + 4z_{0.7} = 90 - 4z_{0.9} + 4z_{0.7}$. Table 3.3 lists the candidate values implemented in the simulation. The scenario of variability considered at the non-test sites is labeled as I, II, and III in Table 3.3. The number of non-test sites is 3 or 9. The sample size at each site is 4, 8, or 12. Different combinations of these factors reflect scenarios of specific interest, which will be addressed in Section 3.3.2 and 3.3.3.

Table 3.3 Simulation parameters for evaluating the power for the test based on the random effects model

Mean	Variance	
	Test site	Non-test sites
$\mu_{k+1} = 90$ or $\mu_{k+1} = 86.97$	$\sigma_e^2 = 4^2$	$\sigma_A^2 = 0$, $\sigma_e^2 = 4^2$ (I)
	$\sigma_e^2 = 4^2$	$\sigma_A^2 = 4^2$, $\sigma_e^2 = 4^2$ (II)
	$\sigma_e^2 = 4^2$	$\sigma_A^2 = 32$, $\sigma_e^2 = 4^2$ (III)

A standard random effects model in Section 3.2.2 is fit to the generated dataset. In each scenario, 1,000 simulations are run. When the calculated tolerance limit is less than the lower standard (here, the lower standard is $L = 90 - 4z_{0.9} = 84.88$), a rejection is recorded. When the true exceedance proportion is 10% at the test site, the rejection ratio indicates the Type I error rate at the boundary of the null hypothesis. When the true exceedance proportion is greater than 10% at the test site, the rejection ratio indicates the power for the test and is used to evaluate the test performance.

3.3 Results

3.3.1 Test performance for the tolerance limit based on the fixed effects model

To describe the general behavior of TL[11], the power formula developed in Section 3.2.6 is applied to the scenario where the region consists of one test site and three non-test sites. Figure 3.2 displays the test performance for the tolerance limit based on a fixed effects model under this scenario. The sample size for each site is the same and the variance across the region is the same. The value on the rejection ratio curve indicates the sample size for the test site.

When the true exceedance proportion is less than 0.1, an increase in the sample size leads to a smaller rejection ratio, i.e., a lower Type I error rate. When the true exceedance proportion is 0.1, TL[11] achieves the pre-specified significance level (having the rejection ratio equal to 0.05). When the true exceedance proportion is greater than 0.1 (i.e., the impairment tends to be severe), the obvious result is that large sample sizes and greater exceedance proportions produce more rejections, i.e., more powerful tests.

Compared to the single site approach, the model-based test (TL[11]) improves the test performance (Figure 3.3). Figure 3.3 plots the difference in the rejection ratio between the model-based test and the single site test. The single site test corresponding to TL[11] uses TS[1] for evaluation (Smith et al., 2003). As the true exceedance proportion increases, the model-based test's performance increases then decreases. When the true exceedance proportion is relatively small, say less than 0.3, larger sample size results in

greater improvement for the model-based test. After a certain exceedance proportion (around 0.5), larger sample size corresponds to smaller improvement.

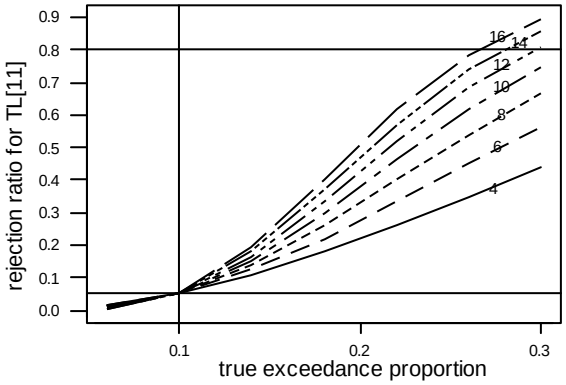

Figure 3.2 Test performance of the tolerance limit based on TL[11] under a fixed effects model. The number of sites in the region is four. Numbers on the curves indicate sample size at the test site.

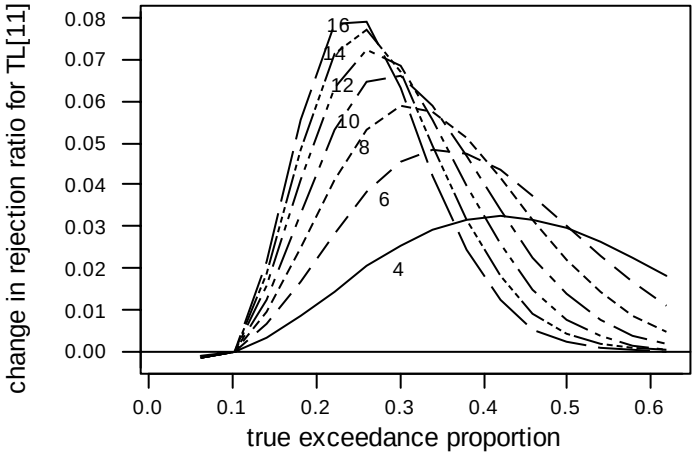

Figure 3.3 Improvement from use of a test based on a fixed effects model relative to a single site test. The number of sites in the region is four. Numbers on the curves indicate sample size at the test site.

3.3.2 Test performance for the tolerance limit based on the random effects model

In this section, the test site is evaluated by using TL[EBLUP]. When there is no between-site variability, all sites in the region are generated from a fixed effects model. When the between-site variability exists, the non-test sites are generated based on a random effects model while the test site is generated using a fixed site mean across all the simulation runs.

When the region consists of 4 sites and all the sites have a baseline mean (90), TL[EBLUP] has a rejection ratio less than 0.05 for all scenarios I, II, and III (defined in Table 3.3). The rejection ratio is around 0.001 for scenario I and II and around 0.003 for scenario III.

When the test site is impaired with the true exceedance proportion of 30% and the three non-test sites have unconditional site means equal to the test site mean, the single site test is expected to have better performance than TL[EBLUP]. This expected behavior comes from two facts in the simulation set-up: 1) TL[EBLUP] has degrees of freedom equal to 3 while the single site test has degrees of freedom equal to 3, 7, or 11, more that of TL[EBLUP] and 2) the TL[EBLUP] has relative to extra variability. Scenario I of TL[EBLUP] is expected to have similar test performance to the single site since there is no between-site variability in the generated dataset and the estimation of this term is expected to be close to zero. Results in Table 3.4 verify this as expected especially for large sample sizes. When the region is generated based on a fixed effects model, "borrowing information" by means of TL[EBLUP] doesn't improve the test power. TL[EBLUP] and the single site test perform similarly. The difference between these two tests (scenario I of TL[EBLUP] and the singe site test) becomes smaller when the sample size increases. When the between-site variability exists (scenario II and III), TL[EBLUP] has lower power than the single site test. When the between-site variability is relatively large (i.e. $\sigma_A^2/\sigma_e^2>1$, in scenario II $\sigma_A^2/\sigma_e^2=2$), the TL[EBLUP] has very low power. The power of scenario II is around 1.5 times of the power of scenario III.

The single site test has better performance for impairment assessment than the test based on the random effects model. Therefore, the test based on the random effects model is generally not recommended.

Table 3.4 Performance of the test based on the random effects model and the EBLUP estimator when the test site is impaired* (p_1 =0.3).

sample size	TL[EBLUP]			SS**
	I	II	III	
4	0.321	0.215	0.160	0.413
8	0.597	0.262	0.154	0.607
12	0.741	0.287	0.176	0.740

*There are four sites in the region.

**SS indicates the single site test (a special case of TL[11], refer to Section 3.2.5).

3.3.3 Factors influencing results for model-based tests

Test properties in terms of error rates vary with factors such as sample size and the number of sites other than the test site. This section evaluates the effects of these factors on power as well as some other factors that are specific to the models.

Sample size and number of non-test sites

For the fixed effects model, increasing the sample size reduces the Type I and Type II error rates. As Figure 3.2 shows, when the test site is truly unimpaired, the rejection ratio decreases with increased sample size, i.e., the Type I error rate becomes lower. When the test site is truly impaired, the rejection ratio increases with the increased sample size, i.e., the Type II error rate becomes lower. For the random effects model, large sample size improves test performance in terms of power as shown in Table 3.4b.

To examine the effect of the number of non-test sites on the test based on the fixed effects model, Table 3.5a and 3.5b describe results using two impairment severities, one below and one above the maximum acceptable exceedance proportion. When the true exceedance proportion is less than 10%, the site is assumed to be unimpaired originally; while the proportion is greater than 10%, the site is treated as impaired. The case where k+1 is equal to 1 represents the single site test (i.e., acceptance sample approach). This case is the baseline for the comparison.

When the sites are originally impaired, the power increases when more sites are pooled together in the fixed effects model. When the sites have low impairment, the false listing rate (i.e., the Type I error rate) decreases with more pooled sites for TL[11]. Tables 3.5a and 3.5b reveal a consistent test behavior pattern for a region with a small number of sites or large number of sites when the sample size is small per site. The effect of the number of non-test sites is consistent with the effect of sample size on the test based on the fixed effects model. Although total sample size may be the same for these two situations, the case with more degrees of freedom will have greater power. For example, the last entry in Table 3.5a and 3.5b has a total of 80 observations but the degrees of freedom are 72 when the number of sites is small and 60 when the number of sites is large.

Table 3.5a Effect of the number of sites on test rejection rate based on the fixed effects model for small sample size and small number of sites*

		TL[11]
True exceedance proportion	k+1	
4%	1	0.0016
	2	0.0013
	4	0.0011
	6	0.0010
	8	0.0009
10%	1	0.0500
	2	0.0500
	4	0.0500
	6	0.0500
	8	0.0500
16%	1	0.1981
	2	0.2071
	4	0.2154
	6	0.2185
	8	0.2201

*The sample size per site is fixed at 10.

Table 3.5b Effect of the number of sites on test rejection rate based on the fixed effects model for small sample size and large number of sites*

True exceedance proportion	k+1	TL[11]
4%	1	0.0058
	10	0.0050
	20	0.0049
10%	1	0.0500
	10	0.0500
	20	0.0500
16%	1	0.1343
	10	0.1413
	20	0.1417

*The sample size per site is fixed at 4.

The effect of the non-test sites on tests based on the random effects model is examined under the simulation framework described in Section 3.2.6 (refer to Table 3.3). The region has four or ten sites. Each site has the same sample size. The true exceedance proportion at the test site is 30%. Simulation results (Table 3.6) indicate that when the test site is impaired and the sample size is not very small (say, at least 8 observations per site), pooling information from more homogeneous sites increases power for the scenario I using the EBLUP test but decreases power for scenario II and III. Further, scenario I using TL[EBLUP] is more powerful than the single site test when more sites with moderate or large sample sizes (say, more than 8 observations per site) are pooled in the test. The results are similar to the test based on the fixed effects model (refer to Figure 3.2). Scenario II is more powerful than scenario III since sites in scenario III have relatively larger between-site variability and thus more uncertainty of site means at the non-test sites.

Table 3.6 Effect of the number of non-test sites on test rejection rate based on the random effects model when the test site is impaired ($p_1 = 0.3$)

Number of non-test sites	Sample size	TL[EBLUP]			SS*
		I	II	III	
3	4	0.332	0.215	0.160	0.413
	8	0.606	0.262	0.154	0.607
	12	0.771	0.287	0.176	0.740
9	4	0.293	0.131	0.072	0.413
	8	0.788	0.148	0.047	0.607
	12	0.945	0.134	0.049	0.740

* The single site test (Refer to Section 3.2.5).

Error structure in the fixed effects model

The error term in the fixed effects model set up in Section 3.2.2 has an independence structure. In practice, it is common to have the correlated measurements within each site. Such correlations affect variance estimation.

Without any prior information, an unstructured variance covariance matrix is suitable for the error term. The variance structure of the response is thus expressed as follows for the case when there are three measurements for each site.

$$Var(\underline{y}_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$$

where $\sigma_{jk} = \sigma_{kj}$ is the covariance between observation *j* and *k* within a site. This type of error structure is not parsimonious. Practical implementation usually requires direct and parsimonious modeling for the within-site correlations (Schabenberger and Pierce, 2002). To illustrate the effect of error structure on model-based tests, my study uses the compound symmetric (CS) and first-order autoregressive (AR(1)) correlation structures to summarize the behavior of within-site disturbances with a small number of parameters. The measurements at one site are thus assumed to be equicorrelated (in the compound symmetry structure) or related immediately to the preceding observation (in the AR(1) structure).

When the within-site correlation is represented by the compound symmetric structure, the variance structure of the response is expressed as follows when there are three measurements for each site.

$$Var(\underline{y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

where ρ is the covariance coefficient and $-1 \le \rho \le 1$. The expectation for the estimated variance in a general case is derived as below, assuming *n* measurements per site. First, note that.

$$E(y_{ij}^2) = [E(y_{ij})]^2 + Var(y_{ij}) = \mu_i^2 + \sigma_i^2$$

$$E(\bar{y}_{i.}^2) = E\left[\left(\frac{\sum_{j=1}^{n} y_{ij}}{n}\right)^2\right] = E\left(\frac{\sum_{j=1}^{n} y_{ij}^2 + 2\sum_{j`>j} y_{ij} y_{ij`}}{n^2}\right)$$

$$= [E(\sum_{j=1}^{n} y_{ij}^2) + n(n-1)E(y_{ij} y_{ij`})]/n^2$$

$$= [n(\mu_i^2 + \sigma_i^2) + n(n-1)(\mu_i^2 + \rho\sigma_i^2)]/n^2$$

$$= [n\mu_i^2 + (n-1)\rho\sigma_i^2 + \sigma_i^2]/n$$

$$E(y_{ij}\bar{y}_{i.}) = E(y_{ij}\sum_{j`=1}^{n} y_{ij`})/n = [E(y_{ij}^2) + (n-1)E(y_{ij} y_{ij`})]/n$$

$$= [\mu_i^2 + \sigma_i^2 + (n-1)(\mu_i^2 + \rho\sigma_i^2)]/n$$

$$= [n\mu_i^2 + \sigma_i^2 + (n-1)\rho\sigma_i^2]/n$$

$$E(s^2) = E\left(\frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2}{(k+1)(n-1)}\right) = \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n} E(y_{ij}^2 + \bar{y}_{i.}^2 - 2y_{ij}\bar{y}_{i.})}{(k+1)(n-1)}$$

$$= \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}[\mu_i^2 + \sigma_i^2 + \frac{n\mu_i^2 + (n-1)\rho\sigma_i^2 + \sigma_i^2}{n} - 2\frac{n\mu_i^2 + \sigma_i^2 + (n-1)\rho\sigma_i^2}{n}]}{(k+1)(n-1)}$$

$$= \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}\frac{(n-1)(1-\rho)}{n}\sigma_i^2}{(k+1)(n-1)}$$

$$= (1-\rho)\sigma^2$$

Thus, the estimated variance is biased. When the measurements are positively correlated ($0 < \rho < 1$), the variance is underestimated and the limits based on the fixed effects model are inflated, which leads to more powerful tests. When the measurements are negatively correlated ($-1 < \rho < 0$), the variance is overestimated and the limits are shrunken, which results in less powerful tests. Water quality measurements are expected to have positive correlations.

When the correlations within a site decrease with temporal separation of measurements, the AR(1) structure is commonly used to model the within-site correlation. The fixed effects model with time series is thus written as

$$y_{it} = \mu_i + s_{it} \text{ and } s_{it} = \phi\, s_{i(t-1)} + e_{it}, |\rho| < 1$$

where $e_{it} \overset{iid}{\sim} N(0, \sigma_e)$. The s_{it}'s follow an AR(1) process, which has two basic properties: $\text{cov}(s_{i(j-w)}, e_{it}) = 0$ and $\text{cov}(e_{i(j-w)}, e_{it}) = 0, \forall w > 0$. At the initial time point, $s_{i1} = \frac{e_{i1}}{\sqrt{1-\phi^2}}$, where ϕ is the autoregressive coefficient (i.e., the lag) and $|\phi| < 1$.

The fixed effects model can also be rewritten as a distributed lag model,

$$y_{it} = \phi\, y_{i(t-1)} + (1-\phi)\mu_i + e_{it}.$$

The variance is estimated as follows.

Based on $s_{it} = \phi\, s_{i(t-1)} + e_{it}$, $Var(s_{it}) = \frac{Var(e_{it})}{1-\phi^2} = \frac{\sigma_e^2}{1-\phi^2}$.

Based on $y_{it} = \mu_i + s_{it}$, $Var(s_{it}) = Var(y_{it} - \mu_i) = Var(y_{it})$.

Thus, $Var(y_{it}) = \frac{\sigma_e^2}{1-\phi^2}$,

$$Cov(y_{it}, y_{i(t-1)}) = Cov(\phi\, y_{i(t-1)} + (1-\phi)\mu_i + e_{it}, y_{i(t-1)}) = \phi Var(y_{it}) = \frac{\phi}{1-\phi^2}\sigma_e^2,$$

and $Var(\underline{y_i}) = \frac{\sigma_e^2}{1-\phi^2}\begin{bmatrix} 1 & \phi & \phi^2 & & \\ \phi & 1 & \phi & \phi^2 & \\ \phi & \phi^2 & 1 & \phi & \phi^2 .... \\ ..... & & & & \end{bmatrix}$.

When there are three measurements for each site, $Var(\underline{y_i}) = \frac{\sigma_e^2}{1-\phi^2}\begin{bmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi & \phi^2 & 1 \end{bmatrix}$. The expectation of the estimated variance is derived as follows.

$$E(y_{ij}^2) = [E(y_{ij})]^2 + Var(y_{ij}) = \mu_i^2 + \frac{\sigma_e^2}{1-\phi^2}$$

$$E(\bar{y}_{i.}^2) = E\left[\left(\frac{\sum_{j=1}^{3} y_{ij}}{3}\right)^2\right] = E\left(\frac{\sum_{j=1}^{3} y_{ij}^2 + 2\sum_{j`>j} y_{ij}y_{ij`}}{3^2}\right)$$

$$= \{3(\mu_i^2 + \frac{\sigma_e^2}{1-\phi^2}) + 2[(\frac{2\phi}{1-\phi^2} + \frac{\phi^2}{1-\phi^2})\sigma_e^2 + 3\mu_i^2]\}/3^2$$

$$= \mu_i^2 + \frac{3+4\phi+2\phi^2}{9(1-\phi^2)}\sigma_e^2$$

For j=1 or 3, $E(y_{ij}\bar{y}_{i.}) = E(y_{ij}\sum_{j`=1}^{3} y_{ij`})/n = [E(y_{ij}^2) + \frac{\phi+\phi^2}{1-\phi^2}\sigma_e^2 + 2\mu_i^2]/3$

$$= \mu_i^2 + \frac{1+\phi+\phi^2}{3(1-\phi^2)}\sigma_e^2$$

For j=2, $E(y_{ij}\bar{y}_{i.}) = E(y_{ij}\sum_{j`=1}^{3} y_{ij`})/n = [E(y_{ij}^2) + \frac{2\phi}{1-\phi^2}\sigma_e^2 + 2\mu_i^2]/3$

$$= \mu_i^2 + \frac{1+2\phi}{3(1-\phi^2)}\sigma_e^2$$

$$E(s^2) = E(\frac{\sum_{i=1}^{k+1}\sum_{j=1}^{3}(y_{ij}-\bar{y}_{i.})^2}{(k+1)(n-1)}) = \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{3}E(y_{ij}^2 + \bar{y}_{i.}^2 - 2y_{ij}\bar{y}_{i.})}{2(k+1)}$$

$$= \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{3}(\mu_i^2 + \frac{\sigma_e^2}{1-\phi^2} + \mu_i^2 + \frac{3+4\phi+2\phi^2}{9(1-\phi^2)}\sigma_e^2) - 2\sum_{i=1}^{k+1}[\mu_i^2 + \frac{1+2\phi}{3(1-\phi^2)}\sigma_e^2 + 2(\mu_i^2 + \frac{1+\phi+\phi^2}{3(1-\phi^2)}\sigma_e^2)]}{2(k+1)}$$

$$= \frac{\phi+3}{3(\phi+1)}\sigma_e^2$$

A general derivation is given in Appendix 1. The preceding derivation reveals that the estimated variance is biased under the AR(1) model. When the lag satisfies $0 < \phi < 1$, the variance is underestimated and the rejection ratio for tests based on the fixed effects model tends to decrease. When the lag satisfies $-1 < \phi < 0$, the variance is overestimated and the limits decrease, which leads to more rejections.

3.3.4 Case study: application to data from Philpott Reservoir

To illustrate the application of the model-based tests in environmental studies, a small dataset is obtained from Philpott Reservoir. This reservoir is nestled in the foothills of Virginia close to the Blue Ridge Mountains. It is managed primarily for flood control and hydroelectric power generation. There is no residential development along its shoreline but there are numerous tourism facilities scattered throughout this area, such as campgrounds, picnic areas, and boat landings (Virginia Department of Games and Inland Fishery). From April 2001 to October 2001, the dissolved oxygen (DO) was measured at four sites monthly (Virginia DEQ). Dissolved oxygen measures the amount of gaseous oxygen dissolved in an aqueous solution. Adequate dissolved oxygen is necessary for good water quality. Stress on biological organisms increases with decreasing dissolved oxygen. The numerical criterion for DO is set at 5.0 (mg/l).

Table 3.7 lists the dataset. The observations in bold and italics are less than the numerical criterion. Without more background information, the rough summary of the data suggests that site3 is impaired and the other three sites are unimpaired. Fixed and random effects models are fit to this dataset. The tests in Table 3.2 as well as the single site approach are carried out. The corresponding results are displayed in Table 3.8. When the site is declared as impaired, the test result is recorded as 1. When the site is declared as unimpaired, the test result is recorded as 0. The test for each site is assumed to be independent. The value in the parentheses indicates the difference between the calculated limit and the lower standard. When the site is truly unimpaired, a large positive value is expected; when the site is truly impaired, a large negative number is expected.

Table 3.7 Dissolved oxygen data collected at Philpott Reservoir (year 2001)

site1	site2	site3	site4
8.97	6.68	6.68	7.21
8.80	6.80	***4.42***	7.90
7.80	7.80	***2.30***	8.50
8.30	8.00	***1.80***	8.40
8.00	8.20	5.00	8.40
8.20	8.00	***0.60***	8.20
10.10	8.70	***0.10***	9.60

Table 3.8 Results of model-based tests for data from Philpott Reservoir*

	SS**	TL[11]	TL[EBLUP]
site1	0(1.94)	0(1.88)	0(4.18)
site2	0(1.08)	0(1.01)	0(3.36)
stie3	1(-3.66)	1(-3.74)	1(-1.21)
site4	0(1.66)	0(1.59)	0(3.91)

*1=reject and declare impairment, 0=do not reject
**SS-the singe site test.

Based on all four tests, site3 may be declared as impaired and the other sites as unimpaired. The difference between limits and the standard reveals the general properties of each approach discussed in preceding sections. For the impaired site, the absolute value of the difference in tests has the ranking as TL[11]>SS>TL[EBLUP]. For the unimpaired sites, the value of the difference in tests has the ranking as TL[EBLUP]> SS>TL[11]. The comparison between the estimated limit and criterion across approaches demonstrates the effect of the quality of pooled sites (non-test sites) on tests. The test based on the fixed effects model has better performance for the impaired site (more likely to reject) but is also more likely to suggest rejection for the unimpaired site compared with the test based on the random effects model. This finding agrees with results in Section 3.3.1 and 3.3.2. Further finding is that the model-based test TL[11] has similar performance as the single site test for the unimpaired sites when comparing the difference between the estimated limit and the numeric criterion.

3.4 Miscellaneous issues

This section discusses several issues associated with the performance of model-based tests. The issues include sample size, flipping hypotheses, multiplicity adjustment, regional heterogeneity, and regional impairment.

3.4.1 Sample size

The effect of the sample size on model-based tests can be examined in terms of power gain. The power gain for model-based tests is defined by $\frac{power(\text{model}) - power(SS)}{power(SS)}$, where $power(\text{model})$ represents the power of model-based tests and $power(SS)$ represents the power of the single site test. Figure 3.4 indicates that when the test is based on a fixed effects model with equal sample sizes across the region, the power gain is present for all observed sample sizes (up to 16) and the gain is up to 16% for a truly impaired site. With increased severity level, the power gain has an arch shape (increases then decreases).

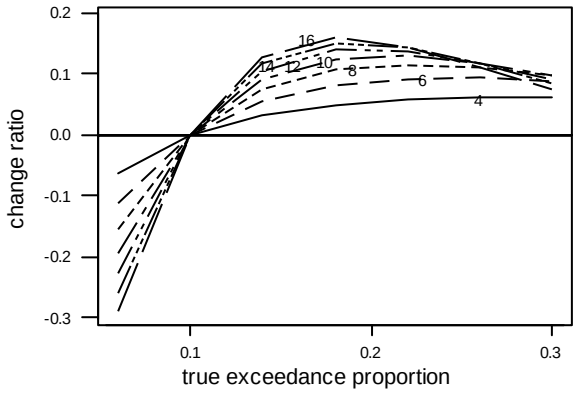

Figure 3.4 Power gain for the model-based test under the fixed effects model. The number on curves represents the sample size of the test site. The region is assumed to have 4 sites with an equal sample size for each site.

In some cases, the test site tends to have a small sample size and the non-test sites have a large sample size. Table 3.9 depicts this scenario for the test based on a fixed effects model using the rejection ratio for performance evaluation. The sample size at the test site is fixed at 4, while the sample size across the non-test sites is equal and is set to either 8 or 12. The number of non-test sites is 3 or 9. The last column in Table 3.9 lists the power gain calculated as $\frac{power(\text{TL}[11]) - power(SS)}{power(SS)}$, where $power(\text{TL}[11])$

represents the power of TL[11]. The conclusions are similar to those from Table 3.5a and Table 3.5b. Pooling homogeneous non-test sites with larger sample size slightly improves the test performance.

Table 3.9 Effect of the number of non-test sites and sample size on tests based on the fixed effects model*

True exceedance proportion	K	Rejection ratio for TL[11]	Power gain
$n_i = 8$			
4%	3	0.0050	-14.5%
	9	0.0049	-15.6%
16%	3	0.1410	5.1%
	9	0.1417	5.6%
$n_i = 12$			
4%	3	0.0050	-15.0%
	9	0.0049	-15.8%
16%	3	0.1415	5.3%
	9	0.1419	5.7%

*The sample size of the test site is fixed at 4 and the sample size of the non-test site is n_i.

Model-based tests are recommended for water quality assessment when small sample sizes are available for assessment. In theory, if a large sample size is possible, increasing sample size for a single site is recommended instead of pooling several sites. Table 3.10 compares the single site test with large sample sizes to the tests based on the fixed effects model with small sample sizes. When the test site is truly impaired and the total measurements at the region are fixed, increasing the sample size at the test site leads to more powerful tests than pooling more sites. This result supports the argument in Chapter 2 that model-based tests for water quality assessment are in the realm of small-area estimation and assessment and the results are the best when the test site has small sample size.

Table 3.10 Rejection ratio comparison between large sample sizes at a single site and small sample sizes at several sites

Region of interest	True exceedance proportion	TL[11]
1 site, n=4	4%	0.00500
	16%	0.13400
1 site, n=16	4%	0.00060
	16%	0.25290
4 sites, n=4	4%	0.00510
	16%	0.13990
1 site, n=10	4%	0.00166
	16%	0.19808
1 site, n=40	4%	0.00003
	16%	0.43580
4 sites, n=10	4%	0.00110
	16%	0.22040

3.4.2 Changing the null and alternative hypotheses

In environmental studies, it has been suggested that the more serious error be to have a claim of no impairment when the site is actually impaired since protecting public health is the top concern of water quality management (Guttorp, 2000). The null hypothesis in my study is thus defined in terms of a null hypothesis of compliance (i.e., unimpairment). Some researchers discuss tests based on switching the null and alternative hypotheses (Kilgour et al., 1998; Guttorp, 2000), i.e., setting non-compliance (impairment) as the null hypothesis. This section uses tests based on the fixed effects model to compare test performance based on flipped null hypotheses.

When the null hypothesis is of non-compliance (i.e., the site is impaired), the hypotheses are $H_0 : p \geq p_0$ (impaired) versus $H_1 : p < p_0$ (unimpaired). The model-based estimated tolerance limit is $\bar{y}_{k+1} - t((k+1)(n-1), 1-\alpha, z_{1-p_0}\sqrt{n})s/\sqrt{n}$ under the balanced fixed effects model. Terms in this limit are defined in Section 3.2.5. When the estimated tolerance limit is less than the lower standard, impairment is claimed, i.e., the null hypothesis is not rejected. Figure 3.5 compares the probability of declaring a site impaired for the flipped hypotheses under the scenario that the region of interest consists of one test site and three non-test sites and the sample size at each site is 4. This comparison can be explained by considering the proportion of samples required for rejection. In the case where the null hypothesis is that the site is not impaired, the null

hypothesis is rejected for proportions greater than the acceptable exceedance proportion, p_0. When the null is that the site is impaired, rejecting the null hypothesis requires a small value for the proportion, i.e., the proportion is somewhat less than p_0. Hence, under the not-impaired hypothesis an acceptable sample has exceedance proportion in form of $p_0 + \delta_1$ while for the other setup an acceptable sample has exceedance proportion equal to $p_0 - \delta_2$, where $0 < \delta_1 < 1$, $0 < \delta_2 < 1$, and both are determined by sample size and the level of the test.

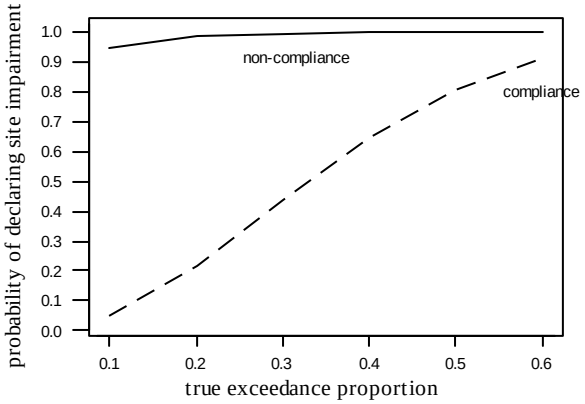

Figure 3.5 Probability of declaring a site impaired using tests based on a fixed effects model.

When the null hypothesis is of non-compliance, the probability of claiming impairment is high (essentially $1-\alpha$) for relatively small exceedance proportions. The model-based test increases this probability more for the null hypothesis of compliance than that for the null hypothesis of non-compliance (Figure 3.6). The ratio in Figure 3.6 is calculated by $\frac{prob(\text{model}) - prob(SS)}{prob(SS)}$, where $prob(\text{model})$ represents the probability of declaring a site impaired for model-based tests and $prob(SS)$ represents the probability of declaring a site impaired for the singe site test. Figure 3.5 is a supportive illustration for different test performance between flipped hypothesis settings. Due to the high probability of declaring a site impaired in the single site test, there is no potential

space for improvement from model-based tests when the null hypothesis is of non-compliance.

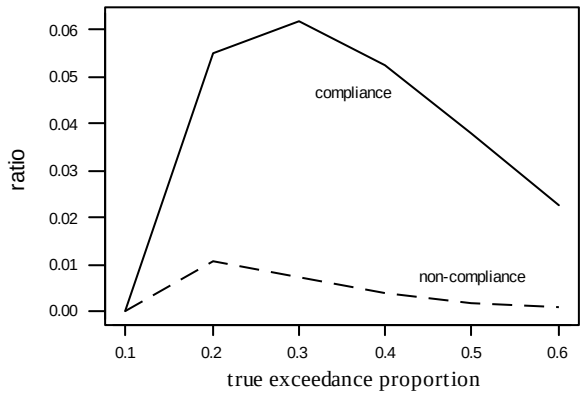

Figure 3.6 Improvement of model-based tests for flipped hypothesis settings.

3.4.3 Multiple tests

In this chapter, the hypothesis testing is conducted for a specific site. When multiple sites are evaluated in a region and the regional impairment is also of interest, the multiplicity adjustment will be suggested. This adjustment is applied in data analyses for many areas but has not been adapted in environmental studies. This section will briefly introduce one multiplicity adjustment strategy (Hochberg's procedure) and illustrate potential effects of multiplicity.

To control the family-wise error rate (FWER), the Bonferroni method is the simplest but has the least power in most cases. The Hochberg method (Hochberg, 1988) is more powerful than the Bonferroni method in terms of controlling FWER not only for independent multiple tests but also for positively correlated test statistics (Dunnett and Tamhane, 1992). Suppose all (k+1) sites in the region are tested. The Hochberg procedure is briefly described below (Hochberg, 1988).

(1) Order the p-value of each test from the smallest to the largest.

(2) Compare the largest p-value to the significance level, α. If it is significant, the all k tests are significant. Stop testing. If it is not significant, the 2nd largest p-value is compared to $\alpha / 2$. If it is significant, then the rest (k-1) tests are significant.

(3) In general, if all previous *(j*-1*)* tests are not significant, the j^{th} largest p-value is compared to α / j. If it is significant, then all the rest of the tests are significant and stop testing; otherwise, go to the next step.

In practice, rejecting all false hypotheses (i.e., the null hypotheses are false and the alternatives are true) is more common. The detailed power calculation can be found in Dunnett and Tamhane's paper (1992).

Following this procedure, a simulation of 10,000 runs is constructed to give a general picture of the effect of multiplicity. Suppose the region of interest has five sites and the marginal power for each test is 80%. The test for each site is either independent or correlated. The correlation between tests is 0 or 0.3. The significance level is 0.05. The overall power is the probability of observing a given number of positive tests (rejecting the null hypothesis). When the number of positive tests (i.e. the minimum of rejection) varies from 0 to 5, the simulation is run to record the rejection ratio based on the Hochberg procedure. For each correlation value (0 or 0.3), 10,000 simulations are run. Results in Table 3.11 demonstrate the significant decrease in the overall power when the number of positive tests increases, which suggests caution be applied in using multiplicity adjustments with model-based tests. Readers can refer to Sidak (1967), Simes (1986), Hochberg (1988), Dunnett and Tamhane (1992, 1995) for further discussion.

Table 3.11 Overall power for Hochberg's multiple testing procedure when five sites are evaluated. Minimum of rejection is the number of positive tests.

Minimum of rejection	Correlation=0	Correlation=0.3
0	0.010	0.053
1	0.990	0.947
2	0.937	0.859
3	0.803	0.749
4	0.590	0.618
5	0.337	0.449

3.4.4 Heterogeneity of mean and variance

In practice, impairment information is often unavailable before evaluation. Water quality varies over the region and the sampled sites may have different degrees of impairment. The distributions of measurements may be completely different for each impairment level. Some simple patterns of heterogeneity are given in Table 3.12 along with terminology to describe the pattern.

Table 3.12 Terminology for different heterogeneity levels

Mean of test site*	Variance of test site*	Terminology
same	same	Regional homogeneity
	different	Local homogeneity-1
different	same	Local homogeneity-2
different	Different	Regional heterogeneity

* Compared to the same parameter at the non-test sites.

When the region of interest has the property of local homogeneity-2 or regional homogeneity (i.e., all the sites come from distributions with the same variance), the heterogeneity level doesn't affect the test result since the different distributions have the same variance. The pooled estimator of variance is still unbiased as discussed in standard textbooks of applied statistics (e.g., Neter et al., 1988). Estimated limits in Table 3.2 use the ordinary estimator or the linear combination of ordinary estimates (in TL[EBLUP]) for the site mean. These estimators of site mean are unbiased. Statistical inference involving model-based tests thus remains valid.

When the sites come from distributions with different variances, i.e., the region is of local homogeneity-1 or regional heterogeneity, the pooled estimator of variance may be biased. The following demonstrates the estimation bias for a simple case under the fixed effects model. Suppose the number of sites in the region is even (i.e., k+1 is even) and half of the sites in the region come from sub-region a. Assume the measurements in the region are distributed as $N(\mu_a, \sigma_a^2)$ with $\mu_a = L + z_{p_a}\sigma_a$. The test site is assumed to be in this sub-region. The other half come from sub-region b and have the distribution

$N(\mu_b, \sigma_b^2)$ with $\mu_b = L + z_{p_b}\sigma_b$. The ratio of σ_b / σ_a affects the model-based estimator. The underlying distribution of the test statistic changes based on this ratio. If the ratio σ_b / σ_a is denoted as ω, the expectation of the estimated variance in the fixed effects model is derived as follows.

$$E(s^2) = E(\frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2}{(k+1)(n-1)}) = \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}E(y_{ij}^2 + \bar{y}_{i.}^2 - 2y_{ij}\bar{y}_{i.})}{(k+1)(n-1)}$$

$$= \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}[\mu_i^2 + \sigma_i^2 + \mu_i^2 + \frac{\sigma_i^2}{n} - 2\frac{n\mu_i^2 + \sigma_i^2}{n}]}{(k+1)(n-1)}$$

$$= \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}\frac{(n-1)\sigma_i^2}{n}}{(k+1)(n-1)} = \frac{\sum_{i=1}^{k+1}\sigma_i^2}{k+1}$$

$$= \frac{\sigma_a^2 + \sigma_b^2}{2} == \frac{(1+\omega^2)\sigma_a^2}{2}$$

In many biological systems, mild to moderate impairment leads to higher variability as well as a larger exceedance proportion. Therefore, if sites in sub-region *a* are not impaired (including the test site) while sites in sub-region *b* are impaired, it is expected that $p_a < p_b$ and $\sigma_a < \sigma_b$. Since the impaired sites have more within-site variability, i.e., $\omega > 1$, the variance estimator is inflated. This inflation leads to an increased rejection ratio and thus a higher Type I error rate. When the test site is truly impaired and some of the pooled sites are unimpaired, the ratio is less than 1, i.e., $\omega < 1$ and the variance is underestimated. This underestimation leads to less rejection and thus lower power.

In practice, the variance pattern is likely to be more complex. One might expect for biological data that unimpaired sites will have similar variance, impaired sites will have different variance, and the variability for severely impaired sites will become small as biota die. These patterns make it difficult to predict general test performance under some scenarios, such as when the test site is truly unimpaired and the pooled sites are also unimpaired but have different variance, or the test site is truly impaired and the pooled sites are more severely impaired. In my study, the fundamental assumption of non-test sites is that they have the same variance as the test site. That is all the sites in the region

are assumed to be from distributions which only differ in means. More complex variance patterns are the subject of future research.

3.4.5 Random effects and regional impairment evaluation

Recently, a concern in monitoring is the proportion of sampled sites in the region declared to be impaired, a regional view of individual sites. To obtain this proportion, people can carry out the individual test for each site and count the number of impairment declarations. This procedure produces an estimator for the impairment proportion based on individual tests, called the "test-based estimator". This estimator can be obtained based on the fixed or random effects model. An alternative approach discussed below uses model-based estimation based on the random effects model. The model set-up in Section 3.2.2 assumes that the error term is the same for all sites in the region, i.e., σ_e is a constant over the region. Define $\mu_L = L + z_{1-p_0}\sigma_e$. A particular site with the mean μ_i is impaired when $\mu_i < \mu_L$. If the random variable A_i has the value a_i for the selected site *i*, site *i* is impaired when $\mu + a_i < \mu_L$. Thus the probability of selecting an impaired site in the region of interest is defined as

$$\begin{aligned}\pi &= \Pr(\mu + A_i < \mu_L) \\ &= \Pr(A_i < \mu_L - \mu) \\ &= \Phi\left(\frac{\mu_L - \mu}{\sigma_A}\right)\end{aligned}$$

based on $A_i \sim N(0, \sigma_A^2)$. A model-based estimator of this probability is

$$\begin{aligned}\hat{\pi} &= \Phi\left(\frac{\hat{\mu}_L - \hat{\mu}}{\hat{\sigma}_A}\right) \\ &= \Phi\left(\frac{L + z_{1-p_0}\hat{\sigma}_e - \hat{\mu}}{\hat{\sigma}_A}\right)\end{aligned}$$

where $\hat{\mu}$ is the estimator for the grand mean μ in the basic random effects model, $y_{ij} = \mu + A_i + e_{ij}$.

The difference between the model-based estimator and test-based estimator is studied using a simulation. Suppose the region consists of four sites. Two of them come from the baseline distribution, $N(90,\ 4^2)$ and the other two follow a distribution of $N(86.97,\ 4^2)$

distribution. In terms of the deviation from the baseline mean, the true regional impairment proportion is 0.5 (i.e., two of the four sites are impaired). In 1,000 simulations, one dataset is generated based on the standard random effects model with the variance ratio of 1 (i.e., $\sigma_A / \sigma_e = 1$) and the other is generated based on the standard fixed effects model. Both of these datasets are fit by a standard random effects model. Each site has 10 observations. The test-based estimator is based on the TL[EBLUP] approach. Results are summarized in Figure 3.7. Evidently the model-based estimator is closer to the true value than the test-based estimator. The regional impairment proportion is considerably underestimated by the test-based estimator. The estimator from FE-test-based is much less than that from RE-test based.

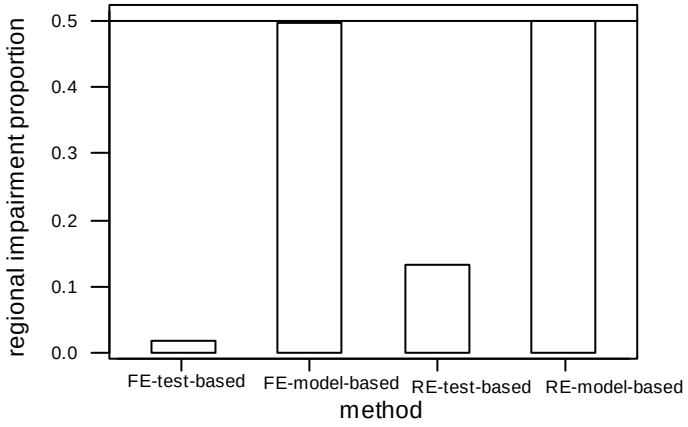

Figure 3.7 Estimated regional impairment proportion when truly 50% impairment in the region. FE and RE indicate that the data set is generated based on a fixed effects model and a random effects model, respectively.

The noticeable difference between test-based and model-based estimators may come from the potential correlation among tests and the low power of individual tests.

3.4.6 Summary

This chapter studies site-specific water quality assessment when the standard is fixed and the available sample size is small for the site of interest. Two model-based tests are

developed that give varying estimates of uncertainty for impairment detection. The results indicate that the test based on the fixed effects model outperforms the single site test when the test site is truly impaired. The test TL[11] is recommended for assessing water quality at sites. The basic random effects model doesn't improve assessment when the sample size is small or there are few non-test sites in the region. The gain in power for the test based on fixed effects models is less than 20%. Sample size, number of non-test sites, and true impairment severity affect test performance. Large sample sizes generally increases test power. The effect of the number of non-test sites on model-based tests is more obvious when the sample size per site is small. When the test site has a small sample size, inclusion of information from non-test sites improves test performance. Increasing sample size of the non-test site will improve the performance but the amount of improvement seems bounded. Even though the standard random effects model doesn't work well for impairment assessment, attention should be paid to effects of the potential factors (e.g., variance ratio) on tests. Although not useful for individual sites, the random effects model helps evaluate regional impairment.

In environmental studies, data with temporal correlation within a site are often encountered. The suitable choice of error structure can correct the estimation bias. The test based on the fixed effects model with complex variance structure (e.g., compound symmetry structure) tends to perform well for positively correlated data. This test is recommended when a small sample is available for assessment. Increasing the sample size at the test site leads to a more powerful test than pooling more sites. When the information of the regional impairment proportion is required, the model-based estimator is recommended. Multiplicity adjustment should be of concern when a regional assessment of impairment is combined with a site-specific assessment.

4. Model-based Assessment Using Prediction Limits

4.1 Introduction

4.1.1 Setting standards with reference conditions

Water quality standards are the foundation of water quality-based monitoring and assessment programs mandated by the Clean Water Act. The process of setting standards requires inputs from many individuals and groups, such as the National Drinking Water Advisory Council (NDWAC), representatives from water utilities, environmental groups, and public interest groups (USEPA, 1990). Even though the results are consensuses of expert opinion, the process involves compromises due to different use-benefit considerations. In addition, there are a lot of different standards definitions, such as national standards, state standards, standards for lakes, and standards for streams. When setting standards, the stability of a standard over time is assumed. This assumption requires substantial investigation of time and resources.

In terms of environmental development, biological standards represent conditions under which "a water body supports a balanced indigenous community of aquatic organisms" (Courtemanch et al., 1989). Therefore, the key of setting standards is to define the baseline condition in which a balanced indigenous community lives and against which test sites can be evaluated. Recent studies suggest that the fixed criterion may not be sufficiently protective in various subregions (Perry and Vanderklein, 1996; USEPA, 2000). To "broaden the implementation of assessment, increase the consistency among different regulatory levels, and improve the success of individual programs", the method of reference sites is documented by EPA programs (Southerland et al., 2005) for setting standards. This method tends to have minimum resource commitment while producing timely and useable information. Moreover, this method helps to define limits of acceptable conditions when specific standards are lacking (Hughes, 1995).

The sites that satisfy certain quality conditions are called reference sites. Several methods can be used to develop these quality reference conditions for defining reference sites: historical data, empirical models, and a consensus of expert opinion in the region of

interest. The reference sites have similar habitat and ecological characteristics with potentially impaired test sites. Reference sites can be used to predict water quality at current or future sites once they are selected for a given region (Kilgour et al., 1998, called Kilgour's work in the remainder).

4.1.2 Kilgour's work

In Kilgrour's paper, the quality of each site is indicated by one observation. The non-test sites are reference sites. The test site is either a reference or a non-reference site. Kilgour's paper describes an *F*-test for determining if one observation from a test site "lies outside of the true acceptable range". They assume the observation at the test site comes from the collection of potentially impaired sites and construct the hypotheses as $H_0 : \mu_r - \mu_{test} = 0$ versus $H_1 : \mu_r - \mu_{test} \neq 0$, where μ_r is the mean of the reference population and μ_{test} is the population mean of potentially impaired sites. The corresponding test statistic is $F = \frac{(\bar{y}_r - y_{test})^2}{(s_r\sqrt{1+1/n_r})^2}$, where $\bar{y}_r$ is the mean of sampled reference sites, s_r is the standard deviation of sampled reference sites, y_{test} is the test site observation, and n_r is the number of reference sites. This test has the general assumption that each site has only one observation. When a single site is assessed by this *F*-test, the power is relatively low if the number of reference sites is small. To detect ecologically important effect size (i.e., the difference between the mean of reference site population and the population mean of potentially impaired observations), they proposed non-central tests corresponding to two null hypotheses, not listing a site and listing a site, respectively. The test statistic of these two non-central tests is the square root of the statistic, $F = \frac{(\bar{y}_r - y_{test})^2}{(s_r\sqrt{1/n_r})^2}$. Power curves show that the test performance of the *F*-test falls between those two non-central tests (refer to Figure 1 in their paper, page 544).

Their *F*-test compares a prediction limit from reference sites to a single observation from a test site, while their non-central tests compare tolerance limits from reference sites to a single observation. However, these three tests all use reference data to construct an interval estimate for a single observation. The *F*-test focuses on prediction of a future

value while the non-central tests focus on estimation of a quantile. In addition, the power in their study is relatively low for detecting small deviations from the reference mean due to small sample sizes. No solution is suggested for improvement in their paper.

My study in this chapter concentrates on the prediction limit in assessment. This differs from Kilgour's study in data structure, significance level, estimation, and format of alternative hypothesis. The dataset in Kilgour's work assumes that a single observation from a site represents the site and all reference sites form a sample from the reference population. In my study, there are multiple observations at each reference site, and reference sites may come from different reference populations. A new observation at a specific reference site of interest will be assessed. This will provide dynamic tracking for reference conditions at a specific site when more observations at this site are available over time. This set-up can be of practical help for updating monitoring programs. Table 4.1 demonstrates the difference in data structure between Kilgour's work and my study. In this table, *i* indicates a site. In Kilgour's work, *i* takes values from 1 to n_r for reference sites and takes the value of *k*+1 for the test site. In my research, *i* takes values from 1 to *k* , n_i indicates the sample size at site *i*, and of interest is $(n_k+1)^{th}$ observation at the k^{th} reference site. In addition, Kilgour's work uses general significance level α (usually 5%) for testing. Practically, certain exceedance proportion is of great concern (USEPA, 1987). It is natural and meaningful to use the maximum acceptable exceedance proportion (p_0) as the significance level, i.e., $\alpha = 0.1$, in estimating the prediction limit. Furthermore, the model-based prediction will be introduced in my study to help solve the issues related to small sample sizes. Lastly, a one-sided prediction limit will be considered in my study instead of two-sided prediction limits.

Model-based estimation and prediction are widely used in small area problems (Rao, 2003) to borrow information from the 'neighborhood'. This chapter will implement this idea in impairment assessment using the estimated criterion from the reference conditions instead of the fixed criterion (in Chapter 3). In this framework, when tolerance limits are of interest, the assessment procedure based on models is similar to that in Chapter 3. Thus the rest of this chapter will only address performance of the model-based prediction. The basic procedure for the model-based prediction in water quality assessment will be

set up under the fixed effects model in Section 4.2. General performance of the model-based prediction will be displayed. More details of this method will be discussed as well as associated practical concerns by using the dataset from Non-coastal Virginia Streams in Section 4.3. Section 4.4 will briefly summarize the findings for the model-based prediction.

Table 4.1 Conceptual comparison of Kilgour's work with the model-based prediction

Kilgour's work (1998)	$\left\{\begin{array}{l}\text{site1} \rightarrow y_1 \\ \text{site 2} \rightarrow y_2 \\ \ldots\ldots \\ \text{site i} \rightarrow y_i \\ \ldots\ldots \\ \text{site } n_r \rightarrow y_{n_r}\end{array}\right\} \Leftarrow \{\text{site } (k+1) \rightarrow y_{k+1}\}$
Model-based prediction	$\text{site1} \rightarrow y_{11}, y_{12}, \ldots, y_{1n_1}$ $\text{site 2} \rightarrow y_{21}, y_{22}, \ldots, y_{2n_2}$ ..... $\text{site i} \rightarrow y_{i1}, y_{i2}, \ldots, y_{in_i}$ ..... $\text{site } k \rightarrow y_{k1}, y_{k2}, \ldots, y_{kn_k} \Leftarrow y_{k(n_k+1)}$

4.2 Model-based prediction

4.2.1 General procedure

The model-based prediction is based on the hypotheses, $H_0 : \mu_r - \mu_{test} = 0$ versus $H_1 : \mu_r - \mu_{test} > 0$, when a lower standard is of interest. Here, μ_r represents the population mean of reference sites and μ_{test} represents the mean of the population for the new observation. With the general form of the fixed effects model set-up in Section 3.2.2, the interest is to predict impairment for the $(n_k+1)^{th}$ observation at the k^{th} reference site. The test statistic is written as

$$t = \frac{\bar{y}_k - y_{test}}{s\sqrt{1+1/n_k}},$$

where $\bar{y}_k = \sum_{j=1}^{n_k} y_{kj}/n$ is the estimator for μ_r, the estimator for μ_{test} is y_{test} (the observation that requires evaluation), and $s = \hat{\sigma}_e = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 / (\sum_{i=1}^{k} n_i - k)}$. n_i is the number of observations at i^{th} site. k is the number of sites in the region of interest. This test statistic follows a t distribution with degrees of freedom equal to $\sum_{i=1}^{k} n_i - k$ under the null hypothesis, as discussed in standard textbooks for applied statistics (e.g., Neter et al., 1988). To determine if the new observation has the same characteristics as the current observations from the reference site, the comparison is carried out between the lower prediction limit $\bar{y}_k - t(\sum_{i=1}^{k} n_i - k, 1-p_0)s\sqrt{1+1/n_k}$ and the new observation y_{test}. If y_{test} is greater than the prediction limit, the new observation is declared unimpaired and the reference site can be treated as reference in the near future. If y_{test} is less than the lower prediction limit, the new observation is declared to be impaired and the reference conditions at this site require investigation. It may be no longer appropriate to treat this site as a reference site.

4.2.2 Performance evaluation

The performance of the model-based prediction can be evaluated by its power. Suppose the true ecological effect size (i.e., the difference between the mean of reference site population and the population mean of potentially impaired observations) is Δ_1, i.e., $\mu_r - \mu_{test} = \Delta_1,\ \Delta_1 > 0$. The power is derived as below.

$$power = \Pr(y_{test} < \bar{y}_k - t(\sum_{i=1}^{k} n_i - k, 1-p_0)s\sqrt{1+1/n_k})$$

$$= \Pr(\frac{\bar{y}_k - y_{test}}{s\sqrt{1+1/n_k}} > t(\sum_{i=1}^{k} n_i - k, 1-p_0))$$

$$= \Pr\left(\frac{\frac{\bar{y}_k - y_{test} - \Delta_1}{\sigma\sqrt{1+1/n_k}} + \frac{\Delta_1}{\sigma\sqrt{1+1/n_k}}}{s/\sigma} > t\left(\sum_{i=1}^{k} n_i - k, 1-p_0\right)\right)$$

$$= \Pr\left(T\left(\sum_{i=1}^{k} n_i - k, \lambda\right) > t\left(\sum_{i=1}^{k} n_i - k, 1-p_0\right)\right)$$

where $\lambda = \frac{\Delta_1}{\sigma\sqrt{1+1/n_k}}$.

With the underlying assumption of a standard normal distribution for the observations, a power curve can be drawn to summarize the performance of model-based prediction. When sample sizes are unequal across the region, only the degrees of freedom matter in the power calculation. The theoretical results will thus be summarized in terms of degrees of freedom instead of the sample size and the number of sites. The sample size at the test site and the total degrees of freedom for the model-based estimation will be addressed in the results.

4.2.3 Results

Using the derivation in Section 4.2.2, the general behavior of the model-based prediction is shown in Figure 4.1 and Figure 4.2. Figure 4.1 gives the power for the single site prediction approach when the sample size varies. This approach is a special case of the model-based prediction, i.e., when other sites don't contribute to the prediction and the variance is estimated from one site. The power plot reveals an increasing trend in power when the sample size becomes larger and the degree of impairment increases. Figure 4.2 displays the scenario when there are 4 observations at the test site and the degrees of freedom vary. With increased degrees of freedom (i.e., more sites or sites with more observations pooled in the model), the increases in power for the model-based prediction are negligible. The more severe the impairment, the more powerful the test with model-based prediction is.

The power improvement from the model-based prediction is also shown in the comparison with the single site prediction approach. Figure 4.3 compares the results from Figure 4.2 to the single site approach, when there are four observations at the test site.

The relative change in power is calculated by $\frac{power1 - power0}{power0}$, where *power1* is the power for model-based prediction, and *power0* is the power for the single site prediction approach. The power gain is an arch with increased impairment severity. When more observations are pooled in the model, the relative change in power becomes larger. The power gain is up to 16% for the observed degrees of freedom and true exceedance proportion. This result is consistent with findings from model-based tests using tolerance limits (refer to Section 3.4.1).

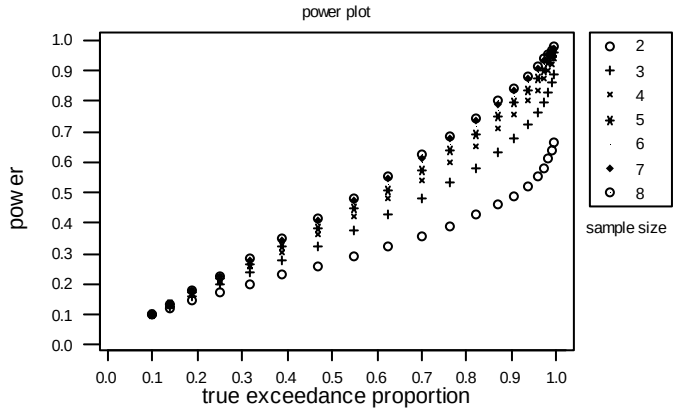

Figure 4.1 Power for test using the single site prediction.

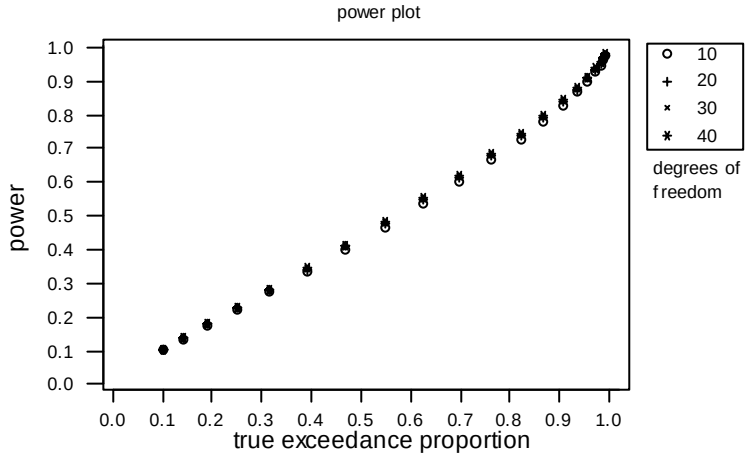

Figure 4.2 Power plot for model-based prediction with 4 observations at the test site.

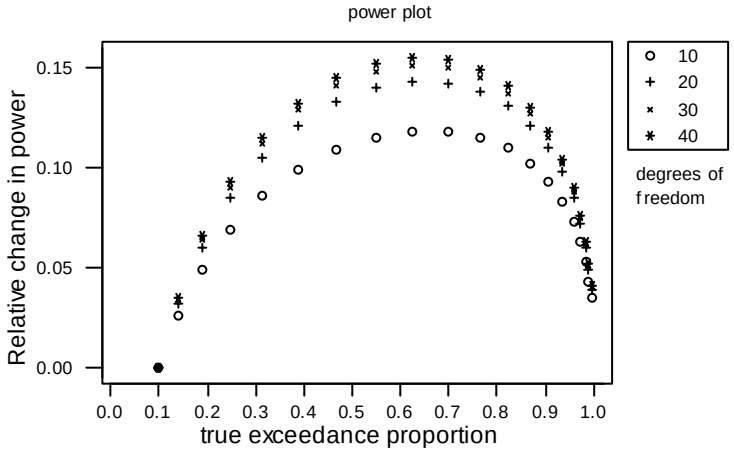

Figure 4.3 Power gain for model-based prediction with 4 observations at the test site.

4.3 Application: data from non-coastal Virginia

This section demonstrates how model-based assessment using the prediction limit may be applied to real data. The applied Virginia dataset was originally used to develop a multimetric macroinvertebrate index for Virginia (VDEQ, 2006). A basic fixed effects model, random effects model, and mixed effects model are fit to reference sites in this dataset. Misclassification of reference sites is addressed for each model. The basic mixed effects model consists of one fixed effect and one random effect as described in standard statistic textbooks (e.g., Neter et al., 1988).

4.3.1 Dataset

The non-coastal Virginia dataset contains information from upstream control sites under the Rapid Bioassessment Protocols (USEPA, 1999), covering 3 ecoregregions (mountain, piedmont, and coastal) and 6 stream orders. To evaluate model-based prediction performance for this application, the stream condition index (SCI) is the response of interest and the study area is narrowed to the mountain ecoregion. When the SCI is greater than 60, the station is viewed as a reference station and labeled as 1. The final cleaned dataset consists of 100 reference stations and 52 non-reference stations. Table 4.2 lists the variables in the final dataset.

Table 4.2 Variables in application dataset

Variable	Range
Stream order	1, 2, 3, 4, 5
Season	Spring, Fall
SCI	22~84
reference	0 – non-reference site (SCI ≤ 60) 1 – reference site (SCI>60)

4.3.2 Impairment detection

Three standard models are fit to the reference sites in the dataset. Table 4.3 displays the components in each model. An independence covariance structure is assumed for the error term. Other basic assumptions are similar to the set-up in Section 3.2.2. The test statistics based on these models are derived in Table 4.4. When the estimated prediction limit is greater than the observed measurement of interest, the observation is claimed as impaired. With this set-up, the FE approach incorrectly claims 1 non-reference site as an unimpaired site (1 out of 52) and 8 reference sites as impaired sites (8 out of 100). The RE1 and M1 approaches incorrectly claim 21 non-reference sites as unimpaired (21 out of 52). The RE2 and M2 approaches misclaim 19 non-reference sites as unimpaired (19 out of 52). Define the false prediction rate for non-reference sites (Type II error) as the proportion of non-reference sites incorrectly declared to be unimpaired. This false prediction rate for models in Table 4.4 is 1.9% for FE, 36.5% for RE1 and M1, and 40.4% for other models. Though the FE approach has 8% false prediction rate for reference sites, with the pre-specified significance level as 10%, it is acceptable to have relatively high false prediction rate for reference sites in the FE approach.

The finding from this application is that the test based on the fixed effects model has a lower Type II error rate than tests based on the random effects or mixed effects models while maintaining acceptable Type I error. Using an empirical best linear unbiased predictor (EBLUP) for site mean estimator slightly improves correct listings.

Table 4.3 Model set-up for application

Label	Model
FE	Fixed effects model: SCI = stream order + error
RE	Random effects model: SCI = stream order + error
M	Mixed effects model: SCI = stream order + season + error Fixed effect – season Random effect - stream order

Table 4.4 Estimation of lower prediction limit

Label	Model
FE	$\bar{y}_k - t(\sum_{i=1}^{k} n_i - k, 1-p_0)s\sqrt{1+1/n_k}$ $s = \hat{\sigma}_e = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2/(\sum_{i=1}^{k} n_i - k)}$
RE1	$\bar{y}_k - t(\sum_{i=1}^{k} n_i - k, 1-p_0)s_m\sqrt{1+1/n_k}$
RE2	$\bar{y}_{k,EBLUP} - t(\sum_{i=1}^{k} n_i - k, 1-p_0)s_m\sqrt{1+1/n_k}$
M1	$\bar{y}_k - t(\sum_{i=1}^{k} n_i - k, 1-p_0)s_m\sqrt{1+1/n_k}$
M2	$\bar{y}_{k,EBLUP} - t(\sum_{i=1}^{k} n_i - k, 1-p_0)s_m\sqrt{1+1/n_k}$
$s_m = \sqrt{q\hat{\sigma}_A^2 + \hat{\sigma}_e^2}$, $q = \frac{(\sum_{i=1}^{k} n_i - k)^2 - \sum_{i=1}^{k} n_i^2}{(\sum_{i=1}^{k} n_i - k)(k-1)}$, $\hat{\sigma}_A^2, \hat{\sigma}_e^2$ vary across models.	

4.4 Closing comments

The model-based prediction discussed in this chapter uses reference sites to determine the standard, while the model-based tests using tolerance limits in Chapter 3 treat the

standard as a fixed value. These two methods correspond to different settings of standards and thus different usage of regional information. They both assess water quality for sites with unknown impairment status. The results in Chapter 3 and Chapter 4 show that model-based tests using tolerance limits and prediction limits have similar test performance. The power for tests using model-based prediction increases with large sample size and severe impairment. The prediction based on the fixed effects model outperforms the single site prediction approach. The power gain is less than 20%.

5. Assessment Using Univariate Regression-based Tests

5.1 Introduction

Ecological assessments provide essential information for water ecosystem protection. Biological/ecological metrics used for impairment assessment include characteristics of the biota that change in some predictable way with increased human influence (Barbour et al., 1996). They are therefore used to indicate biological conditions as well as establish relationships with potential causes of impaired conditions. Chapter 3 and Chapter 4 introduced model-based tests for impairment detection in water quality assessment. Once impairment is detected and the list of impaired water segments is developed under section 303(d) of the Clean Water Act (CWA), the CWA requires localities to establish a priority ranking for water segments on the impairment list and to develop Total Maximum Daily Load (TMDL) programs for these water segments. Identification of the causes of impairment is a critical step in TMDL development. Linking biological conditions with pollutants or stressors causing their impairment is essential before strategies can be developed to remedy the impairment in TMDL studies. By an environmental stressor, I mean any physical or chemical variable in the natural environment that impacts the organism life in that environment in an adverse manner.

Regression analysis methods are often used for modeling relationships between biological variables and stressors (Barbour et al. 1999; Legendre et al., 1998; Reynoldson et al., 2000; Wright et al., 2000). When the regression method is used in the assessment stage, it may be possible to strengthen the test of impairment, especially when there is little information for setting a numerical criterion or the criterion varies with a covariate (Urquhart, 1982; McCormick et al., 2001). The post-assessment evaluation (i.e., the evaluation of restoration following an impairment decision) may also be based on regression methods. After implementing the Total Maximum Daily Load procedure to improve water quality, researchers should re-evaluate the water quality and determine if the restoration process is successful. This re-evaluation also requires knowledge of the linkage between impairment and causes. Regression approaches thus play an important role in evaluating impairment and recovery by modeling biological condition as well as

adjusting biological measures for natural covariates (such as elevation, water temperature).

One currently popular application of regression methods in environmental assessment is the Benthic Assessment of Sediment (BEAST) method used in Canada (Barbour et al. 1999). This method sets criteria by defining reference sites. The reference sites are typically defined by non-biological conditions associated with healthy sites. A site is classified as a reference site only if all of the non-biological criteria are met. Once reference sites are established, the limit around some biological condition at the reference sites determines the space of acceptable conditions and hence an implied criterion, which is used to test impairment of other sites in the region. In this procedure, regression analysis is used to adjust a response by the value of the covariate (i.e., the response is regressed on the covariate instead of stressors). This type of implementation relies on the fact that biological data vary naturally with factors such as elevation, stream gradient, and water temperature. Regression helps neutralize the effect of the covariate when it is not of interest in a study. Though the inclusion of covariates can increase statistical power since it accounts for some of the variability (Urquhart, 1982), no impairment assessment is conducted with this regression technique (McCormick et al., 2001).

Other methods such as generalized additive models (GAMs) may also be used to estimate the response-stressor relationships at test sites (Yuan and Norton, 2003). They "provide a flexible modeling tool for visualizing the relationships between variables" (Hastie, 1992). It is a popular method in ecological studies (e.g., Heegaard et al., 2001). Although other methods can be applied (Bates Prins et al., 2006), the approach here focuses on simple regression models. My study aims at providing a general regression-based test for impairment detection and performance evaluation.

The rest of this chapter consists of four sections. Section 5.2 will set up the basic simple linear regression for regression-based tests and the corresponding test procedure. The boundary of the acceptance region will be defined. Section 5.3 will illustrate the test performance for regression-based tests by a simulation study and a case application, based on different definitions of the boundary of the rejection region. In Section 5.4, the situation where the regressor is not a stressor but a factor associated with the variation in the biological variable (i.e., a covariate) will be addressed by comparing the power for

the covariate-adjusted and ordinary tests. Finally, a brief summary of this chapter will be given in Section 5.5. Extensions of these tests to more complex models will be a topic for further research.

5.2 Methods

The evaluation of impairment is typically described in terms of its nature, magnitude, and spatial/temporal extent. When impairment is defined in terms of a quantified response, such as a count or continuous variable, it is easier to model the status of the biological community with regression methods. When multiple stressors are involved, building response-stressor relationships becomes particularly complex due to the potential contrary directions of stressor effects, data limitation (lack of orthogonality), interactions, and associated interpretation. A simple regression method is implemented here by assuming that each impairment evaluation is based on a continuous biological variable and a single stressor (or a single covariate). The biological variable is regressed on the stressor and the regression results are compared with some acceptance criteria in order to assess the degree of impairment, which is called the regression-based test. The case of multiple responses and stressors will be discussed in Chapter 6. Different from the application of regression in BEAST, my study uses all information to build the regression model instead of just reference sites. Regression methods are directly applied to impairment detection as well as to explain potential impairment. The basic procedure begins with fitting the regression on the stressor by using all available data (which may include potential reference data). The fitted response and related tolerance/prediction interval of the site at the border of ‘good’ and ‘bad’ are then obtained based on the fitted regression. These intervals form a rejection region. A test site is examined to see if it falls in this rejection region.

5.2.1 Model set-up

In this chapter, the biological criterion is assumed to vary as a function of the stressor in a monotonic manner and each measurement represents one of a collection of n sites. There may be one or several sites that are viewed as reference sites. When there is a single response and a single stressor measured at n sites, the regression model is written as

$$\underline{y} = \beta_0 \underline{j} + \beta_1 \underline{x} + \underline{\varepsilon} = X\underline{\beta} + \underline{\varepsilon}$$

where $\underline{y}$ is an $n \times 1$ vector of the response, $\underline{j}$ is an $n \times 1$ vector of 1's, $\underline{x}$ is an $n \times 1$ vector of values of the stressor, $X = [\underline{j}, \underline{x}]$, $\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, and $\underline{\varepsilon} \sim N(\underline{0}, I_n \sigma^2)$. The standard deviation, σ, is treated as unknown. The data are assumed to be collected from *n* sites at one time point. The maximum likelihood estimators of the unknown parameters, $\underline{\beta}$ and σ^2, are $\hat{\underline{\beta}} = (X^T X)^{-1} X^T Y$ and $\hat{\sigma}^2 = \frac{1}{n-2} \underline{y}^T (I - X(X^T X)^{-1} X^T) \underline{y}$. Both are unbiased.

In practice, the biological data are collected over time. When the collection time period is relatively short (e.g., a season of 4 months), the data can be treated as from one point in time. To simplify the discussion of regression-based tests, the potential correlation between the observations/sites is not considered here. All the sites are assumed to be independent of each other.

5.2.2 Test procedure

Suppose impairment is indicated by values below some minimum threshold (i.e., the criterion) of the response and the response is strongly positively associated with the stressor. When more than 10% of the response population exceeds the criterion, *L*, the test site is declared impaired. Therefore, the general hypotheses in terms of the exceedance proportion for the response are $H_0 : p \le p_0$ (unimpaired, don't list) versus $H_1 : p > p_0$ (impaired, list). Here p is the population exceedance proportion for the response at a value of *x*, and p_0 is the acceptable (or maximum) exceedance proportion for the response. Similar to the model-based tests in Chapter 3 and Chapter 4, the estimated tolerance limit and prediction limit from the sample can be used for impairment decisions. However, the regression-based estimators will be the input for estimated limits instead of the ANOVA-based estimators. When the estimated limit is less than the criterion, the test site will be declared as impaired; otherwise, the test site will be declared as unimpaired.

Ideally, when the biological response varies as a function of the stressor in a monotonic manner (for example, pH or DO), it is possible to determine a minimum value

of the stressor that produces no effect or only a slight effect on the biological variable (this is similar to the no-observable effect level in toxicity testing). Suppose the minimum stressor value of interest is x_0 and the fitted response is $\underline{x}_0^T \hat{\underline{\beta}}$, where $\hat{\underline{\beta}} = (X^T X)^{-1} X^T Y$ and $\underline{x}_0 = \begin{pmatrix} 1 \\ x_0 \end{pmatrix}$. The variance of $\underline{x}_0^T \hat{\underline{\beta}}$ is $\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 \sigma^2$. Under the usual regression assumptions, $(n-2)\hat{\sigma}^2 / \sigma^2 \sim \chi^2_{n-2}$. Assume the true response mean is μ, $\mu = \underline{x}_0^T \underline{\beta}$. The test statistic is $\frac{\hat{\mu}_0 - L}{se(\hat{\mu}_0)} = \frac{\underline{x}_0^T \hat{\underline{\beta}} - L}{\sqrt{h\hat{\sigma}^2}} = \frac{\frac{\underline{x}_0^T \hat{\underline{\beta}} - \mu}{\sqrt{h\sigma^2}} + \frac{\mu - L}{\sqrt{h\sigma^2}}}{\sqrt{h\hat{\sigma}^2 / h\sigma^2}}$ with $h = \underline{x}_0^T (X^T X)^{-1} \underline{x}_0$, $\hat{\underline{\beta}}$ and $\hat{\sigma}$ defined in the preceding. This test statistic follows a non-central *t* distribution (as discussed in standard textbooks for applied statistics), $t(df, \lambda)$, with $df = n-2$ and $\lambda = \frac{\mu - L}{\sqrt{h\sigma^2}}$. Under the null hypothesis, the noncentrality λ is $z_{1-p_0} / \sqrt{h}$. The lower tolerance limit of the response at the site with the minimum stressor value is calculated by

$$\underline{x}_0^T \hat{\underline{\beta}} - t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0}}) \sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 \hat{\sigma}^2},$$

where $t(n-2, \alpha, \lambda)$ with $\lambda = \frac{z_{1-p_0}}{\sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0}}$ is the critical coefficient from a non-central *t* distribution with (*n*-2) degrees of freedom and noncentrality parameter λ at the significance level of α.

When a new site is of interest, the prediction limit at the site with the minimum stressor value is

$$\underline{x}_0^T \hat{\underline{\beta}} - t(n-2, 1-p_0) \sqrt{(\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 + \frac{1}{n}) \hat{\sigma}^2},$$

where $t(n-2, 1-p_0)$ is the critical coefficient from a central *t* distribution with (*n*-2) degrees of freedom at the significance level of $1-p_0$. The derivation of the prediction limit is similar as that in Section 4.2.

5.2.3 Boundary of rejection region

The tolerance limit or prediction limit sets the bound on the rejection/acceptance region at a specific stressor value, which limits the rejection area just with respect to the response. If a graph is made using the stressor on the X-axis and the biological response on the Y-axis, then the boundary line is in the vertical direction (Figure 5.1a). If the stressor can only take values that meet the standard of that variable (for example, the standard for dissolved oxygen), setting the boundary in the horizontal direction for protection purposes (this is similar to a no-observable effect level in toxicology) is required. With the extra information from the stressor variable, a broad acceptance region (i.e., a narrow rejection region) can be constructed for making general policies.

In the regression setting, a conservative approach (in terms of unimpairment claims) might use the tolerance/prediction limit of the response at reference sites constructed at the standard of the stressor (which is treated as a fixed value) to define a limit. Alternately, if there is sufficient information from a random sample of reference sites, the mean of the stressor at the reference sites may be calculated as the minimum stressor value of interest, i.e., the criterion of the stressor. Thus the alternative way constructs the limit at the stressor mean of the reference sites, which is likely to be in favor of environment protection. Figure 5.1a conceptually displays the acceptance and rejection region when low values of the response and the stressor indicate 'bad' quality. In this figure, X-axis and Y-axis indicate the stressor and the response, respectively. Generally, increases in stressors are associated with decreases in biological conditions. Figure 5.1a indicates increases in stressors associated with increases in biological conditions. This conversion is made so that the model set-up in Chapter 3 can be directly applied in this chapter and interpretation of test evaluation can be easily understood.

The underlying assumption in Figure 5.1a is that a high value is good (e.g., DO) for the response and stressor. A site is treated as an impaired site (in the tolerance limit approach) if its response value is less than the tolerance limit defined in Section 5.2.2 and the stressor value is less than the minimum stressor value x_0. If the response value is less than the prediction limit and the stressor value is less than the minimum stressor value, the site is declared as impaired in the prediction limit approach. Figure 5.1a indicates that

the acceptance region based on regression tends to claim unimpairment. The chance should be low for reference sites to be declared as impaired.

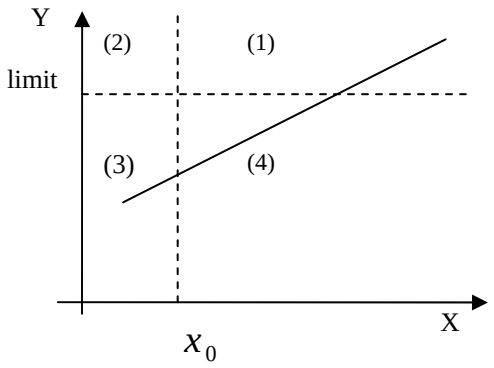

Figure 5.1a Conceptual display of the acceptance and rejection regions. The fourth parts divided by the limit and the x_0 are indicated by (1), (2), (3), and (4). The area in part (1) and (4) indicates the stressor acceptance region; the area in part (1) and (2) indicates the biological acceptance region; the area in part (3) indicates the rejection region with regression adjustment.

Figure 5.1b conceptually displays the rejection regions based on the response limit adjusted by regression when x_0 takes the value of the criterion or the mean of reference sites.

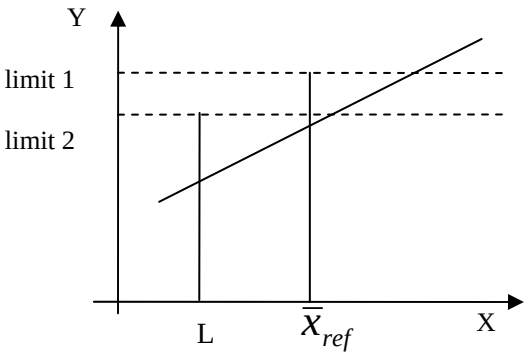

Figure 5.1b Conceptual display of rejection regions based on the adjusted response limit. Limit 1 and limit 2 are the adjusted response limits when x_0 takes the value of the mean of reference sites and the criterion, respectively. The area below each dotted line indicates the rejection region using the adjusted response limit.

To compare test performance of these boundaries, I focus on the scenario that the response is positively correlated with the stressor. The lower the stressor, the lower the response is and the more severe the impairment at the site is. The standard of the stressor is set as x_L. The conservative boundary is obtained when the value of the stressor at the boundary of the acceptance region is x_L. The tolerance/prediction limit is calculated as in the preceding section with $\underline{x}_0$ set to $\begin{pmatrix} 1 \\ x_L \end{pmatrix}$. The protective boundary is based on the value of the stressor at the boundary of the rejection region set as $\bar{x}_{ref}$, where $\bar{x}_{ref}$ is the mean of the stressor at the reference sites. Now $\underline{x}_0$ becomes $\begin{pmatrix} 1 \\ \bar{x}_{ref} \end{pmatrix}$. Table 5.1 lists the boundaries considered in this chapter. In the abbreviations, L refers to the low criterion and R refers to reference. These abbreviations will be used throughout this chapter.

Table 5.1 Definitions of the boundary of the rejection region for tolerance limits (TL) and prediction limits (PL)

X-axis*	Y-axis	Label
$x_0 = x_L$	$\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0}}) \sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 \hat{\sigma}^2}$	LTL
	$\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, 1-p_0) \sqrt{(\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 + \frac{1}{n}) \hat{\sigma}^2}$	LPL
$x_0 = \bar{x}_{ref}$	$\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0}}) \sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 \hat{\sigma}^2}$	RTL
	$\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, 1-p_0) \sqrt{(\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 + \frac{1}{n}) \hat{\sigma}^2}$	RPL

* The minimum value of the stressor, i.e., the value of the stressor at the boundary of the rejection region.

5.2.4 Simulation for performance evaluation

The performance of the methods based on boundaries in Table 5.1 is compared by a simulation study. The simulation mimics the response-stressor relationship between the Hilsenhoff Biotic Index (HBI) and dissolved oxygen (DO). HBI characterizes the overall

pollution tolerance of a benthic macroinvertebrate community (Lenat, 1993). Its value can range from 0 to 10 in a 10 point scale. HBI was originally developed to assess low dissolved oxygen caused by organic loading. Higher HBI values indicate poor water quality and lower values indicate better water quality. Table 5.2 classifies the water quality based on HBI (Hilsenhoff, 1987). To fit the model structure from the preceding section, the response variable is transformed by subtracting the HBI value from 10, which is called NHBI in this chapter. The relationship between NHBI and DO will be directly used in the regression model.

Table 5.2 Water quality classifications for the HBI

HBI value	NHBI value	Water quality	Degree of organic pollution
0-3.50	6.5-10	Excellent	No apparent
3.51-4.50	5.5-6.49	Very good	Slight
4.51-5.50	4.5-5.49	Good	Some
5.51-6.50	3.5-4.49	Fair	Fairly significant
6.51-7.50	2.5-3.49	Fairly poor	Significant
7.51-8.50	1.5-2.49	Poor	Very significant
8.51-10	0-1.49	Very poor	Severe

In the simulation, the region of interest consists of two types of sites, reference and non-reference sites. It is assumed that the stressor doesn't contribute to define reference conditions. The covariance structures at reference and non-reference sites are assumed to be the same. The mean of the response variable at each site indicates water quality. The samples are assumed to come from the multivariate normal distribution having the form, $MVN\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix},\begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}\right)$, where the independent variable X and the response variable Y represent DO and NHBI, respectively. μ_X and μ_Y are the population mean of the stressor DO and the response NHBI, respectively. σ_X^2 and σ_Y^2 are the corresponding population variance. σ_{YX} is equal to σ_{XY}, which represents the covariance between the response and stressor variables. The conditional distribution of the response variable at

the given stressor value $X = x$ is also a normal distribution with the mean $\mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X)$ and standard deviation $\sqrt{\frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2}}$, i.e., $Y \mid X = x \sim N(\mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X), \sqrt{\frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2}})$ (Casella and Berger, 1990).

With this conditional distribution and the consideration of the difference between reference and non-reference sites, datasets consisting of 40 observations/sites and 2 variables are generated in terms of the relative number of sites and the relative site quality (i.e., the site mean). Table 5.3 lays out the 4 cases/scenarios considered in the simulation. The lower standards for the stressor and the response are both set at 5. The distribution, $MVN\left(\begin{pmatrix}6.28\\6.12\end{pmatrix}, \begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$, represents a baseline distribution (i.e., 90% of the sites have DO greater than the standard of 5). The response mean at the non-reference sites is 2 standard deviations away from the baseline mean. The response mean of the reference sites in Case3 (more non-reference sites and better site quality at reference sites) and Case4 (more reference sites and better site quality at reference sites) is 1 standard deviation away from the baseline mean. Case1 (more non-reference sites) indicates the common practical scenario. Case2 (more reference sites) and Case4 focus on the effect of the relative number of reference and non-reference sites. Case3 considers the effect of the unimpairment level. The boundaries defined in Table 5.1 are calculated for each site in each case. If the DO is less than the boundary and the NHBI is less than a specific limit, the test site is declared as impaired, i.e., classified as a 'bad' site. The assessment is carried out independently for each site. The number of sites which are claimed 'bad' among reference sites and non-reference sites will be counted separately. The proportion of 'bad' sites in each group comes from 10,000 simulations. A recommended method is expected to have a small proportion of 'bad' sites for reference sites and a large proportion of 'bad' sites for non-reference sites.

Table 5.3 Simulation scenarios in terms of the number of sites and corresponding distributions

Scenario*	Number of sites	Distribution
Case1	10 reference sites	$MVN\left(\begin{pmatrix}6.28\\6.12\end{pmatrix},\begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$
	30 non-reference sites	$MVN\left(\begin{pmatrix}4.28\\4.52\end{pmatrix},\begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$
Case2	30 reference sites	$MVN\left(\begin{pmatrix}6.28\\6.12\end{pmatrix},\begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$
	10 non-reference sites	$MVN\left(\begin{pmatrix}4.28\\4.52\end{pmatrix},\begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$
Case3	10 reference sites	$MVN\left(\begin{pmatrix}7.28\\6.92\end{pmatrix},\begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$
	30 non-reference sites	$MVN\left(\begin{pmatrix}4.28\\4.52\end{pmatrix},\begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$
Case4	30 reference sites	$MVN\left(\begin{pmatrix}7.28\\6.92\end{pmatrix},\begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$
	10 non-reference sites	$MVN\left(\begin{pmatrix}4.28\\4.52\end{pmatrix},\begin{pmatrix}1 & 0.64\\0.64 & 0.64\end{pmatrix}\right)$

* Each case provides a pair of distributions for reference and non-reference sites using a total sample size of 40.

5.3 Results and discussion

5.3.1 Simulation results

Using the limit and the corresponding stressor value as the boundary of the rejection region, the simulation results in Table 5.4 reveal that the rejection region based on the reference information is conservative in claiming unimpairment. When there are more reference sites in the region than non-reference sites, the rejection region based on the use of numerical criterion for the stressor classifies more non-reference sites as impaired sites than the situation when the reference sites are fewer than the non-reference sites

(comparing Case1 with Case2 for LTL and LPL). When the reference sites have fairly good quality, the results of impairment evaluation are not different from the case when the reference sites are at the baseline (comparing Case1 with Case3 for LTL, LPL, RTL and RPL) with the set-up. In addition, it is obvious that the approach using the stressor mean at reference sites as the boundary of the rejection region tends to declare impairment compared with the approach using the stressor standard. The prediction limit approach tends to have higher misclassification for reference sites compared with the tolerance limit approach when the stressor mean at reference sites is the boundary of the rejection region.

Table 5.4 Proportion of sites claimed to be impaired from reference and non-reference sites

	Method	From non-reference sites (high is good)	From reference sites (low is good)
Case1	LTL	0.37	0.00
	LPL	0.51	0.00
	RTL	0.97	0.20
	RPL	0.97	0.30
Case2	LTL	0.52	0.00
	LPL	0.52	0.00
	RTL	0.90	0.23
	RPL	1.00	0.33
Case3	LTL	0.37	0.00
	LPL	0.51	0.00
	RTL	1.00	0.30
	RPL	1.00	0.30
Case4	LTL	0.52	0.00
	LPL	0.52	0.00
	RTL	1.00	0.23
	RPL	1.00	0.33

5.3.2 Heterogeneity

The simulation in Section 5.2.4 assumes a homogeneous covariance structure for the region, which produces data with the relationship shown in Figure 5.2. The information at reference sites doesn't affect the estimation of variance parameters, consequently the

response-stressor relationship. Empirical evidence suggests that the variability among the non-reference sites is usually larger than that for the reference sites under moderate impairment and the response-stressor relationship among good sites may be different from that among bad sites. In this section, the correlation between the response and the stressor at non-reference sites is assumed to be high, around 0.8. When the sites are unimpaired, the response-stressor relationship is assumed to be weak. To address the correlation effect on impairment assessment, the correlations at the reference and non-reference sites may be assumed to be the same in some cases, around 0.4.

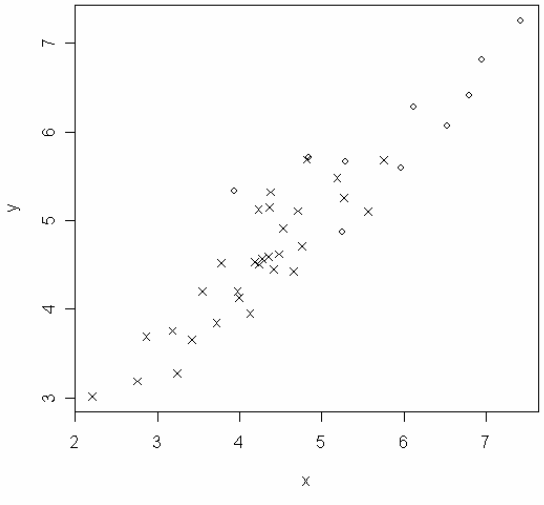

Figure 5.2 Simulated data for the reference and non-reference sites with the same correlation and covariance structure. The X-axis and Y-axis represent the stressor and the response, respectively. The cross and circle indicate non-reference and reference sites, respectively.

Two heterogeneity scenarios are discussed here. In the first scenario, the sites have the same response-stressor relationship but different covariance structure. In the other scenario, the covariance structure and the correlation over the sites are both different for the two types of sites. Table 5.5 details the scenarios in terms of the distribution and the number of reference sites. Case5 and Case6 represent the scenario of sites with the same response-stressor relationship but different covariance structure. Case7 and Case8 represent the scenario of sites with different response-stressor relationship and covariance

structure. Test approaches associated with Table 5.1 are applied to the 4 cases of interest. Figure 5.3 demonstrates scenarios for Case5 and Case7. The correlation between the response and the stressor is fixed at 0.4 for the reference sites. The correlation between the response and stressors at the non-reference sites is either 0.4 or 0.8.

Table 5.5 Simulation framework for the effect of heterogeneity

	Number of sites	Distribution
Case5	10 reference sites	$MVN\left(\begin{pmatrix}6.28\\6.12\end{pmatrix},\begin{pmatrix}1 & 0.32\\0.32 & 0.64\end{pmatrix}\right)$
	30 non-reference sites	$MVN\left(\begin{pmatrix}4.28\\4.52\end{pmatrix},\begin{pmatrix}4 & 1.44\\1.44 & 3.24\end{pmatrix}\right)$
Case6	30 reference sites	$MVN\left(\begin{pmatrix}6.28\\6.12\end{pmatrix},\begin{pmatrix}1 & 0.32\\0.32 & 0.64\end{pmatrix}\right)$
	10 non-reference sites	$MVN\left(\begin{pmatrix}4.28\\4.52\end{pmatrix},\begin{pmatrix}4 & 1.44\\1.44 & 3.24\end{pmatrix}\right)$
Case7	10 reference sites	$MVN\left(\begin{pmatrix}6.28\\6.12\end{pmatrix},\begin{pmatrix}1 & 0.32\\0.32 & 0.64\end{pmatrix}\right)$
	30 non-reference sites	$MVN\left(\begin{pmatrix}4.28\\4.52\end{pmatrix},\begin{pmatrix}4 & 2.88\\2.88 & 3.24\end{pmatrix}\right)$
Case8	30 reference sites	$MVN\left(\begin{pmatrix}6.28\\6.12\end{pmatrix},\begin{pmatrix}1 & 0.32\\0.32 & 0.64\end{pmatrix}\right)$
	10 non-reference sites	$MVN\left(\begin{pmatrix}4.28\\4.52\end{pmatrix},\begin{pmatrix}4 & 2.88\\2.88 & 3.24\end{pmatrix}\right)$

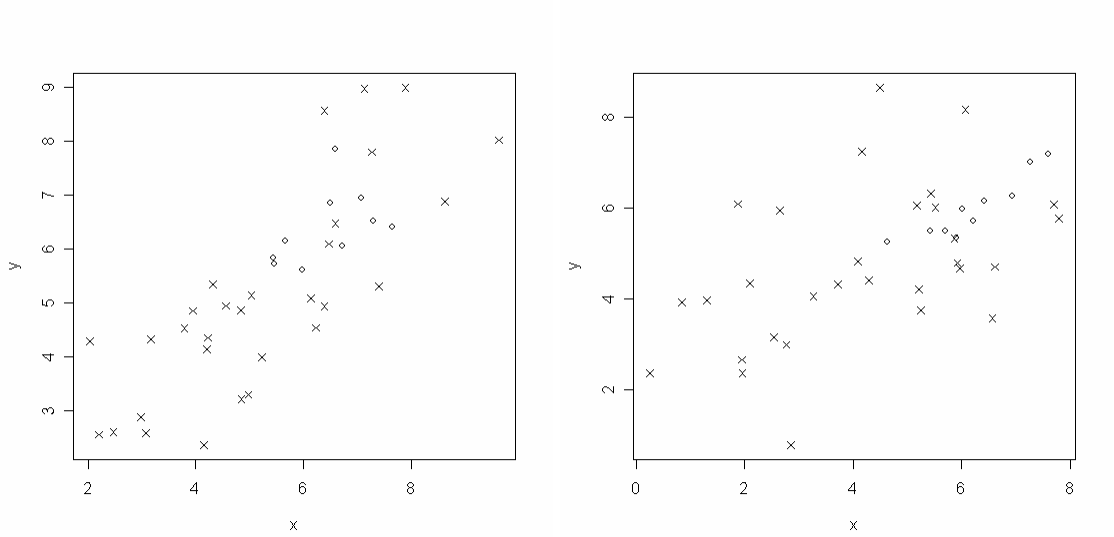

Figure 5.3 Data structures when different correlations/covariance structures exist. The two types of sites in the left graph have different covariance structures but the same correlation. This corresponds to Case5 in Table 5.5. The two types of sites in the right graph differ in correlation and covariance structure. This corresponds to Case7 in Table 5.5. The cross and circle in both graphs indicate non-reference and reference sites, respectively.

Compared with results in Table 5.4, the results in Table 5.6 indicate that different covariance structure for reference and no-reference sites leads to loss of power, i.e., less declarations of impairment for the non-reference sites (Case1 versus Case5). When reference and non-reference sites differ in the correlation and variance structure, more reference sites lead to incorrect classification of reference and non-reference sites when using the RTL approach (comparing Case7 with Case8 in Table 5.6).

Table 5.6 Proportion of sites claimed to be impaired from reference and non-reference sites

	Method	From non-reference sites (high is good)	From reference sites (low is good)
Case5	LTL	0.23	0.00
	LPL	0.40	0.00
	RTL	0.57	0.00
	RPL	0.67	0.00
Case6	LTL	0.30	0.00
	LPL	0.50	0.00
	RTL	0.70	0.03
	RPL	0.70	0.20
Case7	LTL	0.30	0.00
	LPL	0.43	0.00
	RTL	0.63	0.00
	RPL	0.73	0.30
Case8	LTL	0.30	0.00
	LPL	0.50	0.00
	RTL	0.70	0.10
	RPL	0.70	0.30

5.3.3 Application: dataset from the non-coastal Virginia

The Stream Condition Index (SCI) is a composite index for use in flowing streams. “In keeping with the Clean Water Act and current technical guidance from USEPA” (Burton and Gerritsen, 2003), SCI was developed in several states as a metric that can indicate a healthy ecosystem and link the biological indicators with water management and thus protect the water source (for example, Alaska: Major et al., 2002; Virginia: Burton and Gerritsen, 2003; Florida: Frydenborg and Ray, 2005). The SCI for Virginia non-coastal streams (VSCI) is derived from the dataset collected in 1994-2002. The study is described in detail in Burton and Gerritsen's report. The Virginia Department of Environmental Quality (VDEQ) uses the independent probabilistic monitoring data collected in 2001-2004 to validate the VSCI. VSCI contains eight core metrics including HBI (Hilsenhoff Biotic Index, a family biotic index). Some of them decrease with the

stress (such as Total Taxa), some increase (such as HBI). The eight component metrics are commonly treated as stressors in environmental studies while the composite variable (VSCI) is viewed as a direct impairment indicator. Due to the availability of more information of this dataset, regression is thus fit between VSCI and HBI to evaluate water quality at a specific site. In This dataset also includes season, ecoregion, and flag of reference sites. After the removal of missing measurements, the dataset consists of 214 sites that are visited during the spring. The final dataset consists of 202 sites after the obvious outliers are deleted. Figure 5.4 displays the final dataset with 47 reference sites and 155 non-reference sites. The dashed line represents the lower standard for VSCI and HBI.

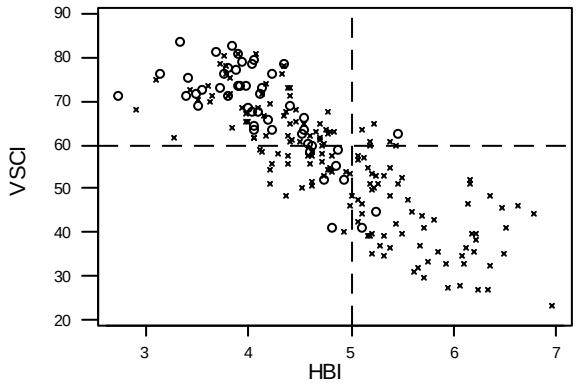

Figure 5.4 Dataset from non-coastal Virginia used for regression-based tests. This dataset was collected in 2001-2004 across three ecoregions. It consists of 47 reference sites and 155 non-reference sites. The cross and circle indicate non-reference and reference sites, respectively. HBI is the stressor and VSCI is the response variable.

Large VSCI values and low HBI values indicate good water quality. The upper criterion for HBI is thus used for defining the acceptance region. Table 5.7 adapts the methods in Table 5.1 for this case study. U in the abbreviations refers to using the upper criterion as x_0. The upper criterion of HBI is set at 5.7 (VDEQ, 2006) and the mean HBI at the reference sites is 4.11. When a site has the HBI value greater than 5.7 and VSCI less than the boundary value of the rejection region (defined in Table 5.7), or the site has

an HBI value greater than 4.11 and VSCI less than the boundary value of the rejection region, the site is classified as impaired. Obviously, the rejection region based on the upper criterion of HBI is expected to claim more sites as impaired than that based on the mean of reference sites due to their relative magnitude. Before using the regression-based test for assessment, a Levene's test for homogeneity of variance is carried out. Though this test suggests the assumption of homogeneous variance structure for the reference and non-reference sites be invalid, the further model specification test (White, 1980) shows that the model is correctly specified, the errors are independent of the regressor, and there is no evidence of heterogeneity. Thus, the assumption of the homogeneous variance structure is kept in the regression-based tests.

Table 5.7 Four boundaries of the rejection region for VDEQ data

X-axis	Y-axis	Label
$x_0 = x_U$ (=5.7)	$\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0}}) \sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 \hat{\sigma}^2}$	UTL
	$\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, 1-p_0) \sqrt{(\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 + \frac{1}{n}) \hat{\sigma}^2}$	UPL
$x_0 = \overline{x}_{ref}$ (=4.11)	$\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0}}) \sqrt{\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 \hat{\sigma}^2}$	RTL
	$\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, 1-p_0) \sqrt{(\underline{x}_0^T (X^T X)^{-1} \underline{x}_0 + \frac{1}{n}) \hat{\sigma}^2}$	RPL

Table 5.8 represents the regression-based test results. Around 50~80% of the non-reference sites are declared as impaired in the prediction limit approach, while the proportion is approximately 10% in the tolerance limit approach. It indicates that regression-based tests using the prediction limit tend to claim impairment. On the other hand, the tolerance limit claims all reference sites as unimpaired. The misclassification rate of non-reference sites in this case study is very high.

Table 5.8 Impairment claims based on four boundaries*

	UTL	UPL	RTL	RPL	Actual
Non-reference (high is good)	0.09(14)*	0.57(88)	0.14(21)	0.75(116)	155
Reference (low is good)	0(0)	0.11(5)	0(0)	0.28(13)	47

* Results in the table are the proportion of sites that are claimed to be impaired for different types of sites. The numbers in parentheses are the corresponding numbers of sites declared as impaired.

To further examine the data, the test based on the fixed effects model is also carried out for each ecoregion. This dataset consists of 3 ecoregions: coast, mountain and Piedmont. In Table 5.9, the number of total sites in each ecoregion is indicated by N and the number of non-reference sites in each ecoregion is indicated by m. A Levene's test for homogeneity of variance indicates that the three ecoregions can be assumed to have the same variance. With this assumption, the TL[11] defined in Section 3.2.5 can be applied to detect impairment for each ecoregion. The lower standard of VSCI is set at 60 (VDEQ, 2006). If the calculated tolerance limit in TL[11] is less than 60, the test ecoregion is declared as impaired. When the ecoregion mean is the least square mean or the adjusted mean by the covariate HBI, the TL[11] approach always declares regional impairment for each ecoregion. When the underlying model for TL[11] has VSCI as the response variable and ecoregion and HBI as the explanatory variables, the TL[11] test has better performance than the regression-based tests for site-specific tests (Table 5.9). This difference lies on the general test behaviors for regression-based tests. Regression-based tests have a broad acceptance region, thus tend to claim unimpairment. In addition, when the proportion of non-reference sites is high (in this VDEQ dataset, 77% sites are non-reference sites), the regression-based tests tend to claim unimpairment (refer to Table 5.4). Furthermore, the potential heterogeneity (though homogeneity is assumed for this dataset) can increase misclassification rate (refer to Section 5.3.2).

As a summary, for this dataset from the non-coastal Virginia, the regression-based tests using two-direction boundaries have high misclassification of non-reference sites. The tests using tolerance limits tend to claim unimpairment while the tests using prediction limits tend to claim impairment. Adjustment for covariate (HBI) improves the performance of tests based on the fixed effects model.

Table 5.9 Impairment claims for ecoregions

Ecoregion	N	m*	UTL	UPL	RTL	RPL	TL[11]
Coast	28	28	9	24	12	26	26
Mountain	77	48	0	16	0	27	47
Piedmont	97	79	5	48	9	63	78
Total	202	155	14	88	21	116	151

*Number of non-reference sites

5.4 Covariate adjustment and reference-based tolerance limits

When simple regression is applied in impairment detection, the estimated tolerance limit is adjusted by the relationship between the biological response and the stressor. The preceding sections focus on the case in which the regressor is a stressor and the impairment decision is made using the limits of the biological variable and the stressor. This decision rule tends to claim unimpairment (using the prediction limit has a tendency to claim impairment compared with using the tolerance limit) due to the broadly defined acceptance region. Adjustment for potential covariates in tests using the ANOVA model may improve test performance as demonstrated by the application in Section 5.3.3. This leads to discussions on alternative use of regression-based tests.

In this section, covariate-adjusted tests are set and evaluated by comparing the covariate-adjusted tolerance limit for the biological variable with the corresponding ordinary tolerance limit in terms of the comparable power. In this comparison, the data consists of one biological response and one covariate (e.g., elevation). Here, the boundary of the rejection region doesn't consider the contribution from the covariate.

With the regression model set in Section 5.2.1, when the impairment is indicated by low values of the biological variable, the lower tolerance limit of the biological variable

at a site with the covariate value of x_0 is $\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{h}})\sqrt{h\hat{\sigma}^2}$, where $\underline{\hat{\beta}}$ and $\hat{\sigma}^2$ are defined in preceding sections and $h = \underline{x}_0^T (X^T X)^{-1} \underline{x}_0$.

The distribution of the response variable conditioned on the covariate is $Y \mid X = x_0 \sim N(\mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x_0 - \mu_X), \frac{\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2})$. The marginal distribution of the response variable is $Y \sim N(\mu_Y, \sigma_Y^2)$. The regression estimator, $\hat{\sigma}^2$, estimates $\frac{\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2}$, i.e., $(1-\eta^2)\sigma_Y^2$ where $\eta^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}$, instead of σ_Y^2. The ordinary limit is $\bar{y} - t(n-1, \alpha, z_{1-p_0}\sqrt{n})s$, where s estimates σ_Y. The variance difference brings the issue of deriving comparable power. The power for the covariate-adjusted test is derived as follow.

$$
\begin{aligned}
power &= \Pr(\underline{x}_0^T \underline{\hat{\beta}} - t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{h}})\sqrt{h\hat{\sigma}^2} < L) \\
&= \Pr(\frac{\underline{x}_0^T \underline{\hat{\beta}} - \mu_1 + \mu_1 - L}{\sqrt{h\hat{\sigma}^2}} < t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{h}})) \\
&= \Pr(\frac{\frac{\underline{x}_0^T \underline{\hat{\beta}} - \mu_1}{\sqrt{h\sigma^2}} + \frac{\mu_1 - L}{\sqrt{h\sigma^2}}}{\hat{\sigma}/\sigma} < t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{h}})) \\
&= \Pr(T(n-2, \lambda) < t(n-2, \alpha, \frac{z_{1-p_0}}{\sqrt{h}}))
\end{aligned}
$$

where $\lambda = \frac{\mu_1 - L}{\sqrt{h\sigma^2}}$, μ_1 is the true site response. With the set-up, $\lambda = z_{1-p_1}\sqrt{n}$, the power for the ordinary test is calculated as $power = \Pr(T(n-1, \lambda) < t(n-1, \alpha, z_{1-p_0}\sqrt{n}))$.

The comparison is valid only when the parameters indicate the same scenario for both tests. With the ordinary test as the comparison base, $\frac{\mu_0 - L}{\sqrt{\sigma^2}} = \frac{\mu_0 - L}{\sqrt{(1-\eta^2)\sigma_Y^2}}$ and $\frac{\mu_1 - L}{\sqrt{\sigma^2}} = \frac{\mu_1 - L}{\sqrt{(1-\eta^2)\sigma_Y^2}}$. The baseline and true exceedance proportion in the covariate-

adjusted test should be adjusted for the variance difference. The power for the covariate-adjusted test thus becomes $power = \Pr(T(n-2,\lambda) < t(n-2,\alpha,\frac{z_{1-p_0}}{\sqrt{h(1-\eta^2)}}))$, where $\lambda = \frac{\mu_1 - L}{\sqrt{h\sigma^2}} = \frac{\mu_1 - L}{\sqrt{h(1-\eta^2)\sigma_Y^2}} = \frac{z_{1-p_1}}{\sqrt{h(1-\eta^2)}}$. The correlation between X and Y, η is assumed known here.

The power depends on the true exceedance proportion, sample size, the correlation, and the covariate value. To illustrate the covariate effect on tests, the covariate is standardized here to be distributed on the scale of [-1,1]. Two scenarios are considered for covariate values. In the first scenario, the covariate only takes two values, -1 and 1. The sample size takes the value of 9, 18, and 27. The other scenario is that the covariate value is equally spaced on [-1, 0) and (0, 1]. The candidate sample size is 10, 20, and 30. In both scenarios, the exceedance proportion is set as 0.1, 0.2, 0.3, or 0.4; the correlation between the response and the covariate is selected from 0, 0.2, 0.4, 0.6, and 0.8.

Table 5.10a to Table 5.10c list the comparison results when the covariate takes three values and the correlation is 0, 0.6, or 0.8. When there is no correlation between the response and the covariate, the test performance is almost identical for the covariate-adjusted and ordinary tests. When the correlation becomes strong, the covariate-adjusted test produces larger power at the mean of the covariate than the ordinary test. This difference is more evident as the sample size increases and the impairment becomes more severe.

Figure 5.5 depicts the comparison in terms of deviation of the covariate value from its mean when there are 10 sites in the region and the true exceedance proportion is 0.3. The covariate-adjusted test at the mean of the covariate is more powerful than that at the end of scales.

Table 5.10a Power for regression-based tests when covariate takes three values and the correlation is 0

True exceedance proportion	Sample size	Power at $x_0 = 0$	Power at $x_0 = 1$ or -1	Power for ordinary test
0.1	9	0.05	0.05	0.05
	18	0.05	0.05	0.05
	27	0.05	0.05	0.05
0.2	9	0.3107	0.2039	0.3166
	18	0.4713	0.3023	0.4763
	27	0.5980	0.3886	0.6021
0.3	9	0.6349	0.4058	0.6449
	18	0.8590	0.6163	0.8636
	27	0.9486	0.7584	0.9505
0.4	9	0.8570	0.6047	0.8647
	18	0.9806	0.8394	0.9818
	27	0.9976	0.9386	0.9978

Table 5.10b Power for regression-based tests when covariate takes three values and the correlation is 0.6

True exceedance proportion	Sample size	Power at $x_0 = 0$	Power at $x_0 = 1$ or -1	Power for ordinary test
0.1	10	0.05	0.05	0.05
	20	0.05	0.05	0.05
	30	0.05	0.05	0.05
0.2	10	0.3642	0.2532	0.3166
	20	0.5523	0.3814	0.4763
	30	0.6890	0.4896	0.6021
0.3	10	0.7291	0.5178	0.6449
	20	0.9257	0.7490	0.8636
	30	0.9810	0.8745	0.9505
0.4	10	0.9256	0.7428	0.8647
	20	0.9952	0.9339	0.9818
	30	0.9997	0.9844	0.9978

Table 5.10c Power for regression-based tests when covariate takes three values and the correlation is 0.8

True exceedance proportion	Sample size	Power at $x_0 = 0$	Power at $x_0 = 1$ or -1	Power for ordinary test
0.1	10	0.05	0.05	0.05
	20	0.05	0.05	0.05
	30	0.05	0.05	0.05
0.2	10	0.4225	0.3237	0.3166
	20	0.6377	0.4912	0.4763
	30	0.7771	0.6210	0.6021
0.3	10	0.8160	0.6590	0.6449
	20	0.9688	0.8781	0.8636
	30	0.9951	0.9590	0.9505
0.4	10	0.9698	0.8766	0.8647
	20	0.9993	0.9858	0.9818
	30	0.9999	0.9985	0.9978

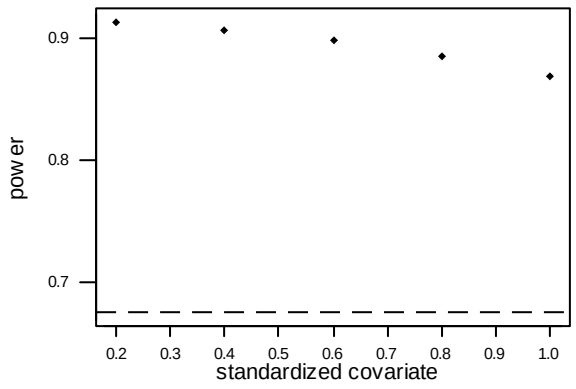

Figure 5.5 Power for covariate-adjusted tests when covariates are equally spaced. The region consists of 10 sites and the true exceedance proportion is 0.3. The dash line represents the power for the ordinary test.

5.5 Summary

Current application of regression methods in environmental studies focuses on modeling the relationship or adjusting biological response variables for covariate(s). There is no formal discussion for impairment assessment based on regression. When impairment is associated with a single stressor it is possible to make an impairment decision based on a

stressor-response relationship. This allows for an approach that uses all of the biological data in the decision process rather than just the data from the reference conditions. There are several ways to implement a test for site impairment, using either a prediction limit or a tolerance limit. Because the prediction interval is narrower than the tolerance limit, the tests based on prediction limits tend to reject more than the tests based on tolerance limits. The regression-based limits must be derived at a specified value of $X = x_0$. Choice of x_0 influences the decision process. Possible values include the criterion for X (x_L) or the mean of X for reference sites. Additionally, a tolerance limit on X using reference site data might be used given sufficient sample size.

When the regression-based tests are applied to real datasets, the misclassification rate is high due to the definition of a broad acceptance region, the proportion of non-reference sites, and the potential heterogeneous variance. Regression-based tests have a broad acceptance region, thus tend to claim unimpairment. In addition, when the proportion of non-reference sites is high, the regression-based tests tend to claim unimpairment. Potential heterogeneity might reduce the power for tests. With these considerations, regression-based tests should be applied with caution. The model assumptions should be examined carefully before using regression-based tests for assessment.

Regression is also useful for adjusting a reference-based limit when the biological condition varies with a covariate. Results indicate that the covariate can improve the decision process when the relationship between the response and the covariate becomes strong.

6. Assessment Using Model-based Clustering

6.1 Introduction

Biological monitoring, or biomonitoring, is the use of biological responses to assess changes in the environment. This type of monitoring programs involves the use of indicators, indicator species, or indicator communities (Phillips and Rainbow, 1993). Generally benthic macroinvertebrates, fish, and/or algae are used as indicators. The use of benthic macroinvertebrates (such as aquatic insects and worms) in the biological assessment of water quality has been a vital and rapidly growing field in the past two decades (Rosenberg and Resh, 2003). There are a large number of species in the family of benthic macroinvertebrates and different stresses produce different macroinvertebrate communities. The data collected in biomonitoring programs are usually large and complicated. Numerous multivariate statistical approaches have been implemented to this complex data to detect biological integrity (Green 1974; ter Braak, 1986; Yuan and Norton, 2003). Most methods involve dimension reduction, which leads to reduced-rank regression in the context of regression.

A further complexity of biomonitoring programs is that ecosystems vary in their biological and environmental features at a wide range of spatial and temporal scales. In many biomonitoring programs, environmental data are collected over large spatial regions. To distinguish ecosystem patterns in different regions, cluster analysis methods are needed for clustering sites in terms of geomorphologic similarity. Model-based clustering is becoming a topic of concerns due to its better classification performance. A few recent applications of model-based clustering include the Bayesian treed model for spatial count data (Denison and Holmes, 2001), ecological regression with spatial partition (Greco et al., 2005), and stressor-response relationship detection with Voronoi tessellations (Bates Prins et al., 2006). Model-based clustering can be divided into three categories: the likelihood method, tree classification, and the randomization approach (Banfield and Raftery, 1993; McLachlan et al., 2002; Okabe et al., 2000).

Using benthic data from the Mid-Atlantic Highlands Assessment (MAHA) with continuous responses to indicate the condition of streams under study, model-based

clustering and tests are developed to predict water quality. The following section will introduce the general technique using redundancy analysis (RDA) with Voronoi tessellations to group data and assess impairment. Section 6.3 will describe the application of this method to MAHA data. Finally, a summary of model-based tests in the multivariate case will be given as well as some comments for future research.

6.2 Methods

In this chapter, random tessellations and redundancy analysis are combined to describe and potentially assess water quality for large ecological datasets. The random tessellation approach is used to generate clusters of sites that may have a similar stressor-response relationship. The within-cluster relationship comes from RDA results. This approach aims at finding the optimal partition or cluster which has the best ability of explaining the relationship. The general procedure starts with determining the number of clusters. Then random seeds (partitions) are generated and Voronoi tessellations (cluster) are formed. RDA is fit to observations within each cluster. The summary statistic (i.e., optimality criterion) is computed for each random partition. The final solution is selected as the partition associated with the optimal value of the summary statistic. Once the optimal clustering result is determined (i.e., the regression model is developed), multivariate tests for predicting the water quality at a new site can be carried out (using the developed regression from the clustering) as described in any standard textbook associated with multivariate regression (e.g., Jobson, 1991). Graphical displays of clustering and test results can help direct interpretation of response-stressor relationships.

6.2.1 Redundancy analysis

Redundancy analysis (RDA) is a variation of the reduced-rank regression (RRR). Based on reducing the dimension of the regression coefficients, the basic form of a reduced-rank regression is expressed as

$$Y = XM + E = (XA)B + E$$

where Y ($n \times m$) and X ($n \times p$) are matrices containing observations on n objects for m responses and p explanatory variables (predictors). M is the $p \times m$ coefficient matrix. A and B are of full rank with rank r ($r \leq \min(m, p)$). E is the error matrix and

$Vec(E) \sim N(\underline{0}, I_n \otimes \Sigma)$, where Σ is the variance-covariance matrix for each row in *E*. RRR aims at minimizing the sum of squared residuals subject to a reduced rank condition. This model implies that *r* linear combinations of explanatory variables are sufficient to model the variation in the response variables. The (*min(p,q)-r*) linear restrictions on *M* are usually unknown and need to be estimated. The estimation is carried out to minimize the norm of $E^T E$ (i.e. $\min_{A,B} \left\| E^T E \right\|$), or minimize the norm of scaled $E^T E$. The common estimation methods include generalized methods of moments, maximum likelihood, and least squares methods.

The analysis of a RRR model is based on the assumption that the coefficient matrix *M* is not of full rank. The elements of *M* are subsequently estimated for a given rank of *M*. The rank of *M* is the number of positive canonical correlations between *X* and *Y*. It should be identified before making any inference from the model. Identifying the rank is equivalent to testing if certain correlations are zero. Frequentist and Bayesian approaches are both available for the rank identification (Anderson, 2002; Geweke, 1996; Corander and Villani, 2001).

When the estimation of RRR is based on minimizing the norm of $E^T E$, the solution to RRR is equivalent to redundancy analysis. The corresponding defining function is

$$(Y - \tilde{Y})^T (\tilde{Y} - XM) = 0,$$

where $\tilde{Y}$ is the fitted value with the assumption that *M* is of full rank. The orthogonal decomposition is $\left\| Y - XM \right\|^2 = \left\| Y - \tilde{Y} \right\|^2 + \left\| \tilde{Y} - XM^* \right\|^2$ for certain M^*. The constraint rank *r* is typically the minimum number of variables of either *Y* or *X*. RDA uses the regression of the standardized *Y* values on the standardized *X* values.

Clustering groups a dataset into subsets (clusters) so that the data in each subset (ideally) share some common properties. In the model-based clustering using RRR, the model for the i^{th} observation in the k^{th} cluster is

$$\underline{y}^{*k}_i = (B^k)^T \underline{x}^{*k}_i + \underline{e}^{*k}_i, \; i=1, 2, ..n$$

where $\underline{y}_i^{*k}$ and $\underline{x}_i^{*k}$ are standardized. B^k is the coefficient matrix. The sign *T* stands for the transpose operator. $\underline{e}_i^{*k}$ is a vector of errors. The fitted values are computed by ordinary least squares, i.e.,

$$\hat{Y} = X^*(X^{*T}X^*)^{-1}X^{*T}Y^*$$

RDA is thus obtained by decomposing the fitted value $\hat{Y}$ using the singular value decomposition (SVD) method. The following eigenvalue equation is solved to compute the canonical variates.

$$\left|S_{\hat{Y}} - \lambda_r I\right| = 0$$

where $S_{\hat{Y}}$ is the covariance matrix of the fitted values $\hat{Y}$. This eigenvalue equation gives the canonical eigenvalues (λ's). The number of nonzero canonical eigenvalues (*r*) is equal to the minimum number of variables of either *Y* or *X*. The sum of the canonical eigenvalues produced by an RDA equals to the amount of variation in *Y* explained by *X*.

6.2.2 Random tessellations

To randomly split data in space, a randomization approach using a two-dimension Voronoi diagram is applied. The Voronoi diagram is sometimes called a Voronoi tessellation. It is a division of space into polyhedral regions of points closest to a fixed set of reference points. Usually the center of a region is set as the reference point (Miller, 1994).

Suppose the dataset consists of *n* observations of *m* responses and *p* potential predictors, i.e., the region (the dataset comes from) consists of *n* sites (Bates Prins et al., 2006). Of interest are *p* stressors and *m* responses in this region. The relationship between *p* stressors (e.g., RH_BKVG, representing bank protective vegetation) and *m* responses (e.g., the Hilsenhoff Biotic index) may represent the structure of a biological community in some way. The two-dimensional Voronoi diagram is created using grid variables, G_1 and G_2. Each observation has a pair of values (g_{1_i}, g_{2_i}) for grid variables. In my study, latitude and longitude are the grid variables, which are not in the set of stressors. For a given number of clusters, *K*, a Voronoi diagram divides the region into polygonal regions. Within each region the response-stressor relationship is modeled. When this

random splitting is repeated enough times (much larger than *n*), the optimal partition of the region can be found. "The optimal partition is declared when the 'averaged' relationship is maximized or a single cluster has the most significant response-stressor relationship" (i.e., the hotspot is selected) (Bates Prins et al., 2006). In the remainder, the term 'cluster' is used to represent cell.

6.2.3 Optimality criterion

The hot-spot optimality criterion is set to detect a single significant response-stressor relationship. This criterion focuses on finding a single cluster which has the most significant response-stressor relationship. Within each partition using the polygonal region, this criterion picks a cluster with the strongest relationship. The selected cluster will represent this partition. The final optimality conclusion is made by finding the maximum over all partitions. The final selected cluster is called the hot-spot region. This criterion is defined by an R-squared like quantity as below.

$$r^2 = \left(\sum_{i=1}^{\min(m,p)} \lambda_i\right) / \left(\sum_{j=1}^{m} \lambda_j\right)$$

where the λ_i's are the constrained eigenvalues (i.e., the eigenvalue of the covariance matrix corresponding to the fitted values, $\hat{Y}$) and the λ_j's are the eigenvalues of the covariance matrix corresponding to the response variables, *Y*.

Different from the hot-spot criterion (i.e., the local optimum criterion), the global optimum criterion tries to achieve the optimality by maximizing the 'averaged' relationship for a partition. For each partition, the 'averaged' statistic (for instance, the averaged R-squared value) is calculated. The maximum is defined over all the partitions. A Bayesian Information like criterion (BIC) can be used for cluster-wise optimization as follows.

$$BIC \approx \sum_{k=1}^{K} \left[n_k \log\left(n_k^{-1} tr(\hat{\Sigma}_k) + \sum_{i=r+1}^{p} \lambda_{ik}^2 \right) - m \log n_k \right]$$

where *p* is the dimension of stressors, *m* is the dimension of response data, n_k is the size of the k^{th} cluster, $\hat{\Sigma}_k$ is the sample covariance matrix for the k^{th} cluster, λ_{ik} is the

eigenvalue associated with the RDA of the response data in the k^{th} cluster, and r indicates the number of nonzero canonical eigenvalues.

6.2.4 Adjustment for reference information

Adjustment for reference conditions can provide an initial sense of deviation from reference conditions thus highlights the separation of reference and non-reference sites. Here, two adjustment methods are implemented: after-adjustment and midway adjustment. With the after-adjustment, sites in the region are standardized by reference information within the final optimal cluster obtained from the random tessellation and RDA. The midway adjustment adjusts sites by reference information within each cluster in every tessellation and gets the final optimal result. The midway adjustment is expected to pick up more information about the relationship.

6.2.5 Impairment prediction

When the full rank regression is written as $Y = XM + E$, where M is the coefficient matrix with full rank (rank=*min(m, p)*). The 100(1-α)% simultaneous prediction intervals at $\underline{x}_0$ ($p \times 1$) are

$$\underline{x}_0^T \underline{\hat{M}}_{(i)} \pm \sqrt{\frac{m(n-p)}{n-p-m+1} F(m, n-p-m+1, 1-\alpha)} \sqrt{1 + \frac{n}{n-p} \underline{x}_0^T (X^T X)^{-1} \underline{x}_0 \hat{\sigma}_{ii}} \ .$$

i=1, 2, …, m.

Here, $\underline{\hat{M}}_{(i)}$ is the i^{th} column of $\hat{M}$ ($\hat{M} = (X^T X)^{-1} X^T Y$). $\hat{\sigma}_{ii}$ is the i^{th} diagonal element of $\hat{\Sigma}_{ee} = (Y - X\hat{M})^T (Y - X\hat{M}) / (n-p)$ (the estimate of covariance matrix of error) (Anderson, 1958).

With the reduced rank model set-up, $Y = (XA)B + E = ZB + E$, the 100(1-α)% simultaneous prediction intervals at $\underline{x}_0$ are derived as

$$\underline{x}_0^T \hat{A} \underline{\hat{B}}_{(i)} \pm \sqrt{\frac{m(n-r)}{n-r-m+1} F(m, n-r-m+1, 1-\alpha)} \sqrt{1 + \frac{n}{n-r} \underline{x}_0^T \hat{A} (\hat{A}^T X^T X \hat{A})^{-1} \hat{A}^T \underline{x}_0 \hat{\sigma}_{ii}}$$

where $\underline{\hat{B}}_{(i)}$ is the i^{th} column of $\hat{B}$. $\hat{\sigma}_{ii}$ is the i^{th} diagonal element of $\hat{\Sigma}_{ee} = (Y - X\hat{A}\hat{B})^T (Y - X\hat{A}\hat{B}) / (n-r)$. $\hat{A}$ and $\hat{B}$ are the MLEs of A and B, respectively.

Determination of $\underline{x}_0$ is important and difficult. Here, $\underline{x}_0$ is set as the mean vector of the reference sites in the optimal cluster or in each cluster form the optimal partition. If all the measurements of the response variables at one site fall in the respective prediction region, the test site is declared as unimpaired. Otherwise, the test site is declared as impaired. In the multivariate case, the boundary of the acceptance region is not set by stressors and responses simultaneously due to the fact that the direction of the impact of the stressors may vary. In Section 6.3.2.5, a probability ellipse is given to describe the reference set using RDA.

6.3 Application in MAHA benthic data

Details of the technique of predicting impairment with model-based clustering will be discussed in this section as this technique is implemented for the Environmental Monitoring and Assessment Program (EMAP) data collected for the Mid-Atlantic Highlands Assessment (MAHA) from 1993 to 1996. As a highly heterogeneous region, the MAHA area spans 6 states and 13 different ecoregions (Bryce et al. 1999). Streams in this region run through forests, wetlands, residential areas and agricultural areas (USEPA, 2006). Environmental management is needed to maintain the health of plants and animals that inhabit these streams. The assessment of current ecological resources can provide more knowledge of current ecosystem status and provide a basis for future actions. Since benthic invertebrate taxa provide a good spatial signal of what has occurred at the site (Bailey, 1994), the benthic data will be the focus of this analysis.

6.3.1 MAHA Data

From 1993 to 1996 benthic assemblage data were collected from wadeable streams in the Mid-Atlantic Highlands region as part of the USEPA's Environmental Monitoring and Assessment Programs. Stream sites were selected with a probabilistic sampling design that allowed regional estimates of stream condition (Overton et al., 1991; Herlihy et al., 2000). This dataset contains physical, chemical, and landscape variables measured at sampled stream segments.

The original benthic data are reduced to include only the non-hand-picked samples and the samples from riffle habitat units. In addition, only the first visit to each site is

kept. This produces a standard benthic dataset of 772 observations. Furthermore, to study the stressor-response relationship, I restrict my interest to the first-order through third order streams. Nine predictors related to watershed land use characteristics are selected based on expert opinion. Eleven response variables are picked for describing richness, balance, tolerance, and trophic properties of benthic species. Any missing value is removed from the dataset. These manipulations yield a dataset of 349 observations (i.e., 349 sites), 9 predictor variables, and 11 response variables (Table 6.1, Table 6.2, and Figure 6.1). The label for each variable is given in Table 6.2. For extremely skewed variables, a transformation is applied (McCormick et al., 2001; Yuan and Norton, 2003). The log transformation is applied to SIMPSON and HBI; SHRDPIND and SCRPPIND are transformed by $\log(Y+1)$. Square-root transformations are suggested for the percentage variables, URBAN, RD_DEN, AG, FOREST, and MINE.

Criteria proposed by Waite et al. (2000) are applied to identify the least-disturbed reference sites. Sites are designated as reference if they satisfy all the following criteria: Gran acid neutralizing capacity (ANC) ≥ 50ueq/L, Chloride (CL, indicating general watershed disturbance) < 100 ueq/L (3.5 mg/L), Sulfate (SO4, indicating acid mine drainage) ≤ 400ueq/L (19.2 mg/L), total Nitrogen (NTL) < 750 ug/L and total Phosphorus (PTL) < 20 ug/L (NTL and PTL indicating nutrient enrichment), and the Rapid Bioassessment Protocols (RBP, an habitat index) > 180 (out of a possible 240). Each criterion is designed to screen out sites that are impaired by different stressors. Applying these criteria yields a total of 62 reference sites in the dataset (solid circles in Figure 6.1).

Table 6.1 Variables used in model-based clustering

Type of variables	List of variables
Grid variables	Latitude, Longitude
Predictor variables (stressor)	URBAN, RD_DEN,POPDENKM,AG,FOREST,MINE, RH_GRAZ,RH_RIPVG,RH_BKVG
Response variables	Richness: TOTLRICH,EPT_PTAX Trophic: SHRDPIND,SCRPPIND Balance: HPRIME,SIMPSON Tolerance:INTLRICH,INTLPIND,PLECPIND,HBI,FACLPIND

Table 6.2 Labels for variables used in model-based clustering

Variable	Label
URBAN	% Watershed of Urbanization
RD_DEN	Road Density (m/ha)
POPDENKM	Population Density (persons/sq km)
AG	% Watershed of Agriculture
FOREST	% Watershed of Forest
MINE	% Watershed of Mining Activity
RH_GRAZ	Lack of Vegetative Grazing Disturbance score
RH_RIPVG	Width of Riparian Vegetation Zone Score
RH_BKVG	Bank protective vegetation score
TOTLRICH	Total Distinct Taxa Richness
EPT_PTAX	EPT % Distinct Taxa
SHRDPIND	Shredder % Individuals
SCRPPIND	Scraper % Individuals
HPRIME	Shannon Diversity
SIMPSON	Simpson Index
INTLRICH	Intolerant Distinct Taxa Richness
INTLPIND	Intolerant % Individuals
PLECPIND	Plecoptera % Individuals
HBI	Hilsenhoff Biotic Index
FACLPIND	Facultative % Individuals

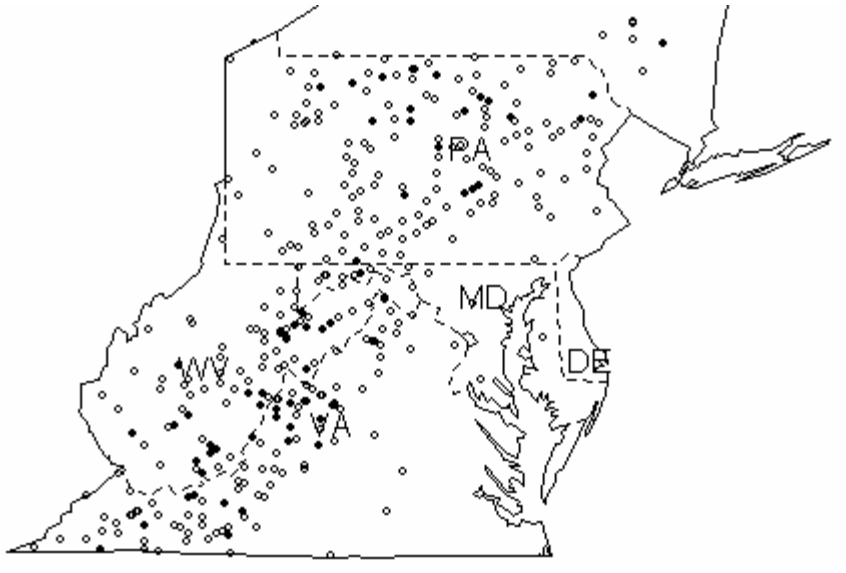

Figure 6.1 Location of 349 stream sites in the Mid-Atlantic Highlands. Solid circles indicate reference sites.

6.3.2 Methods and results

The procedure of model-based clustering and tests consists of selecting groups of sites, performing an RDA, determining the optimal cluster, and predicting site impairment. Details are discussed as follows.

6.3.2.1 Finding the optimal cluster

The choice of the number of clusters, *k*, depends on many factors, such as the response-stressor model used in the clustering (Bates Prins et al., 2006). A reasonable range for *k* can be determined *a priori* with expert opinion. By consulting biologists and aquatic entomologists, my study sets 70 as the minimum number of observations per cluster in a tessellation. With the total sample size (349 observations in the MAHA data), the maximum possible number of clusters is 4. The hotspot criterion (refer to Section 6.2.3) is used to find the optimal cluster. In this scenario, the partition of the region into an optimal set of clusters is not the focus. Rather the focus is to find the cluster with the maximum proportion of inertia. For MAHA benthic data, the size of the optimal cluster for partitions with different numbers of clusters is almost the same. I vary the number of clusters to more precisely geographically define the optimal cluster.

With the prespecified minimum cluster size, Figure 6.2 reveals that the optimum occurs when there are four clusters for the midway adjustment and after-adjustment analyses. The optimal inertia for the after-adjustment analysis actually indicates the inertia of the optimal cluster before adjusted by reference conditions. Note that the optimal inertia tends to increase with a larger number of clusters for this dataset. In practice, the association between the optimal inertia and the number of clusters may not be monotonic. This association depends on the response-stressor relationship and the optimum criterion (Bates Prins et al., 2006). More clusters don't necessarily lead to larger optimal values in model-based clustering. In this MAHA dataset, the optimal cluster picked by the after-adjustment and midway adjustment is similar. The optimal cluster consists of 70 and 72 sites for the after-adjustment and midway adjustment, respectively.

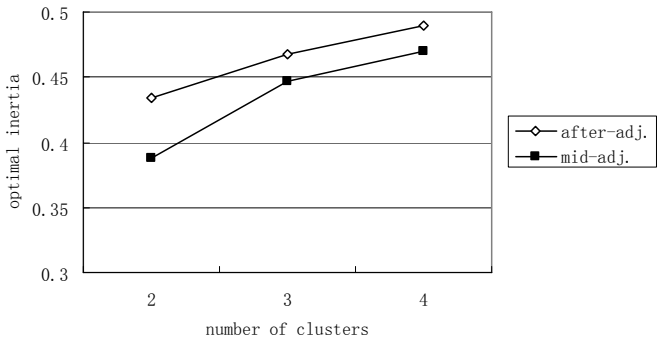

Figure 6.2 Clustering results with the after-adjustment and midway adjustment.

6.3.2.2 Improvement in detecting relationships with model-based clustering

Table 6.3 summarizes the R-Square quantities for different clustering as given. Results reveal that the maximum R-square increases from 0.1624 (from the single model for the whole dataset) to 0.4705 (for the model using four clusters). Figure 6.3 displays the geographical location (in terms of longitude and latitude) of the sites in the cluster with the maximum R-square. These sites are labeled as 2 in the figure. To evaluate individual metrics, a separate regression is fit to each metric. Results are summarized in Figure 6.4. Improvements occur for many metrics but especially for EPT_PTAX, INTLPIND, and

HBI. The adjusted R-squares indicate strong relationships between stressors and these response metrics.

Table 6.3 Summary of clustering results based on R-square criterion using mid-way adjustment

Number of clusters	Cluster	Number of observations	R-square like value	Max R-square
1	1	349	**0.1624**	0.1624
2	1	275	0.1859	0.3873
	2	74	**0.3873**	
3	1	174	0.2237	0.4464
	2	74	**0.4464**	
	3	101	0.2124	
4	1	102	0.2198	0.4705
	2	72	**0.4705**	
	3	83	0.2333	
	4	92	0.3324	

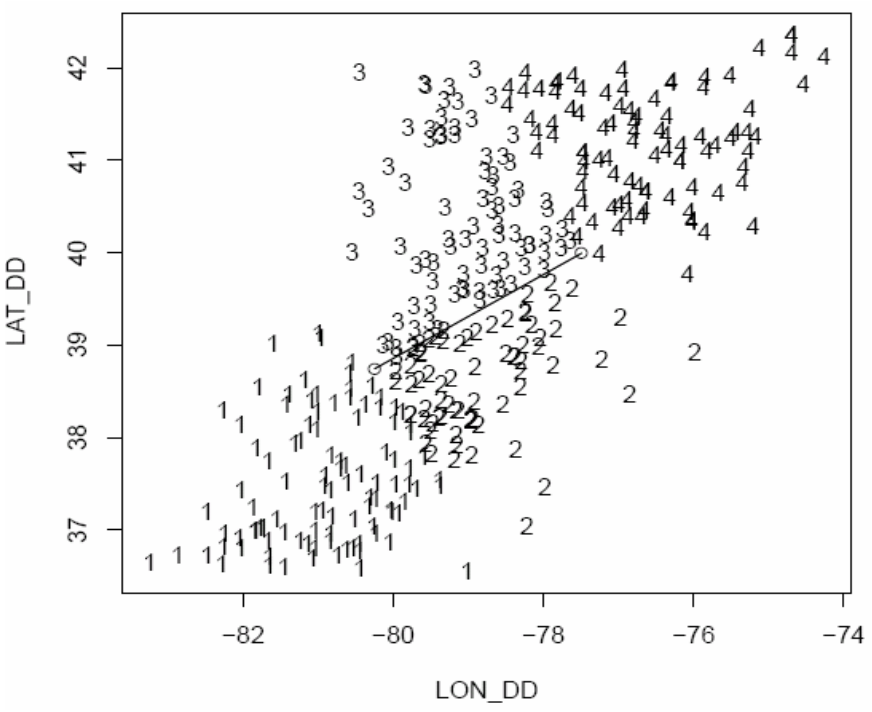

Figure 6.3 Geographical locations of sites in the cluster with maximum R-square. Label 2 indicates the sites in the optimal cluster.

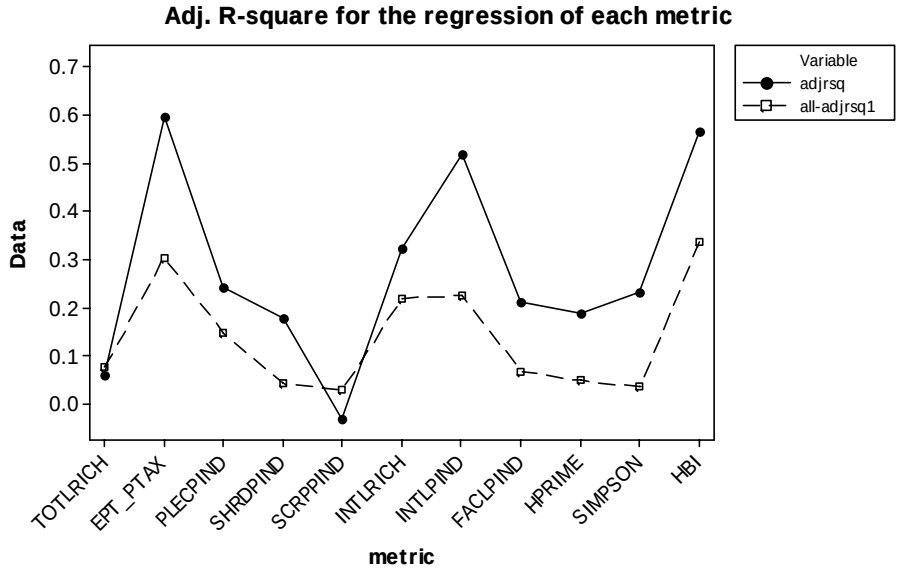

Figure 6.4 Comparison of adjusted R-square of the multiple regression for the optimal cluster and the whole dataset. The solid circle (solid line) and the square (dashed line) indicate the adjusted R-square for the optimal cluster and the whole dataset, respectively.

6.3.2.3 Comparison of the after-adjustment and midway adjustment

In the after-adjustment analysis, the optimal cluster is determined then sites are adjusted by reference information. The within-cluster relationship is shown in Figure 6.5. As a comparison, the RDA biplot from the midway adjustment analysis is displayed in Figure 6.6. This adjustment results indicate a slightly stronger relationship. The RDA output from the software package R indicates that the after-adjustment analysis only explain 37% of the variation in the responses by the relationship (versus 47% from midway adjustment).

Although the variance explained is quite different for the two analyses, the graphs indicate a high degree of similarity. Both display contrasting relationships between population variables (POPDENKM, RD_DEN, URBAN, AG) and habitat variables (FOREST, RH_GRAZ, RH_BKVG, RH_RIPVG). The midway adjustment method tends to balance the habitat variables while the after-adjustment method gives more weight to the FOREST variable.

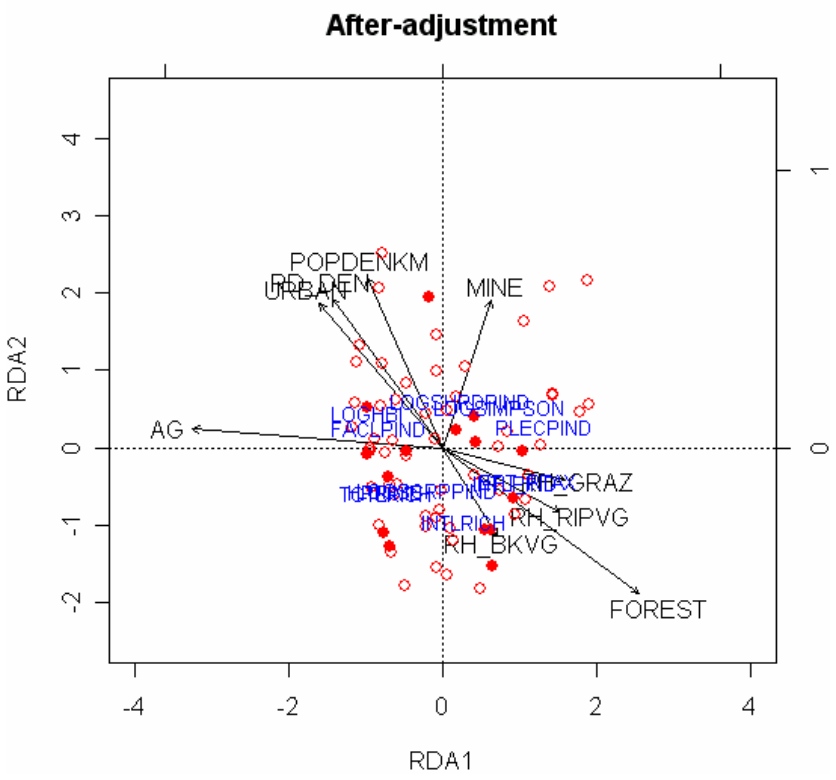

Figure 6.5 Biplot of RDA results for the optimal cluster with the after-adjustment. Solid circles indicate reference sites.

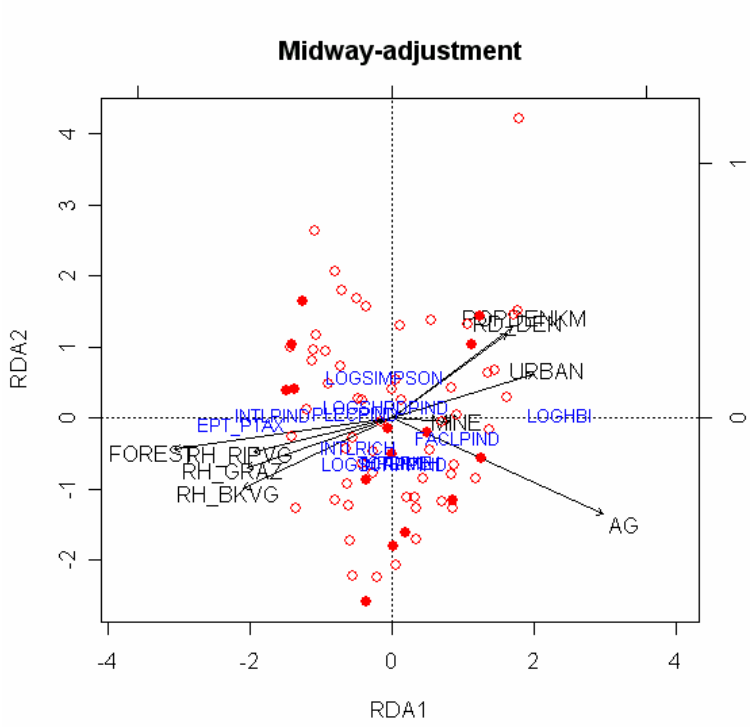

Figure 6.6 Biplot of RDA results for the optimal cluster from the midway adjustment. Solid circles indicate reference sites.

Figure 6.6 suggests that the variables may be combined to visualize relationships. Figure 6.7 displays the contrast of the three standardized dominant benthic variables (LOGHBI - EPT_PTAX - INTLRIND) and the contrast of dominant environmental variables (POPDENKM + RD_DEN + URBAN - FOREST - RH_GRAZ - RH_BKVG - RH_RIPVG). A linear regression is fitted after deleting 7 obvious outliers. The R-square is around 28%, which is smaller than that from model-based clustering. However, this consideration provides an alternative to depict the underlying relationship in a simpler and more interpretable way.

The separation of non-reference and reference sites is not distinct for the midway adjustment. It may be because the criteria for defining reference sites mostly involve chemical variables while the relationship for the clustering involves landscape and metric variables. Figure 6.8 gives the results based on the stressor variables which are used as reference criteria. Now some non-reference and reference sites cluster together. There is a

gradient of sites on the acidification variables (ANC, SO4, and CL) and the nutrient variables (PTL, NTL).

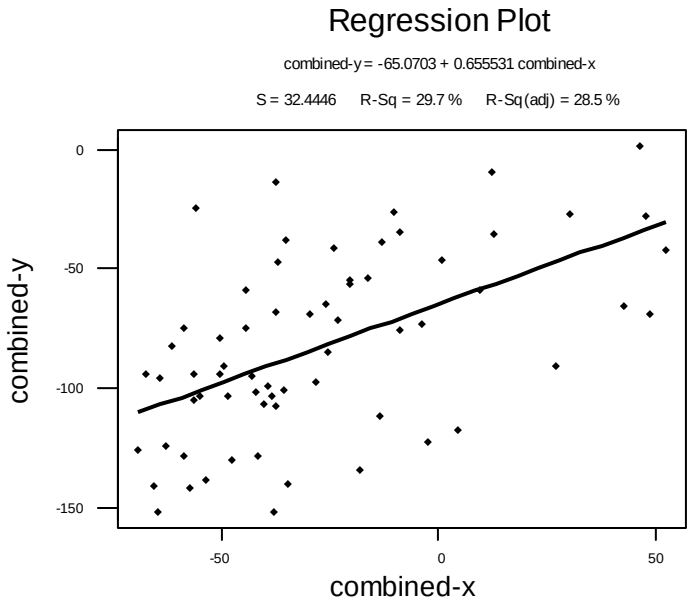

Figure 6.7 Linear regression between the combined response and stressor variables.

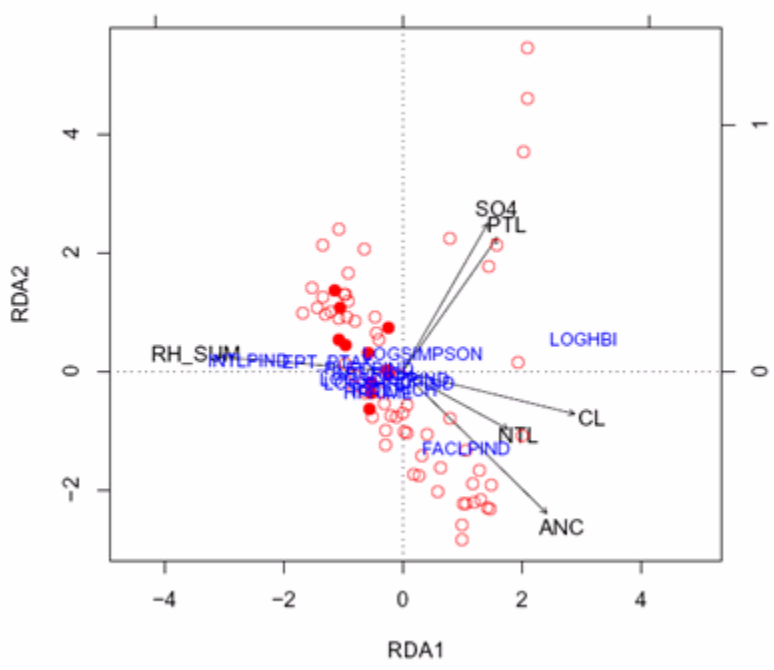

Figure 6.8 Biplot of benthic metrics and chemical variables for the optimal cluster. Solid circles indicate reference sites.

Table 6.4 lists the biplot scores for metrics and stressors for the optimal cluster, which reveals the difference between the two adjustment methods in detail. The values in bold and italic are the largest at each RDA axis. AG and FOREST are likely to have high correlation. This correlation essentially defines a 'new' variable that is the difference between AG and FOREST. This relationship is indicated in the midway adjustment but not in the after-adjustment analysis. The order of adjustment from reference sites affects the display of the response-stressor relationship. Making adjustment with reference information before multivariate analysis leads to an easier interpretation of relationship as expected. LOGHBI and EPT_PTAX dominate the horizontal axis (RDA1) in the RDA biplot. Table 6.5 gives the regression results from fitting individual responses on all stressors. Only a few stressors have significant relationship with LOGHBI or EPT_PTAX (at the significance level of 0.05).

Table 6.4 Biplot scores of RDA for the optimal cluster

	midway adjustment		after-adjustment	
Response	RDA1	RDA2	RDA1	RDA2
TOTLRICH	0.1090	-0.6197	-0.7179	-0.5819
EPT_PTAX	***-2.1240***	-0.1088	***1.0931***	-0.4535
PLECPIND	-0.5070	0.0641	***1.3077***	0.2699
INTLRICH	-0.4699	-0.4126	0.2658	***-0.9555***
INTLPIND	**-1.6738**	0.0612	0.9302	-0.4523
FACLPIND	0.9413	-0.2791	-0.8225	0.2583
HPRIME	0.1006	-0.6190	-0.7391	-0.5636
LOGSHRDPIND	-0.0784	0.1696	0.2375	0.6080
LOGSCRPPIND	-0.1010	-0.6279	-0.2038	-0.5531
LOGSIMPSON	-0.0863	0.5882	0.7374	0.5318
LOGHBI	***2.3586***	0.0437	-0.9828	0.4616
Stressor	RDA1	RDA2	RDA1	RDA2
RD_DEN	0.4516	0.3325	-0.3997	***0.5375***
POPDENKM	0.4744	0.3593	-0.2727	***0.6121***
RH_GRAZ	-0.5698	-0.1940	0.4409	-0.1170
RH_RIPVG	-0.5434	-0.1335	0.4214	-0.2315
RH_BKVG	-0.5837	-0.2764	0.1943	-0.3187
URBAN	0.5551	0.1734	-0.4440	***0.5191***
AG	***0.8316***	-0.3773	***-0.9014***	0.0667
FOREST	***-0.8570***	-0.1219	***0.7089***	***-0.5248***
MINE	0.2316	-0.0126	0.1717	***0.5287***

Table 6.5 Regression results of individual responses on all stressors

Response	Stressor	Estimate	P-value
LOGHBI	Intercept	0.809534	0.721576
(R^2=57.99%)*	RD_DEN	0.000665	0.973987
	POPDENKM	0.005870	***0.021404***
	RH_GRAZ	-0.053560	0.405088
	RH_RIPVG	0.033128	0.486923
	RH_BKVG	-0.143490	***0.012492***
	URBAN	0.077490	0.715506
	AG	0.372864	***0.004003***
	FOREST	0.116570	0.563733
	MINE	0.361888	0.729485
EPT_PTAX	Intercept	-2.091000	0.273037
(R^2=61.66%)	RD_DEN	-0.006580	0.699812
	POPDENKM	-0.005990	***0.005432***
	RH_GRAZ	0.012899	0.809985
	RH_RIPVG	-0.062850	0.117617
	RH_BKVG	0.155984	***0.001400***
	URBAN	0.204334	0.252618
	AG	-0.303620	***0.004992***
	FOREST	0.112514	0.505310
	MINE	-0.210280	0.810027

6.3.2.4 Impairment prediction

Statistical inference of reduced-rank regression can be used to assess impairment in the optimal cluster.

The basic model set-up for RRR is

$$Y = XM + E = (XA)B + E$$

where Y is a $n \times m$ response matrix, X is a $n \times p$ predictor matrix, M is a $p \times m$ coefficient matrix, A and B are of full rank with rank r ($r \le \min(m, p)$) and $Vec(E) \sim N(\underline{0}, I_n \otimes \Sigma)$. Here, *n* is equal to 72, *m* is 11 and *p* is 9.

To examine the reduced-rank structure of the regression coefficient matrix (*M*), the rank of the coefficient matrix first needs to be determined. The likelihood ratio test statistic (LRT) with the correction factor (Reinsel and Velu, 1998) for testing $H_0 : rank(M) \le r$ is

$$[n - (m + p + 1)/2] \sum_{j=r+1}^{m} \ln(1 + \hat{\lambda}_j^2) \underset{A}{\overset{H_0}{\sim}} \chi^2_{(m-r)(p-r)}$$

where λ_j^2 is the eigenvalue of the matrix $\Gamma^{1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Gamma^{1/2}$, where Σ_{YX} is the sample covariance matrix. Γ is the scaling matrix. Generally there are two choices for Γ.

(1) $\Gamma = \tilde{\Sigma}_{ee}^{-1}$, the inverse of the unrestricted maximum likelihood estimate (MLE),

$\Gamma = [(1/n)(Y - X\hat{M})^T(Y - X\hat{M})]^{-1}$ with $\hat{M} = (X^T X)^{-1} X^T Y$;

(2) $\Gamma = \Sigma_{YY}^{-1} = (Y^T Y/n)^{-1}$, the inverse of the sample covariance matrix.

They produce the same MLE of the coefficient matrix *M*, though they differ in the estimators of two full rank matrices, *A* and *B*. Additional rank selection is given by the test, $H_0 : rank(M) = r$ versus $H_1 : rank(M) = r+1$. The corresponding test statistic is $[n-1-(m+p+1)/2]\ln(1+\hat{\lambda}_{r+1}^2) \underset{A}{\overset{H_0}{\sim}} \chi_1^2$.

The squared canonical correlation ρ_j^2 is related to the above eigenvalues λ_j^2 by $\lambda_j^2 = \rho_j^2/(1-\rho_j^2)\ \rho_j^2$. When the scaling matrix is Σ_{YY}^{-1}, the ρ_j^2's are the eigenvalues of $\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$, which has computational advantage.

Based on the graphical RDA results in Section 6.3.2.3, the hypothesized value of *r* is set at 3. The results are given in Table 6.6 (Note that in calculating the sample covariance matrices Σ_{YX} and Σ_{XX}, the data matrices *Y* and *X* are adjusted for sample means). It appears that the possibility that the rank is either two or three can be entertained. Only estimation results for the situation of the rank equal to 3 will be presented in detail here.

Table 6.6 Results of likelihood ratio tests for the rank of the coefficient matrix

Eigenvalue, λ_j^2	Rank, *r*	LRT*	d.f.	Critical value (with α=5%)
	0	221.61	99	123.22
3.9195	1	123.62	80	101.88
0.9449	2	82.72	63	82.53
0.7620	3	47.88	48	65.17

*LRT=likelihood ratio test

The normalized eigenvectors of the matrix $\tilde{\Sigma}_{ee}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\tilde{\Sigma}_{ee}^{-1/2}$ corresponding to the three largest eigenvalues, λ_1^2, λ_2^2, and λ_3^2 are used to obtain the MLE of *M*. The normalized eigenvectors are

V_1^T = (-0.082, -0.219, -0.337, -0.298, 0.271, -0.009, 0.144, 0.351, 0.183, -0.007, 0.701)

V_2^T = (0.337, -0.681, 0.372, 0.251, -0.001, -0.129, 0.318, -0.151, 0.131, -0.235, 0.085)

and

V_3^T = (0.591, 0.471, -0.007, -0.306, -0.171, -0.347, 0.217, -0.126, 0.333, -0.072, 0.075),

respectively. The MLE of *A* is $\hat{A} = \Sigma_{XX}^{-1}\Sigma_{XY}\tilde{\Sigma}_{ee}^{-1/2}V_{(3)}$ and the MLE of *B* is $\hat{B} = V_{(3)}^T\tilde{\Sigma}_{ee}^{-1/2}$ with $V_{(3)} = [V_1, V_2, V_3]$. The corresponding MLE of the coefficient matrix *M* ($M = AB$) is $\overline{M} = \Sigma_{XX}^{-1}\Sigma_{XY}\tilde{\Sigma}_{ee}^{-1/2}V_{(3)}V_{(3)}^T\tilde{\Sigma}_{ee}^{-1/2}$. The final reduced-rank estimate of the regression coefficient matrix is thus written as follows.

−0. 212	−0. 071	0. 049	0. 265	0. 097	0. 113	−0. 070	0. 012	−0. 023	−0. 083	0. 027
−0. 020	−0. 014	0. 028	0. 045	0. 029	0. 025	−0. 027	−0. 013	−0. 007	−0. 023	0. 006
−0. 455	−0. 215	0. 118	0. 618	0. 344	0. 316	−0. 140	0. 058	−0. 010	−0. 117	0. 145
0. 035	−0. 027	0. 151	0. 085	0. 100	0. 066	−0. 134	−0. 102	−0. 033	−0. 102	0. 008
0. 689	0. 338	−0. 691	−1. 276	−0. 652	−0. 622	0. 724	0. 311	0. 217	0. 688	−0. 096
1. 477	0. 791	−0. 748	−2. 305	−1. 392	−1. 222	0. 786	0. 039	0. 114	0. 642	−0. 507
0. 293	−0. 155	1. 292	0. 682	0. 687	0. 473	−1. 175	−0. 926	−0. 341	−0. 973	−0. 046
−0. 593	−0. 137	0. 517	0. 947	0. 213	0. 343	−0. 606	−0. 310	−0. 269	−0. 701	−0. 121
3. 994	−0. 269	−0. 042	−3. 357	1. 829	−0. 097	1. 083	0. 164	1. 274	2. 643	1. 599

The approximate unbiased estimate of the error covariance matrix from the reduced-rank analysis is given as

$$\Sigma_{ee} = (Y - X\overline{M})^T(Y - X\overline{M})/(n-p).$$

With the set up in Section 6.2.5, simultaneous prediction intervals are given by

$$\underline{x}_0^T\hat{A}\hat{\underline{B}}_{(i)} \pm \sqrt{\frac{m(n-r)}{n-r-m+1}F(m, n-r-m+1, 1-\alpha)}\sqrt{1+\frac{n}{n-r}\underline{x}_0^T\hat{A}(\hat{A}^T X^T X\hat{A})^{-1}\hat{A}^T\underline{x}_0\hat{\sigma}_{ii}}$$

where $\hat{\underline{B}}_{(i)}$ is the *i*th column of $\hat{B}$ ($\hat{B} = V_{(3)}^T\tilde{\Sigma}_{ee}^{-1/2}$). $\hat{\sigma}_{ii}$ is the *i*th diagonal element of $\hat{\Sigma}_{ee} = (Y - X\hat{A}\hat{B})^T(Y - X\hat{A}\hat{B})/(n-r)$ and $\hat{A} = \Sigma_{XX}^{-1}\Sigma_{XY}\tilde{\Sigma}_{ee}^{-1/2}V_{(3)}$.

In the optimal cluster, 11 reference sites are pre-defined. The mean vector of these reference sites is defined as $\underline{x}_0$ for calculating prediction interval bounds,

$\underline{x}_0^T$ = (11.073, 3.575, 16.454, 14.545, 16.636, 0.100, 2.999, 8.987, 0.028).

Results in Table 6.7 indicate that most of the prediction intervals based on the reduced rank regression are slightly wider than those based on full rank regression. The difference mainly lies in the lower prediction bound. The lower prediction bound based on the reduced-rank regression is smaller than that for the full rank regression.

If the values of all response variables at one site fall in the prediction intervals for the reference mean, the site will be declared as unimpaired. The full rank regression claims that all the sites in the optimal cluster are unimpaired, while the reduced rank regression claims slightly fewer sites (66 sites out of 72 sites) to be unimpaired. The reason that so many reference sites are 'mislabeled' may be because non-reference and reference sites are combined together thus the reference conditions at theses sites may be masked. In addition, the determination of reference sites affects the clustering thus test results.

Table 6.7 Simultaneous prediction intervals based the full rank and reduced rank regressions

	full rank			reduced-rank		
Variables	UPL	LPL	width	UPL	LPL	width
TOTLRICH	4.789	-5.450	10.239	3.099	-7.324	10.423
EPT_PTAX	4.977	-5.369	10.346	4.288	-5.828	10.116
PLECPIND	5.041	-5.162	10.204	7.217	-3.736	10.954
INTLRICH	4.726	-5.495	10.221	9.303	-2.319	11.622
INTLPIND	4.863	-5.470	10.334	6.742	-4.049	10.792
FACLPIND	5.358	-4.959	10.317	6.672	-3.648	10.321
HPRIME	4.983	-5.265	10.249	3.455	-7.345	10.800
LOGSHRDPIND	5.181	-5.047	10.228	4.347	-6.014	10.361
LOGSCRPPIND	4.932	-5.320	10.252	4.344	-5.759	10.103
LOGSIMPSON	5.206	-5.036	10.242	3.310	-7.408	10.719
LOGHBI	5.453	-4.967	10.420	5.099	-5.078	10.178

6.3.2.5 Graphical display of the acceptance region

A prediction interval ellipse, which defines the graphical acceptance region, can be added to the ordination graph when there are two RDA axes in the biplot. The procedure for obtaining this graphical acceptance region consists of four steps.

Step 1: Get $\underline{\bar{y}}_{ref}$, the mean vector of the response variables at the reference sites in the optimal cluster.

Step 2: Locate the center of the ellipse by multiplying $\underline{\bar{y}}_{ref}$ with the RDA scores of response variables at the two major RDA axis, $\underline{l}_1$, $\underline{l}_2$, to form the center of the ellipse, ($\underline{l}_1^T \underline{\bar{y}}_{ref}$, $\underline{l}_2^T \underline{\bar{y}}_{ref}$).

Step 3: Specify the boundaries of the ellipse by multiplying the vectors of confidence interval bounds, $\underline{UPL}$, $\underline{LPL}$ with the RDA scores of response variables at the two major RDA axis, $\underline{l}_1$, $\underline{l}_2$. In this application, $\underline{UPL}$, $\underline{LPL}$ can be obtained from Table 6.6. For instance, $\underline{UPL}$ is set as follows.

$\underline{UPL}^T$ = (3.099, 4.288, 7.217, 9.303, 6.742, 6.672, 3.455, 4.347, 4.344, 3.310, 5.099).

The endpoints of the ellipse axes are ($\underline{UPL}^T \underline{l}_1$, $\underline{l}_2^T \underline{\bar{y}}_{ref}$), ($\underline{LPL}^T \underline{l}_1$, $\underline{l}_2^T \underline{\bar{y}}_{ref}$), ($\underline{l}_1^T \underline{\bar{y}}_{ref}$, $\underline{UPL}^T \underline{l}_2$), ($\underline{l}_1^T \underline{\bar{y}}_{ref}$, $\underline{LPL}^T \underline{l}_2$).

Step 4: Plot the ellipse with the specified center and axes bounds.

When adding the prediction interval ellipse to the ordination graph, the length of axes is defined by ordination results. If the prediction limits are extremely wide (exceeding the range of data), the prediction interval ellipse may not be displayed on the same graph as the original dataset. Figure 6.9 gives a conceptual ellipse for the reference sites.

The software package, R (2006), is a free software package that provides environment for statistical computing and graphics and is similar to the software package S-Plus. It can produce extra graphical items on the ordination diagrams. The function ‘ordiellipse’ of the package VEGAN draws the probability interval ellipse based on the covariance or correlation matrix of the weighted scores for sites. Though this tool doesn’t give the true prediction interval ellipse, it can help interpret results. The probability interval ellipse for the reference sites (the lower ellipse) in Figure 6.10 almost covers all the sites in the

region, which is close to the conclusion based on the prediction intervals in Table 6.7. The proposed prediction method excludes 6 sites from the unimpaired sites.

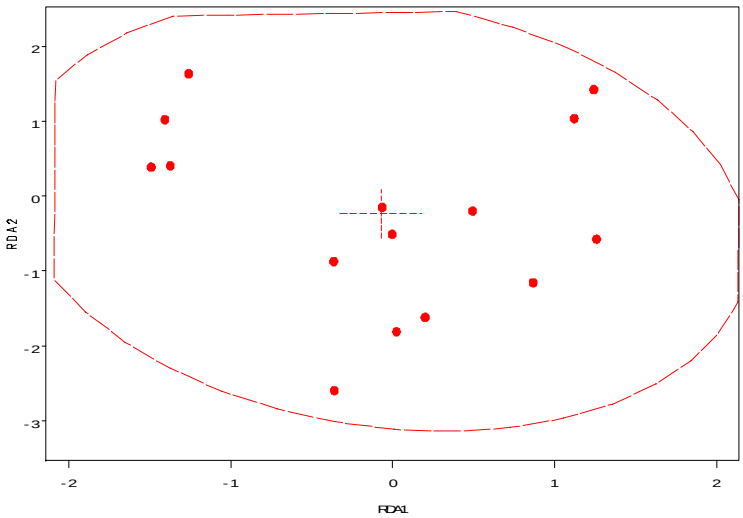

Figure 6.9 Conceptual prediction interval ellipse for the reference sites.

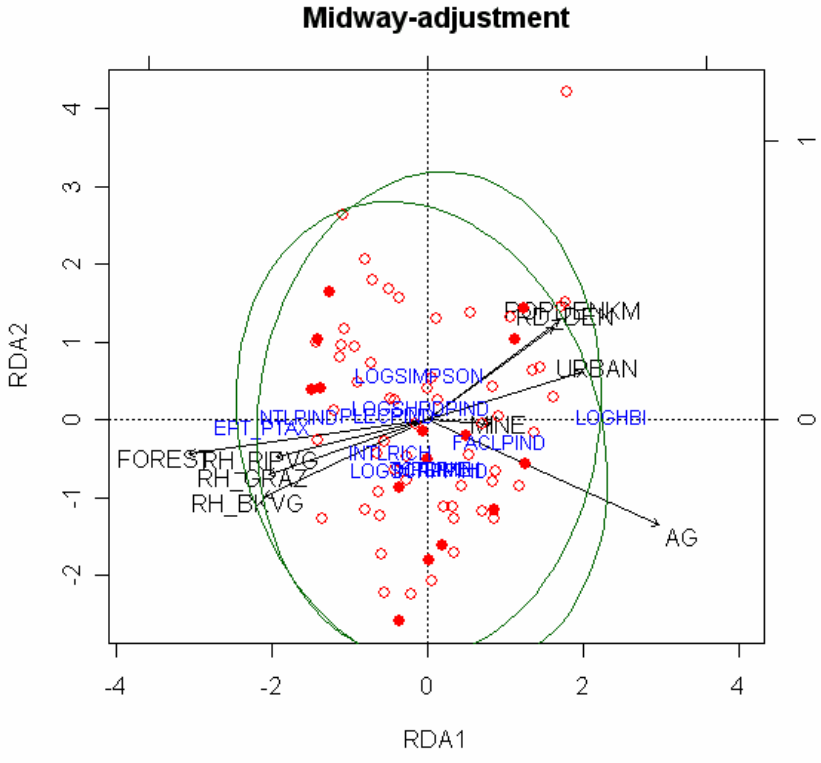

Figure 6.10 Probability ellipses for the optimal cluster. The upper and lower ellipses indicate the 95% probability region for non-reference and reference sites, respectively.

6.4 Concluding remarks

This chapter developed the technique for model-based clustering and tests, and detailed the implementation using a dataset from MAHA. The application shows that the prediction intervals based on the reduced-rank approach tend to be wider than those from the full rank regression. When reduced-rank regression is applied, the tests tend to claim unimpairment, which may lead to misclassify impaired sites as unimpaired. The poor classification results may be related to the way reference sites are determined. If based on defining variables (chemical or habitat stressors), classification is likely to be poor unless the environmental variables in the data are highly correlated with the defining variables.

The MAHA dataset covers a large spatial region. Adjustment by reference sites and model-based clustering help strengthen the response-stressor relationship for some regions. This is expected to help construct more precise tests. Graphical displays of test results provide a quick view of the data and summarize dominant relationships between the biological metrics and the environmental stressors.

In this chapter, the response variables in the MAHA data are continuous. For different data structures, assessment using model-based clustering can be easily extended by using appropriate multivariate analysis methods. For example, when the response variables are the counts of species, canonical correspondence analysis (CCA) will be used to identify the relationship between the distribution patterns of different species and environmental variables. For future research, it is worth setting a completely theoretical framework for regression-based tests in the multivariate case with the consideration of various data structure and different multivariate analysis methods.

7. Summary and Future Research

The primary goal of this research is to evaluate tests using analysis of variance and regression in water quality assessment. Specifically, this research develops model-based tests using tolerance limits and prediction limits (corresponding to different settings of standards), provides general regression-based tests in the univariate case, and proposes model-based clustering in the multivariate case for water quality assessment.

The model-based tests provide a framework for incorporating regional information into site-specific tests. When the standard is fixed and the available sample size is small for the site of interest, the test based on the fixed effects model outperforms the single site test when the test site is highly impaired. The basic random effects model doesn't improve assessment for individual sites but it helps evaluate regional impairment. Among the effective factors in model-based tests, large sample sizes for the test site generally increases test power for most tests; the effect of the number of non-test sites on model-based tests is more obvious when the sample size per site is small. The test based on the fixed effects model is generally recommended for assessment. However, the choice of the underlying model for tests might depend on the sampling strategy and site property. In environmental studies, data with temporal correlation within a site are often encountered. The suitable error structure can be added to the model-based test to account for the autocorrelation. When the information of the regional impairment proportion is in need, the model-based estimator is recommended. A multiplicity adjustment should be of concern when the regional perspective for impairment follows the site-specific assessment.

When impairment is associated with a single stressor, it is possible to make an impairment decision based on a stressor-response relationship. This allows for an approach that uses all the biological data in the decision process rather than just the data from reference conditions. Regression is also useful for adjusting a reference-based limit when biological condition varies with a covariate. The assessment can be improved when a strong relationship exists between the response and the covariate. When dealing with large datasets, adjustment by reference sites and model-based clustering help strengthen the response-stressor relationship, which is expected to help construct more precise tests.

Graphical displays of test results provide a quick view of the data with plenty of information.

There are other issues that might improve tests and can be helpful in application. For example, defining a neighborhood of a site helps determine the degree of similarity between sites. The similarity level provides the criterion for the proportion of pooled information. Should the similarity be evaluated in terms of variables of interest or variables that are not used for assessment, or be evaluated by combining the assessed variables and underlying variables? Such a topic is interesting and useful in water quality assessment. As to models, I restrict current research to the one-way fixed effects model and random effects model. The test performance should be studied when multi-way models or the mixed model is used for more complicated situations. For instance, a basic mixed model might consist of a spatial effect as a fixed effect and a temporal factor as a random effect. A two-way random effects model can include both spatial and temporal factors as random effects. The application of regression-based tests on a real dataset suggests the ANCOVA (analysis of covariance) model. The practical needs determine the model set-up.

When different multivariate analysis methods are used for different data structures, assessment using model-based clustering can be extended and there is an need of a completely theoretical framework for regression-based tests in the multivariate case in future study.

In practice, it is common to have outliers in datasets. When the number of outliers is relatively small (compared to the sample size) and they tend to cluster together, Mahalanobis distance helps identify multivariate outliers through robust estimates of the mean and covariance matrix (Rousseeuw and Van Zomeren, 1990). These types of methods focus on detecting multivariate outliers in one cluster. When the number of outliers is considerable and outliers are spread in multiple clusters (i.e., diffuse outliers are considered), robust clustering methods can be studied in conjunction with outlier identification (Hardin and Rocke, 2004). The testing procedure can thus be constructed correspondingly.

Current reduced-rank regression is restricted to stressor variables. The response variables are often spatial correlated. Dimension reduction can be operated on two sets of

variables, perhaps conditioning on a set of covariates that are not considered to be stressors. Developing the corresponding inference for impairment detection will broaden applications.

Another avenue of research is to compare two procedures, constructing tests based on redundancy analysis (RDA) results or graphically displaying a reference ellipse after multivariate testing. Redundancy analysis can graphically display response-stressor results. If the impairment test is carried out based on canonical axes of RDA, how do people interpret the test results and what's its practical usage? Alternatively, if the aim is to have a graphical display of multivariate testing results, how is the reference ellipse drawn? If reduced-rank regression is used for impairment testing, will the realization of graphical display produce the same result as RDA? These seem to be trivial but any solution will be a big step in application.

Similarity is also a big concern in multivariate studies. Given a site, how will a neighborhood be defined so that the neighborhood can be treated as the optimal group for good display of multivariate analysis? How to define the size of the neighborhood and the neighborhood boundary is also important for future study.

Appendix 1

The fixed effects model with a time series error structure in Section 3.3.3 is given by

$$y_{it} = \mu_i + s_{it} \text{ and } s_{it} = \phi\, s_{i(t-1)} + e_{it}, |\rho| < 1.$$

When there are n measurements ($n > 3$) for each site, the variance structure is written as

$$Var(\underline{y_i}) = \frac{\sigma_e^2}{1-\phi^2}\begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 \ldots\ldots \phi^{n-1} \\ \phi & 1 & \phi & \phi^2 \ldots\ldots \phi^{n-2} \\ \phi & \phi^2 & 1 & \phi \ldots\ldots \phi^{n-3} \\ \ldots & & & \\ \phi & \phi^2 & \ldots\ldots \phi^{n-1} & 1 \end{bmatrix}.$$

The expectation of the estimated variance is derived as follows.

$$E(y_{ij}^2) = [E(y_{ij})]^2 + Var(y_{ij}) = \mu_i^2 + \frac{\sigma_e^2}{1-\phi^2}$$

$$E(\bar{y}_{i.}^2) = E\left[\left(\frac{\sum_{j=1}^{n} y_{ij}}{n}\right)^2\right] = E\left(\frac{\sum_{j=1}^{n} y_{ij}^2 + 2\sum_{j`>j} y_{ij}\, y_{ij`}}{n^2}\right)$$

$$= \{n(\mu_i^2 + \frac{\sigma_e^2}{1-\phi^2}) + 2[\sum_{l=1}^{n-1} \frac{(n-l)\phi^l}{1-\phi^2}\sigma_e^2 + \frac{n(n-1)}{2}\mu_i^2]\}/n^2$$

$$= \mu_i^2 + \frac{n + \sum_{l=1}^{n-1} 2(n-l)\phi^l}{(1-\phi^2)n^2}\sigma_e^2$$

For j=1 or n,

$$E(y_{ij}\bar{y}_{i.}) = E(y_{ij}\sum_{j`=1}^{n} y_{ij`})/n = [E(y_{ij}^2) + \frac{\sum_{l=1}^{n-1}\phi^l}{1-\phi^2}\sigma_e^2 + (n-1)\mu_i^2]/n$$

$$= \mu_i^2 + \frac{1 + \sum_{l=1}^{n-1}\phi^l}{n(1-\phi^2)}\sigma_e^2$$

For j=2 or n-1,

$$E(y_{ij}\bar{y}_{i.}) = E(y_{ij}\sum_{j`=1}^{n} y_{ij`})/n = [E(y_{ij}^2) + \frac{2\phi + \sum_{l=2}^{n-2}\phi^l}{1-\phi^2}\sigma_e^2 + (n-1)\mu_i^2]/n$$

$$= \mu_i^2 + \frac{1 + 2\phi + \sum_{l=2}^{n-2}\phi^l}{n(1-\phi^2)}\sigma_e^2$$

For j=w or n-(w-1), w=1, 2, …, integer(n/2),

$$E(y_{ij}\bar{y}_{i.}) = E(y_{ij}\sum_{j`=1}^{n} y_{ij`})/n = [E(y_{ij}^2) + \frac{2\sum_{l=1}^{w-1}\phi^l + \sum_{l=w}^{n-w}\phi^l}{1-\phi^2}\sigma_e^2 + (n-1)\mu_i^2]/n$$

$$= \mu_i^2 + \frac{1 + 2\sum_{l=1}^{w-1}\phi^l + \sum_{l=w}^{n-w}\phi^l}{n(1-\phi^2)}\sigma_e^2$$

When n is even, the expectation of the estimated variance is expressed as below.

$$E(s^2) = E(\frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2}{(k+1)(n-1)}) = \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}E(y_{ij}^2 + \bar{y}_{i.}^2 - 2y_{ij}\bar{y}_{i.})}{(k+1)(n-1)}$$

$$= \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(\mu_i^2 + \frac{\sigma_e^2}{1-\phi^2} + \mu_i^2 + \frac{n + \sum_{l=1}^{n-1}2(n-l)\phi^l}{(1-\phi^2)n^2}\sigma_e^2) - 2\sum_{i=1}^{k+1}[2\sum_{w=1}^{n/2}(\mu_i^2 + \frac{1 + 2\sum_{l=1}^{w-1}\phi^l + \sum_{l=w}^{n-w}\phi^l}{n(1-\phi^2)}\sigma_e^2)]}{(k+1)(n-1)}$$

$$= \frac{\frac{n + n^2 + \sum_{l=1}^{n-1}2(n-l)\phi^l}{n(1-\phi^2)}\sigma_e^2 - 4\sum_{w=1}^{n/2}(\frac{1 + 2\sum_{l=1}^{w-1}\phi^l + \sum_{l=w}^{n-w}\phi^l}{n(1-\phi^2)}\sigma_e^2)}{(n-1)}.$$

When n is odd and j=$\frac{n-1}{2}+1$,

$$E(y_{ij}\bar{y}_{i.}) = E(y_{ij}\sum_{j`=1}^{n} y_{ij`})/n = [E(y_{ij}^2) + \frac{2\sum_{l=1}^{(n-1)/2}\phi^l}{1-\phi^2}\sigma_e^2 + (n-1)\mu_i^2]/n$$

$$= \mu_i^2 + \frac{1 + 2\sum_{l=1}^{(n-1)/2}\phi^l}{n(1-\phi^2)}\sigma_e^2$$

The expectation of the estimated variance for odd values of *n* is expressed as below.

$$E(s^2) = E(\frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2}{(k+1)(n-1)}) = \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}E(y_{ij}^2 + \bar{y}_{i.}^2 - 2y_{ij}\bar{y}_{i.})}{(k+1)(n-1)}$$

$$= \frac{\sum_{i=1}^{k+1}\sum_{j=1}^{n}(\mu_i^2 + \frac{\sigma_e^2}{1-\phi^2} + \mu_i^2 + \frac{n + \sum_{l=1}^{n-1}2(n-l)\phi^l}{(1-\phi^2)n^2}\sigma_e^2) - 2\sum_{i=1}^{k+1}[\Lambda]}{(k+1)(n-1)}$$

$$= \frac{\frac{n + n^2 + \sum_{l=1}^{n-1}2(n-l)\phi^l}{n(1-\phi^2)}\sigma_e^2 - 4\sum_{w=1}^{(n-1)/2}(\frac{1 + 2\sum_{l=1}^{w-1}\phi^l + \sum_{l=w}^{n-w}\phi^l}{n(1-\phi^2)}\sigma_e^2) - \frac{2 + 4\sum_{l=1}^{(n-1)/2}\phi^l}{n(1-\phi^2)}\sigma_e^2}{(n-1)}$$

where $\Lambda = 2\sum_{w=1}^{(n-1)/2}(\mu_i^2 + \frac{1 + 2\sum_{l=1}^{w-1}\phi^l + \sum_{l=w}^{n-w}\phi^l}{n(1-\phi^2)}\sigma_e^2) + \mu_i^2 + \frac{1 + 2\sum_{l=1}^{(n-1)/2}\phi^l}{n(1-\phi^2)}\sigma_e^2$

Bibliography

Anderson, T.W. (2002). “Reduced Rank Regression in Cointegrated Models”. *Journal of Econometrics* 106: 203-216.

Bailey, R.C., Day, K.E., Norris, R.H. and Reynoldson, T.B. (1994). “Macroinvertebrate Community Structure and Sediment Bioassay Results from Nearshore Areas of North American Great Lakes”. *Journal of Great Lakes Resource* 21(1): 42-53.

Bailey, R.C., Norris, R.H. and Reynoldson, T.B. (2005). *Bioassessment of Freshwater Ecosystems: Using the Reference Condition Approach*. Springer: New York.

Banfield, J.D. and Raftery, E.A. (1993). “Model-Based Gaussian and Non-Gaussian Clustering”. *Biometrics* 49(3): 803-821.

Barbour, M.T., Gerritsen, J., Griffith, G.E., Frydenborg, R., McCarron, E., White, J.S. and Bastian, M.L. (1996). “A Framework for Biological Criteria for Florida Streams Using Benthic Macroinvertebrates”. *Journal of North American Benthological Society* 15(2): 185-211.

Barbour, M.T., Gerritsen, J., Snyder, B.D. and Stribling, J.B. (1999). *Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish*. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water: Washington, D.C.

Barnett, V. and O’Hagan, A. (1997). *Setting Environmental Standards*. Chapman and Hall: London.

Bartsch, A.F. and Ingram, W.M. (1966). “Biological Analysis of Water Pollution in North America”. *Verh International Verein Limnology* 16: 786-800.

Bates Prins, S.C., Smith, E.P., Angermeier, P.L. and Yagow, E.R. (2006). *Clustering Using Stressor-response Relationships with Discussion on Optimal Criteria*. Draft.

Batiuk, R.A., Orth, R.J., Moore, K.A., Dennison, W.C., Stevenson, J. C., Staver, L.W., Carter, V., Rybicki, N.B., Hickman, R.E., Kollar, S., Bieber, S. and Heasly, P. (1992). *Chesapeake Bay Submerged Aquatic Vegetation Habitat Requirements and Restoration Targets: A Technical Synthesis*. EPA: Annapolis, MD.

Bell, P.A. and Carolan, A.M. (1998). *Trend Estimation for Small Areas: from a Continuing Survey with Controlled Sample Overlap*. Working paper.

Bhaumik, D.K. and Kulkarni, P.M. (1996). "A Simple and Exact Method of Constructing Tolerance Intervals for the One-Way ANOVA with Random Effects". *The American Statistician* 50 (4): 319-323.

Bryce, S.A., Larsen, D.P., Hughes, R.M. and Kaufmann, P.R. (1999). "Assessing the Relative Risks to Aquatic Ecosystems in the Mid-Appalachian Region of the United States". *Journal of the American Water Resources Association* 35: 23-36.

Burton, J. and Gerritsen, J. (2003). *A Stream Condition Index for Virginia Non-coastal Streams.* Tetra Tech, Inc. for USEPA Region 3 Environmental Services Division.

Casella, G. and Berger, R.L. (1990). *Statistical Inference*. Duxbury: Belmont.

Corander, J. and Villani, M. (2001). *Bayesian Assessment of Dimensionality in Multivariate Reduced Rank Regression*. English and Swedish, Statistika Institution Technical report.

Courtemanch, D.L., Davies, S.P. and Laverty, E.B. (1989). "Incorporation of Biological Information in Water Quality Planning". *Environmental Management* 13: 35-41.

Das, K., Jiang, J. and Rao, J.N.K. (2004). "Mean Squared Error of Empirical Predictor". *The Annals of Statistics* 32(2): 818-840.

Davis, W.S. and Simon, T.P. (1995). *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making*. Lewis Publishers: Boca Raton.

Denison, D.G.T. and Holmes, C.C. (2001). "Bayesian Partitioning for Estimating Disease Risk". *Biometrics* 57:143-149.

Duan, Y., Ye, K. and Smith, E.P. (2006). "Evaluating Water Quality Using Power Priors to Incorporate Historical Information". *Environmetrics* 17(1): 95-106.

Dunnett, C. W. and Tamhane, A. C. (1992). "A Step-up Multiple Test Procedure". *Journal of the American Statistical Association* 87:162-170.

Dunnett, C. W. and Tamhane, A. C. (1995). "Step-up Multiple Testing of Parameters with Unequally Correlated Estimates". *Biometrics* 51:217-227.

Fay, R. E. and Herriot, R. A. (1979). "Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data". *Journal of the American Statistical Association* 74: 269-277.

Frydenborg, R. and Ray, D. (2005). “Use of Florida's Biological and Habitat Assessment Procedures to Evaluate Sediment Impacts”. *The American Society of Agricultural and Biological Engineers*, paper number 052198, 2005 ASAE Annual Meeting.

Geweke, J. (1996). “Bayesian Reduced Rank Regression in Econometrics”. *Journal of Econometrics* 75:121-146.

Gibbons, R.D. (1994). *Statistical Methods for Groundwater Monitoring.* Wiley: New York.

Gibbons, R.D. (2003). “A Statistical Approach for Performing Water Quality Impairment Assessments”. *Journal of the American Water Resources Association* August: 841-849.

Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold: New York

Gilbert, R.O., LeGore, T. and O’Brien, R.F. (1996). *An Overview of Methods for Evaluating the Attainment of Cleanup Standards for Soils, Solid Media, and Groundwater, EPA Volumes 1, 2 and 3*. EPA technical report.

Greco, F.P., Lawson, A.B., Cocchi, D. and Temples, T. (2005). “Some Interpolation Estimators in Environmental Risk Assessment for Spatially Misaligned Health Data”. *Environmental and Ecological Statistics* 12:379-395.

Green, R.H. (1974). “Multivariate Niche Analysis with Temporally Varying Environmental Factors”. *Ecology* 55(1): 73-83.

Guttorp, P. (2000). *Setting Environmental Standards: A Statistician's Perspective*. NRCSE Technical Report Series, NRCSE-TRS No. 048.

Hardin, J. and Rocke, D. (2004). “Outlier Detection in Multiple Cluster Setting Using the Minimum Covariance Determinant Estimator”. *Computational Statistics and Data Analysis* 44: 625-638.

Hastie, T. J. (1992). *Generalized additive models*. Chapter 7 in Chambers, S. J. M. and Hastie, T. J. (editors), Statistical Models. Wadsworth & Brooks/Cole: California.

Heegaard, E., Birks, H.H., Gibson, C.E., Smith, S.J. and Wolfe-Murphy, S. (2001). “Species-environmental Relationships of Aquatic Macrophytes in Northern Ireland”. *Aquatic Botany* 70: 175–223.

Helsel, D.R. and Hirsch, R.M. (2000). *Statistical Methods in Water Resources*. Elsevier Science Pub Company: New York.

Herlihy, A.T., Larsen, D.P., Paulsen, S.G., Urquhart, N.S. and Rosenbaum, B.J. (2000). "Designing a Spatially Balanced, Randomized Site Selection Process for Regional Stream Surveys: the EMAP Mid-Atlantic Pilot Study". *Environmental Monitoring and Assessment* 63: 95-113.

Hilsenhoff, W.L. (1987). "An Improved Biotic Index of Organic Stream Pollution". *Great Lakes Entomology* 20:31-39.

Hochberg, Y. (1988). "A Sharper Bonferroni Procedure for Multiple Tests of Significance". *Biometrika* 75:800-802.

Hughes, R.M. (1995). *Defining acceptable biological status by comparing with reference conditions*. Pages 31-47 in Davis, W.S. and Simon, T.P. (editors), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers: Washington, D.C.

Hughes, R.M., Howlin, S. and Kaufmann, P. R. (2004). "A Biointegrity Index for Coldwater Streams of Western Oregon and Washington". *Transactions of the American Fisheries Society* 133:301-312.

Hulting, F. L and Harville, D. A. (1991). "Some Bayesian and Non-Bayesian Procedures for the Analysis of Comparative Experiments and for Small-Area Estimation: Computational Aspects, Frequentist Properties, and Relationships". *Journal of the American Statistical Association* 86: 557-568.

Kilgour, B.W., Somers, K. M. and Matthews, D.E. (1998). "Using the Normal Range as a Criterion for Ecological Significance in Environmental Monitoring and Assessment". *Ecoscience* 5(4): 542-550.

Legendre, P. and Legendre, L. (1998). *Numerical ecology*. 2nd English edition. Elsevier Science BV: Amsterdam.

Lenat, D.R. (1993). "Water Quality Assessment of Streams Using a Qualitative Collection Method for Macroinvertebrates". *Journal of North American Benthological Society* 72(3): 279-290.

Lin, P., Meeter, D. and Niu, X. (2000). *A Nonparametric Procedure for Listing and Delisting Impaired Waters Based on Criterion Exceedance*. Technical report. Florida department of environmental protection.

Lipkovich, I.A. (2002). *Bayesian Model Averaging and Variable Selection in Multivariate Ecological Models.* Doctoral dissertation of Virginia Tech.

Mackenthun, K.M. and Ingram, W.M. (1967). *Biological Associated Problems in Freshwater Environments*. U. S. Government Printing Office: Washington, D. C.

Major E.B., Rinella, D.J. and Bogan, D.L.(2002). *2001 Alaska Biological Monitoring and Water Quality Assessment Program Report*. Alaska Department of Environmental Conservation, Division of Air and Water Quality.

Makarenkov, V. and Legendre, P. (2002). "Nonlinear Redundancy Analysis and Canonical Correspondence Analysis Based on Polynomial Regression". *Ecology* 83(4): 1146-1161.

McBride, G.B. and Ellis, J.C. (2001). "Confidence of Compliance: a Bayesian Approach for Percentile Standards". *Water Research* 35(5): 1117–1124.

McCormick, F.H., Hughes, R.M., Kaufmann, P.R., Peck, D.V., Stoddard, J. L. and Herlihy, A.T. (2001). "Development of an Index of Biotic Integrity for the Mid-Atlantic Highlands Region". *Transactions of the American Fisheries Society* 130: 857-877.

McLachlan, G.J., Bean, R.W. and Peel, D. (2002). "A Mixture Model-based Approach to the Clustering of Microarray Expression Data". *Bioinformatics* 18(3): 413-422.

National Research Council, Committee on Restoration of Aquatic Ecosystems (1992). *Restoration of Aquatic Ecosystems--Science, Technology, and Public Policy*. National Academy Press: Washington, D.C.

Neter, J., Wasserman, W. and Whitmore, G.A. (1988). *Applied Statistics*. Allyn and Bacon: Boston.

Okabe, A., Boots, B., Sugihara, K. and Chiu, N.S.(2000). *Spatial Tessellations - Concepts and Applications of Voronoi Diagram*s. Wiley: New York.

Overton, W.S., Stevens, D.L. and White, D. (1991). *Design Report for EMAP*. USEPA, Corvallis, Oregon.

Owen, D.B. (1968). “A Survey of Properties and Applications of the Noncentral t-Distribution”. *Technometrics* 10:445-478.

Perry, J. and Vanderklein, E. (1996). *Water Quality: Management of a Natural Resource*. Blackwell Science: Malden.

Phillips, D.J.H. and Rainbow, P.S. (1993). *Biomonitoring of Trace Aquatic Contaminants*. Elsevier Applied Science: New York.

Posthuma, L., Suter, G.W. II and Trass, T.P. (2001). *Species Sensitivity Distributions in Ecotoxicology*. CRC Press: Boca Raton.

Prasad, N.G.N. and Rao, J.N.K (1990). “The Estimation of Mean Squared Errors of Small Area Estimators”. *Journal of the American Statistical Association* 85: 163-171.

R version 2.4.0 (2006). Copyright (C) The R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rao, C.R. (1964). “The Use and Interpretation of Principal Component Analysis in Applied Research”. *Sankhya A* 26: 329-358.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley: New York.

Reinsel, G.C. and Velu, R.P. (1998). *Multivariate Reduced-rank Regression: Theory and Applications*. Springer: New York.

Reynoldson, T.B., Day, K.E. and Pascoe, T. (2000). *The Development of the BEAST: a Predictive Approach for Assessing Sediment Quality in the North American Great Lakes*. Page 165-80 in Wright, J.F., Sutcliffe, D.W. and Furse, M.T. (editors), Assessing the Biological Quality of Freshwaters: RIVPACS and Other Techniques. The Ferry House: Ambleside.

Rosenberg, D.M. and Resh, V.H. (2003). *Freshwater Biomonitoring and Benthic Macroinvertebrates*. Springer: New York.

Rousseeuw, P.J. and Van Zomeren, B.C. (1990). “Unmasking Multivariate Outliers and Leverage Points”. *Journal of the American Statistical Association* 85: 633-651.

SAS Version 8.2 (1999). Copyright (C) SAS Institue Inc. Cary, NC, USA.

Schabenberger, O. and Pierce, F.J. (2002). *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press: Boca Raton.

Schaible, W.L. and Casady, R.J. (1994). “Comment on “Small area estimation: An appraisal” by Ghosh, M. and Rao, J.N.K.”. *Statistical Sciences* 9: 55-93.

Shabman, L. and Smith, E.P. (2003). "Implications of Applying Statistically Based Procedures for Water Quality Assessment". *Journal of Water Resources Planning and Management* ASCE/July/August: 330-336.

Sidak, Z. (1967). "Rectangular Confidence Regions for Means of Multivariate Normal Distributions". *Journal of the American Statistical Association* 62:626-633.

Simes, R. J. (1986). "An Improved Bonferroni Procedure for Multiple Tests of Significance". *Biometrika* 73:751-754.

Smith, E.P., Ye, K., Hughes, C. and Shabman, L. (2001). "Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act". *Environmental Science and Technology* 35(3): 606-612.

Smith, E.P., Zahran, A., Mahmoud M. and Ye, K. (2003). "Evaluation of Water Quality Using Acceptance Sampling by Variables". *Environmetric* 14:373-386.

Smith, R.W. (2002). "The Use of Random-model Tolerance Intervals in Environmental Monitoring and Regulation". *Journal of Agricultural, Biological, and Environmental Statistics* 7(1): 74-94.

Taylor, W.A. (1995). *Guide to Acceptance Sample*. ISBN 0-9635122-0-X.

Ter Braak, C. J. F. (1986). "Canonical Correspondence Analysis: a New Eigenvector Technique for Multivariate Direct Gradient Analysis". *Ecology* 67(5): 1167-1179.

Ter Braak, C. J. F. (1994). "Biplots in Reduced-rank Regression". *Biometrical Journal* 36(8): 983-1003.

Urquhart, N.S. (1982). "Adjustment in Covariance when One Factor Affects the Covariate". *Biometrics* 38:651-660

Urquhart, N.S., Overton, W.S. and Birkes, D.S. (1993). "*Comparing Sampling Designs for the Monitoring of Ecological Status and Trends: Impact of Temporal Patterns*". Page 71-85 in Barnett, V. and Turkman, K.F. (editors), Statistics for the Environment. Wiley: New York.

Urquhart, N.S. and Kincaid, T.M. (1999). "Designs for Detecting Trend from Repeated Surveys of Ecological Resources". *Journal of Agricultural, Biological, and Environmental Statistics* 4(4): 404-414.

USEPA (1987). http://www.epa.gov/r5water/cwa.htm. *Clean Water Act*.

USEPA (1989). Statistical Guidance Document.

USEPA (1990). “National Primary and Secondary Drinking Water Regulations: Synthetic Organic Chemicals and Inorganic Chemicals; Proposed Rule”. *Federal Register* 55(143): 30370.

USEPA (1991). *Technical Support Document for Water Quality-based Toxics Control.* EPA Office of Water: Washington, D.C.

USEPA (1992). Statistical Guidance Document.

USEPA (1997). *EMAP Research Strategy*. EPA Office of Research and Development: Washington, D.C.

USEPA (2000). Statistical Guidance Document.

USEPA (2000). http://www.epa.gov/waterscience/standards/about/crit.htm.Water Quality Standards.

USEPA, Chesapeake Bay Program (2000). *Ambient Water Quality Criteria: For Dissolved Oxygen, Water Clarity and Chlorophyll a for Chesapeake Bay and its Tidal Tributaries*.

USEPA (2006). *MAHA Streams Assessment.* EPA Office of Research and Development: Washington, D.C.

USEPA Water Quality Standard. http://www.epa.gov/waterscience/standards/.

VDEQ (2006). “Using Probabilistic Monitoring Data to Validate the Non-Coastal Virginia Stream Condition Index”. *Virginia Department of Environmental Quality Technical Bulletin* WQA/2006-001.

Waite, I.R., Herlihy, A.T., Larsen, D.P. and Klemm, D.J. (2000). “Comparing the Strengths of Geographic and Nongeographic Classifications of Stream Benthic Macroinvertebrates in the Mid-Atlantic Highlands, USA”. *Journal of the North American Benthological Society* 19: 429-441.

Weerahandi, S. (1993). “Generalized Confidence Intervals”. *Journal of the American Statistical Association* 88: 899-905.

White, H. (1980). “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity”. *Econometrics* 48: 817 -838.

Wright, J.F., Sutcliffe, D.W. and Furse, M.T. (2000). *Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Technique.* Freshwater Biological Association: Ambleside.

Ye, K. and Smith, E.P. (2002). “A Bayesian Approach to Evaluating Site Impairment”. *Environmental and Ecological Statistics* 9:379-392.

Yoder, C.O. and Rankin, E.T. (1995). “The Role of Biological Criteria in Water Quality Monitoring, Assessment and Regulation”. *Ohio EPA Technical Report* MAS/1995-1-3.

Yuan, L.L. and Norton, S.B. (2003). “Comparing Responses of Macroinvertebrate Metrics to Increasing Stress”. *Journal of North American Benthological Society* 22(2): 308-322.