Data Analysis Considerations in Producing 'Comparable' Information for Water Quality Management Purposes

Lindsay Martin Griffith, Assistant Engineer, Brown and Caldwell
Robert C. Ward, Professor, Colorado State University
Graham B. McBride, Principal Scientist, NIWA (National Institute of Water and Atmospheric Research), Hamilton, New Zealand
Jim C. Loftis, Professor, Colorado State University

A report based on thesis material prepared by the lead author as part of the requirement for the degree of Master of Science from Colorado State University

February 2001

ABSTRACT

Water quality monitoring is being used in local, regional, and national scales to measure how water quality variables behave in the natural environment. A common problem, which arises from monitoring, is how to relate information contained in data to the information needed by water resource management for decision-making. This is generally attempted through statistical analysis of the monitoring data. However, how the selection of methods with which to routinely analyze the data affects the quality and comparability of information produced is not as well understood as may first appear.

To help understand the connectivity between the selection of methods for routine data analysis and the information produced to support management, the following three tasks were performed.

- An examination of the methods that are currently being used to analyze water quality monitoring data, including published criticisms of them.

- An exploration of how the selection of methods to analyze water quality data can impact the comparability of information used for water quality management purposes.

- Development of options by which data analysis methods employed in water quality management can be made more transparent and auditable.

These tasks were accomplished through a literature review of texts, guidance and journals related to water quality. Then, the common analysis methods found were applied to portions of a river water quality dataset from New Zealand. The purpose of this was to establish how information changes as analysis methods change, and to determine if the information produced from different analysis methods is comparable.

The results of the literature review and data analysis are then discussed and recommendations are made addressing problems with current data analysis procedures. Options are listed through which to begin solving these problems and produce better information for water quality management.

It was found that null hypothesis testing is the most popular method through which to produce information, yet assumptions and hypotheses are loosely explained and alternatives rarely explored to determine the validity and comparability of the results. Other data analysis methods (using graphical, non-null hypotheses or Bayesian methods) that might be more appropriate for producing more comparable information are discussed, along with recommendations for further research and cooperative efforts to establish water quality data analysis protocols for producing information for management.

TABLE OF CONTENTS

Introduction

Measuring water quality conditions, as a means of defining water quality problems and developing solutions, has been a part of water pollution control efforts since the mid-1900s. States established water pollution control policies and institutional arrangements to control water pollution as well as designed the means to measure the quality of the water to implement the policies. Passage of the Federal Water Quality Act of 1965 initiated legally defined requirements for states to ‘monitor’ water quality as part of an enhanced federal role in the nation’s water quality management efforts.

Before the federally required, state-based monitoring programs could mature, a major change in the United States’ approach to water quality management occurred with passage of the Federal Water Pollution Control Act Amendments of 1972 (commonly referred to as the Clean Water Act). While appearing to be an update of existing law, the 1972 Act revolutionized water quality management in the U.S. Management of water quality now involved the issuing of discharge permits that, in cases of water-quality limited streams, required large volumes of information about water quality conditions in the impacted water body. Furthermore, lists of water bodies not meeting standards were required to be assembled (Section 303d) and periodic assessments of a state’s water quality conditions had to be prepared and submitted to the U.S. Environmental Protection Agency (Section 305b). At the same time the new law required that discharge permits, where the receiving water was not water-quality limited, be ‘technology based’. Considerable effort, in the early years of implementing the new ‘Clean Water Act’, was consumed with controlling ‘point sources’ of pollution. It has only been in the last 10 years that attention has returned to the ‘water-quality’ limited water bodies in such a way as to highlight anew the need for extensive data and information about the nation’s water quality conditions and the impacts of ‘non-point sources’ of pollution.

Today, defensible information based on sound water quality data is becoming increasingly important as numerous lawsuits are directing renewed nationwide attention to the cleanup of water quality problems through the development of total maximum daily loads (TMDLs) for section 303(d) (GAO, 2000).

In order to evaluate the status of their waters, and comply with 303(d) and 305(b) reporting requirements of the Clean Water Act, states and other entities have collected water quality data and prepared water quality assessments. However, there is a view that the assessments and reporting of this data do not provide indisputable information about the true quality of the nation’s waters (PEER, 1999; GAO, 2000). “All too often, monitoring projects are initiated with a minimum of forethought, and result in a collection of poorly-documented data which are never analyzed, [and if they are] provide little or any feedback to resource managers, and contribute little or nothing to our understanding of the systems being monitored” (MacDonald, 1994).

This raises the question: how should data analysis methods be chosen to produce information from the data and increase our understanding of the monitored system? Answering this question often raises concerns about the validity of the assumptions that are implicit in most statistical analysis procedures, thus calling into question the appropriateness of the analysis procedures chosen. The *ad hoc* selection of data analysis methods also hurts the validity of the results and the comparability of the information produced. Another, more common, concern is that if the analysis methods are not determined prior to the collection of data, then the analyst has freedom to choose the methods that will produce the most favorable outcome (e.g., by adjusting hypothesis tests' significance level).[1]

Purpose

The purpose of this report is to review the current statistical analysis procedures used by a variety of monitoring entities to produce water quality information and to provide insight into the issues surrounding the difficult task of selecting methods to analyze water quality data in support of management's 'legal' monitoring requirements.

More specifically, the following sections will: (1) inventory the data analysis methods that are currently being employed to analyze water quality monitoring data, as well as the criticisms of current methods; (2) explore how the selection of methods to analyze water quality data can impact the comparability (i.e., similarity or suitability for comparison) of information used for water quality management purposes, and; (3) offer options by which data analysis methods employed in water quality management can be made more transparent and auditable (i.e., the methods can be reviewed, easily understood, and verified).

These tasks will be accomplished through a literature review of texts, guidance documents, and journals related to water quality monitoring. Then, the more common data analysis methods found will be applied to a portion of the New Zealand Water Quality River Network data set. The purpose of this is to establish how information changes as analysis methods change, and to determine if the information produced from different analysis methods is comparable. The results of the literature review and data analysis comparisons will be discussed and recommendations made for addressing potential problems resulting from the production of 'non-comparable' information within water quality management efforts in the United States.

Scope

Data analysis, from a water quality management perspective, can be approached from one of two directions: (1) production of information from transparent and auditable data analysis protocols that are comparable over time and space; or (2) exploration of an existing data set to see "what the data say" about water quality conditions in a water body. Statistical methods are used in both situations, but in different ways. This study addresses the first approach, i.e., routine data analysis according to a stated protocol, but realizes that the use of statistics in water quality management often mixes the two.

An argument that often falls out of the above confusion is that there should never be "recommendations" of routine analysis methods, as this censors the methods that might be used for exploratory data analysis. However, the data analysis methods discussed in this report will be limited to those methods that are used routinely by water quality management to assess water

[1] Throughout this report we use "null hypothesis test" rather than "significance test". These terms are most usually used interchangeably, but this is unfortunate (Goodman, 1993). One of last century's greatest statisticians (R.A. Fisher) coined the term "significance test" in the context of not having an alternative hypothesis and never "accepting" the null hypothesis. Other notable statisticians (Neyman and Pearson) used "hypothesis test" in the context of having an alternative hypothesis (the complement of the null) and allowing the possibility of accepting the null—but for making decisions, not for making scientific inference.

quality for: (1) temporal trends, (2) differences in populations (e.g., upstream/downstream differences and step trends), and (3) standards violations. These are the three types of information that are most often utilized in legally mandated management efforts (Ward *et al.*, 1990), and which can be used to interpret the quality of the water for regulatory, economic and legal purposes. Therefore, statistical methods used in modeling analyses (including multivariate analyses, time-series analyses and multiple regression techniques) were not included in this research, as these are used more often in exploratory studies and/or as predictive tools.

Part I. Current Routine Water Quality Data Analysis Methods

Criticisms of Water Quality Assessments

Water quality monitoring, and its attendant data analysis processes, are the primary means through which information about our nation's waters is developed. The manner in which monitoring is conducted and data analyzed has been the subject of considerable discussion, debate and criticism over the years (Ward, 1996). The latest two reports along this line are PEER (1999) and GAO (2000). Such criticism is not limited to water quality monitoring. The methods through which data in the medical and behavioral sciences are analyzed and interpreted are increasingly questioned (i.e., Nunnally, 1960; Carver, 1978; Berger & Berry, 1988; Fleiss, 1987; Chow & Liu, 1992; Goodman, 1993;Royall 1997; Veiland & Hodge 1998, Johnson, 1999) and some of these criticisms are now being applied to the means employed to traditionally analyze and interpret water quality data.

A recent report by an anonymous group of EPA and other agency employees criticizes the water quality assessments made by the States. It states that "inconsistencies in the amounts of waters monitored or evaluated as well as variations in how impairment and designated use attainment are measured, produce a hodgepodge of information that is of little value in determining national water quality trends or comparing water quality among individual States" (PEER, 1999).

Another report produced by the U.S. General Accounting Office reaches similar conclusions about the validity the EPA's *National Water Quality Inventory*, a compilation of all state water quality assessments [305(b) and 303(d) reports]. GAO (2000) states that this report can not meaningfully compare information across states because of considerable variation in: (1) the way states select their monitoring sites; (2) the kinds of tests states perform and how the results of these tests are interpreted; (3) the methods used to determine causes and sources of pollution; and (4) the analytical methods chosen to evaluate water quality (i.e., chemical, physical, or biological properties of water). "By aggregating these states' data, EPA is implicitly suggesting that these data can, in fact, be compared and in doing so is increasing the likelihood that the data will be misused or misinterpreted" (GAO, 2000).

While 15 recommendations are made in the PEER (1999) report to improve the 305(b) reports produced by states, as well as several recommendations by GAO (2000) to improve the usefulness of the *National Water Quality Inventory*, no recommendation is made in either report about how to improve the quality of information produced from states' monitoring systems. One key to such improvement involves better connecting the data analysis and interpretation methods with the management information sought. Though data analysis methods are rarely questioned, there are a small number of researchers and academics questioning the methods used to produce water quality information. The following review compiles the arguments brought forth by these critiques.

Similar to PEER (1999), a report of the Virginia Water Quality Academic Advisory Committee (Shabman *et al*., 1998) makes 17 recommendations to the Virginia Department of Environmental Quality to meet the General Assembly's Water Quality Monitoring, Information and Restoration Act requirements. These recommendations cover the water quality assessments used for 303(d) and 305(b) reporting. However, several of the recommendations directly address the statistical analysis methods used to produce information from water quality monitoring systems. Although it never recommends specific analysis methods for trend detection, in general the report recommends "improved explanations of current use of statistical inference procedures". It recommends incorporating the relationship to flow in the analyses for trends, and critiques current EPA recommended procedures for determining standards compliance, arguing an improved statistical procedure for determining this information. (Shabman *et al*., 1998)

A third recommendation from the Virginia committee is that the statistical power (i.e., sensitivity) of various temporal sampling patterns should be carefully reviewed in order to design a monitoring program which will optimize analysis opportunities (Shabman *et al*., 1998). This is a common theme in statistics, and more criticisms of testing without considering power will be discussed below. It is important to realize that when using common null hypothesis tests, obtaining a *statistically* significant result does not necessarily imply that one has found an *environmentally* significant result.[2]

The process through which water quality information is produced has become more targeted in the academic field in recent years. Many researchers are criticizing the appropriateness of the statistical procedures used to produce water quality information. From discrediting specific methods for inappropriate use, to rejecting entire categories of methods for inappropriate theory, the typical standard data analysis methods are increasingly being examined in an effort to improve information produced from monitoring.

The EPA Guidance (EPA, 1989 & 1992) prompted one critique of incorrect use of methods for Statistical Analysis of Groundwater Monitoring at RCRA (Resource Conservation and Recovery Act) sites. In this guidance it is recommended that for a data set with large numbers of non-detects, Poisson prediction limits and Poisson tolerance limits be used. Loftis *et al*. (1999) find that neither the Poisson distribution nor associated tolerance or prediction limits should be used with concentration data.

Another type of criticism is the issue of statistical power in monitoring design. "Many have noted the lack of attention paid to statistical power in research and monitoring programs" (Santillo *et al*., 1998). Statistical power is defined as the probability of detecting an effect where one exists. To be useful, the analysis tests chosen should have good power to detect environmentally important effects or trends.

Standard statistical procedures minimize Type I errors (error of a false positive), by specifying the significance level (α) before the tests are performed. However, efforts to minimize Type I errors lead to increases in Type II errors (denoted as β), an error of accepting the tested hypothesis when it is actually false. Power is therefore defined as $1 - \beta$, the probability of rejecting the null hypothesis when it is truly false.

This lack of attention to power considerations draws doubts to the capability of many monitoring programs to properly detect trends, because too few data points are available to give the analysis adequate power to detect important trends (Santillo *et al*., 1998). On the flip side of this argument is the fact that as databases grow in size, tests become too powerful, detecting ever-

[2] A statistically significant result is obtained when a sufficiently small "p-value" is obtained (i.e., $p \leq \alpha$, where α is the specified significance level). The p-value is defined as the probability of getting data at least as extreme as was obtained if the tested hypothesis were true (in the one-sided case the hypothesis is assumed to be only just true). This probability is typically obtained from tables prepared from standard results of mathematical statistics.

smaller differences, leading to unimportant differences turning out to be judged as statistically significant (McBride, 1999a).

Some promote the practice of using a power analysis after statistical tests have been applied (e.g., Zar, 1984), in order to determine the 'sensitivity' of the analysis method. Another publication (Johnson, 1999), made available on the Internet by the Northern Prairie Wildlife Research Center and the USGS, is critical of power analysis. As mentioned above, power analysis is used to determine the sample size needed to have a specified probability (power) of declaring as significant a particular difference or effect (Johnson, 1999). However, when power is determined *after* a test has been performed to guard against wrongly declaring the null hypothesis to be true, he claims that the results can be misleading. This retrospective power analysis, estimated with the actual data used and the observed effect size, is meaningless, as a high p-value will result in a low estimated power (Johnson, 1999). Power analysis programs, however, assume the input values for effect and variance are known, rather than estimated, so they may give misleadingly high estimates of power. The author states that the questions about the likely size of true effects can be better addressed with confidence intervals than retrospective power analysis (Johnson, 1999).

The criticism with potentially the most far-reaching force implies that null hypothesis testing can be inappropriate for environmental data. Such testing is the category of statistical analyses that examines a null hypothesis (positing no effect or trend *whatsoever*) or its alternative (positing a non-zero effect or trend), and determines if data constitute significant evidence against the null (via the p-value, as explained in footnote 2). "Unfortunately, when applied in a cookbook fashion, such significance tests do not extract the maximum amount of information available from the data" (McBride, Loftis & Adkins, 1993).

McBride *et al.* (1993) claim that null hypothesis testing has three problems, which are applicable in environmental monitoring:

1. A conclusion that there is a significant result can often be reached merely by collecting enough samples (increasing sample size increases chance of rejecting the null);
2. A statistically significant result is not necessarily practically significant; and
3. Reports of the presence or absence of significant differences for multiple tests are not comparable unless identical sample sizes are used.

For the past several years, the use of hypothesis testing in the medical profession has been questioned. The argument has been made that comparing p-values with "arbitrary" significance values (typically $\alpha = 0.05$) does not objectively prove that the data are displaying a characteristic that is not merely chance. In fact, it has been suggested by certain statisticians that p-values are "startlingly prone" to attribute significance to fluke results (Matthews, 1998). Discussions have been raised over the "evidential value" of a p-value, and what it really means in terms of proving anything (Gibbons & Pratt, 1975; Berger, 1986; Goodman & Royall, 1988; Schervish, 1996; Royall, 1997; Veiland & Hodge, 1998). Those with less knowledge of statistical theory mistakenly confuse it with the Type I error of hypothesis testing (α), and this link between the two has become standard, but misleading practice (Goodman, 1993). Some data analysts are questioning the appropriateness of using p-values at all with hypothesis testing (i.e., Berger, 1986; Berger & Berry, 1988; Goodman, 1993; Royall, 1997; Matthews, 1998), favoring instead a Bayesian or likelihood approach.[3]

[3] Null hypothesis testing uses "classical statistics", wherein probability is defined as frequency of events *in the long run*. This class of methods can therefore only contemplate probability of data given an assumed hypothesis (e.g., as in footnote 2, the p-value is the probability of gaining data at least as extreme as was

The water quality and biology fields are also addressing the confusion over using *p*-values to support significant findings. Johnson (1999) states that: (1) the *p*-value is often used as the probability that the results obtained were due to chance, (2) 1-*p* is often used as the "reliability" of the result, and (3) *p* is the probability that the null hypothesis is true.

"Unfortunately, all of these conclusions are wrong. The *p*-value is the probability of the observed data or more extreme data, given that the null hypothesis is true, the assumed model is correct, and the sampling done randomly" (Johnson, 1999). Determining which outcomes of an experiment or survey are more extreme than the observed one, so a *p*-value can be calculated, may require knowledge of the intentions of the investigator (i.e., the stopping rule) (Berger & Berry, 1988). "Hence, *p*, the outcome of a statistical hypothesis test, depends on results that were not obtained, that is, something that did not happen, and what the intentions of the investigator were" (Johnson, 1999). Such information and intentions are often not easily obtained.

Another common mistake in hypothesis testing is the notion that null hypotheses can be "accepted". But failing to reject a null hypothesis does not prove that it is true (Zar, 1984; Johnson, 1999). Especially with small samples, one must be careful not to accept the null hypothesis, as this is most probably merely a reflection of the lack of power (Johnson, 1999). Even more arbitrary is the designation that a result is "significant" if the *p*-value falls below some cut-off value, usually given as the maximum acceptable Type I error risk, α. This means that for tests using $\alpha = 5\%$, a *p*-value of 0.049 is significant for a one-sided test, whereas a *p*-value of 0.051 is not (Johnson, 1999). Such a minor difference can be deceptive, as it is derived from tests whose assumptions are often only approximately met (Preece, 1990).

Null hypotheses state that some parameter equals zero, or that some set of parameters are all numerically equal. Such hypotheses are almost invariably known to be false before any data are collected (Berkson, 1938; Savage, 1957; Johnson, 1995). If these hypotheses are not rejected, it is usually because sample size is too small (Nunnally, 1960) and so power is too low (Johnson, 1999). This is why such null hypotheses should never be "accepted".

In the field of drug testing, it has been agreed that testing a null hypothesis between means (which is standard practice in water quality data analysis) is not appropriate, as it is evident that the probability of rejecting the null hypothesis tends always to increase with sample size (Chow & Liu, 1992—the *p*-value grows smaller as the number of samples increases). A solution to this problem was suggested by Good (1982), who proposed that *p*-values be standardized to a sample size of 100, by replacing the *p*-value with $p\sqrt{(n/100)}$, where *n* is the "sample size" (i.e., the number of data).

An even more pertinent question would be: why test a null hypothesis at all, if it seems virtually impossible for two different drugs to have the same effect (McBride, 1998)? It has become common practice in drug testing to test whether or not a difference between means/medians might be within a prescribed interval, instead of exactly zero (Chow & Liu,

obtained *if* the tested hypothesis is true). Bayesian methods invert this, to calculate the probability of a hypothesis given the data obtained. In doing so they use Bayes' rule, whereby a prior belief (in the form of a probability distribution) is updated by the data to obtain a posterior probability distribution. Hence, this probability contains some personalistic content, with which some statisticians are uncomfortable. Likelihood methods use a form of Bayesian analysis restricted to the relative merits of competing hypotheses, rather than their absolute probabilities.

1992). As a consequence the *p*-value does not necessarily keep on decreasing as the number of samples is increased, and it is valid to infer that the tested hypothesis could be accepted.

Water quality guidance documents, such as the EPA's for statistical analysis of monitoring data at RCRA sites (1989, 1992) often recommend hypothesis testing, such as ANOVA. This type of test can be stated as the following: For the time period given, are the means of a water quality variable equal in all the wells sampled? Or is one or more different from the others? McBride *et al.* (1993) point out that as in drug testing, we know in advance there will be differences, so why perform the test at all? If there exists a statistically significant difference, this may not translate to a practical significant difference from a management point of view unless power is properly considered (not the norm).

McBride (1999a) explores this option further. He states that a recurring issue in statistical analysis has been the failure to use power analysis to select an appropriate sample size so as to minimize the risk either of failing to detect important differences or of detecting the unimportant. "Advocates of power analysis have been increasing in environmental science and management. However, there is discomfort with tests becoming too powerful, i.e., as sample size increases, tests of point hypotheses will tend to detect ever-smaller differences. One response is to de-emphasize the role of tests and rely on confidence intervals." However, McBride (1999a) chooses to support interval testing as a solution to the inappropriateness of testing a point null hypothesis.

Such problems, as discussed above, have led to a "significant test controversy" (Morrison & Henkel, 1970; Harlow *et al*., 1997) in the social and behavioral sciences, as well as water quality and biology, with the following remedial measures proposed:

1. Abandonment of testing hypotheses about differences in favor of estimation of differences (Oakes, 1986);
2. Use of interval tests (McBride, 1999a); and
3. Using a combination of estimation and testing with greater emphasis on statistical power in the design of monitoring systems and interpretation of significant test results (Millard, 1987).

McBride *et al.* (1993) suggest that the entrenchment of hypothesis testing in the environmental field makes its abandonment unrealistic, but do make several other recommendations related to those in the social and behavioral sciences. One recommendation supports the emphasis on statistical power, stating that both types of errors (Type I and Type II) should be considered when designing a sampling program. "In this way one can seek to have a higher probability of detecting a difference of practical significance (because Type II error is related to the difference in means), corresponding to a particular effect size (chosen by the analyzer), as well as a low probability of raising false alarms".

Another recommendation is to rely more on interval estimation rather than hypothesis testing. "In trend detection, more information is conveyed by plotting a trend line with confidence limits through a time series than by simplistic yes/no of hypothesis testing."

The final recommendation by McBride *et al*. (1993) refers to interval testing, in which the analysts test whether or not the difference in means is greater than some prescribed interval. "An advantage of this test is that the analyst must state the difference of practical significance to management, also the failure to reject the null no longer induces complacency". This is because the results now mean something, ecologically and environmentally.

Despite its criticisms, null hypothesis testing is still widely used and accepted to develop information from all sorts of data, especially in the water quality field. This prevalence will be demonstrated in the next section. Despite its drawbacks, some advocate more appropriate types of hypothesis testing (i.e., the discussion of interval testing by McBride *et al*., 1993), as well as greater attention to the details of the test, including power analysis, sample size and stating the hypothesis. All of these discussions and criticisms help to illustrate the need for more careful attention paid to the selection of analysis methods when the ultimate goal is defensible and comparable information.

Current Water Quality Data Analysis Procedures

In this section, current practice and "state-of-the-art" procedures used to analyze water quality data for information purposes are examined. The review of literature focuses on the use of statistics to produce information, not summary statistics. This information, as discussed below, is limited to common information needed by management, i.e., temporal trends, differences in populations, and standards compliance. The extent of the review covers the major entities involved in water quality monitoring assessments, including the United States Geological Survey (USGS), U.S. Environmental Protection Agency (EPA), private groups and academia, and determines if there are 'standards' emerging for the analysis of water quality data, as a whole or within organizational structures. The review covers environmental statistics textbooks, agency publications, water quality reports from state environmental agencies, and refereed journals.

When beginning this literature review it was thought there might be *de facto* standards for data analysis developing in the water quality field. Use of the term *standard* is not meant to imply that there is an established set of statistically based data analysis methods that have been reviewed and recommended for all water quality monitoring situations. However, this section will attempt to establish that there are certain methods being used time and time again by a variety of monitoring entities, depending on the type of information sought. Conclusions will address whether or not *de facto* data analysis standards are emerging in the analysis of water quality data.

Recommended Guidance for Statistical Analysis of Water Quality Data

The first step in trying to establish whether *de facto* standard procedures exist was to search for guidance, or widely available and accepted protocols, for water quality data analysis. In the search for guidance on data analysis methods, it appears that no major entity has established a set of comprehensive *standard* data analysis methods, or methods with which to interpret results of data analysis into information for management.

There exist several textbooks that directly address statistical analysis procedures for environmental data (e.g., Gilbert, 1987; Ward *et al*., 1990; Helsel & Hirsch, 1992). These texts provide numerous options for analyzing data, often categorized by the information needed (in statistical terms). The inclusion or omission of certain methods in the texts might be viewed as a type of guidance, yet none of these methods outline protocols through which to infer information for management decision making from the analysis results.

The USGS has no published defined guidance for analysis of water quality data, but does have the largest collection of published water-quality assessments. In these studies, authors often site USGS publications as the basis for selecting data analysis methods. For example, Helsel & Hirsch (1992), the textbook mentioned above, is commonly cited as a reference for using the

Seasonal Kendall test for detecting trend. In Hirsch (1988), the Hodges-Lehmann class of estimators is found to be robust in comparison to other nonparametric and moment based estimators for determining the magnitude of changes of various constituents between two time periods (step trends). By the fact that they are commonly cited in many USGS water quality studies, these types of publications serve as guidance for water quality data analysis in the USGS.

In an academic study, Montgomery & Reckhow (1984) recommend certain techniques for detecting trends in lake water quality, and go on to recommend these procedures for other water bodies as well. Another academic study, Montgomery & Loftis (1987), explored the applicability of the *t*-test for detecting trends in water quality variables. The results of this study "suggest that the *t*-test is robust for non-normal distributions if the distributions have the same shape and sample sizes are equal". It is also robust for unequal variances if the sample sizes are equal. If either of these considerations is not met, as well as the presence of serial dependence or seasonality, then the *t*-test is not a robust test to detect a step trend. Another non-agency study, Harcum *et al*. (1992), recommends using the Seasonal Kendall and Mann-Kendall tests for trend detection, depending on the data attributes.

Using a study conducted in New Zealand to determine effects of alluvial gold mining operations on benthic invertebrate communities, McBride (1998) demonstrated that traditional point hypothesis tests may not provide satisfactory answers to questions of environmental impact, because they might not be asking or addressing the right questions. Using the theories of interval testing, it is possible to set-up the data analysis in two different ways, one with a hypothesis that the differences between population means are equivalent (within a prescribed interval), or one in which they are inequivalent (beyond that interval). The information produced from using each of these hypotheses is very different, and reflects an emphasis or non-emphasis on environmental protection, a key point to environmental management. Testing the hypothesis that the streams are equivalent (which is therefore not a *null* hypothesis) protects the environmental user's risk, resting the "burden of proof" on the monitoring system to show that an impact has occurred. However, the latter approach of testing a null hypothesis of inequivalence is a more "precautionary" approach, assuming the stream has been impacted, unless proven otherwise (McBride, 1998). This study serves as guidance by demonstrating the importance of complete understanding of the implications behind each hypothesis to management decision-making, as well as the importance of determining the test hypothesis before analysis, as information can change depending on the structure of the hypothesis.

A type of graphical display that has become more widely recommended and used in data analysis is the box plot. McGill *et al*. (1978) describes three variants of the box plot display, which are used in exploratory data analysis and visual summaries. Although the authors explain that the user's personal preference is the best criterion for interpretation, this article suggests that graphical displays of data "provide insight into the meaning of the data without the possibility of misinterpretation due to unwarranted assumptions".

The largest collection of guidance for data analysis was found in publications by the U.S. Environmental Protection Agency. Guidance has been published by the EPA for the states' submittal of 303(d) lists and 305(b) reports. Numerical and narrative criteria to determine use support are recommended in the biannual guidelines, however no specific statistical or scientifically defensible data analysis methods appear to be endorsed by the organization for the information required in these reports.

EPA appears to publish guidance that helps the states and other reporting entities compile and interpret information to support specific EPA rules and programs (e.g., *Information*

Collection Rule: Draft Data Analysis Plan: EPA, 1997b; *The Monitoring Guidance for the National Estuary Program*: EPA, 1992; *Monitoring Guidance for Determining the Effectiveness of Nonpoint Source Controls*: EPA, 1997c; and *Statistical Analysis of Groundwater Monitoring Data at RCRA (Resource Conservation Recovery Act) Facilities*: EPA, 1989;1992).

EPA also has research publications that can be viewed as recommendations for particular methods. In Loftis (1989), seven statistical tests for trend were evaluated under various conditions and performance was compared using actual significance level and power. The evaluations resulted in the following recommendation by the authors: for annual sampling use the Mann-Kendall test for trend, and for seasonal sampling, use either the Seasonal Kendall test or the Analysis of Covariance (ANOCOV) on ranks test. A guidance document for determining improvements from agricultural non-point source control programs was developed and published by North Carolina State University for the EPA (Spooner *et al*., 1985). These authors give recommendations on monitoring design, appropriate hypotheses, data requirements, assumptions, and testing procedures.

With the exceptions discussed above, attempts to produce standard sets of guidance procedures for water quality data analysis are relatively few and uncoordinated between agencies. To illustrate, in the field of groundwater monitoring, Adkins (1992) states that "due to the wide variety of information needs and site conditions, it is impractical to expect a single data analysis protocol to be suitable for all groundwater quality monitoring systems…[and that] no generally acceptable design framework for the development of groundwater quality data analysis protocols exists today". Therefore, instead of producing a guidance recommending specific analysis procedures, Adkins (1992) presents a framework for individual development of groundwater quality data analysis protocols, a positive step towards making information more comparable.

Peer Reviewed Water Quality Assessments

Although general 'standard' methods for water quality monitoring analysis may not be published, it is hypothesized that they are established through common practice, especially within organizations and types of monitoring entities.

This section reviews the current use of statistics, beyond guidance, in the water quality field. To gain an overview of the use of statistics, recent issues of five major environmental refereed journals were examined: Journal of the American Water Resources Association, Environmental Monitoring and Assessment, Environmental Management, Water Resources Research, and Marine Pollution Bulletin. The peer-reviewed studies included here are limited to those that sought information related to environmental management: temporal trends, differences in population (including upstream/downstream differences, before/after differences, and spatial differences), and standards compliance.

Trend Analyses

Most trend analyses were performed with non-parametric tests for trend in order to avoid complications in the data set and assumptions of normality, and, to make the tests more robust. The most popular method was the Seasonal Kendall test (seasonal extension of the nonparametric Mann-Kendall test) for monotonic trend, used in 12 out of the 19 studies where trend was determined (highlighted in gray, Table I). It is especially popular with USGS studies. The USGS is also very thorough about performing the test on both the original data and flow-adjusted concentrations, but only if a strong correlation exists between concentration and flow. All trend detection studies reviewed are summarized in Table I.

Differences in Populations

There were a greater variety of tests chosen to determine differences in population. Three major groups of analyses prevailed: (1) using Signed Rank, Rank Sum or variations of those procedures, (2) using cluster type analyses, and (3) using ANOVA or variations. The most popular tests were the Wilcoxon Rank-sum/Mann-Whitney test or its extension for more than 2 populations, the Kruskal-Wallis test (8 out of 20 studies reviewed, light gray highlight in Table II) and the Analysis of Variance test (ANOVA used in 5 out of 20 studies, dark gray highlight in Table II). Most studies tested for normality before choosing a hypothesis test, though some just assumed nonparametric statistics should be used. Almost all the tests used were for nonparametric distributed data. With the exception of Dennehy *et al*. (1995), no hypotheses were stated. But it was evident by the testing that all performed a hypothesis test with a point null hypothesis of the means/medians between groups being equal. The USGS studies seemed to prefer the Wilcoxon Rank-Sum (Berndt, 1996; Abeyta & Roybal, 1996) or Kruskal-Wallis test (Abeyta & Roybal, 1996; Dennehy *et al*., 1995; McMahon & Harned, 1998; Mueller, 1995). All of the studies reviewed are summarized in Table II.

Standards Compliance

Determination of standards compliance was not commonly sought via statistical tests in the research type assessments reviewed (see Table III for summary of assessments which involved standards compliance). Therefore, part of this literature review attempts to describe how states generate this information for their 303(d) and 305(b) reporting requirements, especially in light of the current 303(d) listings and Total Maximum Daily Load (TMDL) debate. Many states do not publish their assessment methodologies, so personal communication via the phone and/or email was the primary venue through which such information was gathered. The purpose was to try and establish if there are common methods used by the states for their water quality assessments, not to document every detail of every state's assessment methodology. The following states responded: *New York, New Jersey, Region III (Delaware, Pennsylvania, Maryland, Virginia, West Virginia, District of Columbia)*, *Oklahoma, Arizona, Hawaii, Virginia, Kentucky, California, South Carolina, Florida, Tennessee, North Carolina, Alabama.* It was found that documented analysis methods or statistical tests are rarely used to determine use support assessments or standards violations. Often only simple "percentage of standard exceedences" is used to assess a water body, along with subjective evaluation of the waterbody according to narrative criteria. For summaries of each state's methodologies, refer to Martin (2000).

Table I: Water Quality Assessments Involving Trend Detection

Author	*Monitoring Entity*	*Distribution Assumption*	*Actual Hypothesis Stated*	*Test Used*
Clow & Mast (1999)	USGS	NP	None stated	Seasonal Kendall Tau or Mann-Kendall
Baldys, *et al.* (1995)	USGS	NP	Null hypothesis of no significant trend	FAC Seasonal Kendall Tau or Mann-Kendall
Mattraw, *et al.* (1987)	USGS, NPS and SFWMD	NP	None stated	FAC Seasonal Kendall Tau or Mann-Kendall
Rinella (1986)	USGS	NP	None stated	FAC Seasonal Kendall Tau or Mann-Kendall
Berndt (1996)	USGS	NP	None stated	Seasonal Kendall Tau or Mann-Kendall
Mueller (1995)	USGS	NP	None stated	FAC Seasonal Kendall Tau or Mann-Kendall
Mueller (1990)	USGS	NP	None stated	FAC Seasonal Kendall Tau or Mann-Kendall
Snyder *et al.* (1998)	Academia	NP, Parametric	Null = no tendency for one sampling location to have nutrients greater than another location	Duncan's new multiple range test (Ott, 1988) - test of the difference in means of multiple populations, % reduction of means
Stoddard *et al.* (1998)	EPA, Academia, Vermont DEC	NP	None stated	SKT, Analysis of Chi-squares and meta-analysis
Pinsky *et al.* (1997)	EPA, Academia	NP, Parametric	None stated	Auto-regressive first order process, comparing means/medians
Takita (1998)	Susquehanna	NP	None needed	Double mass comparison
Havens *et al.* (1996)	SFWMD	Parametric	None stated	Satterthwaite's corrected t-test
Dennehy *et al.* (1995)	USGS	NP	Null states that no trend exists	LOWESS (to highlight patterns), FAC SKT
Butler (1996)	USGS	NP, Parametric, Parametric, NP	Null means there is no trend or no sig. diff between means/medians	FAC SKT (periodic & monthly), FAC LR (annual), Step Trend two sample t-tests, Wilcoxon Rank Sum
Smith, *et al.* (1987)	USGS	NP	None stated	SKT and FAC SKT
Vaill & Butler (1999)	USGS	NP	Null hypothesis of no trend	monotonic trends: SKT and FAC SKT, Sen Slope estimator, Lowess to determine in what part of the record the trend occurred. Step trends: Parametric 2-sample t-test and NP Wilcoxon rank-sum test applied to raw data
Heiskary, *et al.* (1994)	Minnesota Pollution Control Agency	NP	Null hypothesis of no trend	Kendall's tau-b (Gilbert, 1987)
Lavenstein & Daskalakis (1998)	NOAA	NP	None stated	Kendall-tau test for linear correlation
Brown *et al.* (1998)	NOAA	NP	None Stated	Spearman-rank Correlation method, meta-analysis

Table II: Water Quality Assessments Involving Differences in Populations

Author	*Monitoring Entity*	*Distribution Assumption*	*Actual Hypothesis Stated*	*Test Used*
Younos *et al*. (1998)	VWRRC, Academia	NP	None stated	Wilcoxon Test (Hollander & Wolfe 73)
Arthur, *et al.* (1998)	Academia	NP	None stated	Wilcoxon Signed Rank
Berndt (1996)	USGS	NP	None stated	Wilcoxon Rank-Sum
Pinsky *et al*. (1997)	EPA, Academia	NP, Parametric	None stated	Wilcoxon Rank-Sum, Chi-Square test of hypothesis of equal proportions in population
Abeyta & Roybal (1996)	USGS	NP, NP, NP, Parametric	None stated	Wilcoxon Rank-Sum, Kruskal-Wallis, ANOVA, ANOVA & paired t-tests
Sample *et al*. (1998)	USDA NRCS	NP, NP, NP	None stated	Rank Sum, Signed Rank, Hodges-Lehmann Estimator
McMahon & Harned (1998)	USGS	NP	None stated	Kruskal-Wallis, and Tukey's Multiple Comparison
Mueller (1995)	USGS	NP	None stated	Kruskal-Wallis
Koebel, *et al.* (1999)	SFWMD	NP, NP	None stated	TSS, Turbidity, Nutrients - Kruskal-Wallis, Dunn's test, ANOVA & paired t-tests
Momen *et al*. (1997)	Academia	Parametric, Parametric	None stated	Tukey's multiple comparison for mean separation, ANOVA (temporal and spatial)
Takita (1998)	Susquehanna	NP	None needed	Plotted Annual Loads vs. Discharge Ratio
Dennehy *et al*. (1995)	USGS	NP	Null states that no difference exists	Kruskal-Wallis test
Snyder *et al*. (1998)	Academia	NP?	None stated	Friedman's test (Gilbert, 1987), Cluster Analysis (Davis, 1986), Cross-Correlation Analysis
Stoe (1998)	Susquehanna	Parametric?	None stated	PCA, Cluster analysis, Habitat Assessment scores and Biological Condition scores
Nimmo *et al*. (1998)	USGS, EPA, Academia, CDOW	Parametric	None stated	ANOVA & paired t-tests, Student-Newman-Keuls method of separating means
Colman & Clark (1994)	USGS	NP	None stated	ANOVA
Rinella (1986)	USGS	NP	None stated	Tukey's multiple comparison
Kennedy (1995)	TxDOT, North Central Texas COG	NP	None stated	Kruskal-Wallis test, Mann-Whitney test
Kress, *et al.* (1998)	Israel Oceanographic and Limnological Research	Parametric	None stated	GLM least squares, t-test, Mann-Whitney a-parametric test
Brown *et al*. (1998)	NOAA	NP	None stated	GT2 multiple comparison method

Table III: Water Quality Assessments Involving Standards Compliance

Author	*Monitoring Entity*	*Distribution*	*Hypothesis Stated*	*Test Used*
Berndt (1996)	USGS	NP	None stated	% exceedence of MCL, highest means reported
Lapp *et al*. (1998)	Academia	NP	None stated	observed mean does not exceed DW standard in Canada
Nimmo *et al*. (1998)	USGS, EPA, Academia, CDOW	Parametric	None stated	average concentrations compared to chronic 4-day aquatic life criterion (USEPA)

Literature Review Summary

This review indicates that many types of analyses are being used to provide information about water quality. The first major conclusion is that although there are some who criticize null hypothesis testing, this type of analysis is alive and well in the field of water quality. It is interesting to note that although it does seem to be popular, as evidenced by its inclusion in guidance documents and water quality studies, the actual hypothesis tested is rarely reported, despite recommendations to the contrary in many of the guidance documents (Gilbert, 1987; Ward *et al*., 1990; Helsel & Hirsch, 1992; Montgomery & Reckhow, 1984; EPA, 1992; EPA, 1997c).

With a few exceptions (Heiskary *et al*., 1994; Momen *et al*., 1997; EPA, 1992; EPA, 1997c), the power of hypothesis testing is not considered. The weight of evidence in making a decision about trends or differences in populations relies solely on the acceptable Type I error (α) and the obtained p-value.

The literature review does not support the conclusion that there exist *de facto* standards for data analysis. The review of refereed journals found a large variety of graphical, statistical, and estimation analysis techniques. EPA provides many types of guidance for different regulatory programs, yet the analysis recommendations differ between programs, and efforts do not seem to be coordinated between programs. It *was* apparent that specific methods were preferred by the USGS for trend detection (Seasonal Kendall test) and differences in populations (Wilcoxon Rank-Sum/Kruskal-Wallis and ANOVA).

The major commonalties to all the data analyses performed was that with a few exceptions: (1) justification was rarely given for choosing a certain test beyond the data being parametric or nonparametric, (2) the hypothesis tested was rarely stated, (3) alternative analysis methods, if explored, were not reported, and (4) the power (or sensitivity) of the hypothesis test was rarely calculated.

Given the extremely wide array of data analysis methods being employed in producing information about water quality conditions, there is little reason to expect that comparable information is being produced in support of water quality management decision-making.

Part II. Information Comparability of Data Analysis Methods

The previous section was dedicated to compilation of information in order to determine how water quality data are being analyzed for information purposes. Recent criticisms of statistical hypothesis testing have questioned the main process through which information is produced from water quality data, i.e., hypothesis testing. Nevertheless, the literature review of current practice established that using hypothesis testing (often called "significance testing") is accepted in texts, guidance documents, and water quality studies published in refereed journals.

The literature review also establishes that there are a wide variety of methods that are available for data analysis. Many times, those who are analyzing water quality data are not statisticians, and rely on these texts, guidance documents, and observations of previous studies to select the analysis methods.

The purpose of this section is to document the connections between selection of data analysis methods and the comparability of the information produced. Using a high quality data set provided by the New Zealand National Institute of Water & Atmospheric Research (NIWA), several different analysis methods were performed in the areas of trend detection, differences in populations, and standards compliance. The results of the different methods within each area were compared in order to illustrate how information changes depending on the analysis methods used.

Three statistical packages were utilized in the data analysis procedures. WQStat Plus™ (Version 1.56, developed by Intelligent Decision Technologies Ltd., Loveland, Colorado) was chosen for its inclusion of nonparametric procedures, easy flow-adjustment and water quality data analysis focus. Minitab™ (Release 12, developed by Minitab Inc.) was chosen because of its broad base of statistical procedures, both parametric and nonparametric. MS-Excel™ (part of the Microsoft Office package) was also used for its basic statistical functions and ease of data manipulation (the data used was originally received in MS-Excel™ format). Comparison of results of like tests between statistical packages should also help to demonstrate the variability of information.

Demonstrating Various Statistical Methods on New Zealand Data

The New Zealand River Network data set was chosen for analysis because of its high quality and accessibility. The data record is from a 77 river-site monitoring network distributed throughout New Zealand's North and South Islands (Smith *et al*., 1996). The monitoring network's design is well documented and the network has been operated consistently over its 10-year life with excellent quality control procedures in place. The data was readily made available, in an easy to use format (MS-Excel™ Spreadsheets) for purposes of this study (refer to Martin, 2000 for actual data used in this study).

The format of the New Zealand data allowed for easy transition to data analysis, a reason that this particular set was chosen. The New Zealand data was accompanied by meta-data that described the monitoring sites, how the samples were collected and analyzed, and all other ancillary data that would be of use to a data analyst (i.e., dates and units of measurement). Censored data (e.g., nondetects) were not used in this data, as all concentrations were reported. A few sites had missing data for certain dates, which were represented with a period (.) in the appropriate worksheet cell.

The only manipulation required for importation of the data into WQStat Plus™ and Minitab™, was cutting and pasting of the data columns into the appropriate format for the respective software. The required formats were described in the software user manuals.

Selection of Three Sites and Constituents for Data Analysis

Not all sites or constituents of the River Network were analyzed as part of this demonstration. Sites and constituents were chosen upon review of Smith *et al.* (1996) and with input from Graham McBride, NIWA, Hamilton, New Zealand. Descriptions of the sites were provided in the appendices of the New Zealand data set (Bryers, 1999) and are given below. For purposes of this study four data records at four sites were selected, as follows:

A. Site HM4 for BOD_5. This site is on the Waikato River, and is located downstream of the catchment area. It has potential impacts from agriculture, paper and pulp industries, and has additional inputs from Hamilton, Ngaruawahia, Huntly, thermal power stations, swamps, pasture and coal mining. (Bryers, 1999) The New Zealand Trends paper (Smith *et al.*, 1996) failed to detect any trends for BOD_5 after the first 5 years at this site.

B. Site RO2 for NH_4 analysis.[4] This site is on the Tarawera River, a major river in the area, downstream of major pulp and paper industries and exotic forest plantations. There is agricultural pasture in the valley. (Bryers, 1999) The New Zealand Trends paper (Smith *et al.*, 1996) showed an upward trend in NH4 at the 5% level (i.e., using a significance level of $\alpha = 5\%$) for the first 5 years at this site.

C. Site RO1 for NH_4 analysis. This site will only be used in the differences in population analysis. RO1 is upstream of site RO2 (above) on the Tarawera River. Between the two sites are potential environmental impacts from pulp mills (especially the Tasman Pulp and Paper Company's Kraft pulp mill), farming, a town (Kawerau), and a geothermal area (Bryers, 1999). This site was used as an upstream site for differences in population's analysis only.

D. Site HM6 for NO_x data.[5] This site is not downstream of any urban sources, but is a major tributary of the Waihou River. It contains or will contain discharges from several large gold mining operations as well as agricultural impacts from some pasture usage. (Bryers, 1999) The New Zealand Trends paper (Smith *et al.*, 1996) showed an upward trend of NO3 at the ($p<5\%$) level after the first 5 years.

In order to illustrate the importance of distribution assumption in hypothesis testing, it was each data set was tested for normality. This was accomplished using the Chi-Squared Goodness-of-Fit Procedure in WQStat Plus™ (IDT, 1998), which tests the following hypothesis:

$\mathbf{H_0}$: the data are normally distributed *vs.* $\mathbf{H_A}$: the data are not normally distributed

Flow adjustment of the raw data was performed only in WQStat Plus™, as this was the only package that had the ability to directly approximate the flow-adjusted concentrations. This

[4] Ammoniacal nitrogen, i.e., $NH_4 = NH_3\text{-N} + NH_4^+\text{-N}$

[5] Oxidized nitrogen, i.e., $NO_x = NO_2\text{-N}$ and $NO_3\text{-N}$

procedure was used to help determine how flow can affect or change the information produced from the monitoring data. Flow adjusted concentrations (FAC) were used in normality testing, and trend detection testing in order to give an indication of how information can change when flow effects are taken into account.

Statistical Methods Used to Determine Trends

Analysis of the New Zealand data set for trends includes data from all ten years. As a means of additional quality control on the information being produced, analysis of the first 5 years was compared to the same analysis performed by a study published after the first 5 years of New Zealand's monitoring effort, entitled *Trends in New Zealand's National River Water Quality Network* (Smith *et al.*, 1996). The second 5-year data was also analyzed separately, as well as a comparison of both 5-year analyses to an analysis of the 10-year data. Analyses were performed on raw data and flow-adjusted concentrations (FAC). The following statistical trend tests were performed, both testing the following hypothesis.

- Mann-Kendall Test/Sen Slope Estimator – WQStat Plus™.
- Seasonal Kendall Test – WQStat Plus™. The Seasonal Kendall Test was also used to test for trends in flow data at the three sites chosen: HM4, RO2 and HM6 for both 10-year data and each 5-year data set. This was performed to help in interpretation of the flow-adjusted trend results.

Both tests apply to these hypotheses:

$\mathbf{H_0}$: No trend exists over time *vs.* $\mathbf{H_A}$: An upward or downward trend exists over time

Statistical Methods Used to Determine Differences in Populations

The difference in population analysis was performed between the first 5-year and second 5-year data sets for the sites HM4, HM6 and RO2, as well as a test between sites RO1 (upstream) and RO2 (downstream) for NH_4. The following tests, listed below, were performed for comparability of results. For further demonstration of comparability of results, the two-sample *t*-test was performed in both MS-Excel™ and Minitab™, and the Mann-Whitney test was performed in WQStat Plus™ and Minitab™.

- Two sample parametric *t*-test—MS-Excel™ and Minitab™. The *t*-test assuming equal variances was performed for site HM4, HM6 and RO2 data. For comparison of RO1 to RO2, the *t*-test for unequal variances was used, based on *F*-test for equal variances results (see Martin, 2000). The following hypotheses are tested:

  $\mathbf{H_0}$: $\mu_x = \mu_y$ (the means for groups *x* and *y* are identical) *vs.*
  $\mathbf{H_A}$: $\mu_x \neq \mu_y$ (the means for groups *x* and *y* are not equal).

- Mann-Whitney test – WQStat Plus™ and Minitab™. This nonparametric method tests the following hypotheses:

H_0: the medians of two populations are equal *vs.*
H_A: the medians of the two populations are not equal.[6]

- Interval Tests – MS-Excel™. This is a slightly more elaborate form of a parametric *t*-test. Interval tests are largely used in the pharmaceutical industry involved in drug-testing analyses (Chow & Liu, 1992). The hypothesis for an interval test can take one of two forms, one testing for equivalence between groups, and one testing for inequivalence (McBride, 1999a). The difference between these two tests is that in the equivalence test, the tested hypothesis is that the populations are ecologically equivalent (in which case the difference is no greater than the interval), whereas in the inequivalence test, the tested hypothesis is that they are not (in which case the difference *is* greater than the interval). The former hypothesis adopts the conservative stance commonly used in science; the latter is precautionary. Both tests recognize that the means will be different, but not necessarily equivalent. (McBride, 1999a)

  The interval chosen for these tests in this analysis was ±20% of the mean of the upstream or background data. While this was arbitrarily chosen, the estimates provided in McBride (1998) served as a guide for the magnitude (in which the ±20% figure was nominated by a benthic ecologist—Dr J. Quinn, NIWA). The purpose is to illustrate how different data analysis methods affect information. Establishing an equivalence interval requires knowledge of the behavior and effect of each constituent in the environment, something which is beyond the scope of this study.

  A highly detailed explanation of the development of this type of testing used for environmental data can be found in McBride (1999a). The algorithm through which the tests were performed in MS-Excel™ can be found in Martin (2000). These procedures use the following hypotheses for testing the equivalence hypothesis (McBride, 1999a), where d_L and d_U are the lower and upper bounds of the equivalence interval and $d = \mu_x - \mu_y$ is the actual difference in means:

  H_0: $d_L \le d \le d_U$ (the difference in means lies within the equivalence interval),
  H_A: $d < d_L$ or $d > d_U$ (the difference in means lies beyond the equivalence interval).

  The hypotheses for an inequivalence test is:

  H_0: $d \le d_L$ or $d \ge d_U$ (the difference in means lies beyond the equivalence interval),
  H_A: $d_L \le d \le d_U$ (the difference in means lies within the equivalence interval).[7]

Statistical Methods Used to Determine Compliance (Standards Violations)

For these tests the New Zealand standard for BOD_5 was compared to the data for BOD_5 from site HM4. Although the country has few national numerical standards, 2 or 3 ppm is often the accepted limit set by waste load allocations (McBride, 1999b). The data set for site HM4

[6] Note that with stricter assumptions (i.e., distributions are identical in shape but shifter in location), this is also a test on means (not just medians, Conover 1980:217), as in WQStat Plus.
[7] Some authors use "<" in place of "≤", and vice versa. Similarly for ">" and "≥". This is of no consequence for continuous variables, and the probability of equality is zero.

never exceeded 3 ppm, so for the purposes of this illustration, the excursion limit was set at 2 ppm. The following estimations (not hypothesis testing) were used (all using WQStat Plus™):

- Proportions
- Tolerance Limits
- Tolerance Interval
- Confidence Interval
- Prediction Limits

Results of Data Analysis

The following section examines the results of applying the methods discussed above. Particular attention is paid to comparing the differences in results (i.e., information) that are consequences of changing the analysis method. It is the lack of comparable information resulting from arbitrary selection of data analysis methods that is the focus of the results presentation.

Testing for Normality

All data sets were tested for normality in order to interpret the resulting information from parametric and nonparametric hypothesis tests. Raw vs. flow-adjusted concentrations (FAC) affected the outcome of this test (Table 4). Most data sets tested failed to reject the null hypothesis that they were normally distributed. However, as discussed earlier, failure to reject a null hypothesis does not prove that it is true. This is why there is a question as to whether these data are normally distributed or not (a common problem in water quality data analysis). (see Martin, 2000 for detailed results)

Table 4: Normality Testing Results

Site_Constituent	Hypothesis Test Result	Conclusion
RO1_NH_4 (raw)	Reject the null hypothesis	Not normal
RO2_NH_4 (raw)	Fail to reject the null	Cannot prove normal
RO2_NH_4 (FAC)	Fail to reject the null	Cannot prove normal
HM4_BOD_5 (raw)	Reject the null hypothesis	Not normal
HM4_BOD_5 (FAC)	Fail to reject the null	Cannot prove normal
HM6_NO_x (raw)	Fail to reject the null	Cannot prove normal
HM6_ NO_x (FAC)	Fail to reject the null	Cannot prove normal

Results for Trend Detection

This analysis compared the Mann-Kendall/Sen's Slope Estimator (MK) for trend with the Seasonal Kendall (SKT) test on 10-year data, raw and flow-adjusted (FAC), as well as the 1st and 2nd 5-year data. All calculations were performed using WQStat Plus™ (see Tables 5-7).

Table 5: Trend Detection Results for Site HM6, Constituent NO_x

Data	Test	Results	Slope Estimate
10 yr – flow	*SKT*	*⇓ ($p < 0.2$)*	*-0.11 units/year*
10 yr – raw	MK	Fail to reject null of no trend	3.0 units/year
10 yr – raw	SKT	Fail to reject null of no trend	1.9 units/year
10 yr - FAC	MK	⇑ ($p < 0.05$)	11.1 units/year
10 yr - FAC	SKT	⇑ ($p < 0.1$)	9.0 units/year
1st 5 yr – flow	*SKT*	*⇓ ($p < 0.05$)*	*-0.88 units/year*
1st 5 yr – raw	MK	Fail to reject null of no trend	-9.4 units/year
1st 5 yr – raw	SKT	⇓ ($p < 0.2$)	-28.3 units/year
1st 5 yr - FAC	MK	⇑ ($p < 0.05$)	36.8 units/year
1st 5 yr - FAC	SKT	⇑ ($p < 0.1$)	28.4 units/year
2nd 5 yr – flow	*SKT*	*Fail to reject null of no trend*	*-0.13 units/year*
2nd 5 yr - raw	MK	Fail to reject null of no trend	8.0 units/year
2nd 5 yr - raw	SKT	Fail to reject null of no trend	20.1 units/year
2nd 5 yr - FAC	MK	⇑ ($p < 0.1$)	27.2 units/year
2nd 5 yr - FAC	SKT	⇑ ($p < 0.2$)	23.3 units/year

Statistically significant results on flow values are italicized.

In Table 5, findings are similar for both tests, but not exactly so. It is often standard practice to choose an acceptable significance level as α = 0.05 (stated in WQStat Plus as “95% Confidence Level”). If that were the case in this analysis, only the Mann-Kendall test would have detected any trends in the 10-year flow-adjusted concentrations and the 1st 5-year flow-adjusted concentrations. WQStat™ gives results for various alphas up to 0.2 (i.e., minimum 80% confidence) and so allows the user to see the alpha giving a significant result. These results illustrate that findings can change by choosing a significance level (α) after results are obtained.

Flow-adjusted concentrations changed the outcome of the trend test upon examination of the trendline in the time series plot and in the 1st 5-year hypothesis test, as the direction changed from downward to upward trend. The slope estimators seem to have similar (i.e., comparable) results (see Martin, 2000 for Trend Analysis results). It is interesting to note that where a downward trend in flow existed, so did an upward trend in constituent concentration in flow-adjusted concentrations, but not exclusively. This finding could aid in the interpretation of the temporal behavior of the constituent.

Table 6: Trend Detection Results for Site HM4, Constituent BOD_5

Data	Test	Results	Slope Estimate
10 yr – flow	*SKT*	*Fail to reject null of no trend*	*-4.69 units/year*
10 yr – raw	MK	⇓ ($p < 0.01$)	-0.033 units/year
10 yr – raw	SKT	⇓ ($p < 0.05$)	-0.033 units/year
10 yr – FAC	MK	⇓ ($p < 0.01$)	-0.034 units/year
10 yr – FAC	SKT	⇓ ($p < 0.05$)	-0.036 units/year
1st 5 yr – flow	*SKT*	*⇓ ($p < 0.05$)*	*-45.2 units/year*
1st 5 yr – raw	MK	Fail to reject null of no trend	-0.016 units/year
1st 5 yr – raw	SKT	Fail to reject null of no trend	0.0 units/year
1st 5 yr – FAC	MK	Fail to reject null of no trend	-0.039 units/year
1st 5 yr – FAC	SKT	Fail to reject null of no trend	-0.031 units/year
2nd 5 yr – flow	*SKT*	*Fail to reject null of no trend*	*1.21 units/year*
2nd 5 yr – raw	MK	Fail to reject null of no trend	-0.028 units/year
2nd 5 yr – raw	SKT	Fail to reject null of no trend	-0.046 units/year
2nd 5 yr – FAC	MK	Fail to reject null of no trend	-0.025 units/year
2nd 5 yr – FAC	SKT	Fail to reject null of no trend	-0.044 units/year

Note that the slopes for the MK test on the raw data for the first and last five years (-0.016 and -0.028 units/year) do not straddle the 10-year result (-0.033 units per year). This is not an error; one may easily demonstrate that such straddling does not always occur.

These findings (i.e., Table 6) illustrate how hypothesis tests are more likely to detect a trend as sample size increases, a phenomenon common to all the tests performed in this chapter. No trend was detected in either 5 years of data, but was detected in the 10-year data. However, because the sample sizes are different, results (i.e. the *p*-value) from these two sample sets can not be directly compared. (See Martin, 2000 for complete results) Determination of flow trend did not reveal anything about flow-adjusted constituent behavior.

Table 7: Trend Detection Results for Site RO2, Constituent NH4

Data	Test	Results	Slope Estimate
10 yr – flow	*SKT*	*Fail to reject null of no trend*	*0.01 units/year*
10 yr – raw	MK	⇑ ($p < 0.1$)	1 unit/year
10 yr – raw	SKT	⇑ ($p < 0.05$)	1 unit/year
10 yr – FAC	MK	⇑ ($p < 0.05$)	1 unit/year
10 yr – FAC	SKT	⇑ ($p < 0.05$)	1 unit/year
1st 5 yr – flow	*SKT*	*⇓ ($p < 0.05$)*	*-2 units/year*
1st 5 yr – raw	MK	⇑ ($p < 0.01$)	7 units/year
1st 5 yr – raw	SKT	⇑ ($p < 0.05$)	7 units/year
1st 5 yr – FAC	MK	⇑ ($p < 0.01$)	6 units/year
1st 5 yr – FAC	SKT	⇑ ($p < 0.05$)	5 units/year
2nd 5 yr – flow	*SKT*	*⇑ ($p < 0.05$)*	*2 units/year*
2nd 5 yr – raw	MK	⇓ ($p < 0.01$)	-5 units/year
2nd 5 yr – raw	SKT	⇓ ($p < 0.05$)	-4 units/year
2nd 5 yr – FAC	MK	⇓ ($p < 0.05$)	-4 units/year
2nd 5 yr – FAC	SKT	⇓ ($p < 0.1$)	-3 units/year

These findings (i.e., Table 7) illustrate how an upward trend in the first half of the constituent data record and a downward trend in the second half of the record might reconcile itself. The upward trend was stronger than the downward trend, and so was detected in the overall 10-year data. Again, the level of detection of trend was different for both tests. However, in this RO2 analysis, the Seasonal Kendall test was more sensitive in the 10-year HM4 data set, as opposed to results for site HM6, in which the Mann-Kendall test seemed more sensitive. The Mann-Kendall test detected a trend at a smaller alpha level in both the 1st and 2nd 5-year raw and flow-adjusted concentrations. This is not surprising, as the Mann-Kendall test is sensitive to any type of temporal dependence and is detecting seasonality as trend. The slope estimates are very comparable at this site. (See Martin, 2000 for complete results) Again, a downward trend in flow correlated to an upward constituent trend, and vice versa, in both the raw and flow-adjusted constituent concentrations.

It is obvious from these results that as methods change, so can the outcome for trend analyses. This emphasizes the importance of understanding the assumptions of the analysis method (i.e. is the data seasonally dependent? If so, use the more appropriate Seasonal Kendall test), and the impact of sample size and significance level on detecting a true ‘long-term’ trend.

Results for Differences in Populations Analysis

This series of analyses compared the first 5-year data to the second 5-year data for BOD_5 (site HM4), NO_x (site HM6) and NH_4 (site RO2). To illustrate an analysis for spatial differences, a comparison was made between upstream and downstream NH_4 values for sites RO1 and RO2.

Table 8: Differences in Population Results for Site HM4, Constituent BOD_5

Test	Results
MS-Excel™ *t*-test (1st 5-yrs vs. 2nd 5-yrs)	Significant Difference ($p = 0.019$)
Minitab™ *t*-test (1st 5-yrs vs. 2nd 5-yrs)	Significant Difference ($p = 0.019$)
Equivalence Interval test	Equivalent (2nd 5-yrs within interval of ±20% of 1st 5-yrs mean – $p < 0.05$)
Inequivalence Interval test	Inequivalent (2nd 5-yrs *not* within interval of ±20% of 1st 5-yrs mean – $p < 0.05$)
Minitab™ Mann-Whitney	Significant difference ($p = 0.0148$)
WQStat™ Plus Mann-Whitney	Significant difference ($p < 0.05$)

These findings (i.e., Table 8) vary depending on alpha level, test and hypothesis. This illustrates how important assumptions of distribution and hypothesis are when testing, as well as selection of an acceptable Type I error (α). Again it illustrates that choosing the significance level after results are obtained can change the information obtained.

Minitab™ and WQStat Plus™ gave comparable results for the Mann-Whitney test. (Note: WQStat only detects a significant difference for the two-tailed test.) In general the results from different statistical packages are comparable, though results are presented differently in each one.

At the beginning of this section it was found that the raw data for site HM4_BOD_5 are not normally distributed. This could mean that a parametric *t*-test is not appropriate, and a nonparametric procedure could be more powerful. Therefore, the best information from this analysis may come from the Mann-Whitney test. (See Martin, 2000 for detailed Differences in Populations results)

Table 9: Differences in Population Results for Site HM6, Constituent NO_x

Test	Result
MS-Excel™ t-test (1st 5-yrs vs. 2nd 5-yrs)	Fail to reject the null of equal means
Minitab™ t-test (1st 5-yrs vs. 2nd 5-yrs)	Fail to reject the null of equal means
Equivalence Interval test	Fail to reject the null of equivalence
Inequivalence Interval test	Rejected the null of inequivalence (2nd 5-yrs *within* interval of ±20% of 1st 5-yrs mean – $p < 0.05$)
Minitab™ Mann-Whitney	Fail to reject the null of equal medians
WQStat™ Plus Mann-Whitney	Fail to reject the null of equal means

All of the tests in Table 9 failed to reject the null hypotheses of equal central tendency between the first and second 5-year data. This data also failed to reject the null of normal distribution, so the *t*-tests are likely to be more powerful tests of the difference in the two populations. However, failure to reject the null of equal means in the standard *t*-test does not prove that they are equal. The best information in this analysis comes from the equivalence test with the null hypothesis that the two populations are *inequivalent.* Rejection of this null indicates that the second 5-year data lay within an interval of ±20% of the first 5-year data mean, making them equivalent. Of course, this is supposing that the ±20% change is an ecologically acceptable change in NO_x. (See Martin, 2000 for complete results)

Table 10: Differences in Population Results for Site RO2, Constituent NH_4

Test	Result
MS-Excel™ *t*-test (1st 5-yrs vs. 2nd 5-yrs)	Fail to reject the null of equal means
Minitab™ *t*-test (1st 5-yrs vs. 2nd 5-yrs)	Fail to reject the null of equal means
Equivalence Interval test	Fail to reject the null of equivalence
Inequivalence Interval test	Rejected the null of inequivalence (2nd 5-yrs *within* interval of ±20% of 1st 5-yrs mean – $p < 0.05$)
Minitab™ Mann-Whitney	Fail to reject the null of equal medians ($p = 0.259$)
WQStat™ Plus Mann-Whitney	Fail to reject the null of equal means

These NH_4 data failed to reject the null of normal distribution (Table 10), so the *t*-test is an appropriate and powerful test. However, as in the analysis at the previous site (HM6), failure to reject the null of equal means does not prove that the means are in fact exactly equal. Again the best information comes from the equivalence test with the null hypothesis that the two populations are *inequivalent.* Rejection of this null proves at the 95% level that the mean of the second 5-year NH_4 data lies within an interval of ±20% of the first 5-year NH_4 data mean, making them equivalent. (See Martin, 2000 for complete results)

Table 11: Analysis of Differences Between NH_4 at RO1 and RO2

Test	Result
MS-Excel™ *t*-test (1st 5-yrs vs. 2nd 5-yrs)	Significant Difference ($p \leq 0.001$)
Minitab™ *t*-test (1st 5-yrs vs. 2nd 5-yrs)	Significant Difference ($p \leq 0.001$)
Equivalence Interval test	Rejected the null of equivalence (RO2 *not* within interval of ±20% of RO1 – 95% confidence)
Inequivalence Interval test	Fail to reject the null of inequivalence
Minitab™ Mann-Whitney	Significant Difference ($p \leq 0.001$)
WQStat™ Plus Mann-Whitney	Significant Difference ($p \leq 0.01$)

This analysis (Table 11) shows that when the concentration differences are large between populations, distribution assumptions, hypotheses and alphas do not have a great affect on the results. Although NH_4 at site RO2 failed to reject the null of normal distribution, the Mann-Whitney test is most appropriate because NH_4 at site RO1 is not normally distributed. (See Martin, 2000 for complete results)

Results for Standards Compliance

Table 12: Standards Compliance Results for Site HM4, Constituent BOD_5

Test	Compliance Results
Proportion Estimate – raw	3.3% excursions (0,7%) CI
Parametric Tolerance Limit – raw	Exceeded limit
Nonparametric Tolerance Limit – raw	Compliant
Parametric Tolerance Interval – raw	Compliant
Nonparametric Tolerance Interval – raw	Exceeded limit
Parametric Prediction Limit – raw	Exceeded limit
Nonparametric Prediction Limit – raw	Compliant
Parametric Confidence Interval for the mean - raw	Compliant
Nonparametric Confidence Interval for the median - raw	Compliant

“CI” means 95% confidence interval.

Each of these analyses (Table 12) gives different kinds of information about the data. The most straightforward is the proportion estimate, which tells exactly the proportion of excursion, along with a confidence interval so that the data can be representative of not only the sample, but also the population as a whole. These findings show that 3.3% of the data exceeded the excursion, and that up to 7% exceedance can be expected at the 95% confidence level.

The other procedures’ outcomes (Tolerance Limit, Tolerance Interval, Prediction Limit and Confidence Interval) were highly influenced by the distribution assumption. The raw BOD_5 data was shown to be not normal in the “Testing for Normality” section, so the nonparametric results are more appropriate in assessing compliance. The Tolerance Limit/Interval and Prediction Limit procedures are more appropriate for determining if a single sample exceeds a compliance limit or interval based on background data. Whereas the Confidence Interval is more appropriate for determining if the mean/median of a population exceeds a standard that is based on central tendency. The variety of results again illustrates the non-comparability of information produced from different analysis methods. It is important to note that all of these methods were performed on raw concentrations, which is the basis for most water quality standards. However, with the upcoming emphasis on TMDLs, mass loadings is another important measure of pollutant levels which could easily be incorporated into these methods, and not be dependent on flow. In this way, pollutant levels in streams with differing flows could be compared.

Discussion of Data Analysis Methods Selection

The previous sections have established that: (1) there are a large variety of methods employed in water quality data analysis to produce information; (2) hypothesis testing is by far the most popular type of analysis used to interpret water quality monitoring data (used in 17 of 19 Trend Studies and 16 of 20 Differences in Population Studies), and; (3) many of these common methods, when applied to one set of data, do not produce comparable results.

When completing a water quality assessment, it is usually assumed that the analyst will make an independent decision based on his or her interpretation of the data and information needs, after the data are collected. This fact introduces considerable uncertainty into the analysis of water quality data and results in non-comparable information. This raises concerns about the actual management decision, stemming from the information on which it was based. If there is a lack of confidence in the methods used to produce information for management, then there will be a lack of confidence in the ultimate decision as well. The only way to instill confidence in the management decision is to remove the concerns over the process through which information for the decision was created.

'Standard' Data Analysis Methods?

This issue raises the question: Is it feasible to develop a set of 'standard' water quality data analysis methods for specific forms of management information (i.e., trends, differences, standards compliance) that can produce comparable information that is defensible? The simple answer is yes, as this question is not new to water quality management. "Perhaps the best way to ensure that data collected during different studies are comparable is to encourage all investigators to use standardized sampling and analysis protocols whenever possible" (Becker & Armstrong, 1988). Currently there are professionals in the field who have been charged with determining which sampling and laboratory analysis methods result in comparable information (see Methods and Data Comparability Board of the National Water Quality Monitoring Council; http://wi.water.usgs.gov/pmethods). This is an especially pertinent issue as the interest in data sharing continues to rise.

This suggestion is not made without reservation. A natural conflict stems from the need to obtain comparable information, and permitting site-specific conditions to be considered in how data are analyzed and interpreted. The answer to this issue is not readily apparent, nor are professionals studying the problem and its solutions. At present, the discussions of 'appropriate' use of statistics in water quality monitoring tend to be within various water-management related agencies. The literature review in Chapter III clearly illustrates that some agencies have produced guidance for data analysis over the years, yet without much coordination within or outside of the agency. The National Water Quality Monitoring Council is currently facing the issue described here, and exploring the mechanisms that could help monitoring systems produce comparable information.

Several issues besides the methods selection itself will need to be addressed. Although some advise to the contrary (Ward *et al.*, 1986), many analysts select the analysis methods after examining the data and its distribution. In fact, this is recommended by existing guidance (i.e., Montgomery & Reckhow, 1984; Chatfield, 1985). The latter author recommends the following process: (1) Clarify the objectives of the investigation; (2) Collect the data in an appropriate way; (3) Investigate the structure and quality of the data; (4) Carry out an initial examination of the data; (5) Select and carry out an appropriate formal statistical analysis; (6) Compare the findings with previous results or acquire further data if necessary; and (7) Interpret and communicate the results." If 'standard' data analysis methods are developed, should they follow this same line of thinking?

There are good arguments for both sides of this issue. Choosing the analysis method before examining the data allows for impartial agreement and approval of the process by all interested parties without the bias of data appearance. However, choosing the method after analysis allows for selection of the most scientifically appropriate methods for the type of data gathered, without prior assumptions, but also allows for post-hoc selection of α, which, as illustrated in the previous section, can greatly influence the results. This issue in and of itself begs the assistance of professionals who are knowledgeable about water management to provide guidance for data analysis protocols.

Another topic that develops from the suggestion of standardizing data analysis methods deals with the extent that the analyst is allowed to produce information that directly relates to the management decision-making. Most management decision-makers are not statisticians. Should results of analyses only be presented (such as a rejection of a null hypothesis and obtained *p*-value), or an interpretation in terms of meaning presented as well? Should management be allowed to decipher statistical results, without the bias of the analyst? Guidance is needed for these questions to be resolved. Only those involved in water management know the expertise of their colleagues in understanding these scientific issues. Comprehension will vary among managers, and so may the role of the analyst in interpreting information produced from the data analysis. EPA (1998) dealt with this issue in the development of their *Guidelines for Ecological Risk Assessment*. The following process was recommended: "To ensure mutual understanding between risk assessor [i.e., analysts] and managers, a good risk characterization will express results clearly, articulate major assumptions and uncertainties, identify reasonable alternative interpretations, and separate scientific conclusions from policy judgments. Risk managers use risk assessment results, along with other factors (e.g., economic or other legal concerns), in making risk management decisions and as a basis for communicating risks to interested parties and the general public."

Finally, the question that directly pertains to the work presented in this paper is: What would these 'standard' data analysis methods look like? With the exception of a few estimation and graphical procedures, the methods used in the previous chapter were all based on the statistical theory of null hypothesis testing, which the beginning of Part I established is "under fire" in some parts of the scientific world. It is easy to see in the results of the New Zealand data analysis that information changes depending on the method selection, but why? The answer lies in several flaws of applying hypothesis testing to environmental (observational) data.

One flaw, which is rarely understood, is that results based on *p*-values from tests with different sample sizes are not comparable. Another flaw lies in the dependence of results on the validity of the assumptions (i.e. is the data serially dependent or normally distributed?). However, the greatest of these flaws, which has been mentioned previously, is that the resource managers and analysts of water quality monitoring data are often not statisticians, and so are repeatedly choosing analysis methods without a thorough understanding of the underlying assumptions, meaning of test parameters, or interpretation of results. Johnson (1999) states, “While many of the arguments against hypothesis tests stem from their misuse, rather than intrinsic value, I believe that one of their intrinsic problems is that they encourage misuse”.

Why Use Null Hypothesis Testing?

Nester (1996) suggests several reasons why null hypothesis tests are so widely used: (1) they appear to be objective and exact; (2) they are readily available and easily invoked in many commercial statistics packages; (3) everyone else seems to use them; (4) students, statisticians and scientists are taught to use them; and (5) some journals and editors and thesis supervisors demand them. The research in the previous chapters validates these claims. Yet perhaps the best explanation of why null hypothesis testing is so popular rests on the foundation of the scientific method. Under that method, a theory is postulated, which generates predictions, or hypotheses. Experiments and surveys are conducted to ‘test’ the hypothesis. The results of the experiment either refute the hypothesis, indicating that the theory is incorrect, or do not refute the hypothesis, letting the theory stand. Null hypothesis testing *appears* to be in harmony with this. But there are strong arguments that in fact it is not. This is particularly because the tested hypothesis, being null, does not correspond to the competing hypotheses science would wish to consider. Because these hypotheses are known often to be a priori false (Johnson, 1999) we can obtain conflicting and confusing signals about the validity of the hypotheses we actually wish to test.

So why test null hypotheses at all? McBride (2000) states that comparison of p-values for tests with similar numbers of samples does provide an elegant way of ranking the importance of differences measured, if sample sizes are identical. He also acknowledges that in constructing models, p-values are most useful in determining important explanatory variables in statistical models. However, this is more a function of exploratory data analysis, and not data analysis that better connects water quality information to management decision-making.

One answer would be that a statistical test could be only one factor in evidence of interpretation of the data. In this way, a single rejection of a point null hypothesis, or a p-value, would not be the only information leading to a management decision. Other pieces of information would need to be gathered to either support or refute the findings of the statistical test. EPA (1998) has produced guidance for ecological risk assessment that follows this type of process.

“Ecological risk assessment evaluates the likelihood that adverse ecological effect may occur or are occurring as a result of exposure to one or more stressors. It is a flexible process for organizing and analyzing data, information, assumptions and uncertainties. Ecological risk assessment provides a critical element for environmental decision making by giving risk managers an approach for considering available scientific information along with the other factors they need to consider (e.g., social, political, legal or economic), in selecting a course of action.” (EPA, 1998)

There exist alternatives to statistical testing that can provide scientifically defensible information to management about the quality of the water being monitored (e.g., likelihood methods, randomization tests, Bayesian methods). It is not within the scope of this report to provide great detail about analysis alternatives, but the following section will outline some of the other pieces of information that could accompany or even replace statistical tests in order to make the information more comparable and meaningful to management.

Data Analysis Tools to Make Information More Comparable

There are many procedures that can be applied along with statistical tests in order to give more meaning to the results beyond the *p*-value. It might be assumed that these procedures are already mandatory for statistical analysis of water quality data, yet the literature review in the previous section suggests that they are not. The following is a list of procedures and methods that could aid in the interpretation of water quality data, especially in combination with hypothesis testing.

- Graph the data (e.g. time-series, Q-Q plots, box plots, etc…) for visual interpretation and to aid in developing assumptions (i.e. correlation, seasonality).
- Test data for normality to aid in methods selection (parametric vs. nonparametric methods)
- Use flow-adjusted concentrations where appropriate (i.e. trends)
- Consider power when choosing tests and determining sample size (but realize that power increases with sample size, and many samples can make a test too "sensitive")
- Use estimation and confidence interval techniques in lieu of hypothesis testing, or as a compliment to the results

There also exist alternative analysis methods to hypothesis testing that, although not as prevalent, may provide more comparable results and more pertinent information for management decision making. More thorough explanations of each type of method, listed below, are provided in Martin (2000).

- Meta-analysis
- Interval testing (discussed in Part I)
- Decision Theory
- Likelihood ratios
- Bayesian Methods

Comparable Information in Other Fields of Data Collection

One excellent example of the goal for the water quality field is the area of weather reporting. Atmospheric scientists have developed, from a large list of variables and processes, a graphical interpretation of weather conditions that conveys instantly to the user the current state of the weather, what has occurred in the past, and what is likely to happen in the future. The importance of weather in our immediate lives has perhaps been the impetus to create consensus in atmospheric condition assessment. These weather interpretations are transparent, comparable and auditable, as they are standardized and accepted to convey the best information upon which to act.

Another example is the area of economic reporting. Several different indicators and indexes have been developed to aid in interpretation of the daily/monthly/yearly flux of the economy. Graphics, in the form of time series plots of these indexes, are used to convey understanding of trends in various sectors of the economy (Ward, 1998). For example, the Dow Jones Index has become an accepted 'standard' method for reporting a type of economic information upon which management and business decisions are based.

"The indicators and indices have been developed through well-documented and reviewed protocols. This is not to say that there are not disagreements over how the indices are computed, but it does reflect these debates occurring away from day-to-day reporting of the information" (Ward, 1998). "In other words, the science that underpins economic reporting is well developed and documented in protocols that are established on their scientific merit and not their particular outcome" (Ward, 1998).

The above section has outlined just a few of the analysis alternatives that can either replace, or supplement statistical data analysis methods. However, the entrenchment of hypothesis testing in the scientific world, combined with the plethora of analysis alternatives, make it difficult for data analysts to produce comparable information from water quality data analysis.

The subject of this discussion has focused on developing 'standard' guidance for data analysis methods, and how some methods might improve the comparability of information from monitoring. It is obvious that there are many 'right' methods for analysis, yet management is often missing comparable information for decision-making. Management needs information that is dependable, concise, comparable and bias-free in order to make fair and auditable decisions regarding the environment. Arguments about the process through which the information underlying management decision-making was created can only be eliminated through acquisition of comparable information in a manner that is transparent and auditable. Does this call for the development of 'standard' analysis methods?

Development of 'standard' protocols for water quality data analysis is suggested as a means to help this field mature to the same point of confidence about information for management decision-making as observed in weather and economic reporting. This, in turn, could perhaps bring the water quality field closer to the public, allowing water quality monitoring information to be broadly examined, and increasing public support for monitoring efforts.

Summary

The previous sections have fulfilled the tasks outlined in the introduction: (1) to examine the data analysis methods that are currently being used to analyze water quality monitoring data, as well as the criticisms of using those types of methods; (2) to explore how the selection of methods to analyze water quality data can impact the comparability of information used for water quality management purposes, and; (3) to offer options by which data analysis methods employed in water quality management can be made more transparent and auditable.

These tasks were accomplished through a literature review of criticisms of current data analysis methods, as well as texts, guidance and journals dealing with water quality assessments. Then, the common statistical analysis methods found were applied to the New Zealand Water Quality River Network data set. The purpose being to establish how information changes as analysis methods change, and to determine if the information produced from different data analysis methods was comparable. The results of the literature review and data analysis were then discussed, highlighting problems with the prevalent use of hypothesis testing in the water quality field. Part II further discussed options through which to begin solving these problems and produce comparable information for water quality management decision-making.

Conclusions

For several years it has been known or suspected that current methods for producing information from water quality data are subject to misuse and inappropriate application. Lack of statistical knowledge has caused poorly planned method selection and results that are not always comparable. This thesis has documented the problems associated with data analysis method selection for water quality monitoring, in an effort to provide problem definition as the first step in creating a solution. The process of documenting these problems has led to the conclusions discussed below:

1. Reviewing literature on water quality monitoring reveals the frequent use of a common class of statistical procedures (i.e., hypothesis testing) to produce information about water quality from the raw data. The majority of reviewed analysis methods use the concept of "statistical significance" to validate the information produced, be it comparison of means/medians (e.g., upstream/downstream averages), or evaluation of trends, or detection of extremes. It is with these methods that most of our knowledge about the water quality of our nation has been derived. From government monitoring projects to private monitoring studies, it appears from the literature review (Part I) that despite recent efforts to provide auditable information, data analysis procedures are often loosely planned and documented and statistical results rarely explained. Except for a few studies of water quality statistics (Harcum *et al.*, 1992; Hirsch, 1988; Montgomery & Reckhow, 1984, Montgomery & Loftis, 1987; Loftis, 1989; McBride, 1998, 1999a), alternative analysis methods with which to compare results are never explored, their significance is rarely explained, and information, once produced, never questioned, just reported as is. Of course discussions that led up to publication, if they questioned the methods, are rarely shared with the reader.

2. Through EPA's requirements for State 303(d) reports and 305(b) listing of impaired waters, it is apparent that the vision is being developed to create monitoring systems that will produce information that will answer basic questions about our nation's water quality. But when reviewing state assessment methodologies and other water quality studies, it is evident that the analysis procedures fall short of providing indisputable information due to the fact that the assessments are often based on subjective narrative criteria or relatively small monitoring data sets, and lack broadly peer-reviewed and agreed upon data analysis methods.

3. Although the methods selected to produce water quality information are generally being used correctly, they may not be universally accepted, or appropriate for the type of information about the environment that is needed. The availability of numerous analysis procedures means that methods selected to produce the same type of information (i.e., trends) may be different, resulting in a non-comparable basis for the same management decisions.

4. Because hypothesis-testing methods have been available and widely accepted, their appropriateness has been rarely questioned in the field of water quality monitoring. An argument that is at the forefront of the medical sciences is whether to use hypothesis testing at all (Berger & Berry, 1988; Loftus, 1991; Chow & Liu, 1992; Royall, 1992; Royall 1997). The value of these discussions in medicine is that they illustrate to other scientific fields that there are concerns with creating valid information using hypothesis

testing methods for data analysis (McBride *et al*, 1993; McBride 1998, 1999a; Johnson, 1999).

5. The solution to producing more valid information for management decision-making depends on the appropriateness of the methods chosen for the type of questions being asked, and the comparability of these methods with other, similar assessments. Many of the supplemental and alternative methods to hypothesis testing discussed in the previous chapters (e.g. graphical, estimation, Bayesian methods) could be utilized to aid in the interpretation of monitoring data, data which is influenced by so many unknown variables that interpretation is often difficult. The use of new methods that are more appropriate in creating scientifically defensible information is becoming more common in the medical field (Chow & Liu, 1992). However, these methods have not managed to effectively infiltrate water quality monitoring. Medical and epidemiological studies have shown that the use of methods such as meta-analyses, Bayesian statistics, and equivalence testing can produce more objective and valid information from the data than standard hypothesis testing. These alternatives, as well as others, need to be explored for applicability to water quality data analysis, in an effort to produce more comparable information from monitoring.

6. Solutions to the problems documented in this research may not come through common analysis methods, but instead require a deeper understanding of statistical theory, closer connections to the use of the information (i.e., management input), as well as new thinking about data analysis procedures. These considerations in the development of ‘standard’ water quality data analysis protocols will help to ensure that the procedures are transparent and auditable, and that results are comparable.

Recommendations

The following recommendations are suggested to help further the endeavor of providing better data analysis methods through which to produce information for management decision-making. These suggestions could be fulfilled through further academic study, interagency cooperative efforts (e.g., state and national water quality monitoring councils), or through a single entity taking the lead in providing guidance for water quality data analysis.

1. The subjects explored in this thesis established that there are many methods available for analysis and interpretation of water quality data. Not only are there classical hypothesis testing methods, but estimation, likelihood and Bayesian methods, to name a few. It was beyond the scope of this thesis to explore the applicability of these methods to water quality data and compare the results with those from hypothesis testing, but such an examination could prove very useful.

2. If null hypothesis testing is to continue to be the main venue through which water quality data are interpreted, better attention must be paid to distribution assumptions, flow-adjustment, and power analysis. The first two are easily handled, but the third, power analysis, is a complex subject. Power can be used to determine effective sample sizes to detect a significant difference fairly easily. However, calculation of the power of certain tests given a sample size can be complicated for parametric statistics, and even more so for nonparametric. It is important to note that, with the exception of interval testing, power tends to increase with sample size, so that trends tend to be detected more often merely because the number of samples has increased. This poses a substantial problem for accumulating data programs. Furthermore, in performing power analyses a consensus is needed on the magnitude of effect (or impact or trend) that is important to be detected (the same information as is needed to define the interval for an equivalence test). Eliciting such information can be difficult. Power analysis tools (software, internet calculators) can aid greatly, but a broad review of these tools for comparability of results must first take place in order to ensure quality of results.

3. The recent development of protocols for biological monitoring and assessment methodologies could prove to be the most informative way to assess water quality. These methods are relatively new, and so have not been scrutinized like the methods used to interpret chemical data. Although not discussed in this paper, many of the same statistical issues discussed in this thesis apply to biological data as well. The movement towards establishing broadly peer-reviewed methods for data analysis is impending, and all avenues of analysis methods should be thoroughly explored.

The bottom line is that the application of science, individually administered, is not going to make data analysis any easier, or results more comparable. There are too many variables involved, and too many methods through which to explore data. Nevertheless, if management requires accepted, scientifically defensible methods that produce comparable results upon which to base their decisions, consensus about what those methods should be is highly desirable. Several documents have been developed for standard methods for sampling protocols and laboratory analysis. Following this trend, it seems only natural to develop harmonized methods of data analysis as well. As discussed in the Scope section of Part I, this should only include methods used for management decision-making. Exploratory data analysis employed by researchers needs to remain untethered and flexible.

This is an issue that can only partially be resolved through science. Research can establish that there are common methods being used, compare the results obtained with differing methods, and document that there are problems with current data analysis procedures. But the decision-makers who are knowledgeable about monitoring resources, costs, and consequences of individual decisions will need to be the ones who, through a fair and open process, develop a guidance of acceptable methods for water quality monitoring data analysis.

Acknowledgements

The lead author wishes to sincerely thank John Klein and the U.S. Geological Survey for generating interest and funding for this research effort. As a co-sponsor for the National Water Quality Monitoring Council, this organization recognized the need for comparable information derived from water quality monitoring data and used by management for decision-making. The development of this report, and the thesis upon which it is based, was aided not only by the members of the council, but also by the expertise and enthusiasm of the three co-authors, Dr. Robert Ward, Mr. Graham McBride, and Dr. Jim Loftis.

References

Abeyta, C.G. and R.G. Roybal. 1996. Ground-water quality, water year 1995, and statistical analysis of ground-water quality data, water years 1994-95, at the chromic acid pit site, U.S. Army Air Defense Artillery Center and Fort Bliss, El Paso, Texas, *U.S. Geological Survey Water-Resources Investigations Report 96-4211*, U.S. Geological Survey.

Adkins, N.C. 1992. A framework for development of data analysis protocols for groundwater quality monitoring systems. Ph.D. Thesis, Colorado State University.

Arthur, M.A., Coltharp, G.B. and D.L. Brown. 1998. Effects of best management practices on forest streamwater quality in Eastern Kentucky. *Journal of the American Water Resources Associatio*n*, Paper No. 97106* 34(3):481-495.

Baldys, S., L.K. Ham and K.D. Fossum. 1995. Summary statistics and trend analysis of water quality data at sites in the Gila River Basin, New Mexico and Arizona, *U.S. Geological Survey Water-Resources Investigations Report 95-4083*, U.S. Geological Survey.

Becker, D.S. and J.W. Armstrong. 1988. Development of regionally standardized protocols for marine environmental studies. *Marine Pollution Bulletin* 19(7):310-313.

Berger, J.O. and D.A. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76:159-165.

Berger, J.O. 1986. Are P-values reasonable measures of accuracy? in *Pacific Statistical Congress*, eds. I. S. Francis, B. F. J. Manly, F. C. Lam, Elsevier, pp. 21–27.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* 33:526-542.

Berndt, M.P. 1996. Ground-water quality assessment of the Georgia-Florida Coastal Plain study unit – analysis of available information on nutrients, 1972-92, *U.S. Geological Survey Water Resources Investigations Report 95-4039*, U.S. Geological Survey, Tallahassee, Florida.

Brown, D.W., McCain, B.B. Horness, B.H. Sloan, C.A., Tilbury K.L., Pierce S.M., Burrows, D.G., Chan, Sin-Lam, Landahl, J.T. and M.M. Krahn. 1999. Status, correlations, and temporal trends of chemical contaminants in fish and sediment from selected sites on the Pacific Coast of the U.S.A. *Marine Pollution Bulletin* 37(1-2):67-85.

Bryers, G.G. 1999. National Rivers Water Quality Network: Data and Meta-Data. National Institute of Water and Atmospheric Research, Hamilton, New Zealand.

Butler, D. L. 1996. Trend analysis of selected water quality data associated with salinity control projects in the Grand Valley, Lower Gunnison River Basin, and at Meeker Dome, Western Colorado, *U.S. Geological Survey Water-Resources Investigations Report 95-4274*, U.S. Geological Survey.

Carver, R.P. 1978. The case against statistical testing. *Harvard Educational Review* 48:378-399.

Chatfield, C. 1985. The initial examination of data. *Journal of the Royal Statistical Society*, Series A 148:214-253.

Chow, S.C. and J.P. Liu. 1992. *Design and Analysis of Bioavailability and Bioequivalence Studies.* Marcel Dekker, New York.

Clow, D.W. and M.A. Mast. 1999. Long-term trends in stream water and precipitation chemistry at five headwater basins in the Northeastern United States. *Water Resources Research* 35(2):541-554.

Colman, J.A. and S.F. Clark. 1994. Geochemical data on concentrations of inorganic constituents and polychlorinated biphenyl congeners in streambed sediments in tributaries to Lake Champlain in New York, Vermont, and Quebec, 1992, *U.S. Geological Survey Open-File Report 94-472*, U.S. Geological Survey.

Dennehy, K.F. Litke, D.W., McMahon, P.B., Heiny, J.S. and C.M. Tate. 1995. Water quality assessment of the South Platte River Basin, Colorado, Nebraska, and Wyoming – analysis of available nutrient, suspended-sediment, and pesticide data, water years 1980-92, *U.S. Geological Survey Water-Resources Investigations Report 94-4095*, U.S. Geological Survey.

EPA. 1989. 1992 addendum. Statistical analysis of ground-water monitoring data at RCRA facilities: interim final guidance. U. S. Environmental Protection Agency Office of Solid Waste, Washington D.C.

EPA. 1992. Monitoring guidance for the national estuary program, *EPA-842-B-92-004*. U. S. Environmental Protection Agency Office of Water, Washington D.C.

EPA. 1997b. Information collection rule: draft data analysis plan. U.S. Environmental Protection Agency Office of Water, Washington D.C.

EPA. 1997c. Monitoring guidance for determining the effectiveness of nonpoint source controls*, EPA-841-B-96-004*. U.S. Environmental Protection Agency Office of Water, Washington D.C.

EPA. 1998. Report of the federal advisory committee on the total maximum daily load (TMDL) program*, EPA 100-R-98-006*. The National Advisory Council for Environmental Policy and Technology, U. S. Environmental Protection Agency Office of the Administrator.

Fleiss, J.L. 1987. Some thoughts on two-tailed tests. *Controlled Clinical Trials* 8:394.

GAO. 2000. Water quality: key EPA and state decisions limited by inconsistent and incomplete data, *GAO/RCED-00-54.* U.S. General Accounting Office.

Gibbons, J. D., and Pratt, J. W. 1975. *P*-values: interpretation and methodology. *The American Statistician*, 29, 20–25.

Gilbert, R.O. 1987. *Statistical methods for environmental pollution monitoring*. Van Nostrand Reinhold, New York.

Good, I.J. 1982. Standardized tail-area probabilities. *Journal of Statistical Computation and Simulation* 16:65-66.

Goodman, S. N. and Royall, R. 1988. Evidence and scientific research. *American Journal of Public Health*, 78(12), 1568–1574.

Goodman, S. N. 1993. P-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* 137(5):485-496.

Harcum, J.B., J.C. Loftis, and R.C. Ward. 1992. Selecting trend tests for water quality series with serial correlation and missing values. *Water Resources Bulletin* 28(3):469-478.

Harlow, L.L.; S.A.Muliak, and J.H Steiger (eds) 1997. *What If There Were No Significance Tests?* Lawrence Erlbaum, Mahwah, New Jersey.

Havens, K.E., Flaig, E.G., James, R.T., Lostal, S. and D. Muszick. 1996. Results of a program to control phosphorus from dairy operations in South-Central Florida, USA. *Environmental Management* 20(4):585-593.

Heiskary, S., Lindbloom, J. and C.B. Wilson. 1994. Detecting water quality trends with citizen volunteer data, Minnesota Pollution Control Agency. *Lake and Reservoir Management* 9(1):4-9.

Helsel, D.R. and R.M. Hirsch. 1992. *Statistical methods in water resources: studies in environmental science 49.* U.S. Geological Survey Water Resources Division, Reston, Virginia. Elvesier, New York.

IDT. 1998. *WQStat Plus™ User's Guide*. Copyright: Intelligent Decision Technologies, Ltd. Longmont, Colorado 80501.

Johnson, D.H. 1995. Statistical sirens: the allure of nonparametrics. *Ecology* 76:1998-2000.[8]

Johnson, D.H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63(3):763-772.

Kennedy, K. 1995. A statistical analysis of TxDOT highway storm water runoff: comparisons with the existing North Central Texas municipal storm water database. North Central Texas Council of Governments.

Koebel, J.W. Jr., Jones, B.L. and D.A. Arrington. 1999. Restoration of the Kissimmee River, Florida: water quality impacts from canal backfilling. *Environmental Monitoring and Assessment* 57:85-107.

Kress, N., Hornung, H. and B. Herut. 1998. Concentrations of Hg, Cd, Cu, Zn, Fe and Mn in deep sea benthic fauna from the southeastern Mediterranean Sea: a comparison study between fauna collected at a pristine area and at two waste disposal sites. *Marine Pollution Bulletin* 36(11):911-921.

[8] www.npwrc.usgs.gov/resource/1999/statsig/statsig

Lapp, P., Madramootoo, C.A., Enright, P., Papineau, F. and J. Perrone. 1998. Water quality of an intensive agricultural watershed in Quebec. *Journal of the American Water Resources Association, Paper No. 97033* 34(2):427-437.

Lavenstein, G.G. and K.D. Daskalakis. 1998. U.S. long-term coastal contaminant temporal trends determined from mollusk monitoring programs, 1965-1993. *Marine Pollution Bulletin* 37(1-2):6-13.

Loftis, J.C. 1989. An evaluation of trend detection techniques for use in water quality monitoring programs, *EPA-600-3-89-037*. U.S. Environmental Protection Agency Office of Research and Development, Environmental Research Laboratory, Corvallis, Oregon.

Loftis, J.C., Iyer, H.K. and H.J. Baker. 1999. Rethinking Poisson-based statistics for groundwater quality monitoring. *Groundwater* 37(2):275-280.

Loftus, G.R. 1991. On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology* 36:102-105.

MacDonald, L.H. 1994. Developing a monitoring project. *Journal of Soil and Water Conservation* May-June:221-227.

Martin, J.D. and C.G. Crawford. 1987. Statistical analysis of surface-water quality data in and near the mining region of southwestern Indiana, 1957-80, *U.S. Geological Survey Water-Supply Paper 2291*, U.S. Geological Survey.

Martin, L.M. 2000. The Role of Data Analysis Methods Selection and Documentation in Producing Comparable Information to Support Water Quality Management. Unpublished MS Thesis, Chemical and Bioresource Engineering Department, Colorado State University, Fort Collins, Colorado, 80523.

Matthews, R. 1998. *Lying science.* Sunday Star-Times, Auckland, New Zealand. September 27.

Mattraw, H.C., D.J. Scheidt and A.C. Federico. 1987. Analysis of trends in water quality data for water conservation area 3A, the Everglades, Florida, *U.S. Geological Survey Water-Resources Investigations Report 87-4142*, U.S. Geological Survey.

McBride, G.B., Loftis, J.C. and N.C. Adkins. 1993. What do significance tests really tell us about the environment? *Environmental Management* 17(4):423-432. (Errata in *18(1)*:317)

McBride, G.B. 1998. Statistical methods: when differences are equivalent. *Water & Atmosphere* 6(4):21-23. National Institute of Water and Atmospheric Research, New Zealand.

McBride, G.B. 1999a. Equivalence tests can enhance environmental science and management. *Australia & New Zealand Journal of Statistics* 41(1):19-29.

McBride, G.B. 1999b. National Institute of Water and Atmospheric Research, Hamilton, New Zealand. Thesis consultation. Personal Communication.

McBride, G.B. 2000. Some issues in statistical inference. National Institute of Water and Atmospheric Research, Hamilton, New Zealand.[9]

McGill, R., Tukey, J.W. and W.A. Larsen. 1978. Variations of box plots. *The American Statistician* 32(1).

McMahon, G. and D.A. Harned. 1998. Effect of environmental setting on sediment, nitrogen, and phosphorus concentration in Albemarle-Pamlico Drainage Basin, North Carolina and Virginia, USA. *Environmental Management* 22(6):887-903.

Momen, B., Eichler, L.W., Boylen, C.W. and J.P. Zehr. 1997. Are recent watershed disturbances associated with temporal and spatial changes in water quality of Lake George, New York USA? *Environmental Management* 21(5):725-732.

Montgomery, R.H. and J.C. Loftis. 1987. Applicability of the t-test for detecting trends in water quality variables. *Water Resources Bulletin, Paper No. 86130*, 23(4).

Montgomery, R.H. and K.H. Reckhow. 1984. Techniques for detecting trends in lake water quality. *Water Resources Bulletin, Paper No. 82105* 20(1):43-52.

Morrison, D.E. and R.E. Henkel. 1970. *The Significance Test Controversy—A Reader.* Aldine, Chicago.

Mueller, D.K. 1990. Analysis of water quality data and sampling programs at selected sites in North-Central Colorado, *U.S. Geological Survey Water-Resources Investigations Report 90-4005,* U.S. Geological Survey.

Mueller, D.K. 1995. Nutrients in ground water and surface water of the United States: an analysis of data through 1992, *U.S. Geological Survey Water-Resources Investigations Report 95-4031*, U.S. Geological Survey.

Nester, M.R. 1996. An applied statistician's creed. *Applied Statistics* 45:401-410.

Nimmo, D.R., Willox, M.J., LaFrancois, T.D., Chapman, P.L., Brinkman, S.F. and J.C. Greene. 1998. Effects of metal mining and milling on boundary waters of Yellowstone National Park, USA. *Environmental Management* 22(6):913-926.

Nunnally, J.C. 1960. The place of statistics in psychology. *Educational and Psychological Measurement* 20:641-650.

PEER: Public Employees for Environmental Responsibility. *Murky Waters: Official Water Quality Reports are All Wet, An Inside Look at EPA's Implementation of the Clean Water Act*. May 1999.[10]

Pinsky, P., Lorber, M., Johnson, K., Kross, B., Burmeister, L., Wilkins, A. and G. Hallberg. 1997. A study of the temporal variability of atrazine in private well water, part II: analysis of data. *Environmental Monitoring and Assessment* 47(2):197-221.

[9] www.niwa.cri.nz/pgsf/stats/news.html
[10] www.peer.org/execsum.html

Preece, D.A. 1990. R.A. Fisher and experimental design: a review. *Biometrics* 46:925-935.

Rinella, J.F. 1986. Analysis of fixed-station water quality data in the Umpqua River Basin, Oregon, *U.S. Geological Survey Water Resources Investigations Report 85-4253*, U.S. Geological Survey.

Royall, R.M. 1992. The elusive concept of statistical evidence. J.M. Bernardo, J.O. Berger, A.P. David and A. Smith (eds), *Bayesian Statistics* 4:405-418. Oxford University Press, New York.

Royall, R. M. 1997. *Statistical Evidence: a Likelihood Paradigm*, London: Chapman and Hall.

Sample, L.J., Steichen, J. and K.R. Kelley Jr. 1998. Water quality impacts from low water fords on military training lands. *Journal of the American Water Resources Association, Paper No. 96165* 34(4):939-949.

Santillo, D., Stringe, R.L., Johnston, P.A. and J. Tickner. 1998. The precautionary principle: protecting against failures of scientific method and risk assessment. *Marine Pollution Bulletin* 36(12):939-950.

Savage, I.R. 1957. Nonparametric statistics. *Journal of the American Statistical Association* 52:331-344.

Schervish, M. J. 1996. *P* values: what they are and what they are not, *The American Statistician*, 50(3), 203–206.

Shabman, L.A., Hershner, C., Kator, H.I., Smith, E.P., Smock, L.A., Younos, T., Shaw, L.Y and C.E. Zipper. 1998. Report of the Water Quality Academic Advisory Committee. *Virginia Water Resources Research Center Report No. SR8-1998.*[11]

Smith, D.G., McBride, G.B., Bryers, G.G, Wisse, J and D.F.J. Mink. 1996. Trends in New Zealand's national river water quality network. *New Zealand Journal of Marine and Freshwater Research* 30:485-500.

Smith, R.A., R.B. Alexander and M.G. Wolman. 1987. Analysis and interpretation of water-quality trends in major U.S. rivers, 1974-81, *USGS Water Supply Paper 2307*. United States Government Printing Office.

Snyder, N.J., Mostaghimi, S., Berry, D.E., Reneau, R.B., Hong, S., McClellen, P.W. and E.P. Smith. 1998. Impact of riparian forest buffers on agricultural nonpoint source pollution. *Journal of the American Water Resources Association, Paper No. 96132* 34(2):385-395.

Spooner, J., Maas, R.P., Dressing, S.A., Smolen, M.D. and F.J. Humenik. 1985. Appropriate designs for documenting water quality improvements from agricultural NPS control programs, *EPA 440/5-85-001*. U.S Environmental Protection Agency Perspectives on Nonpoint Source Pollution:30-34.

[11] www.vwrrc.vt.edu/publications/special.htm

Stoddard, J.L., Driscoll, C.T., Kahl, J.S. and J.H. Kellogg. 1998. A regional analysis of lake acidification trends for the northeastern U.S. 1982-1994. *Environmental Monitoring and Assessment* 51:399-413.

Stoe, T.W. 1998. Water quality and biological assessment of the Wiconisco Creek watershed. *Susquehanna River Basin Commission Publication No. 193*.

Takita, C.S. 1998. Nutrients and suspended sediment transported in the Susquehanna River Basin, 1994-96, and loading trends, calendar years 1985-96. *Susquehanna River Basin Commission Publication No.194*.

Vaill, J.E. and D.L. Butler. 1999. Streamflow and dissolved-solids trends, through 1996, in the Colorado River Basin upstream from Lake Powell – Colorado, Utah and Wyoming, *USGS Water Resources Investigations Report 99-4097*, U.S. Geological Survey, Denver, Colorado.

Veiland, V. J., and Hodge, S. E. 1998. Book Reviews: Statistical Evidence: A Likelihood Paradigm. By Richard Royall. *American Journal of Human Genetics*, 63, 283–289.[12]

Ward, R.C. 1996. Water Quality Monitoring: Where's the Beef? *Water Resources Bulletin* 32(4).

Ward, R.C. 1998. Management and Monitoring of Water Quality: CB/CE 545 Fall Class Notes. Colorado State University.

Ward, R.C., J.C. Loftis and G.B. McBride. 1986. The "data-rich but information poor" syndrome in water quality modeling. *Environmental Management* 10(3):291-297.

Ward, R.C., J.C. Loftis and G.B. McBride. 1990. *Design of Water Quality Monitoring Systems*. Van Nostrand Reinhold, New York.

Younos, T.M., Mendez, A., Collins, E.R. and B.B. Ross. 1998. Effects of a dairy loafing lot-buffer stream on stream water quality. *Journal of the American Water Resources Association, Paper No. 97040* 34(5):1061-1069.

Zar, J.H. 1984. *Biostatistical Analysis*. McGraw-Hill, Englewood Cliffs, NJ.

[12] www.journals.uchicago.edu/AJHG/journal/issues/v63n1/980002/980002.text.html