De Facto Data Analysis Methods for Goal Oriented Monitoring: What does current practice tell us?

Lindsay Melissa Martin

Lindsay is a graduate student at Colorado State University, currently pursuing her M.S. degree in Agricultural and Bioresource Engineering. Her thesis research deals with statistical analysis of water quality monitoring data, and is funded by the USGS. She grew up in Bryan, Texas and earned her B.S. degree in Agricultural Engineering from Texas A&M University.

Reviewing literature on water quality monitoring reveals the commonality of using statistical procedures to produce information about the water quality from the raw data. These statistical methods use the concepts of "statistical significance" (i.e. p-values) to validate the information produced, be it comparison of means/medians (e.g. upstream/downstream averages), evaluation of trends, or determination of standards compliance. These significance testing procedures are accepted throughout the field as the appropriate methods through which to draw conclusions from the monitoring data.

The purpose of this paper will be to review the current statistical analysis methods used in water quality monitoring, and establish a connection between commonly selected methods and information sought from the monitoring program. It is hypothesized that *de facto* standards for data analysis exist, through the use of specific procedures for specific information needs. The review will establish whether or not this hypothesis is correct.

Introduction

A classic definition of the word monitor is "to watch, observe, or check, especially for a special purpose" (Webster's New Collegiate Dictionary, 1977). Water quality "monitoring" is more than checking to make sure water quality standards are not violated. Monitoring is the process of seeking information about the behavior of water quality variables in the environment (e.g. average conditions, trends, and extremes) (Ward et al., 1986). "Monitoring is performed in support of water quality management and is universally recognized as indispensable for effective management" (Ward et al., 1986).

A common problem, which arises from monitoring, is how to relate information contained in data to the information needed by management for decision-making. For example, if a legal goal from the Clean Water Act is to restore and maintain the nation's water quality, then what information about water quality variables can be used to inform the public and water managers if water quality has been maintained or improved?

A common answer to this problem is to use statistical data analysis methods to produce information from the water quality data. The field of statistics provides an organized approach to quantify the unavoidable uncertainties about the inferences drawn from water quality data (Ward, 1998). Snedecor and Cochran (1980) define statistics as a field that deals with collecting, analyzing, and drawing conclusions from data, and it has been accepted for several years that water quality monitoring is gradually becoming recognized as a statistical sampling procedure (Ward and Loftis, 1983).

The purpose of this paper is to examine current practice and "state-of-the-art" procedures used to analyze water quality data for information purposes. The review of literature focuses on the use of statistics in literature 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 established "standards" of monitoring 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 that 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 statistical analysis methods that have been reviewed and recommended for all water quality monitoring situations. However, this paper will attempt to establish that there are certain methods that are 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 through which to interpret results from data analysis into information for management.

There exist several textbooks that directly address statistical analysis procedures for environmental data (e.g. Helsel and Hirsch, 1992; Gilbert, 1987; Ward et al., 1990). 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 researcher's publications in their data analysis. For example, Helsel and 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 and 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 and 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-tau and Mann-Kendall test 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. 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 null hypothesis that the streams are equivalent 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 is becoming 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 305(b) reports and 303(d) lists. 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.

This literature review found that the EPA mainly publishes guidance that helps the states and other reporting entities compile and interpret information to support 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).

The EPA also has research publications that can be viewed as endorsements for particular methods. In Loftis et al. (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 Covariances (ANOCOV) on ranks test. A guidance document for determining improvements from agricultural nonpoint 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.

The next step of the literature review was to determine what the actual current use of statistics is in water quality data analysis. 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.

Peer Reviewed Water Quality Assessments

This section serves to establish the current use of statistics, beyond guidance, in the water quality field. To gain a comprehensive view 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 making the tests more robust. The most popular analysis was the Seasonal Kendall Tau (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 studies reviewed which dealt with trend detection 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 significance 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 given. But it was evident by the testing that all performed a significance 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 and Roybal, 1996) or Kruskal-Wallis test (Abeyta and Roybal, 1996; McMahon and Harned, 1998; Mueller, 1995; Dennehy et al., 1995). All of the studies reviewed are summarized in II.

Standards Compliance

Determination of standards compliance was not commonly sought via statistical tests in the research type assessments that were reviewed (see Table III for summary of assessments which involved standards compliance). Therefore, part of this literature review attempted to describe how states generate this information for their 305(b) and 303(d) 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. 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.

Table I: Water Quality Assessments Involving Trend Detection

Author	*Monitoring Entity*	*Distribution Assumption*	*Actual Hypothesis Stated*	*Test Used*
Clow and Mast (1999)	USGS	NP	None stated	Season Kendall Tau or Mann-Kendall
Baldys, Ham and Fossum (1995)	USGS	NP	Null hypothesis of no significant trend	FAC Season Kendall Tau or Mann-Kendall
Mattraw, Scheidt and Federico (1987)	USGS, NPS and SFWMD	NP	None stated	FAC Season Kendall Tau or Mann-Kendall
Rinella (1986)	USGS	NP	None stated	FAC Season Kendall Tau or Mann-Kendall
Berndt (1996)	USGS	NP	None stated	Season Kendall Tau or Mann-Kendall
Mueller (1995)	USGS	NP	None stated	FAC Season Kendall Tau or Mann-Kendall
Mueller (1990)	USGS	NP	None stated	FAC Season 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 diff. 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	Satterwaite's 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, Alexander and Wolman (1987)	USGS	NP	None stated	SKT and FAC SKT
Vaill and 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, Lindbloom and Wilson (1994)	Minnesota Pollution Control Agency	NP	Null hypothesis of no trend	Kendall's tau-b (Gilbert, 1987)
Lavenstein and 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 (Hollender & Wolfe 73)
Arthur, Coltharp and Brown (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 and 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 and Harned (1998)	USGS	NP	None stated	Kruskal-Wallis, and Tukey's Multiple Comparison
Mueller (1995)	USGS	NP	None stated	Kruskal-Wallis
Koebel, Jones and Arrington (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 and 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, Hornung and Herut (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)
Bexfield and Anderholm (1997)	USGS	?	None stated	compared daily and quartile concentrations to standards

State Determinations of Designated Use Support

New York: Judgements are made on use support according to narrative criteria established by the state. New York stated that "the bulk of Priority Waterbody List (PWL) information is reflective of *evaluation* as opposed to *monitoring* efforts. This report did not qualify how the area of effect (i.e. stream miles) is determined for each segment reported. They are currently implementing a rotating basin approach for future assessments. (NYS Department of Environmental Conservation, 1998)

New Jersey: Judgements on use support are qualified by monitoring data and criteria developed by the state. No statistical tests are used. However, the protocol for determining use support is documented thoroughly. For example: for recreational use support, data collected over 5 years was compared to the NJ Surface Water Quality Standard criteria for fresh water streams, and use support determined according to the criteria listed below in Table III.10.

Table III.10: New Jersey Recreational Use Support Criteria

Use Support	**Assessment Criteria**
Full Support	The fecal coliform geometric avg. was <200 MPN/100ml and <10% of individual samples exceeded 400 MPN/100ml
Partial Support	Fecal coliform geo. Avg. was <200 MPN/100ml but >10% of samples exceeded 400 MPN/100ml
No support	Fecal coliform geo. Avg. >200 MPN/100 ml and >10% of samples exceeded 400 MPN/100ml

New Jersey also established its miles affected according to the criteria that the number of miles is the distance between the two monitoring points plus 1000 feet upstream. Other use support designations and trends were reported, but no protocol was documented for their determination. (NJ Department of Environmental Protection, 1998)

Region III (Delaware, Pennsylvania, Maryland, Virginia, West Virginia, District of Columbia): Criteria for use support assessment are those recommended by the EPA for 305(b) reports. Some states use biology to determine use support, following the EPA's Rapid Bioassessment Protocol. "By and large, simple percentages of standard violations are used to make a judgement call for water body assessments" (Barath, 2000).

Oklahoma: This state delineates all of their criteria for use support determination, with most criteria being comparisons of monitored data to standards. For example: Oklahoma uses the EPA recommendations for numerical parameters (full support = <10% violations, partial support = >11% but <25% violations, and no support = >25% violations). At least ten samples are required for this determination in streams, and 20 vertical profiles in lakes. However, fewer can be used if exceedence is assured. Any monitoring site shall not represent more than 10 wadable stream miles, or a lake area more than 250 surface acres. (Oklahoma Water Resources Board, 1999)

Arizona: No trends are evaluated, and no statistical tests are used. The use support criteria are enumerated from Arizona DEQ (2000). Arizona also uses macroinvertebrate-based bioassessment criteria to determine use, generally following EPA's guidelines. However, this Index of Biological Integrity (IBI) is not statistically based, it uses a scoring system and percentiles. No water body assessed as partially supporting or non-supporting based solely on biocriteria will be placed on the state's 303(d) list prior to identification and cause of the impairment, as it could be the results of natural phenomenon. (Marsh, 2000)

California: Individual regions do not provide information about how they determine use support. The only known protocol is for Los Angeles, which uses the criteria recommended by the USEPA. (Richard, 2000)

Hawaii: Use support is determined partially by comparing bacteria and chemical water quality data to state standards. For those categories that don't have applicable state standards, narrative criteria were created for judgement decisions instead of numerical/statistical based decisions. (Teruya, 2000)

Virginia: Criteria for use support is enumerated by the state. The actual numerical/narrative decision protocol follows the EPA recommended criteria for use support determinations. Assessment decisions are based on both monitored and evaluated data. Virginia also sets protocols for determining affected areas, e.g. stating that no station shall represent more than 10 miles of wadable stream. This determination is a judgement-based decision taking several enumerated factors into account. (Virginia Department of Environmental Quality, 1999)

South Carolina: This state uses the EPA's recommended assessment criteria for 305(b) reporting. (Kirkland, 2000)

Florida: As a portion of Florida's efforts, the state has adopted an Environmental Mapping and Assessment Program (EMAP) type of statistical analysis. The goal is to determine the overall conditions of water bodies within a geographical area. For example, the state will make statements such as, "with a confidence level of .90, the median value for NO3 in small lakes in north central Florida is (say) 1.3 mg/l plus or minus 0.4 mg/l. The state has been broken into 20 geographical units based on hydrologic drainage basins. These analyses will be performed for six resources. They are confined ground water, unconfined ground water, small lakes, large lakes, high order streams and low order streams. A sister organization in the state is conducting a similar analysis for Florida's estuaries. (Copeland, 2000)

Tennessee: This state generally follows the EPA's recommendations for use assessments, but has some discretion in the "magnitude and duration" of water quality standard violations. (Denton, 2000)

North Carolina: Use support for 305(b) and 303(d) listing are based on monitored and evaluated data, with more confidence placed on monitored data. Biological indexes and physical/chemical data are used to determine use support, similar to the procedures Arizona uses. However, biological data/indexes take precedence over chemical/physical data when determining use support. (Swanek, 2000)

Kentucky: Kentucky's approach is a combination of targeted sites and random survey sites, mainly using biological data to determine use support. Many of their water quality stations are at sites also sampled biologically. However, there are a few sites, mainly large rivers, where only water quality data are collected and from which use assessments are made. The state has just embarked on an intensive watershed monitoring program in 1998, in which the first 5-year watershed cycle will concentrate primarily on a broad picture of water quality in the state. (VanArsdall, 2000)

In this watershed cycle, the state will sample approximately 350-400 random sites over the 5-year watershed cycle, concentrating on 1 to 3 major river basins each year. The watershed will be sampled for macroinvertebrates and habitat. These samples will allow the state to extrapolate aquatic life use to most miles of wadable streams from a 1:100,000 scale hydrologic network. (VanArsdall, 2000)

Kentucky does no random survey water quality sampling because of inadequate resources. For targeted water quality sampling, the fixed statewide network consists of 71 sites located at the downstream reaches of 8-

digit cataloging units, mid-unit in the 8-digit watersheds, influent to major reservoirs, and major tributaries. These are sampled bimonthly except when they fall into the watershed cycle, and then they are sampled more frequently for one year. In the rotating watershed water quality network, the state will sample about 30 sites each year that fill in the hydrologic gaps in the fixed network by picking up most of the 5th order waterhsheds. Some are also sited for other purposes such as predominant land use, TMDLs, least impacted, etc… Sampling frequency at these sites depends on the objective of the particular site. (VanArsdall, 2000)

Because of help from other federal and state agencies, Kentucky has much more biological sampling resources at their disposal, and these resources are used for targeted biological sampling. They are able to sample most 4th order streams for at least one assemblage and habitat. This informs the state which basins have problems that need to be addressed by later sampling and mitigation activities. Over the 5-year watershed cycle, this targeted biological sampling will total over 1000 sites. (VanArsdall, 2000)

Alabama: This state follows the EPA recommended assessment criteria (percentages for chemical data). If there exists a large data set it is considered "monitored" data for assessment. "For example, 5 month (June-October), once-a-month sampling is considered *monitored*, but if the field personnel sample any less than this it would be considered *evaluated* data." Alabama is also developing specific site criteria for biological, physical/chemical, and habitat data, as well as criteria for determination of miles/acres affected. However, as of yet, Alabama does not have a state methodology for judging biology index/metrices results. (Reif, 2000)

Conclusions

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 significance testing, this type of analysis is alive and well in the field of water quality. It is interesting to note that although hypothesis testing seems to be popular, as evidenced by its inclusion in guidance documents and water quality studies, the actual hypothesis tested is never reported, despite recommendations to the contrary in many of the guidance documents (Gilbert, 1987; Ward et al., 1990; Helsel and Hirsch, 1992; Montgomery and Reckhow, 1984; EPA, 1992; EPA, 1997c).

With a few exceptions (Heiskary, Lindbloom and Wilson, 1994; Momen et al., 1997; EPA, 1992; EPA, 1997c), the power of significance 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 obtained significance level (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. The 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 significance test was never 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.