7 Data Analysis

By J. Spooner, J.B. Harcum, D.W. Meals, S.A. Dressing, and R.P. Richards

7.1 Introduction

This chapter of the guidance examines options for planning and analyzing data collected in nonpoint source watershed studies. The emphasis of this chapter is on projects at the watershed or subwatershed level, although evaluation of individual BMPs is also addressed. These analysis approaches complement the watershed project design considerations discussed in section 2.4 of this guidance.

Specifically, this chapter discusses the following topics:

- Exploratory data analysis
- Data transformations that might be necessary to prepare data for valid statistical analysis
- Methods to deal with extreme values, censored data, and missing data
- Data analysis methods for water quality problem assessment
- Data analysis methods for project planning
- Data analysis methods for assessing BMP or watershed project effectiveness
- Techniques for load estimation

The reader may wish to refer to chapter 4 (Data Analysis) of the 1997 guidance (USEPA 1997b) which was written largely to provide a primer on statistical methods for analysis of data generated by nonpoint source watershed projects. The 1997 guidance addresses various topics on statistical analysis in considerable detail, including estimation and hypothesis testing, characteristics of environmental data, and basic descriptive statistics. In addition, the 1997 guidance compares parametric and nonparametric tests, recommends appropriate methods for routine analyses, and provides numerous examples of the application of various statistical tests. Additional resources for data analysis approaches are also available in various *Tech Notes* and other publications (see References).

7.2 Overview of Statistical Methods

A wide range of parametric and nonparametric methods exists for analyzing environmental data. In some cases, graphical methods will be suitable to meet analysis objectives; more rigorous statistical analysis approaches may be best otherwise. This section provides a brief overview and summary of key features of these various methods. Readers should consult the 1997 guidance (USEPA 1997b) and additional sources (e.g., statistics textbooks and software packages) for greater detail.

Recommended statistical methods are summarized in Table 7-1 through Table 7-6 based on watershed project phase or need because experience indicates that this type of grouping will be practical for many involved in such efforts. Methods in these tables are recommended, but the tables do not include all possible alternative approaches. Additional discussion and illustrative examples follow in sections 7.3 through 7.8. Because of its importance to many watershed projects, especially those addressing TMDLs,

pollutant load estimation is addressed separately in section 7.9. While most of the methods described in this chapter are more commonly applied to water chemistry, flow, and precipitation data, many can also be applied to biological data as well. Recommended approaches for analyzing biological data are described in detail in chapter 4, and some examples are also provided in this chapter.

7.2.1 Exploratory Data Analysis and Data Transformations

It is often necessary to work with a mix of information and data during the initial stages of watershed projects. A major first task involves gathering and organizing available information and data, followed by an initial examination of the data to help identify water quality problems, pollutants, sources, and pathways. Exploratory data analysis techniques are well suited to this project phase, and should also be applied as a first step to all data subsequently collected by the project. Exploratory data analysis is also a critical first step in beginning to analyze water quality data from watershed projects that are underway, before undertaking more complex analysis.

Exploratory data analysis provides basic information about the data record, including the data distribution and an assessment of missing and extreme values. The presence of autocorrelation and seasonal cycles should also be evaluated. EDA can also be useful to examine clusters in the data or relationships between variables and/or sample locations.

Table 7-1 summarizes exploratory data analysis methods by analytical objective. The type of method (parametric, nonparametric, graphical), basic data requirements (e.g., distribution, independence), and major cautions and concerns are also included in the table.

Table 7-1. Exploratory data analysis methods (see discussion, section 7.3)

Analytical Objective	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
Describe behavior of variable(s)	Univariate statistics (e.g., range, mean, median, interquartile range, variance)	P, N	Minimal	Mean is sensitive to extreme values; median may be preferred measure of central tendency.
Evaluate distribution and assumptions of independence and constant variance	Plots (histogram, probability, lag-n autocorrelation, cumulative distribution functions); skewness, kurtosis; Durbin-Watson statistic to detect presence of autoregressive lag 1 pattern; Shapiro-Wilk test; Kolmogorov-Smirnov test	P, N, G	Minimal to moderate	Data transformations to satisfy likely statistical testing assumptions should be examined. Autocorrelation functions (ACF) which examine auto correlation at each lag require equal time-space data and appropriate software.
Identify extreme values and anomalies	Plots (e.g., time series, boxplots) Compute frequency or proportion of observations exceeding threshold value; cumulative frequency or duration plots	G, P, N	Minimal	Outliers should not be deleted if error cannot be confirmed.

Analytical Objective	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
Observe seasonal or other cycles	Plots (time series, seasonal boxplots)	G	Minimal	More intensive techniques are generally required to confirm and quantify trends.
	Examination of autocorrelation pattern			Use software that can generate autocorrelation function (ACF) graphs (see section 7.3.6).
Find clusters or groupings	Cluster analysis, principal components analysis, canonical correspondence analysis, discriminant function analysis	P, N, G		Factors determining groupings may be difficult to discern or interpret.
Preliminary comparison of two or more locations or time periods	Boxplots	G	Minimal	Visual comparisons should be confirmed by numerical tests.
Examine relationships between variables	Correlation, regression	P	Data must be normally distributed to apply parametric analysis	Graphical analysis should be used to confirm and understand numerical correlation coefficient. Correlation does not guarantee causation.
	Spearman's rho or rank correlation coefficient	N	Can be used when both independent and dependent variables are ordinal or when one variable is ordinal and the other is continuous	
	Bivariate scatterplots LOWESS smoother	G	Minimal	Visual comparisons should be confirmed by numerical tests.

*Key to Method Type: G = Graphical, N = Nonparametric, P = Parametric

Table 7-2 summarizes methods that can be applied to adjust (e.g., transform) data based on the requirements of methods (e.g., normal distribution required for parametric analyses) to be used in the next phase of data analysis. This table also identifies methods that can be used to address problems caused by unexpected events, including washed out monitoring equipment, floods, droughts, ice, failed BMP implementation plans, and equipment and laboratory errors.

Table 7-2. Methods for adjusting data for subsequent analysis (see discussion, section 7.3)

Analytical Objective	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
Obtain a normal distribution (for parametric approaches)	$\log_{10}$ and $\log_e$ (ln) are most commonly used transformations in water resources	P	Original data values must be positive and non-zero.	Other transformations (e.g., Box-Cox) may be required to achieve normal distribution. Very small numbers and legitimate zero values may require a different transformation (e.g., $\log_{10}$(value + n). Transformations will not correct issues of independence. Back-transformations may be difficult to interpret.
	arc-sine square root transformation	P	Used for proportions	
	If distribution assumptions cannot be met, adopt methods resistant to errors in results caused by deviations from the assumption of normality	N	Minimal	Nonparametric procedures may still have other assumptions that must be met for usage. If distributional assumptions can be met, then parametric tools tend to be more powerful.
Accommodate extreme values	Use methods resistant to errors in results caused by extreme values such as: nonparametric trend tests or frequency analyses	N, G	Moderate	If the data are missing due to right censoring (too high to measure), techniques discussed in section 7.4 should be considered.
	Data stratification (e.g., by seasons, base flow, storm, and floods)		Moderate	
	Use covariate/ explanatory variable such as flow to help 'explain' the influence of extreme values			
	Utilize log transformed data to minimize skewness caused by the extreme values	P	Minimal	
Manage missing data	Data aggregation to create uniform time intervals by averaging or using the median value	P	Minimal	Missing values are ignored in most nonparametric and parametric tests; however, some tests require equal spacing of observations. Data aggregation to accommodate missing data or changes in data frequency must be done with care.
	Estimate missing values based upon regression relationship from other sites or events		Regression relationship with data from similar basin (e.g., flow). Sometimes it may also be appropriate to use the flow/concentration relationship at the same station to estimate missing concentration data	Only use when the data meet the assumptions for regression analysis and the sample size is large enough that the regression relationship is reliable.

Analytical Objective	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
Adjust for autocorrelation	Aggregation of data to less frequent observations	N	Minimal	Aggregation must be consistent (e.g., monthly mean of n daily observations), not mix of different sample frequencies.
	Use of parametric time series analysis techniques available in many statistical software tests	P	Generally equally time-space data observations	Software may correct for both autocorrelation and seasonality.
	Adjust the standard error for the trend (difference or slope) to accommodate for the reduced effective degrees of freedom		Need to calculate the autocorrelation coefficient at lag 1 for this adjustment (see section 7.3.6)	
Adjust for seasonality or other cycles	Use non-parametric trend tests that adjust for seasonality		Generally the month of year is needed for the input data set	
	Add explanatory variables that 'explain' the season affect		e.g., add data columns representing seasonal components for seasonal cycle (e.g., sin/cos terms) or monthly indicator variables	
	Use time series models that incorporate a lag term(s) to incorporate for seasonal cycles into statistical models			

*Key to Method Type: G = Graphical, N = Nonparametric, P = Parametric

7.2.2 Dealing with Censored Data

Censored values are usually associated with limitations of measurement or sample analysis, and are commonly reported as results below or above measurement capacity of the available analytical equipment. Table 7-3 summarizes techniques to use when dealing with censored data.

Table 7-3. Methods to deal with censored data (see discussion, section 7.4)

Analytical Objective	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
Accommodate censored data (i.e., values less than detection or reporting limits)	Use parametric (e.g., maximum likelihood estimation (MLE) and robust regression on order statistics (ROS)) or nonparametric procedures designed to accommodate censored data.	P, N, G	Knowledge about analytical detection limits, practical quantitation limits, and data reporting conventions is required to interpret the meaning of censored data.	Although common, substitution of half the detection limit is not recommended as more robust tools are readily available.

*Key to Method Type: G = Graphical, N = Nonparametric, P = Parametric

7.2.3 Data Analysis for Water Quality Problem Assessment

Problem assessment is generally considered the first phase of a watershed project. Data analysis at this stage typically involves using historical data to assess whether water quality standards are being met or whether designated beneficial uses of waters are threatened, and the causes (e.g., pollutants) and sources of identified problems. More refined problem assessment will include determination of pollutant pathways and critical areas needing restoration or BMPs. Methods to support these types of analyses are summarized in Table 7-4.

Table 7-4. Data analysis methods for problem assessment (see discussion, section 7.5)

Analytical Objective	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
Summarize existing conditions	Univariate statistics (e.g., mean, median, range, variance, interquartile range) for different sampling locations, time series analysis for long-term trends and seasonality, and regression analysis comparing pollutant concentrations or loads to hydraulic variables	P,N	Minimal to moderate	To compare locations within or across watersheds, data from different locations must be consistent and comparable (e.g., synoptic survey, multiple sampling stations).
	Boxplots and/or time series plots for different sampling locations	G		
Assess compliance with water quality standards	Identification of extreme values with boxplots or time series plots; calculation of means (arithmetic or geometric) over specific time period(s)	P	Minimal to moderate	Criteria for determining impairment vary (e.g., single observation exceedance vs. geometric mean over n observations); both monitoring program and data analysis must be tailored to regulatory requirements.
	Frequency or probability plots, duration curves	G		
Identify major pollutant sources	Correlation or regression analysis or Kendall's Tau for monotonic association of water quality constituent(s) vs. subwatershed characteristic(s) (e.g., total P concentration vs. manured acres)	P, N, G	Concurrent data from monitored subwatersheds: subwatershed land use and/or management data	Correlation does not guarantee causation; consider transport and other pollutant delivery mechanisms.
	Compare boxplots or bivariate scatterplots from monitored subwatersheds with distinctive land use and/or management; ANCOVA analysis	G, P		
Define critical areas	t-Test, ANOVA, Kruskall-Wallis, cluster analysis to identify significant differences in pollutant concentration/load among multiple sampling points	P, N	Concurrent data from monitored subwatersheds:parametric or nonparametric analysis can be used depending on data distribution	Conditions determining pollutant generation (e.g., storm event, season, management schedule) must be considered in drawing conclusions about critical areas. Modeling may be useful.

*Key to Method Type: G = Graphical, N = Nonparametric, P = Parametric

7.2.4 Project Planning Data Analysis

Project planning involves both land treatment and monitoring design. Decisions regarding project duration, BMP and restoration needs and scheduling, and implementation tracking and monitoring should all be supported by information and appropriate analysis. The quality of information available will vary from project to project. In many cases, the analysis and decisions will have to rely on historical data (perhaps collected for other purposes) or on data from other sites in the region. The methods summarized in Table 7-5 are recommended to assist with various aspects of project planning.

Table 7-5. Data analysis methods for project planning (see discussion, section 7.6)

Analytical Objective	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
Determine pollutant reductions needed to meet water quality objectives	Mass balance/TMDL Receiving waterbody relationships Load-duration curves Reference watershed	P, G		
Estimate BMP treatment needs	Compare estimated pollutant reduction efficiencies of planned BMPs with reductions needed	P	Appropriate local or published values on BMP pollutant reduction efficiencies	Published efficiencies do not generally account for interactions in multiple-BMP systems or pollutant transport or delivery issues beyond edge of field/BMP site. Modeling may be a better approach.
Estimate minimum detectable change (MDC)	MDC calculation (Spooner et al. 2011a)	P	Mean and variance of water quality variable(s) of interest; parameters of planned monitoring program (e.g., sampling frequency)	If MDC is larger than anticipated response to treatment, may need to re-evaluate extent of planned land treatment and/or duration of water quality monitoring. If data are unavailable from subject watershed, data from elsewhere must be used.
Locate monitoring stations	Identify major pollutant sources, critical areas as in Table 7.5 if data are available	P	Concurrent data from subwatersheds (e.g., from a synoptic survey)	Conditions determining pollutant generation (e.g., storm event, season, management schedule) must be considered.
	Target land areas of particular land use/management and/or expected treatment implementation	G	Land use and management data, estimates of treatment adoption	Station location depends on many other factors, including project objectives, monitoring design, and site requirements.

*Key to Method Type: G = Graphical, N = Nonparametric, P = Parametric

7.2.5 BMP and Project Effectiveness Data Analysis

Table 7-6 includes recommended methods for assessing the effectiveness of BMPs and watershed projects. In general, the analytical objective of both kinds of efforts is to document change in pollutant concentrations or loads or both in response to BMP implementation. These methods are linked to monitoring designs that are described in section 2.4. Methods for assessing BMP and project effectiveness using biological data are presented in chapter 4.

Table 7-6. Data analysis methods for assessing BMP or watershed project effectiveness (see discussion, sections 7.7 and 7.8)

Analytical Objective	Monitoring Design Used	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
BMP efficiency	Plot	ANOVA	P	Data must meet assumptions for parametric statistics to apply; otherwise use nonparametric test	Plot data may not easily extrapolate to field or watershed scale.
		Kruskal-Wallis	N		
	Input/output	Paired t-Test, Wilcoxon, or Mann-Whitney tests of input vs. output EMCs (Event Mean Concentrations) or loads	P, N	Data must meet assumptions for parametric statistics to apply; otherwise use nonparametric test	Representing change in load or concentrations as a percent reduction may not be representative for low input concentrations or loads.
		Effluent probability	N		
Watershed project effectiveness	Paired watershed	ANCOVA, paired t-Test, Wilcoxon Rank Sum, Mann-Whitney	P, N	Data from control and treatment watersheds must exhibit significant linear relationship. Conditions (e.g., precipitation, discharge) must be in similar range during calibration and treatment periods.	Quality of relationship between control and treatment watersheds determines level of change that can be detected. Addition of covariates to paired regression model may improve ability to document response to treatment.
	Above/below-Before/after	t-Test of input vs. output EMCs or loads, ANCOVA, Wilcoxon Rank Sum, Mann-Whitney	P, N	Data must meet assumptions for parametric statistics to apply; otherwise use nonparametric test	Change in pollutant concentration or load measured at the below station may be difficult to detect if concentrations or loads at the above station are high.
	Single Watershed Monotonic Trend	Linear regression on time Multiple linear regression on time and covariates Linear regression on time, covariates, and periodic functions	P	Numerous techniques are available, depending on objectives, available data on covariates, seasonality	Trend analysis is most effective with data sampled consistently at fixed locations and fixed time intervals for period sufficient to overlap seasonal or management cycles that do not represent real trends. Covariates such as stream flow, season, etc. are essential to assist with isolating trends due to BMPs.
		Mann-Kendall Mann-Kendall on residuals from regression on covariates Seasonal Kendall	N	Numerous techniques are available, depending on objectives, available data on covariates, seasonality	

Analytical Objective	Monitoring Design Used	Recommended Method	Method Type*	Data Requirements	Major Cautions and Concerns
	Single Watershed Step Trend	t-Test before and after step, Wilcoxon Rank Sum, Mann-Whitney	P, N	Data must meet assumptions for parametric statistics to apply; otherwise use nonparametric test	Selection of step change point in time must be made *a priori* and related to watershed activities, e.g., onset of treatment. Covariates such as stream flow, season, etc. are essential to assist with isolating trends due to BMPs.
	Multiple watersheds	t-Test or Wilcoxon Rank-Sum test ANOVA or Kruskal-Wallis test Regression analysis	P, N	Data must meet assumptions for parametric statistics to apply; otherwise use nonparametric test	Watersheds need to fall into 2 groups (e.g., treated and untreated) for t-Test or Wilcoxon Rank-Sum test. For more than two groups use ANOVA or Kruskal-Wallis.
		Boxplots of results from watershed groupings (e.g., treated/untreated)	G	Minimal	Visual comparisons should be confirmed by numerical tests.
	Linking land treatment to water quality changes	Correlation, regression of pollutant concentration or load on land treatment metric(s)	P, N	Requires quantitative monitoring data on land treatment. Use of explanatory variables (e.g., precipitation, animal populations) may strengthen analysis.	Water quality and land treatment data must be collected on comparable spatial and temporal scales. Monitored pollutants must match pollutants addressed by implemented BMPs.

*Key to Method Type: G = Graphical, N = Nonparametric, P = Parametric

7.2.6 Practice Datasets

This chapter presents a wide range of parametric and nonparametric methods, including several illustrative examples. Because practice is the best way to learn how to apply these methods, example datasets and eight problems are provided to allow readers to test their skills. Using their own statistics software, readers are encouraged to apply the tests indicated in Table 7-7 to the example datasets listed in the fourth column. The objective and statistical tests are listed in the second and third columns of the table. The specific problems and the answers are given in the files identified in the last column.

Table 7-7. Practice datasets

Problem Number	Objective	Test	Dataset in Sampledata.xlsx	Problem and Answer File
1	Test for conformance to normal distribution	Graphical, skewness, kurtosis, Shapiro-Wilk, Kolmogorov	1	normality.pdf
2	Characterize data	Descriptive statistics	1	description.pdf
3	Compare two groups	t-Test	1	2groups.pdf
		Wilcoxon/Kruskal-Wallace		
4	Compare input/output for a BMP	Paired t-Test	2	pairedtests.pdf
		Wilcoxon Rank Sum Test		
5	Compare three groups	ANOVA	1	3groups.pdf
		Kruskal-Wallace		
6	Examine relationships between variables/stations	Correlation	1	correlationregress.pdf
		Simple linear regression		
7	Assess change due to treatment in paired-watershed design	ANCOVA	1	pairedancova.pdf
8	Calculate MDC for a single station	Minimum detectable change	3	mdc.pdf

All files are available at: https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/monitoring-and-evaluating-nonpoint-source-watershed

7.3 Exploratory Data Analysis (EDA) and Data Adjustment

After a monitoring program is up and running, it is never too soon to begin to evaluate the data. Basic data evaluation should not wait until the end of the project or when a report is due; regular examination of the data should be part of ongoing project activities. A carefully designed monitoring program will have the right kind of data, collected at appropriate times and locations to achieve the objectives, and a plan for analyzing the data.

Describing and summarizing the data in a way that conveys their important characteristics is one purpose of EDA. When deciding how to analyze any data set, it is essential to consider the characteristics of the data themselves. Evaluation of characteristics like non-normal distribution and autocorrelation will help determine the appropriate statistical analysis. Some common characteristics of water quantity and quality data (Helsel and Hirsch 2002) include:

- A lower bound of zero – no negative values are possible.
- Presence of outliers, extreme low or high values that occur infrequently, but usually somewhere in the data set (outliers on the high side are common).
- Skewed distribution, due to outliers or influential data.
- Non-normal distribution.
- Censored data – concentration data reported below some detection limit or above a certain value.
- Strong seasonal patterns.
- Autocorrelation – consecutive observations strongly correlated with each other.

- Dependence on other uncontrolled or unmeasured variables – values strongly co-vary with such variables as streamflow, precipitation, or sediment grain size.

As such, the overall goal of data exploration is to uncover the underlying structure of a data set and set the stage for more detailed analysis, including hypothesis testing. Specific objectives for data exploration might include:

- To find potential problems with data quality such as data entry error, lab or collection errors
- To find extreme values and potential anomalies
- To describe the behavior of one or more variables
- To test distribution and assumptions of independence and constant variance
- To see cycles and trends
- To find clusters or groupings
- To make preliminary comparisons of two or more locations or time periods
- To examine relationships between variables

At the start, check the data for conformity with original plans and QA/QC procedures. Use the approved project Quality Assurance Project Plan (QAPP) as a guide; see section 8.3 for details on preparing a QAPP. A key part of EDA is to verify the data entered in the data sets are valid and not anomalies due to data entry, lab, or collection errors.

Understanding how the data behave with respect to such features as distribution(s), cycles, clusters, seasonality, and autocorrelation assists with selecting the appropriate statistical tests to evaluate achievement of project goals. Data analysis to address project goals will involve more thorough statistical analysis that will be guided by understanding of the data set through EDA.

A secondary reason for doing exploratory data analysis is to start to make sense of the data actually collected. The purpose of EDA is to get a feel for the data, develop ideas about what it can tell, and how to draw some preliminary conclusions. EDA is similar to detective work – sifting through all the facts, looking for clues, and putting the pieces together to find suggestions of meaning in the data.

This process of data exploration differs from traditional hypothesis testing. Testing of hypotheses always requires some initial assumption or prediction about the data, such as "The BMP will reduce phosphorus loads." Although formulating and testing hypotheses is the foundation of good data analysis, the first pass through of the data should not be too narrowly focused on testing a single idea. Hypothesis testing is discussed in section 7.6.1, which focuses on data analysis for project planning. EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow in favor of the more direct approach of allowing the data themselves to reveal their underlying structure. EDA uses a variety of techniques, both numerical and graphical, to open-mindedly search for new, perhaps unexpected, insights into the data. Approaches to EDA for aquatic system biological data have been described by EPA as part of the Causal Analysis/Diagnosis - Decision Information System (CADDIS) (USEPA 2010).

Data exploration is a necessary first step in analyzing monitoring data. Unless initial exploration reveals indications of patterns and relationships, there is unlikely to be something for further analysis to confirm.

J. W. Tukey (1977), the founder of exploratory data analysis, said, "EDA can never be the whole story, but nothing else can serve as the …first step."

For more information, refer to *Tech Notes 1: Monitoring Data Exploring Your Data, The First Step* (Meals and Dressing 2005).

7.3.1 Steps in Data Exploration

Data exploration is a process of probing more deeply into the dataset, while being careful to stay organized and avoid errors. Here are some typical steps in the process of EDA (modified from Jambu 1991), although not all of them may apply to every situation.

1. **Data management.** In the process of working with the data, files will be created. These files should be updated, checked, and validated at regular intervals. The importance of data screening and validation cannot be overemphasized. This should always be done before embarking on specific analyses, plotting, or other procedures. Be as sure as possible that the data are free from entry errors, typos, and other mistakes before proceeding.
2. **One-dimensional analysis.** The first step in really exploring the data is often to simply describe or summarize the information one variable at a time, independent of other variables. This can be done using basic statistics on range, central tendency, and variability, or with simple graphs like histograms, pie charts, or time plots. This kind of information is always useful to put data in context, even though more intensive statistical analysis will be pursued later.
3. **Two-dimensional analysis.** Relationships between two variables are often of great interest, especially if there is a meaningful connection suspected (such as between suspended sediment and phosphorus) or cause and effect process (such as between rainfall and streamflow). Relationships between two sampling locations (such as treatment and control watersheds) or between two time periods (like spring snowmelt and summer) are often of interest. Graphical techniques like scatter plots and numerical techniques like correlation are often used for this purpose.

Because graphs summarize data in ways that describe essential information more quickly and completely than do tables of numbers, graphics are important diagnostic tools for exploring the data. There is no single statistical tool that is as powerful as a well-chosen graph (Chambers et al. 1983). Enormous amounts of quantitative information can be conveyed by graphs and the human eye-brain system is capable of quickly summarizing information, simultaneously appreciating overall patterns and minute details. Graphs will also be essential in ultimately conveying project results to others. With computers and software available today, there are no real constraints to graphing data as part of EDA. Graphical display options are described in section 4.3 of the 1997 guidance (USEPA 1997b).

There are more advanced steps in data exploration including analysis of multiple variables and cluster analysis (section 7.3.8). Also, see chapter 4 of the 1997 guidance (USEPA 1997b) for background on some of these methods.

The project goals and the type of monitoring should guide exploration. If monitoring occurs at a single point while upstream BMPs are implemented gradually, trends may be of the greatest interest. If sampling for phosphorus above and below a land treatment area, a comparison of phosphorus concentrations at the two stations might be necessary. For an erosion problem, a relationship between streamflow and suspended solids concentrations before and after land treatment might be of interest.

The following sections present some specific techniques for exploring data.

7.3.2 Describe Key Variable Characteristics

In most cases, the data should be examined to summarize key characteristics and to determine if the data satisfy statistical assumptions required for parametric statistical analyses. Data that do not meet parametric statistical assumptions should be transformed or nonparametric tests should be used. Key characteristics that are meaningful include central tendency, variability, and distribution.

7.3.2.1 Central Tendency

- The **mean** is computed as the sum of all values divided by the number of values. The mean is probably the most common data summary technique in use; however, an extreme value (either high or low) has much greater influence on the mean than does a more 'typical' value. Because of this sensitivity to extremes, the mean may not be the best summary of the central tendency of the data.
- The **median,** or 50th percentile, is the central value of the distribution when the data are ranked in numerical order. The median is the data value for which half of the observations are higher and half are lower. Because it is determined by the order of observations, the median is only slightly affected by the magnitude of a single extreme value. When a summary value is desired that is not strongly influenced by a few extremes, the median is preferable to the mean.

Both the mean and median should be calculated for comparison.

7.3.2.2 Variability

- The sample **variance**, and its square root the **standard deviation**, are the most common measures of the spread (dispersion) of a set of data. These statistics are computed using the squares of the difference between each data point and the mean, so that outliers influence their magnitudes dramatically. In data sets with major outliers, the variance and standard deviation may suggest a much greater spread than exists for the majority of the data. This is a good reason to supplement numerical statistics with graphical analysis.
- The **coefficient of variation (CV)**, defined as the standard deviation divided by the mean, is a relative measure of the variability (spread) of the data. The CV is sometimes expressed as a percent, with larger values indicating higher variability around the mean. Comparing the CV of two data groups can suggest their relative variability.
- The **interquartile range** (IQR) is defined as the 75th percentile minus the 25th percentile. Because it measures the range of the central 50 percent of the data, it is not influenced at all by the 25 percent of the data on either end and is relatively insensitive to outliers.

7.3.2.3 Skewness

Water resources data are usually skewed, meaning that the data values are not symmetric around the mean or median, as extreme values extend out farther in one direction. Streamflow data, for example, are typically right-skewed because of occasional high-flow events (Figure 7-1). When data are skewed, the mean is not equal to the median, but is pulled toward the long tail of the distribution by the effects of the extreme values. The standard deviation is also inflated by the extreme values. Because highly skewed data restrict the ability to use hypothesis tests that assume the data have a normal distribution, it is useful to evaluate the skewness of the data. The **coefficient of skewness** (g) is a common measure of skewness; a right-skewed distribution has a positive g and a left-skewed distribution has a negative g. There are multiple measures of skewness with varying possible ranges. Interpretation of skewness values calculated by Excel, for example, is aided by estimating the standard error of skewness with the following simplified[1] equation for large (<5 percent difference from true value for n≥30) samples (Elliott 2012):

$$Standard\ Error = \sqrt{6/n}$$

where n is the sample size. For n=24, the standard error of skewness is 0.5 using the simplified equation. A skewness value of more than twice this amount (i.e., less than -1 or greater than 1 in this case) indicates a skewed distribution, but a value between -1 and 1 is not proof that the data are normally distributed. Other tests such as goodness-of-fit tests (below) must also be performed to determine if the distribution is normal.

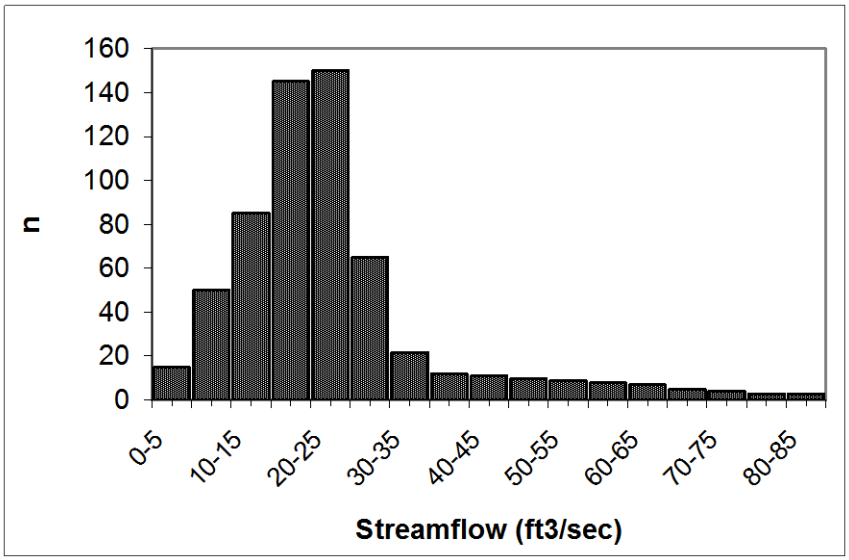


Figure 7-1. Right-skewed distribution

[1] The true standard error of skewness is calculated as: $\sqrt{6n(n-1)/(n-2)(n+1)(n+3)}$

7.3.2.4 Data Distribution

Many common statistical techniques for hypothesis-testing (parametric tests) require, among other characteristics, that the data be normally distributed. It is common practice to apply tests such as the Shapiro-Wilk test or the Kolmogorov-Smirnov (KS) test to evaluate the normality of the data; both of these tests are commonly available in statistical software. The probability plot correlation coefficient (PPCC) can also be used to test for normality. PPCC is essentially a correlation coefficient between the data values and their normal score (i.e., data on probability paper) and the interpretation of the PPCC is similar to that for the correlation coefficient r. This procedure is outlined by Helsel and Hirsch (2002) in section 4.4 and in Appendix Table B.3 which gives critical values for accepting/rejecting the normal assumption.

Histograms are familiar graphs, where bars are drawn whose height represents the number or fraction of observations falling into one of several categories or intervals (see Figure 7-1). Histograms are useful for depicting the shape or symmetry of a data set, especially whether the data appear to be skewed. However, histogram appearance depends strongly on the number of categories selected for the plot. For this reason, histograms are most useful to show data that have natural categories or groupings, such as fish numbers by species, but are more problematic for data measured on a continuous scale such as streamflow or phosphorus concentration.

Quantile plots (also called cumulative frequency plots) show the percentiles of the data distribution. Many statistics packages calculate and plot frequency distributions; instructions for manually constructing a quantile plot can be found in Helsel and Hirsch (2002) and other statistics textbooks. Quantile plots show many important data characteristics, such as the median or the percent of observations less than or greater than some critical threshold or frequency. With experience, an analyst can discern information about the spread and skewness of the data. Figure 7-2 shows a quantile plot of *E. coli* bacteria in a stream; the frequency of violation of the Vermont water quality standard can be easily seen (the standard was exceeded ~65 percent of the time). Flow and load duration curves (see section 7.9.3) are useful tools for visualizing the distribution of streamflows or pollutant loads across a full range of conditions.

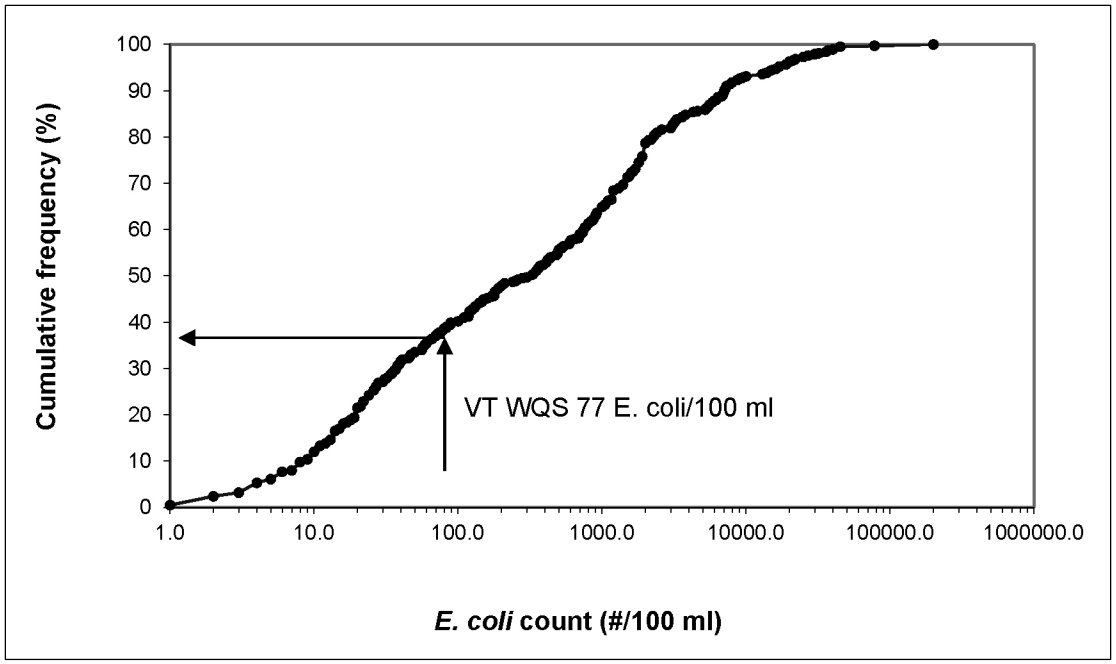


Figure 7-2. Quantile plot or cumulative frequency plot of *E. coli* data, Berry Brook, 1996 (Meals 2001)

A boxplot presents a schematic of essential data characteristics in a simple and direct way: central tendency (median), spread (interquartile range), skewness (relative size of the box halves), and the presence of outliers are all indicated in a simple picture. There are many variations and styles of boxplots, but the standard boxplot (Figure 7-3) consists of a rectangle spanning the 25th and 75th percentiles, split by a line representing the median. Whiskers extend vertically to encompass the range of most of the data (e.g., the 5th and 95th percentiles), and outliers beyond this range are shown by dots or other symbols. The definition of whiskers and outliers may differ among graphing programs; standard definitions can be found in statistics textbooks (e.g., Cleveland 1993; Helsel and Hirsch 2002). When boxplots are presented, the definitions of the rectangle, whiskers, and outlier symbols should be clearly specified.

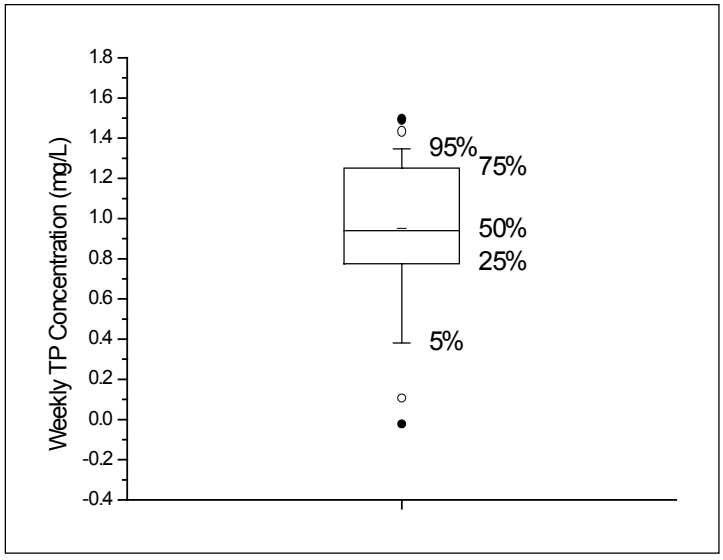


Figure 7-3. Boxplot of weekly TP concentration, Samsonville Brook, 1995 (Meals 2001)

7.3.2.5 Transformations to Handle Non-normal Data with Parametric Statistical Tests

Evaluations conducted thus far may suggest that the data do not conform to a normal distribution. In cases where it is desirable or convenient to use statistical tools that require normally distributed data sets or have a constant variance, transformation may reduce skewness and result in a data set that is more normally distributed. Transformation is simply defined as applying the same mathematical operation to all records in the dataset. Helsel and Hirsch (2002) provide a summary of common transformations. Statistical software packages will often come with Box-Cox transformation tools that allow the analyst to identify the best transformation for achieving normality, although logarithmic (e.g., $\log_{10}$ or $\log_e$) transformation is certainly the most common strategy (Box and Cox 1964). Regardless of which transformation is used, the data analyst should verify that the transformation results in a dataset that satisfies applicable assumptions.

Subsequent analysis of log-transformed data must be done with care, as quantities such as mean and variance calculated on the transformed scale are often biased when transformed back to the original scale. The geometric mean (the mean of the log-transformed data back-transformed to the arithmetic scale), for example, differs from the mean of the untransformed distribution. Furthermore, results of statistical analysis may be more difficult to understand or interpret when expressed on the transformed scale. Typically, when analysis is performed on the log transformed data, the final statistical results are converted to express the results as a percentage change (see Spooner et al. 2011a for additional details on this approach).

Do not assume that a transformation will solve all the problems with the data distribution. Always test the characteristics of the transformed data set again. Violations of the assumption of a normal distribution can lead to incorrect conclusions about the data when parametric tests are used in subsequent hypothesis testing. With that said, some parametric trend tests are robust to some deviation from normality. From a practical standpoint it is best to be consistent. For example, if a log transformation is merited for TP concentrations at most locations in a particular data set, then log transforming all TP for all site locations is a practical course of action.

If transformed data cannot satisfy the assumptions of parametric statistical analysis, consider nonparametric techniques for data analysis. With regard to hypothesis testing, there are a host of nonparametric tests that are robust against non-normality. These tests are often based on the ranks of the data and the influence of a few extreme values is reduced. However, keep in mind that while the normality assumption is relaxed, nonparametric tests have other assumptions (constant variance and independence of data observations) that must be met for their results to be valid. If distributional assumptions can be met, then parametric tools tend to be more powerful. Many nonparametric procedures are described in section 4.11.3 in the 1997 guidance and recommended in Table 7-1 through Table 7-6.

7.3.3 Examination for Extreme, Outlier, Missing, or Anomalous Values

7.3.3.1 Extremes and Outliers

Extreme values are frequently encountered in NPS monitoring efforts and include the exceptionally high and low flow values associated with floods and droughts, respectively. Suspended sediment concentrations may be exceptionally high during spring runoff when cropland fields are bare or when streambank slumping occurs. Very low pesticide levels may be observed with increasing time elapsed since application on cropland. In some cases, the extremes may be more important for water quality than are typical conditions. For example, the extreme values in some lake variables (e.g., Secchi disc readings,

turbidity, and pH), the duration of the extreme values, and the season may be the dominant influence on the extent to which lakes support designated beneficial uses. In streams, it is often the extreme low dissolved oxygen condition that determines the character of the biological community the stream can support. Extreme concentrations in toxic contaminants such as pesticides may also be more important than the mean values with respect to acute toxicity to aquatic biota. Nevertheless, extreme concentrations can have an inordinate effect on some statistical analyses, and the analyst must consider these issues when selecting data analysis tools.

On the other hand, outliers can result from measurement or recording errors and this should be the first thing checked (e.g., check lab and field logs). *If no error can be found, an outlier should never be rejected just because it appears unusual or extreme. All samples considered valid after exploratory analysis contain information that should be considered when analyzing monitoring data.* Different subsets of the same dataset may reveal varying aspects of the condition of the water resource. For example, extreme conditions may be most important when considering violations of water quality standards or load allocations from a watershed. Annual or monthly loads may not completely illuminate the severity of a problem, whereas high loads during extreme flow conditions may account for most of the pollutant load. It is commonly observed that the majority of annual pollutant export occurs during a small proportion of the time. Identifying these extremes and understanding the conditions under which they occur may be a key to understanding and interpreting watershed monitoring results.

One approach for identifying and summarizing extreme values is to describe the situation by computing the frequency or proportion of observations exceeding some threshold value (e.g., a water quality criterion). Cumulative frequency or duration plots are also useful to visualize the influence of extreme values on a dataset. In addition, determine whether most or all of the extreme values can be attributed to certain conditions in the watershed (e.g., spring runoff, cropland tillage). In these cases, it might be more useful to stratify the dataset by season or management condition. In this way, monitoring results can be analyzed by season, and values that were "extreme" in the dataset as a whole may be more easily interpreted in their respective season(s).

Histograms can be useful to illustrate exceedances of standards, targets, and goals by setting categories or classes that are outside the standard or target. Quartile plots and boxplots are also useful tools to evaluate the presence of extreme values.

Boxplots can be a useful visual tool for highlighting extreme values in environmental data. They show both the spread and the range of the data. Important values visualized by boxplots include the mean (or the median), and standard error limits (or 25^{th} and 75^{th} percentiles). Values falling outside these 'limits' depict values that are from the tails of the data distribution.

Plotting the data in sequence with date as the horizontal axis are time series plots. Figure 7-4 shows a time series plot of weekly phosphorus concentration data from three stream stations. It is clear that around the middle of the year, something occurred that led to dramatic spikes in P concentration at Station 2, a phenomenon demanding further investigation. Field investigation revealed concentrated overland flow from a new CAFO upstream.

To analyze data sets with extreme values, consider using non-parametric trend tests. If documenting the number or occurrence of extreme values is an objective (e.g., for evaluation of violations of water quality standards or pesticide spikes), frequency analyses are useful. Stratifying the data by seasons or flow conditions (e.g., base flow, storm flows, and flooding) may be helpful in evaluating conditions and trends within each flow regime. Using flow as an explanatory variable/covariate in trend analysis may be helpful

to explain the influence/importance of the extreme values. Using the log transformation often minimizes the skewness caused by the extreme values and enables the use of parametric trend techniques. If the data are missing due to right censoring (too high to measure), techniques discussed in section 7.4 should be considered.

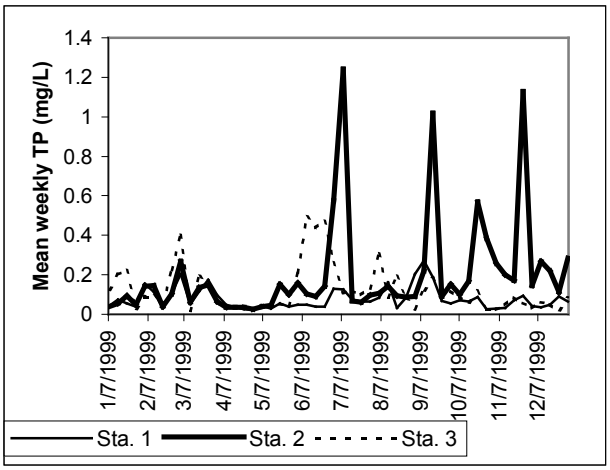


Figure 7-4. Time plot of weekly TP concentration, Godin Brook, 1999 (Meals 2001)

7.3.3.2 Anomalous Values

Plotting the data can also reveal data errors or anomalies. Figure 7-5 shows a time series plot of total Kjeldahl nitrogen (TKN) data collected from three Vermont streams. Something happened around May, 1996 that caused a major shift in TKN concentrations in all three streams. In addition, it is clear that after October, no values less than 0.5 mg/L were recorded. In this case, this shift was not the result of some occurrence in the watersheds, but an artifact of a faulty laboratory instrument, followed by the establishment of a lower detection limit of 0.50 mg/L. Discovery of this fault, while it invalidated a considerable amount of prior data, led to correction of the problem in the lab and saved the project major headaches down the road.

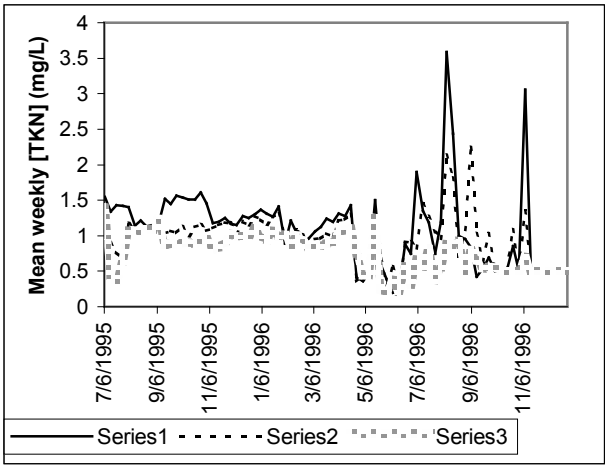


Figure 7-5. Time plot of TKN data from three stream stations, 1995-1996 (Meals 2001)

7.3.3.3 Missing Data

The reality of any watershed project monitoring program is that samples will be missed, equipment will fail or be overwhelmed, droughts and floods will occur, and sample analysis limitations will be exposed, resulting in missing and extreme (both high and low) values. However, if data are missing because of extreme conditions (e.g., streamflow was too high to obtain a measurement or water was so low a sample could not be drawn), then missing data may also represent extreme conditions.

The presence of a few missing values in a data series is not generally a major cause for concern, although some parametric tests (e.g., trend analyses that include autocorrelation errors using time series) require equal spacing of observations[2]. One way to cope with extensive missing data is to aggregate data to a longer, uniform time interval by averaging or using the median value of a group of data points. Daily observations, for example, could be aggregated to weekly means or medians. Such an operation would have an additional potential benefit of reducing autocorrelation (see section 7.3.6). A downside to this approach, however, is a reduced significance level due to fewer degrees of freedom. Do not aggregate data when there is a systematic change in sampling. For example, if the early data were collected as monthly observations and the more recent data were collected as quarterly data, it is not correct to aggregate the monthly data to quarterly averages and then perform analyses. This is because the averaging calculation changes the variability of that portion of the record in comparison to the remainder of the record, resulting in a violation of "identically distributed" assumption of most (including nonparametric) hypothesis tests. In these cases, the analyst will need to subsample from the more intensely monitored data set to best mimic the sampling from the less sampled portion of the data.

For loading analyses that require flow data, it is expected that the missing flow data due to equipment failure could be estimated by evaluating regression relationships with flow from nearby basins. On the other hand, flows that exceed the weir capacity or reach a stage so high that the technician cannot access the site are exceptional events. Certainly one approach to addressing this data gap is to apply the previously mentioned regression relationship with a nearby station. Another approach might be to treat these observations as "greater than the maximum flow" and apply methods appropriate for censored data described in section 7.4.

7.3.4 Examination for Frequencies

For categorical data such as watershed area in different land uses or number of aquatic macroinvertebrates in certain taxonomic groups, data can be effectively summarized as frequencies in histograms or pie charts. Figure 7-6 shows a pie chart of the percent composition of orders of macroinvertebrates in a Vermont stream, clearly indicating that dipterans dominate the community.

[2] Some statistical software such PROC AUTOREG in SAS yield valid trend results with autocorrelated data with missing data points, as long as the input record contains equal spaced time intervals (e.g., weekly).

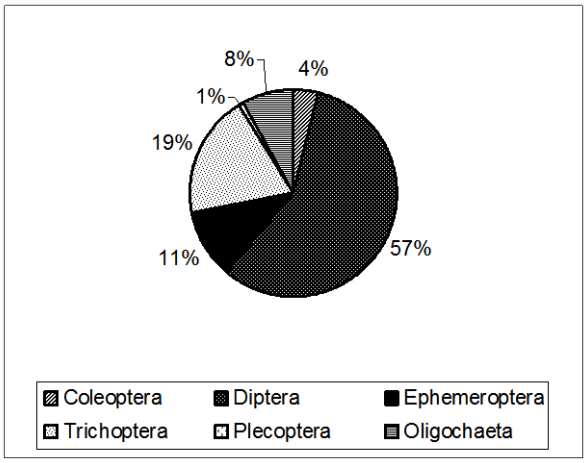


Figure 7-6. Percent composition of the orders of macroinvertebrates, Godin Brook, 2000 (Meals 2001)

7.3.5 Examination for Seasonality or Other Cycles

Monitoring data often consist of a series of observations in time, e.g., weekly samples over a year. One of the first, and the most useful, things to do with any time series data is to plot it. Plotting time series data can provide insight into seasonal patterns, trends, changes, and unexpected events more quickly and easily than tables of numbers.

Figure 7-7 shows a time series plot of *E. coli* counts in a Vermont stream. The extreme range of the counts (five orders of magnitude) and the pronounced seasonal cycle are readily apparent, with the lowest counts occurring during the winter. It is easy to see the times of year when the stream violates the water quality standard for bacteria.

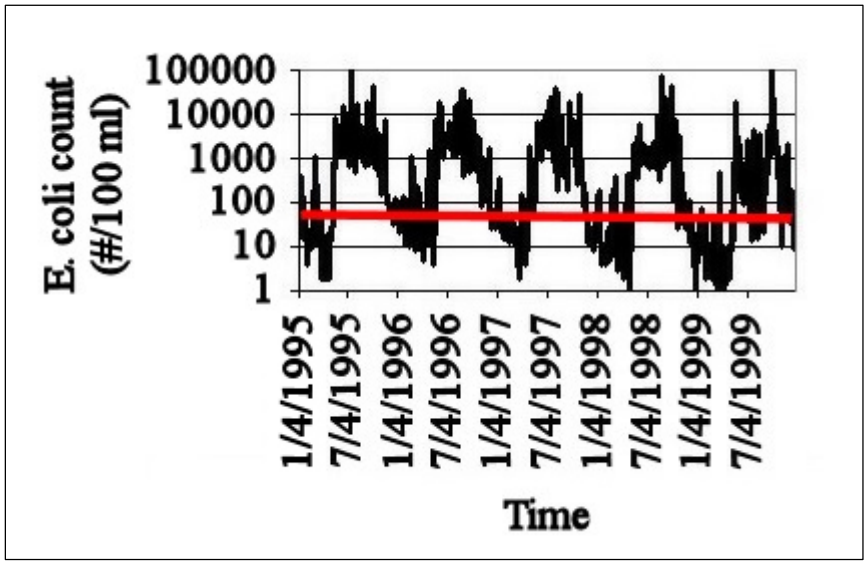


Figure 7-7. Time series plot of weekly *E. coli* counts, Godin Brook, 1995-1999 (Meals 2001). Red line indicates Vermont WQS of 77 *E. coli*/100 ml.

7.3.6 Autocorrelation

Because many hypothesis-testing statistical techniques require that residuals from the statistical tests be independent, it is useful to check the data set for autocorrelation during EDA. Typically, if the data points exhibit autocorrelation, so will the residuals from a statistical test which does not correct for autocorrelation.

Time series data collected through monitoring of water resources often exhibit autocorrelation (also called serial correlation or dependent observations) where the value of an observation is closely related to a previous observation (usually the one immediately before it). Autocorrelation in water quality observations is usually positive in that high values are followed by high values and low values are followed by low values. For example, streamflow data often show autocorrelation, as numerous high wet-weather flows tend to occur in sequence, while low values follow low values during dry periods.

Terms Used in this Section

Lag: the difference in time steps by which one observation comes after another. The lag value is the number of time steps.

Autocorrelation: the correlation between lagged values in a time series (data collected over equal intervals of time, can also be spatial distances)

Correlation Coefficients, ρ_j: a set of correlations for each lag. The autocorrelation coefficient for lag 1 is the correlation between each data in a time series and its previous (lag 1) observation. The autocorrelation coefficient, ρ_j, for lag j is the correlation between each datumin a time series and the observation that lags by j time steps.

Autoregressive: situation where past values (or nearby values for spatial analyses) have an effect on current values. For example, when most of the correlation between the lag variables is between each current value and the immediately preceding value, it is a first-order autoregressive process denoted as AR(1). AR(2) is second order, where previous two values effect the current value, etc. Autoregressive, order 1, AR(1) is common for weekly and monthly water quality samples.

Moving Average: an averaging of a fixed number of consecutive observations, with or without weights. Moving average models are denoted MA(1), MA(2), …MA(q) to indicate the order or maximum lag for consecutive observations that are averaged.

ARIMA (autoregressive integrated moving average) models: time series models that include both autoregressive terms and/or moving average terms

Autocorrelation Function (ACF): the set of correlations (e.g., autocorrelation coefficients) between each value in a series of values (e.g., x_t) and the lagged values within the same series (e.g., x_{t-1}, x_{t-2}, etc.). Alternatively stated, this is the pattern of correlation coefficients vs. lag value. This is generally depicted as a graph of each lag and its autocorrelation coefficient with a standard error bar to help determine the statistical significance of each of the correlation coefficients for each lag. The pattern/shape of the ACF, along with the PACF, is used to assist in determining if the data follow an AR, MA, or ARIMA pattern, and by what order (lag). For example, a seasonal AR(1) series has a large ρ_1, with subsequent ρ_j's trailing off, and a strong seasonal lag correlation.

Partial Autocorrelation Function (PACF): the correlation between two variables, taking into account the relationships of other variables to these two variables. The PACF for an AR(1) series drops to 0 after lag 1).

Autocorrelation usually results in a reduction of the effective sample size (degrees of freedom). It therefore affects statistical trend analyses and their interpretations. As the magnitude of autocorrelation increases, the effective sample size decreases, and the true standard error is therefore greater than if autocorrelation is incorrectly ignored. Adjustment for autocorrelation is needed so that the power of detecting a difference or trend is not incorrectly inflated. For data sets with high autocorrelation, a larger sample size (e.g., longer monitoring duration) than would be necessary in the absence of autocorrelation may be required to correctly detect significant changes or trends.

Autocorrelation is often significant in very frequent data collection, such as that done with recording sensors (e.g., temperature, turbidity). Daily, weekly, and monthly samples also exhibit autocorrelation, but usually to a lesser extent. The time interval between independent samples differs with the water resource and variable. The magnitude of autocorrelation in surface water quality concentrations is usually quite large for samples collected more frequently than monthly (Loftis and Ward 1980a and 1980b, Lettenmaier 1976, Lettenmaier 1978, Whitfield and Woods 1984). Loftis and Ward (1980a and 1980b) verified that some surface water quality samples collected less frequently than once a month may be considered independent if the seasonal variation is removed, although Whitfield (1983) found significant autocorrelation between stream discharge samples taken as much as 60 days apart. Compared to surface water data series, ground water data series tend to retain significant autocorrelation, even with longer sample intervals. Similarly, a ground water data series tends to have greater autocorrelation when compared to surface water data series taken at the same time intervals. This may be due to slower water movement and mixing in ground water as compared to surface waters.

There are numerical techniques to test for autocorrelation, but a simple graphical method can suggest whether data have significant autocorrelation: the lag plot. A lag plot is a graph where each data point is plotted against its predecessor in the time series, i.e., the value for day two and the value for day one are plotted as an x, y pair, then day three, day two, and so on. Different time lags can be examined. A "lag-1" plot uses each data value paired with its immediate predecessor (t2, t1), a "lag-2" plot uses each data value paired with the value observed two steps previously (t3, t1), and so on. Random (independent) data should not exhibit any identifiable structure or pattern in the lag plot. Non-random structure in the lag plot indicates that the underlying data are not random and that autocorrelation may exist. Figure 7-8 shows a lag-1 plot of weekly streamflow data, suggesting that autocorrelation needs to be addressed.

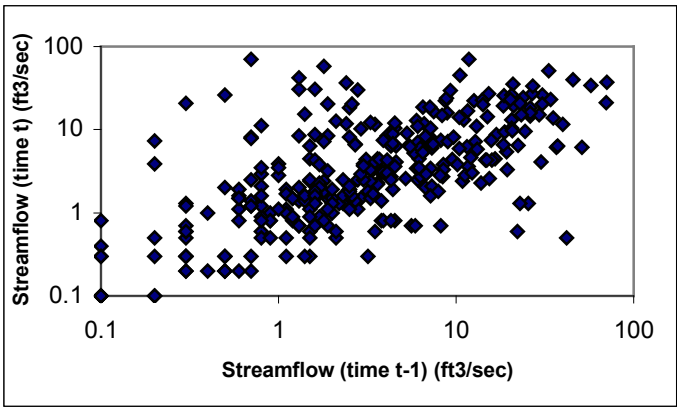


Figure 7-8. Lag-one plot of streamflow observations, Samsonville Brook, 1994 (Meals 2001)

Autocorrelation can be expressed numerically by calculating the correlation between observations separated by j lag time periods. The autocorrelation corresponding to the jth lag is the correlation between the observation at a given time and the observation taken j observation periods earlier. It is denoted by ρ_j or $\rho(j)$. For example, for the first lag (j=1), ρ_1 represents the autocorrelation of data points one time period removed. The time period is a function of the sample frequency and corresponds to the length of time between samples (e.g., daily, weekly, monthly). The range of values for ρ_j is -1 to +1, where +1 represents a perfect positive autocorrelation and -1 is a perfect negative correlation. The sample estimate of autocorrelation is given by r_j (in practice, ρ_j is often used to depict the sample autocorrelation coefficients).

Time series generally exhibit patterns indicated by the pattern of autocorrelation coefficients at various lags. These patterns reveal key characteristics about the data that should be incorporated into subsequent trend analyses. For weekly and less frequent water quality sample collection, the autoregressive, lag 1 or AR(1) data structure is usually appropriate. In this case, most of the autocorrelation can be explained by the correlation between each observation and its previous observation. Moving Average (MA) data structures occur when an observation is only related to the observations up to the lag value (q) and not observations before[3]. Rarely is a MA structure alone useful with water quality samples. However, for some daily or more frequent sampling, a combination of AR and MA data structures become appropriate, known as ARIMA (AutoRegressive Integrated Moving-Average) models.

One common test for autocorrelation is the Durbin-Watson (DW) test. The DW test is appropriately used when the data exhibit first order (lag 1) autoregressive (AR(1)) behavior. AR(1) is common with water quality data collected weekly, biweekly, or monthly. Daily or samples collected more frequently usually exhibit ARIMA autocorrelation structures. Even so, the DW test can be useful to indicate the presence of autocorrelation with such samples as well. The DW test may also be used to test for independence (i.e., the absence of autocorrelation) in the residuals from regression models.

Many statistical software packages offer tools for examining autocorrelation. For example, the Autocorrelation Function (ACF) is the set of all the lag j autocorrelations and is usually depicted as a plot of each lag autocorrelation versus the lag number (Figure 7-9 from Minitab (2016) and Figure 7-10 from JMP (SAS Institute 2016b)) for the same data set. Visual inspection of the ACF is useful to detect the presence of autocorrelation and define the structure of the autocorrelation. Typically, the lag autocorrelation confidence limits (approximately two-standard deviation errors) are also shown on the ACF graphs. This helps analysts determine if the autocorrelation coefficient at lag j is significant. Seasonal patterns show up as cycles in the ACF. As a point of comparison, Figure 7-11 shows a time series plot of independent data (i.e., zero correlation) together with its ACF graph.

Another useful graph is the Partial Autocorrelation Function (PACF) which is included as the last chart in Figure 7-9 and in the last column of Figure 7-10. The PACF is the partial amount of R-square (i.e., correlation) gained due to the additional lag term added to the right hand side of the model (Box and Jenkins 1976). Patterns of the PACF that show dramatic decrease to non-significant values after a lag j, indicate an autoregressive series of order (lag) j. For a qth order moving average model, MA(q), the theoretical ACF function drops off to 0 after lag q with an exponentially decaying PACF value between lag 0 and lag q.

[3] j and q both refer to the number of lags, j for AR and q for MA.

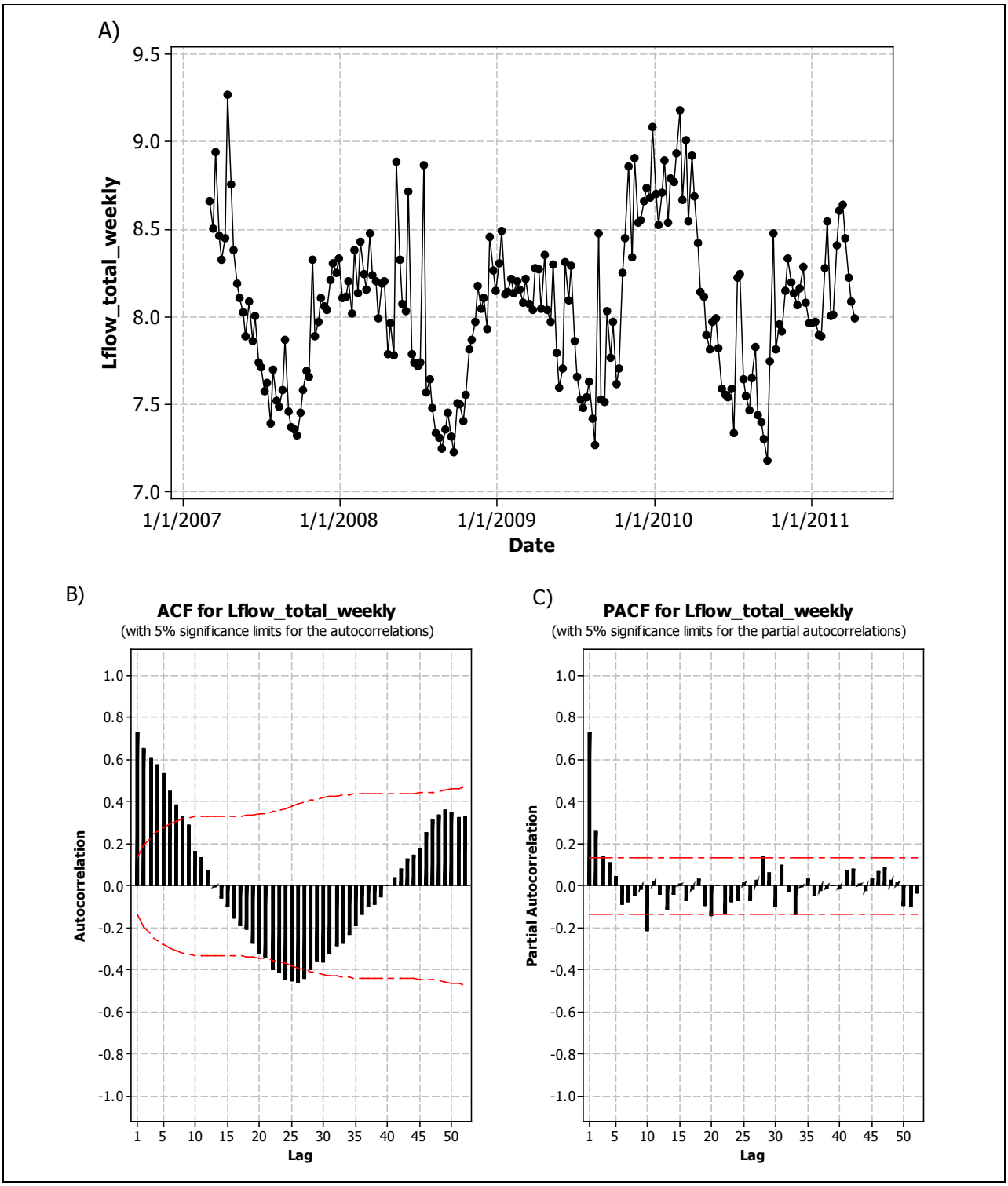


Figure 7-9. A) Time series plot, B) autocorrelation function (ACF) graph, and C) partial autocorrelation function (PACF) graph of Log(10) weekly flow from the Corsica River National Nonpoint Source Monitoring Program Project generated by Minitab. The steps are: Stat > Time Series > Autocorrelation (or Partial Autocorrelation). Identify the time series variable and enter number of lags. Select options for storing ACF, PACF, t statistics, and Ljung-Box Q statistics as desired. Press ok.

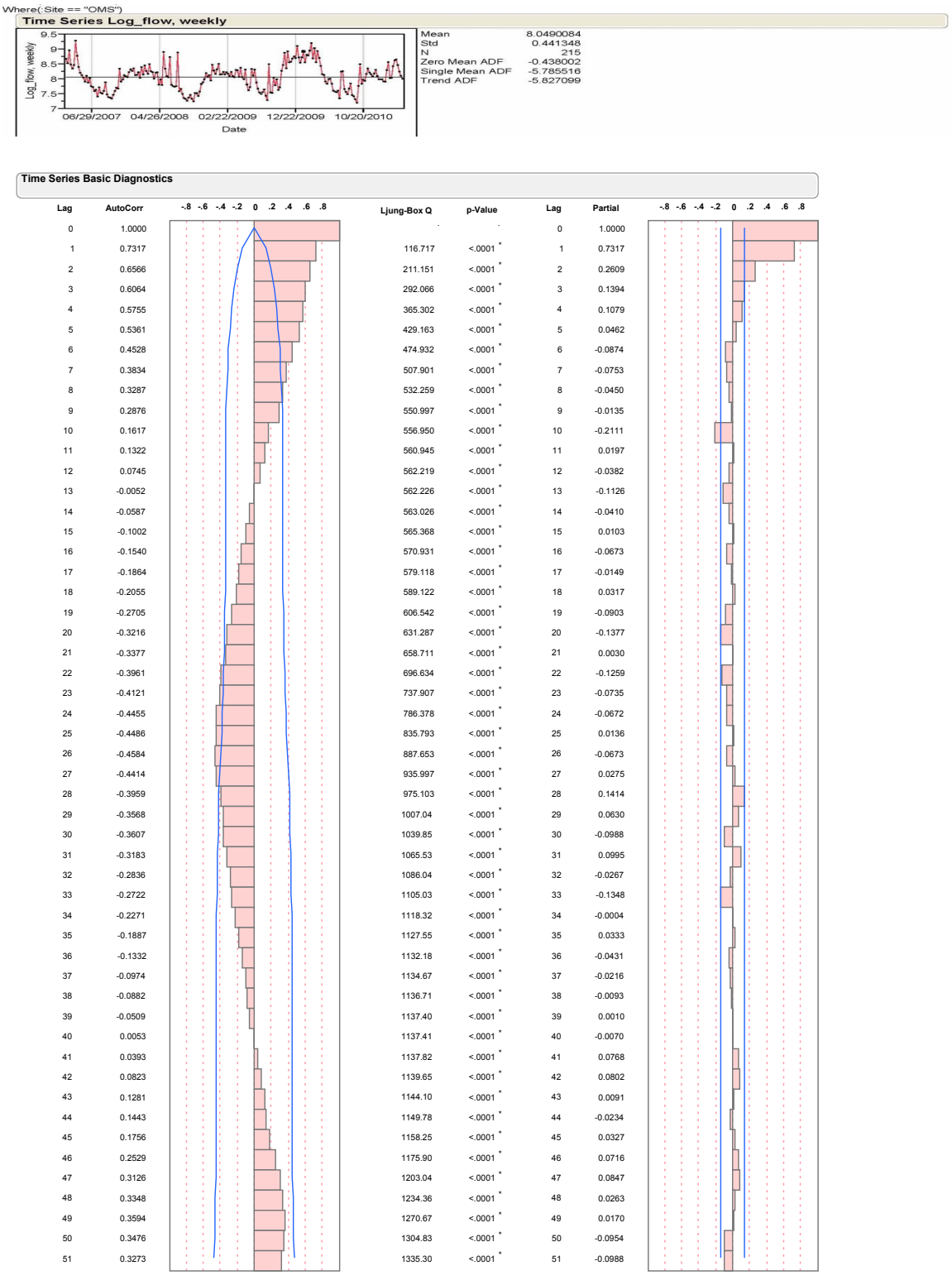


Figure 7-10. Autocorrelation Function (ACF) graph of weekly flow from the Corsica River National Nonpoint Source Monitoring Program Project generated by JMP. The steps are: Click "Analyze" tab, select "Modeling" followed by "Time Series." Select Y time series (LFLOW) and X time series (Date).

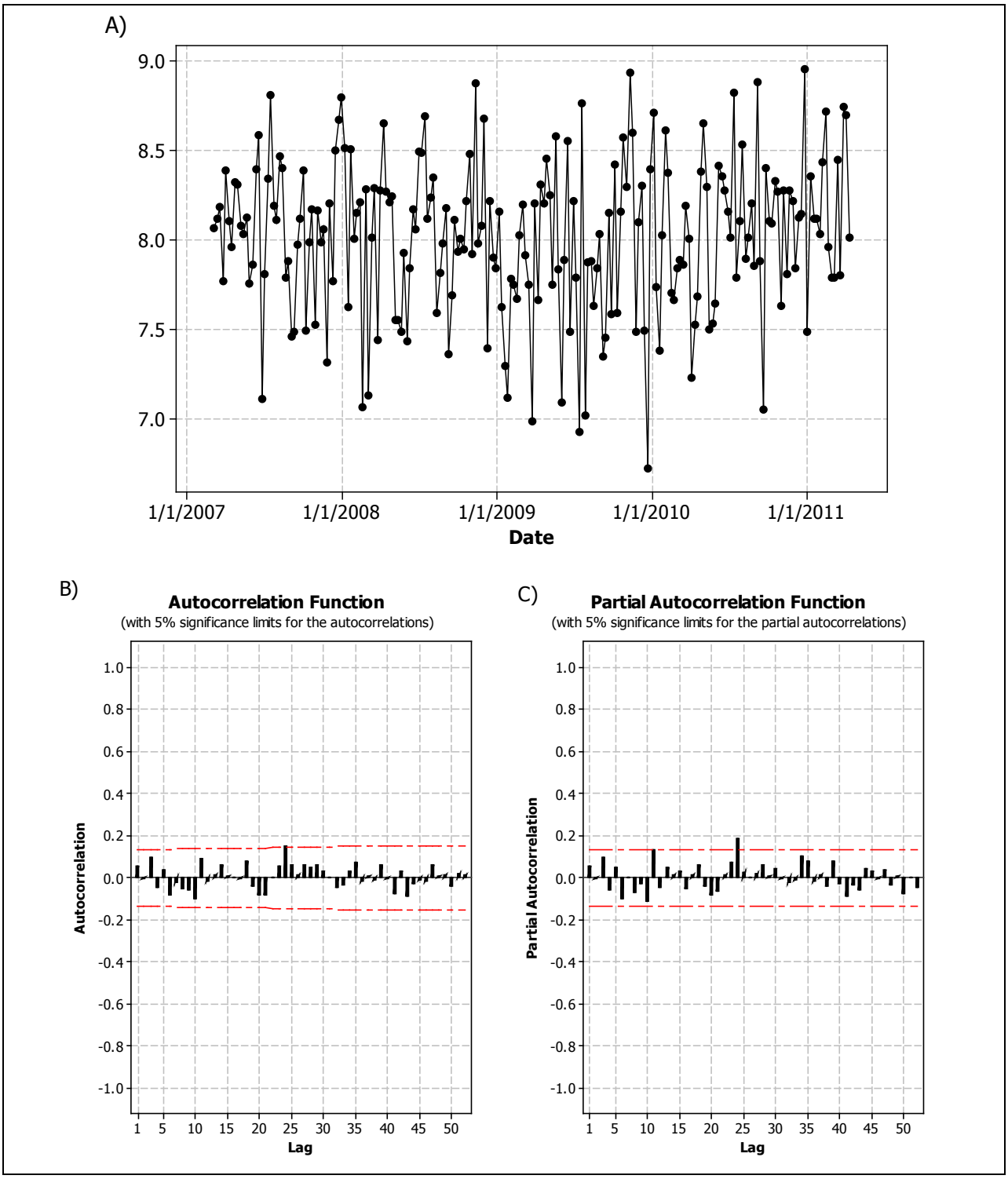


Figure 7-11. A) Time series plot, B) autocorrelation function (ACF) graph, and C) partial autocorrelation function (PACF) graph of data with zero autocorrelation (i.e., independent data with respect to time)

An autoregressive error pattern of order 1, AR(1) means that an observation is correlated with the previous observation. And, because each previous observation is related to the observation prior to it, each observation is related to all past values, but the highest correlation is with the most recent observation. A theoretical[4] AR(1) time series structure is identified by an ACF pattern that trails off bounded by an exponential decay after the first lag and the PACF dropping to 0 after lag 1 (or lag j for higher order AR series).

The patterns for both ACF and PACF in Figure 7-9A and Figure 7-9B are typical of a water quality data set with AR(1) and a strong seasonal pattern (some might argue for an AR(2) in this case which speaks to the fact that interpretation of the patterns is required, with analysts often relying on the preponderance of evidence across monitoring sites). The lag autocorrelations for weekly flow data from the Corsica River (MD) NNPSMP project in these figures do show some significant autocorrelation coefficients. The ρ_j falling outside of the red/blue lines are significant at the 95 percent confidence level. Significant autocorrelation for lag 1, as well as a strong seasonal autocorrelation pattern is evident.

Readers should consult statistics textbooks and software packages for greater detail on this and other methods to test for autocorrelation.

7.3.6.1 Methods to Handle Autocorrelation

Autocorrelation in analysis of time series data can sometimes be reduced by aggregating data over different time periods, such as weekly means rather than daily values. Use of weekly means preserves much of the original information of a daily data series, but separates data points far enough in time so that autocorrelation is reduced. When aggregating data, it is important to use a consistent procedure, e.g., using the weekly mean of 7 daily values for each week in the year, rather than mixing weekly means for some weeks with single grab samples for other weeks. Aggregation has disadvantages including: reducing the degrees of freedom and potential power of a statistical test and dampening out the potentially important high or low data.

Several statistical packages can incorporate a time series error term in the statistical model to address autocorrelation. For example, PROC AUTOREG in SAS (SAS Institute 2016d) can be used for linear regression when the error terms are autoregressive. Similar tools are available in Minitab's time series tools (i.e., Stat > Time Series) or R's statistics package.

Alternatively, if the data exhibit AR(1), which is typical for water quality data collected weekly, biweekly, or monthly, an adjustment can be made to the standard error of the trend (step or slope) terms. The correction factor was derived by Matalas and Langbein (1962) and simplified with a large sample size approximation by Fuller (1976):[5]

$$std.dev._{corrected} = std.dev._{uncorrected} \sqrt{\frac{1+\rho}{1-\rho}}$$

[4] Patterns from water quality sampling data will resemble theoretical patterns but will usually deviate in some way, requiring that the analyst develop a feel for interpreting such graphics.

[5] The exact formula is given by $std.dev._{corrected} = std.dev._{uncorrected} \sqrt{\frac{1+\rho}{1-\rho} - \frac{2}{n}\frac{\rho(1-\rho^n)}{(1-\rho)^2}}$ where n is the sample size.

Where ρ = autocorrelation coefficient at lag 1
Std. dev = the standard deviation of the trend term (e.g., standard error of the difference between mean values between two time periods or standard error of the slope of a linear regression).

See Spooner et al. (2011a) for additional details on this approach.

7.3.6.2 Methods to Handle Autocorrelation Caused by Seasonality

When the data exhibit seasonal cycles, incorporation of explanatory variables can be added to parametric methods to allow for adjustment of seasons. Four common approaches are used. One is to add 1 or 2 cycles by using sine and cosine terms to a linear regression model, for example, as described in *Tech Notes 6: Statistical Analysis for Monotonic Trends* (Meals et al. 2011). This approach assumes that the sine or cosine terms realistically simulate annual or semiannual seasonal cycles.

A second approach is to incorporate seasonality into the time series model. An ARIMA time series model could be used that incorporates a time series model with seasonal lag value ("differencing value" or "d"[6] in an ARIMA model, ARIMA(p,d,q)) corresponding to the length of the seasonal cycle. For example, an annual cycle will appear as a strong positive autocorrelation at lag 12 when the data series consists of monthly values or at lag 4 for quarterly values. As noted above, readers should consult statistics textbooks and software packages for greater detail on ARIMA models.

A third approach is to simply add monthly (or other seasonal) indicators to each observation in the dataset and incorporate these indicator variables in a regression model. The number of indicator variables needed is *S*-1[7]. For example *S*-1 would be 11 when the cycle is annual, but where the same months behave similarly over the years. Each indicator variable (X_1 through X_{11}) is assigned a value of 0 or 1, as indicated below:

X_1 = "1" for "January" but "0" otherwise
X_2 = "1" for "February" but "0" otherwise
…
X_{11} = "1" for November" but "0" otherwise
Note: December values would all be depicted by "0" values for X_1-X_{11}

After the indicator variables are added to the dataset, regress Y_t on the indicator variables and other independent variables (e.g., time).

A fourth approach to address seasonality is to use non-parametric tests that can handle monthly seasonality. The Seasonal Wilcoxon Rank Sum Test or Seasonal Mann-Whitney Rank Sum Test compares two or more groupings (e.g., seasonal t-test or analysis of variance). The Seasonal Kendall Test incorporates seasonal components when testing monotonic trends. Both parametric and non-parametric trend tests are featured in section 7.8.2.4. There is also a variant of the Kendall tau test (seasonal Kendall tau test with serial correlation correction (Hirsch and Slack 1984)) that can handle seasonality while also adjusting for autocorrelation.

[6] *Differencing* is a term used in time series analyses, where *d* is the order of differencing which creates a new time series, W_t, whose values at time t is the difference between x(t) and x(t+d). W_t then becomes the series used in the time series analysis.
[7] Where *S* would represent the number of time periods (e.g., months, seasons).

7.3.7 Examination of Two or More Locations or Time Periods

Comparison of two or more variables with EDA can mean comparing different data sets, such as stream nitrogen concentrations above and below a feedlot or phosphorus concentrations from a control and a treatment watershed, or comparing data from the same site over two different time periods, such as phosphorus loads from calibration vs. treatment periods.

The characteristics that make boxplots useful for summarizing and inspecting a single data set make them even more useful for comparing multiple data groups representing multiple sites or time periods. The essential characteristics of numerous groups of data can be shown in a compact form. Boxplots of multiple data groups can help answer several important questions, such as:

- Is a factor (location, period) significant?
- Does the median appear to differ between groups?
- Does apparent variability differ between groups?
- Are there outliers? Where?

Boxplots are helpful in determining whether central values, spread, symmetry and outliers differ among groups. If the main boxes of two groups, for example, do not substantially overlap on the vertical scale, there may be a reason to suspect that the two groups differ significantly (note that such difference should be tested using quantitative statistical techniques). Interpretation of boxplots can help formulate hypotheses about differences between groups. Figure 7-12 shows a boxplot of total suspended solids concentrations in three Vermont streams. The plot suggests that TSS concentrations may tend to be slightly lower at Station 3 compared to the other two stations; however, because the boxes overlap, it is unlikely that any comparison of medians would result in statistically significant differences.

Inferences about differences between locations or time periods resulting from graphical evaluation of the data must be confirmed by more rigorous hypothesis testing analyses (see sections 7.7 and 7.8).

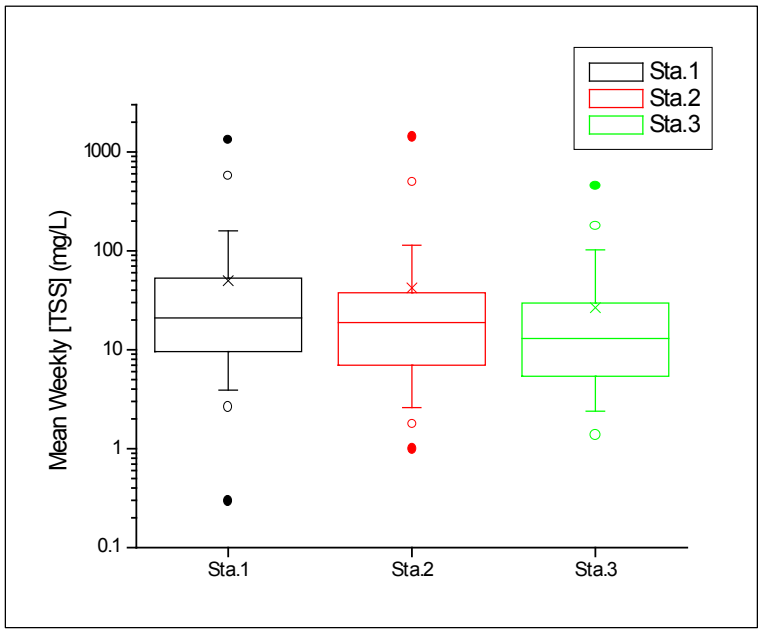


Figure 7-12. Boxplots of TSS concentration for three stream stations, 1998 (Meals 2001)

7.3.8 Examine Relationships between Variables

Looking at how variables relate to each other is a way to begin to consider causality, i.e., is the behavior of one variable the result of action by another. Such ideas can suggest sets of variables to evaluate together. For example, if variable B (e.g., suspended sediment concentration) goes down as variable A (e.g., acres of reduced tillage) goes up, has the BMP program improved water quality? Examination of correlations between different variables observed simultaneously (e.g., SSC and total P or turbidity and SSC) can suggest relationships that might change with BMP programs or indicate where one variable could serve as a surrogate for another. Graphical analysis (e.g, scatterplots of variable A vs. variable B) can suggest meaningful correlations that would need to be confirmed with more rigorous statistical tests.

The two-dimensional scatterplot is one of the most familiar graphical methods for data exploration. It consists of a scatter of points representing the value of one variable plotted against the value of another variable from the same point in time. Scatterplots illustrate the relationship between two variables. They can help reveal if there appears to be any association at all between two variables, whether the relationship is linear, whether different groups of data lie in separate regions of the scatterplot, and whether variability is constant over the full range of data.

Figure 7-13 shows a scatterplot of phosphorus export in a control and a treatment watershed in Vermont. Note that the data are plotted on a log scale to obtain a linear relationship. There is a strong positive association between P in the two streams. This simple scatterplot indicates that it is probably worth proceeding with more rigorous statistical analysis to evaluate calibration between the two watersheds in a paired-watershed design. As with this example, it is common that the relationship between variables is exponential. In such cases, the log transformation allows the relationship to be expressed linearly and evaluated using linear regression.

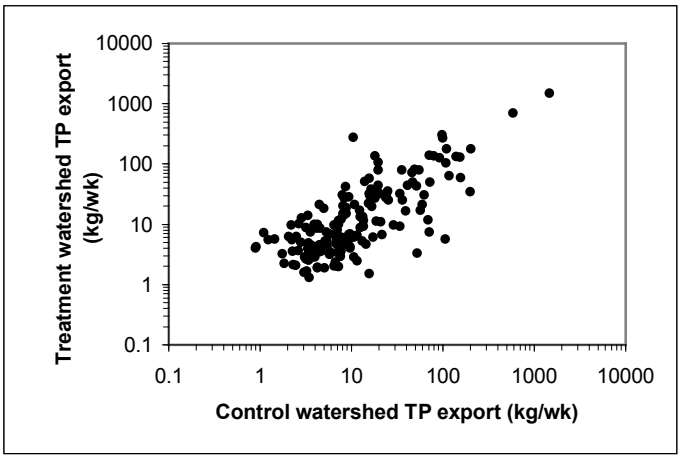


Figure 7-13. Scatterplot of weekly TP export from control and treatment watersheds, calibration period (Meals 2001)

Figure 7-14 shows another scatterplot examining the relationship between streamflow and *E. coli* counts in another Vermont stream. In a nonpoint source situation, a positive association between streamflow and bacteria counts may be expected, as runoff during high flow events might wash bacteria from the land to the stream. In this case, however, it does not require application of advanced statistics to conclude from Figure 7-14 that there is no such association (in fact the correlation coefficient r is close to zero). However, recall that EDA involves an open-minded exploration of many possibilities. In Figure 7-15, the

data points have been distinguished by season. The open circles represent data collected in the summer period and there still appears to be no association between streamflow and *E. coli* counts. The solid circles, representing winter data, now appear to show some positive correlation (r = 0.45) between streamflow and bacteria counts, with high bacteria counts associated with high flows. This picture suggests that something different is happening in winter compared to summer with respect to streamflow and *E. coli* in this watershed, a subject for further investigation.

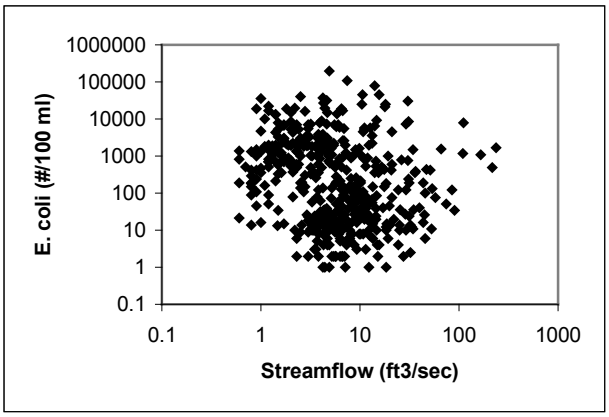


Figure 7-14. Scatterplot of *E. coli* vs. streamflow, Godin Brook, 1995-1998, all data combined (Meals 2001)

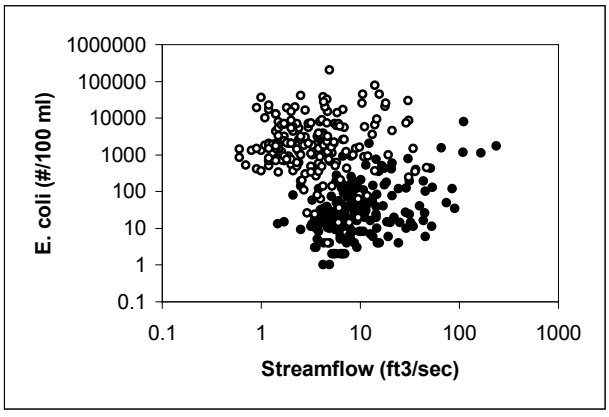


Figure 7-15. Scatterplot of *E. coli* vs. streamflow, Godin Brook, 1995-1998, where solid circles = winter, open circles = summer (Meals 2001)

In looking for correlations in scatterplots, choose the variables carefully. A common mistake is the comparison of variables that are already related by measurement or calculation. An example of such spurious correlation is the comparison of streamflow with load. Because load is calculated as concentration multiplied by flow, a scatterplot of flow vs. load has a built-in correlation that means very little, even though it looks good in a scatterplot. Also remember that correlation does not guarantee causation – just because two variables are correlated does not mean that the variation in one is caused by variation in the other.

There are many numerical techniques available to examine and test the relationship between two or more variables. In EDA, the simplest technique is correlation, which measures the strength of an association between two variables. The most common measure of correlation is Pearson's *r*, also called the *linear correlation coefficient*. If the data lie exactly on a straight line with positive slope, *r* will equal 1; if the data are perfectly random, *r* will equal 0. For Pearson's *r*, both variables should be normally distributed and continuous (Statistics Solutions 2016). The test also assumes a straight-line relationship between the variables and constant variance (homoscedasticity). Pearson's *r* is sensitive to outliers.

Other measures of correlation that are less sensitive to outliers include the nonparametric Kendall's *tau* and Spearman's *rho* (Spearman's rank correlation coefficient). Spearman's *rho* makes no assumptions about the distribution of the data and is an appropriate test when the variables are at least ordinal and the variables are monotonically related (Statistics Solutions 2016). With ordinal variables, the ordering of values is known but the differences between them are not quantified (e.g., Excellent, Good, Fair, Poor).

Measures of correlation are easily calculated by most statistical software packages and are described in chapter 4 of the 1997 guidance (USEPA 1997b). It must be cautioned that whenever a numerical correlation is calculated, the data should also be plotted in a scatterplot and examined visually as described above. Many different patterns can result in the same correlation coefficient. Never compute a correlation coefficient and assume that the data follow a simple linear pattern.

There are several methods of simultaneously evaluating variables that are likely related to each other. Cluster analyses group variables and/or observations into similar categories usually based on an agglomerative hierarchical algorithm which is the most common clustering pattern used in water quality analyses. In this clustering procedure, each observation begins as an individual "cluster." The similarities or distances between these clusters are measured using one of several options, including Euclidian distance and correlation coefficients. The closest two clusters are then merged into a new cluster. Distances are calculated again using the updated set of clusters, and the process repeated until only one cluster remains. The result of this analysis is a sequence of groupings that can be represented in a cluster tree or dendrogram. The analyst can then perform a visual analysis to infer potential groupings and relationships among variables. It is important to note that cluster analysis does not consider multicollinearity between the variables. Cluster analysis conducted as part of EDA might be used to explore and define site or time groupings that would be useful to explore in later analysis.

Other multivariate techniques that can be applied in subsequent analysis include principal components analysis, canonical correlation, and discriminant analysis (SAS Institute 1985). These methods are discussed further in section 7.5.2.5.

7.3.9 Next Steps

Data exploration results (knowledge of how data are distributed, their characteristics, and their relationships) will help illustrate any needs to adjust the data to enable the appropriate subsequent statistical tests. In addition, hypotheses can be refined to facilitate more advanced statistical techniques. section 7.4 describes methods for accounting for censored data. Sections 7.5 through 7.9 present various advanced procedures for analyzing data for a range of purposes. Section 7.10 presents a list of tools and other resources for data analysis.

7.4 Dealing with Censored Data

7.4.1 Types of Censoring

Monitoring programs such as those analyzing for pesticides, metals, or other constituents often present at very low concentrations may report lab results where concentration is below the detection limit of the analysis. Bacteriological tests may report very high results as "too numerous to count" (TNTC). Such data – typically reported as "<" or ">" (left- and right-censored, respectively) some value – are referred to as "censored" data.

Censored values are usually associated with limitations of measurement or sample analysis, and are commonly reported as results below or above measurement capacity of the available analytical equipment. Results that are indistinguishable from a blank sample are normally reported as less than the detection limit (DL). The true values of these *left-censored* observations are considered to lie between zero and the DL. Depending on the laboratory, some results greater than the DL may be identified as less

than the quantitation limit (QL) or reported as a single value and given a data qualifier to indicate the value is less than the QL. Typically, results reported as less than the QL indicate that the analyte was detected (i.e., greater than the detection limit), but at a low enough concentration where the precision was deemed too low to reliably report a single value. These *interval-censored* observations are considered to lie between the DL and QL.

Left- and interval-censored observations are less commonly encountered when working with sediment and nutrients because they are usually present at levels above their QLs. However; left censoring is common when toxics and pesticides are being analyzed.

An example of *right-censoring* includes microbiological analyses with misestimated dilution resulting in TNTC (too numerous to count) and exceedance of flow gage limits during floods. Right-censoring may also be encountered when lakes and estuaries are monitored for light penetration via Secchi depth and the result is reported as *visible on bottom*, i.e., the Secchi disk is observable on the bottom.

Helsel (2012) provides a seminal discussion of varying reporting limits and concerns with some data censoring practices. This guidance recommends that detection limits and quantitation limits be stored with the measurements and each result be clearly qualified to indicate its relation to the DL or QL as appropriate.

7.4.2 Methods for Handling Censored Data

There is no single ideal method for managing censored data in statistical analyses. When comparing various methods, this guidance recommends that analysts use methods that minimize bias and error. Extensive research in water resources as well as other fields of science such as survival analysis (e.g., how long does a cancer patient live after treatment) has considered numerous techniques. One deficiency over the last 20 years has been the lack of readily available tools for widespread use, making many of these tools out of reach for general use. Efforts continue to improve upon the availability of these tools. The most notable is a compilation of methods and recommendations developed by Helsel (2012) with additional information provided at *Practical Stats*. Much of the remaining discussion in this section is derived from Helsel's book (Helsel 2012) and the reader is encouraged to review his book for a more in-depth discussion.

7.4.2.1 Past Methods

With improved tool access, past methods for accommodating censored observations can be avoided. The most notable past method is simple substitution. This involves the replacement of censored observations with zero, ½DL, or DL. Although simple substitution is commonly used (and even recommended) in some state and federal government reports as well as some refereed journal articles, there is no real theoretical justification for this procedure. Substitution may perform poorly compared to other more statistically robust procedures, especially where censored data represent a high proportion of the entire dataset. More egregiously, some reports have simply deleted observations less than the detection limit. Some past researchers have recommended simply reporting the actual measured concentrations even if the concentrations are below the DL (Gilliom et al. 1984). This approach has not gained traction as laboratories are reluctant to implement such a practice, although Porter et al. (1988) suggested that an estimate of the observation error could be reported to better qualify the measurement. While simple substitution might be convenient for initial exploratory analyses using spreadsheet tools, more robust procedures are available and are recommended.

7.4.2.2 Using Probability Distribution Theory to Estimate the Summary Statistics

In environmental sciences, two common methods considered for estimating summary statistics from censored data sets include maximum likelihood estimation (MLE) and robust regression on order statistics (ROS). Both methods ultimately rely on a distributional assumption and both methods allow for multiple detection limits and estimation of confidence intervals. The reader is referred to Helsel (2012) for a more detailed discussion.

MLE uses the uncensored observations, the proportion of censored observations, and a distributional assumption to compute estimates of summary statistics. A lognormal distribution is commonly assumed with water quality data; however, commercial software will usually allow a variety of assumptions to be considered.

The robust ROS procedure (Helsel and Cohn 1988) relies on fitting a regression line to a normal probability plot of the uncensored observations and is applicable for multiple censoring levels. If the uncensored data do not fit a normal distribution, the analyst can transform the uncensored data with lognormal or other appropriate transformation. The process of selecting the best transformation is similar to that if all data were uncensored and diagnostics are typically available in current statistical software. The regression is then used to impute values for the censored data. The imputed and uncensored data are then, if necessary, transformed back to their original data scale, allowing summary statistics to be estimated using standard techniques. Confidence intervals for the mean and standard error estimates can be computed using bootstrapping (e.g., Helsel 2012). In summary for the mean, a random sample (with replacement) is selected from the site data. These data are passed through the robust ROS procedure described above, and a resulting mean is computed. The process of selecting a random sample, implementing the robust ROS procedure and computing a resulting mean is repeated, say, 1,000 times. Confidence limits are then empirically selected from this set of 1,000 means (e.g., the 5^{th} and 95^{th} percentile of these 1,000 means would be the 90 percent confidence interval on the mean).

The MLE tool can be applied to less-thans and TNTC in the same data set. Helsel (2012) provides recommendations for which method to use based on the number of observations and degree of censoring. Notably, no method works well when the degree of censoring exceeds 80 percent. In the situations where the censoring level exceeds 80 percent, Helsel (2012) recommends reporting information on the percent of observations above a meaningful threshold and no further summary statistics. For all summary statistics with censored data, this guidance recommends reporting the maximum detection limit, number of observations, and number of censored observations with all summary statistics.

7.4.2.3 Hypothesis Testing with Censored Data

There are a variety of nonparametric hypothesis tests that can be directly used with raw data sets that have censored observations and generally rely on the rank (or order) of the data. These tests include the Mann-Whitney test (two random samples), Wilcoxon (paired samples), and Kruskal-Wallis (several random samples), and Kendall and Seasonal Kendall tau (monotonic trends). In these tests, censored observations are treated as tied values, no different from cases where ties might occur between uncensored observations. Consider the ordered data set of <1, <1, 1.5, 4, 8, 9, 10, and 10. The two censored observations (of <1) are less than all the other observations, but are treated as tied to each other. The handling of the two "*<1*'s" is no different than the two *10*'s which are both greater than all the other values, but tied with each other. One deficiency of these tests is that they are limited to a single detection limit (e.g., the tests do not have a method to compare "<1" and "<2"). To apply the above nonparametric tests with data sets that have multiple detection limits, the analyst will need to re-censor the data to the

highest detection limit. Note: do not use the previously described ROS procedure to impute values for censored data and then apply one of the nonparametric tests described in this paragraph (or parametric tests), as erroneous results might be computed because the rank of the imputed values were calculated based upon the order of data set entry, which is not related to any true ranking of the actual water quality values.

An alternative approach is to apply MLE regression tools that are designed for multiply censored dependent variables. Similar to simple regression or multiple regression, relationships between singly- or multiply-censored dependent variables can be established with independent variables. Indicator variables can be used to set up groupings to expand the MLE regression tool for comparing two or more groups or seasonal/explanatory adjustments as well.

7.5 Data Analysis for Problem Assessment

7.5.1 Problem Assessment – Important Considerations

One of the most critical steps in controlling NPS pollution is to correctly identify and document the existence of a water quality problem. The water quality problem may be defined either as a threat to or impairment of the designated use of a water resource. Impairments are generally defined and identified as violations of water quality standards (WQS). Water quality standards define the goals for a waterbody by designating its uses, setting criteria to protect those uses, and establishing provisions such as antidegradation policies to protect waterbodies from pollutants. A WQS consists of four basic elements:

1. **A designated use of the water body.** States and Tribes specify appropriate water uses to be achieved and protected, taking into consideration the use and value of the waterbody for public water supply, for protection of fish, shellfish, and wildlife, and for recreational, agricultural, industrial, and navigational purposes. In designating uses for a water body, States and Tribes consider the suitability of a water body for the uses based on the physical, chemical, and biological characteristics of the water body, its geographical setting and scenic qualities, and economic considerations.
2. **Water quality criteria**. Water quality criteria are science-based numeric pollutant concentrations or narrative requirements that, if met, will protect the designated use(s) of the water body. Criteria may be based on physical, chemical, or biological characteristics. Numeric criteria may, for example, establish limits for concentrations of toxic pollutants to protect human health or aquatic life. Narrative criteria stating that a water body must be "free from" toxic contaminants can serve as a basis for limiting the toxicity of waste discharges to aquatic life.
3. **An antidegradation policy.** Water quality standards include an antidegradation policy that maintains and protects existing uses and water quality conditions necessary to support such uses, maintains and protects high quality waters where existing conditions are better than necessary to protect designated uses, and maintains and protects water quality in outstanding national resource waters. Except for certain temporary changes, water quality cannot be lowered in such waters.
4. **General policies.** States and Tribes may adopt policies and provisions regarding implementation of water quality standards, such as mixing zones, variances, and low-flow policies. Such policies are subject to EPA review and approval.

Water quality monitoring to support problem assessment is usually focused on documenting violations of WQS in time (e.g., frequency of exceedance) and space (e.g., geographic extent of exceedance). Water quality data for such purposes may be collected by an ongoing monitoring program (e.g., a state ambient monitoring program) or by a reconnaissance study designed to provide a preliminary, low-cost overview of water quality conditions in the area of interest (see section 2.4.2.1). The EPA ATTAINS database is the repository for information from state integrated reporting (IR) on water quality conditions under sections 305(b), 303(d), and 314 of the Clean Water Act, and the Reach Address Database contains state IR geospatial data. ATTAINS includes state-reported information on support of designated uses in assessed waters, identified causes and sources of impairment, identified impaired waters, and TMDL status.

A detailed discussion of monitoring designs has been presented in chapter 2 of the 1997 guidance (USEPA 1997b). Some designs appropriate for problem assessment have been discussed in section 2.4 of this guidance. In general, monitoring designs appropriate for collecting data to support NPS problem assessment include:

- **Synoptic surveys** designed to determine the magnitude and geographic extent of WQS violations, often used to identify pollutant source areas within a watershed;
- **Above/below monitoring**, wherein a potential pollutant source area is bracketed between upstream and downstream sampling points to assess the impact of the source area on pollutant levels; and
- **Trend monitoring** designed to collect long-term time-series data at one or more watershed sampling points that are useful in determining the frequency and magnitude of exceedance of WQS.

Both above/below (if pre- and post BMP data is collected) and trend monitoring designs can also be applied to other monitoring objectives such as project effectiveness evaluation using permanent monitoring stations equipped with automatic sampling equipment and continuous flow measurement devices.

Grab samples with instantaneous flow measurements for a few sampling events may be sufficient for initial problem assessment and source identification, but monitoring data for problem assessment should include both baseflow and stormwater monitoring necessary to fully characterize the system. Storm sampling is useful for documenting the delivery of pollutants by runoff and overland flow, critical considerations for waters impacted by NPS. Combined with hydrologic data, basic climatic information can be used to evaluate the seasons or times of the year when pollutant levels are highest or lowest and when high flow events, drought, or other factors affect water quality. Note that concentration data alone without concurrent flow or stage data are often of limited utility.

Biological monitoring is used widely in water quality assessments and EPA provides information and links to resources addressing various aspects of the application of aquatic life criteria in water quality assessments. Chapter 4 of this guidance is devoted to biological monitoring. The discussion below, however, emphasizes the use and application of statistical analysis to chemical and physical monitoring data for which there is a greater body of literature. See chapter 7 of *Handbook for Developing Watershed Plans to Restore and Protect Our Waters* (USEPA 2008) for a broad discussion of approaches to assessing water quality problems and identifying causes and sources of those problems using a wide range of information sources.

7.5.2 Data Analysis Approaches

7.5.2.1 Summarize Existing Conditions

In a single stream or subwatershed, one monitoring location may be sufficient for problem assessment. More often, sampling at two or more locations is necessary to evaluate existing conditions of the watershed. Concurrently, sampling at two or more locations can aid in identification of subwatersheds that merit further evaluation for pollution reductions or water resource protections.

When data from different locations in a watershed or different sampling time periods are consistent and comparable (e.g., from a synoptic survey or from multiple watershed stations in the same monitoring regime), a first step is to summarize existing conditions using univariate statistics – mean, median, range, variance, interquartile range – for different sampling locations. If differences over time or flow conditions are evident, it may be useful to group the data into separate baseflow and wet-weather strata or by season. If enough samples have been collected (i.e., at least three), existing water quality can be compared across multiple sites. Visual comparisons between sites can be depicted graphically using boxplots. Figure 7-16 shows a set of boxplots for one year of weekly conductivity data from three small watershed trend stations in Vermont (Meals 2001). Conductivity at site WS1 appears to be substantially lower than that observed at the other two stations; conductivity at WS2 tended to be somewhat higher than that observed at WS3, with more frequent high extreme values. Mean or median values can be compared between two sites using the unpaired Student's t-Test or a nonparametric equivalent such as the Wilcoxon Rank Sum Test (also known as the Mann-Whitney Rank Sum Test). More than two sites can be compared using Analysis of Variance or the Kruskal-Wallis k Sample Test. Adjustments for seasons or hydrologic explanatory variables should be considered by employing appropriate statistical tests such as Analysis of Covariance or the Seasonal Wilcoxon Rank Sum Test (also known as the Mann Whitney Rank Sum Test). If the data between two sites are paired, differences can be tested using the paired Student's t-Test or the Wilcoxon Signed Rank Sum Test. Paired tests are generally more powerful and should be used when enabled by collecting samples at the same time period at two sites.

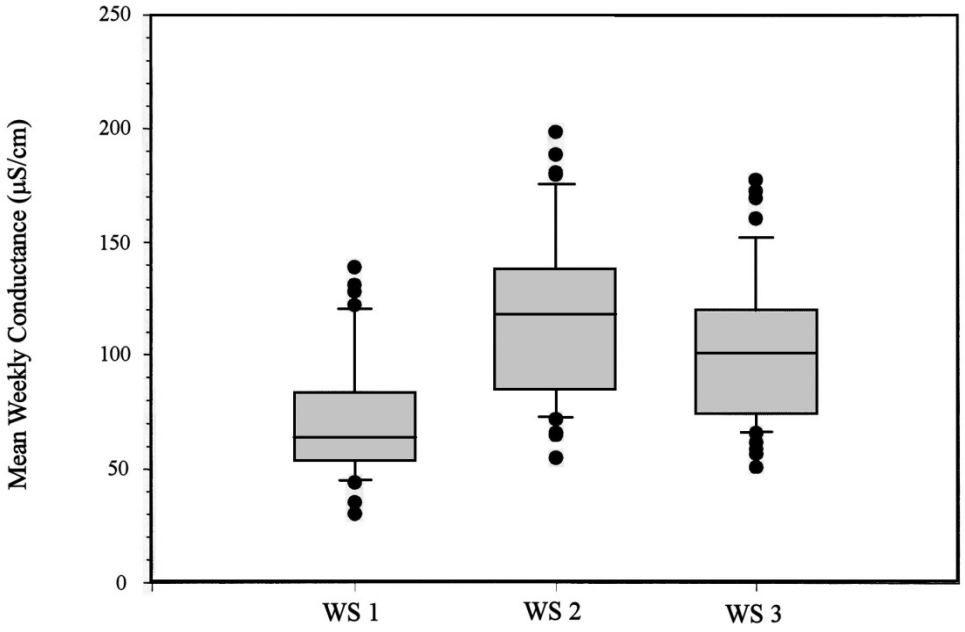


Figure 7-16. Boxplots of conductivity at three Vermont monitoring stations, October 1999 – September 2000 (Meals 2001)

Time series plots can visually reveal relationships over time and between locations. Figure 7-7 from section 7.3, for example, shows very clearly the seasonal cycle in *E. coli* counts in a Vermont stream, and Figure 7-4 reveals a different behavior at Station 2 compared to other stations regarding P concentrations. Time series statistical analyses can reveal autocorrelation and seasonality (see section 7.3.6).

Regression analysis between variables of primary interest (e.g., pollutant concentration/loads) and explanatory variables such as stream discharge can assist in documenting hydraulic relationships at a single monitoring location or between subwatersheds. Establishing relationships among variables can be very helpful in project planning as well. Scientists involved in the Upper Grande Ronde (OR) NNMP project, for example, explored relationships between fish and environmental factors via multivariate analysis and found that management and restoration activities that focus on reducing the maximum annual stream temperature would be the most effective in creating stream conditions that support salmonids (Drake 1999).

7.5.2.2 Assess Compliance with Water Quality Standards

Water quality data can be evaluated for violation of water quality standards (WQS). Note that specific requirements for documenting impairment in a regulatory sense may vary by circumstance. For some states and for some pollutants, a single observation exceeding a WQS may be sufficient to designate impairment. In other cases, determination of impairment must be based on violation of a WQS over a defined period of time or number of observations. A WQS for bacteria to support shellfishing may, for example, be based on a geometric mean of a number of different samples collected over a 30-day period, rather than on a single sample. Sanitary surveys in North Carolina, for example, include a shoreline survey to identify potential pollutant sources, a hydrographic and meteorological survey, and a

bacteriological survey (NCDENR 2016). Both the monitoring program and data analysis must be tailored to the regulatory requirements that apply to the watershed under study.

A data series should be plotted and the pattern evaluated for exceedance of WQS; plots of a time series at a single station or boxplot of multiple stations can be examined. Figure 7-17 shows how a time series plot can illustrate both the frequency and magnitude of violations of WQS. The dashed line represents the water quality criterion for chronic exposure; all of the observations exceed that level. The red line marks the acute criterion and shows that several observations exceeded that concentration. Moreover, most of the excursions above the acute criterion occurred around April, suggesting a seasonal aspect to the impairment. This kind of pattern may support inferences about pollutant source activity.

One way to evaluate the frequency or probability of violating WQS is to use probability plots or duration curves. Figure 7-18 shows a cumulative frequency plot of three years of *E. coli* data from a Vermont agricultural watershed (Meals 2001). In this case, it can be seen that compliance with the Vermont WQS of 77 cfu/100 ml *E. coli* occurred about 36 percent of the time and the stream was therefore considered impaired for *E. coli* about 64 percent of the time. If the USEPA criterion of 235 cfu/100 ml were applied, the stream would be in compliance with that criterion about 48 percent of the time and impaired about 52 percent of the time.

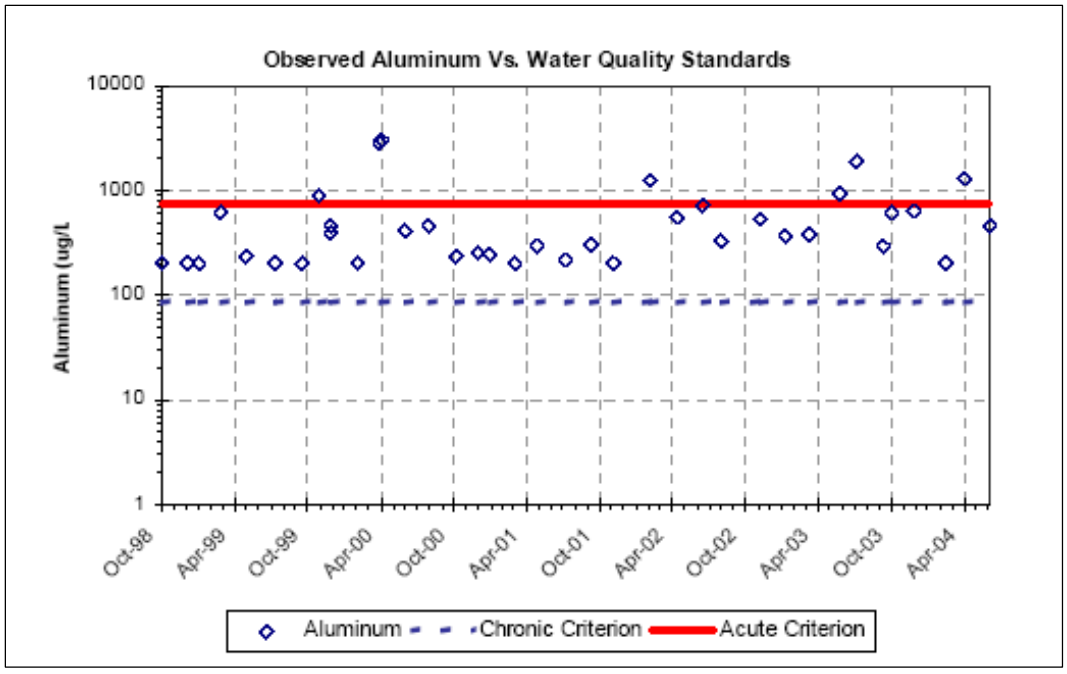


Figure 7-17. Example time series plot of observed aluminum concentrations compared to water quality criteria

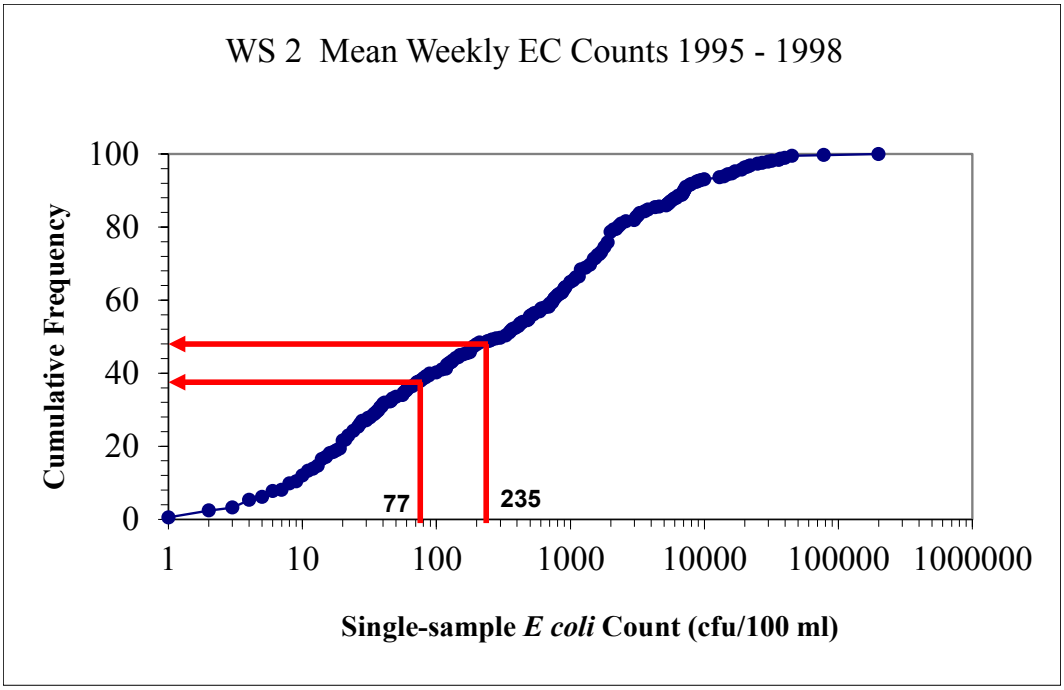


Figure 7-18. Cumulative frequency plot of three years of *E. coli* data from a Vermont stream (adapted from Meals 2001). Red lines represent frequency of observations at or below the VT WQS of 77 cfu/100 ml and the frequency of observations at or below the EPA criterion of 235 cfu/100 ml.

7.5.2.3 Identify Major Pollutant Sources

Cost-effective treatment of watersheds to address the pollutants and other causes of water quality problems requires knowledge of the sources contributing to the problems. Commonly used approaches to identifying and characterizing sources use both water quality and land-based information at varying levels of detail and quality (USEPA 2008). This section describes methods for analyzing water quality and associated monitoring data to characterize and aid in the prioritization of pollutant sources as part of the watershed planning process. See section 4.4.5 for an example of using biological monitoring in the Lake Allatoona/Upper Etowah River (GA) watershed.

Data from a synoptic survey or from regular monitoring of several subwatersheds combined with data on land use, management, or other land-based characteristics can inform understanding of major pollutant sources in a watershed. Correlation or regression analysis can be applied to explore relationships between pollutant concentrations and subwatershed characteristics, e.g., total P (TP) concentrations vs. manured cropland or suspended sediment concentration vs. cropland in cover crops. Annual mean or median values for pollutant concentrations could be compared to annual data on land use/management activities because concentrations will vary widely between individual events against land characteristics that are relatively constant within a single year or crop season. However, this simplification will not reveal seasonal and hydrologic variability in water quality or responses to short term land use changes such as animal numbers or fertilization. Where suitable knowledge of land use or land management is available, it may be more useful to provide water quality summary data for different periods that reflect distinctly different land use/management conditions (e.g., after spring manure applications vs. remainder of the year) during the monitoring period.

Boxplots or bivariate scatterplots can be compared between monitoring sites that reflect distinctive land use or management, thereby suggesting important pollutant source activities. If sufficient data from different subwatersheds or sampling stations exist, analysis of variance (ANOVA), or the nonparametric Kruskal-Wallis k Sample test can be used to test for significant differences in pollutant concentrations between sites and then compare these findings to differences in land use between the drainage areas sampled (graphical or tabular summaries). Analysis of covariance (ANCOVA) should be considered in cases where data are sufficient to test for differences among sites or seasons with adjustment for covariates such as precipitation or flow. See sections 4.6 and 4.8 of the 1997 guidance (USEPA 1997b) for a discussion of ANOVA and ANCOVA.

If flow data are available with concentration data, load estimates can be calculated to compare the magnitudes of pollutant sources (see section 7.9 for load estimation methods). The spatial and temporal resolution possible for load estimates will be determined by the number and location of sampling sites and the time frame and frequency of sampling events, respectively. Source-specific or subwatershed loads will generally be more helpful than loads at the watershed outlet, and in many cases seasonal loads or a classification of event vs. baseflow loads will be very helpful in the watershed project planning phase (see section 7.6).

It should be noted that correlation does not guarantee causation. Specifics of pollutant source activity and transport/delivery mechanisms must be considered to focus in on causation. Time of travel studies for various points in the watershed, for example, can be helpful in better characterizing the relationship between various sources or subwatersheds and downstream water quality. USGS describes methods for measuring time of travel (Kilpatrick and Wilson 1989).

7.5.2.4 Define Critical Areas

Data collected in the problem assessment phase can be used to help define critical source areas for pollutants, knowledge that is key to understanding the watershed, prioritizing land treatment, and evaluating project effectiveness. With concurrent data from monitored subwatersheds or tributaries (e.g., from a synoptic survey), statistical tests such as the Student's t Test or ANOVA can be used to identify significant differences in pollutant concentration or load among multiple sampling points. Such data can be displayed graphically in a map to show watershed regions that may be major contributors of pollutants. Figure 7-19, for example, shows a map of NO_2+NO_3-N concentrations from an April, 2003 synoptic survey in the Corsica River (MD) watershed (Primrose 2003). Nitrate/nitrite concentrations were found to be excessive in four subwatersheds, high in sixteen, and moderately elevated in seventeen others. Benchmarks for determining excessive/high/moderate or similar categories can be based on numeric water quality criteria or reference watershed values. If flow data were also available, it would be possible to estimate loads and compare subwatersheds on the basis of absolute (e.g., kg TP) or areal (e.g., kg TP/ha) loads. Figure 4-3 of section 4.4.5 illustrates how biological monitoring data from the Lake Allatoona/Upper Etowah River (GA) watershed were used for site-specific assessments of biological condition.

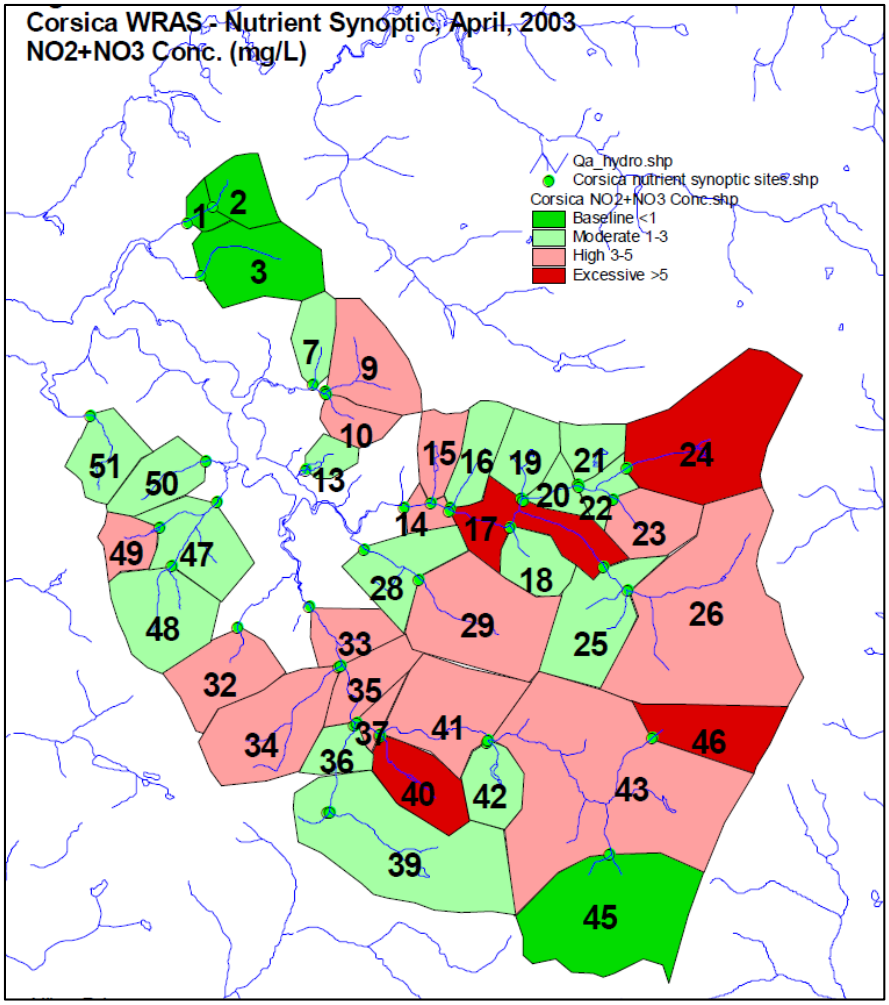


Figure 7-19. Map of synoptic sampling results from 41 stations in the Corsica River Watershed (Maryland) for NO_2+NO_3-N concentration (Primrose 2003). Pink and red shaded subwatersheds represent drainage areas contributing high (3-5 mg/L) and excessive (>5 mg/L) NO_2+NO_3-N concentrations, respectively.

Assessment of critical areas using a small set of water quality data has some limitations. Conditions determining pollutant generation (e.g., storm event, season, management schedules) must be considered in drawing conclusions about critical areas. Data collected during the active crop growth season may show a very different situation from data collected in winter, although for source identification purposes, it may be preferable to sample during the most critical times of year. The data mapped in Figure 7-19, for example, were collected in April, during or immediately following the spring planting and fertilizer application season when N losses from recently applied fertilizers might be expected to be high. Secondly, the spatial resolution of source area identification is limited by the resolution of the sampling network. Detailed site evaluation and/or modeling may be required to identify critical source areas on a finer scale.

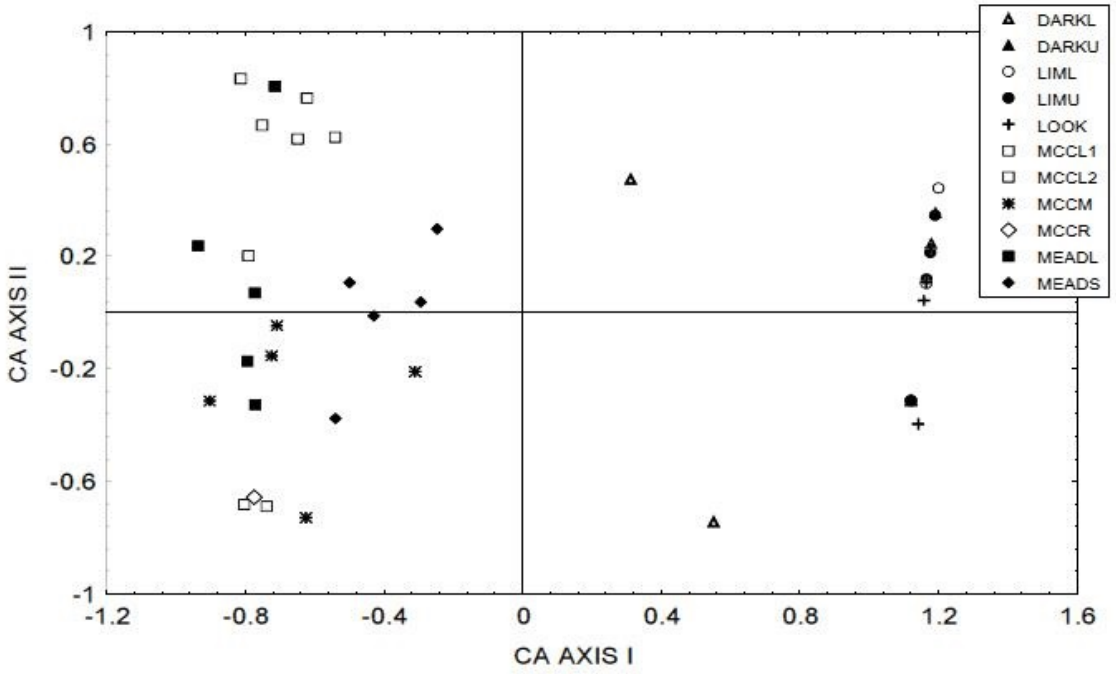
Another problem with using only a small set of water quality samples to determine critical areas is that some sources are by default removed from consideration. For example, the role of streambanks and stream channels in delivering sediment and sediment-bound pollutants such as P is often only partially understood at the beginning of watershed projects. The Sycamore Creek (MI) NNMP project, for example, focused on no-till and continuous cover to reduce sediment loads, but later concluded that the stream channel stabilization implemented in one subwatershed must have been at least as important as no-

till in reducing suspended solids loads (Suppnick 1999). Solutions to sedimentation problems in Lake Pittsfield (IL) progressed from an initial emphasis on no-till, terraces, and waterways (1979-1985), to numerous water and sediment control basins and a single large sedimentation basin (1992-1996), and then to stream restoration using stone weirs and streambank vegetation (1998) when it was learned that massive bank erosion was increasing sediment yield (Roseboom et al. 1999). See section 4.4.5 for a detailed example of using biological monitoring in the Lake Allatoona/Upper Etowah River (GA) watershed.

7.5.2.5 Additional Approaches

In most cases, projects in the planning phase have limited information with which to perform statistical analyses, particularly advanced procedures. Where such data exist, however, multivariate statistical procedures such as factor analysis, principal component analysis, canonical correlation analysis, and cluster and discriminant analysis can be used to define (and perhaps subsequently adjust for) complex relationships among variables such as precipitation, flow, season, land use, or agricultural activities that influence NPS problems. Spatial and temporal patterns can be revealed with these techniques. Scatterplots of ordination scores can be a useful method to summarize multivariate datasets and visualize spatial and temporal patterns.

Ordination techniques can also be powerful during the EDA phase when looking for patterns and structure in the data. The upper Grande Ronde basin project, for example, used correlation and canonical correspondence analysis to determine which environmental variables are largely responsible for differences in fish assemblages between reference and impaired sites (Drake 1999). Figure 7-20 shows a correspondence analysis plot showing intermediate/impaired sites and reference sites ordinating on the left and right side of the origin (Drake 1999). Scatterplots such as Figure 7-20 can be a useful way to summarize multivariate datasets and visualize these spatial and temporal patterns. With such variables identified, the next step was applying principal component analysis to determine if these variables could be used to track stream improvements over time. These statistical procedures are discussed briefly below. The reader is referred to statistics textbooks and other resources for additional information. Further, it is recommended that these procedures are performed by or in consultation with a trained statistician.



DARKL; Dark Canyon Creek – lower site
DARKU; Dark Canyon Creek– Upper site
LIML; Limber Jim Creek– Lower site
LIMU; Limber Jim Creek– Upper site
LOOK; Lookout Creek
MCCL1 & 2; McCoy Creek – Lower site 1 & 2
MCCM; McCoy Creek – Middle site
MCCR; McCoy Creek – Restored reach
MEADL; Meadow Creek – Lower site
MEADS; Meadow Creek at Starkey

Figure 7-20. Correspondence analysis biplot of Grande Ronde fish data (Drake, 1999)

Principal component analysis (PCA) is a multivariate technique for examining linear relationships among several quantitative variables, particularly when the variables are correlated to each other. PCA can be used to determine the relative importance of each independent variable and determine the relationship among several variables. Given a data set with p numeric variables, p principal components or factors can be computed. Each principal component (or factor) is a synthesized variable that is a linear combination of the original variables (SAS Institute 1985). The first principal component explains the most variance in the original data, while the second principal component is uncorrelated with (i.e., orthogonal to or statistically independent from) the first principal component and explains the next greatest proportion of the remaining variance. This process is continued until there are p statistically independent principal components that explain as much of the variance as possible. The results of PCA can often be enhanced through factor analysis, which is a procedure that can be used to identify a small number of factors that explain the relationships among the original variables. One important aspect of factor analysis is the ability to transform the factors (i.e., reconfigure the linear combinations of original variables) from PCA so that they make more sense scientifically. The SAS procedures PROC PRINCOMP and PROC FACTOR can be used for these analyses (SAS Institute 2010).

Principal component analyses and factor analysis can be used in regression analysis to reduce the number of variables or degree of freedoms (d.f.) by using a subset of the principal components (factors) that explain the majority of the variance of the data set instead of using all of the original variables. This essentially reduces the degrees of freedom used, but incorporates most of the information from each of

the explanatory variables, hence increasing the validity and power of the regression analysis. Using PCA to incorporate many explanatory variables into a regression model is superior to other techniques that arbitrarily drop explanatory (X) variables; those may incorrectly drop the more important variables due to multicollinearity between the X's. In principle, PCA and factor analysis could be beneficial to projects in a number of other ways, including helping investigators focus problem assessments on the most important indicators and stressors, aiding in the selection of water quality and land use/treatment variables to be used in the monitoring program, and guiding BMPs toward the most important pollutant sources.

Canonical correlation analysis (CCA[8]) is a technique for analyzing the relationship between two sets of multiple variables (e.g., a set of nutrient variables and a set of biomass-related variables). This multivariate approach examines said relationship "by finding a small number of linear combinations from each set of variables that have the highest possible between-set correlations" (SAS Institute 1985). These linear combinations of variables from each set are synthetic variables called 'canonical variables' and the coefficients of the linear combinations (which are similar to Pearson r) are referred to as the 'canonical weights' (SAS Institute 1985). The first canonical correlation is the correlation between the canonical variables from each set that maximizes the correlation value in accounting for as much as possible of the variance in the variable sets. The second canonical correlation is between a second set of canonical variables, is uncorrelated with the first canonical variables, and produces the second highest correlation coefficient. Additional correlations are established until all variance is explained or the maximum number of canonical correlations has been used (i.e., the number of variables in the smaller set). As such, the canonical variables are similar to principal components in summarizing total variation (SAS Institute 1985).

In simple terms, CCA can be used in problem assessment to look for relationships between sets of grouped variables to help better understand existing water quality problems or the relationships between land use/management variables (e.g., imperviousness, acreage receiving manure) and pollution variables (e.g., discharge, pollutant concentrations) to help guide decisions on BMP selection and placement. There are several output statistics (e.g., significance, correlations, coefficients) in CCA, and the reader is referred to statistical textbooks and other sources for additional details. It should be noted, however, that while many correlations may be output from a specific analysis, only the strongest correlations should be considered for interpretation.

Discriminant analysis is used to assess relationships between a categorical (grouping) variable (e.g., presence or absence of a fish species) and multiple quantitative (predictor) variables (e.g., pH, temperature, D.O.). The category options (e.g., present or absent) are assigned a priori—normally verification of the a priori grouping is performed during discriminant function analysis. Discriminant analysis can be used to verify the observational groupings defined by each cluster (see section 7.3.8) or other defined grouping based on the values of the quantitative variables. This type of analysis is referred to as `classificatory discriminant analysis' and is probably the most common application of discriminant analysis in water quality research. The SAS procedures DISCRIM (parametric) and NEIGHBOR (nonparametric) can be used to perform classificatory discriminant analyses (SAS Institute 1985).

Discriminant analyses can also be used to define a subset of quantitative variables that best describes the differences among the groups; see, for example, the SAS procedure STEPDISC (SAS Institute 1985). Canonical discriminant analysis is equivalent to canonical analysis described above except that a set of quantitative variables is related to a set of classification variables (SAS Institute 1985). Principal

[8] Canonical correspondence analysis is also often abbreviated as CCA.

component analysis is used as an intermediate step in the calculation of the canonical variables. The SAS procedure CANDISC can be used to perform canonical discriminant analyses (SAS Institute 1985).

Cluster and discriminant analyses can be used to understand and adjust for relationships among water variables. For example, spatial heterogeneity and homogeneity can be revealed. This may be necessary to study the transport of a pollutant in a system or to remove the spatial component in order to detect changes over time.

In many cases, watershed projects use simulation models to help with problem assessment and planning. Water quality models that include land use/land treatment and are calibrated using water quality data from the watershed or similar watershed(s) can also assist with identification of critical pollutant sources. The reader is referred to USEPA's watershed project planning guide (USEPA 2008) and TMDL modeling website for additional information on water quality models.

7.6 Data Analysis for Project Planning

Existing data or data collected specifically in support of a developing watershed project may play important roles in project planning, including determination of land treatment needs and design of a water quality monitoring program. These and other aspects of watershed planning are addressed in detail in *Handbook for Developing Watershed Plans to Restore and Protect Our Waters* (USEPA 2008).

7.6.1 Estimation and Hypothesis Testing

Project planning – including setting clear project goals – should result in the articulation of hypotheses that can be tested using appropriate statistical tests. The hypothesis must be stated in quantitative terms that can be adequately addressed by statistical analyses and must be directly related to the stated water quality monitoring goals.

The *null hypothesis* (H_o) is a specific hypothesis about a population that is being tested by analyzing the collected sample data. In water quality studies, the null hypothesis is generally a statement of no change, no trend over time or space, or no relationship(s). In contrast, the *alternative hypothesis* (H_a or H_1) is generally the opposite of the null, e.g., a statistically significant change, a trend over time or space, a relationship between 2 or more variables.

The general approach to hypothesis testing is to:

1. State the null and alternative hypotheses. For example:
   - H_o – There is no statistically significant trend over 10 years in TP at the subwatershed stream outlet
   - H_a – There is a statistically significant trend over 10 years in TP at the subwatershed stream outlet
2. Determine a parameter (e.g., mean, median, slope/trend over time) that would provide a point estimate to test if the sample data follow a distribution that would be expected if the null hypothesis was true, or more importantly, to test if there is evidence that the data come from an alternative population.
3. Design a sampling plan that would collect data to test if there is statistical evidence to reject the null hypothesis and accept the alternative hypothesis.

4. Analyze the sample data to calculate the sample point estimate and its confidence interval based upon the collected data variability.
5. Compare the confidence interval to the point estimate under the null hypothesis to determine if there is statistical evidence to reject the null and accept the alternative hypothesis (e.g., statistical evidence that a trend has occurred over time).

It should be noted that if the null hypothesis is not rejected, it is inappropriate to state that the null hypothesis is accepted. Instead, failure to reject the null or failure to detect significant differences or trends is the proper way to state such results. Failure to reject the null could be due to high sample variability, low sample size, or no real differences or trends. The chance of documenting a true difference or trend with statistical significance is improved by increasing sample frequency and longevity, and by using a monitoring design that will isolate the change/trend, while accounting for some of the high variability in data values observed in natural water quality systems. Effective monitoring designs are described in chapters 2-4.

There are two types of errors in hypothesis testing:

1. Type I: The null hypothesis (H_o) is rejected when H_0 is really true.
2. Type II: The null hypothesis (H_o) is not rejected when H_0 is really false.

The probability of making a Type I error is equal to the significance level (α). The probability of a Type II error is β. The power of a test (1- β) is the probability of correctly rejecting Ho when Ho is false. While the significance level is often taken for granted to be 0.05, a different value might be more appropriate for some NPS studies.

7.6.2 Determine Pollutant Reductions Needed

To set goals for a watershed project, it is important to estimate the pollutant reduction required to meet water quality objectives, usually to meet WQS. There are several approaches to developing such estimates:

- **Mass balance/TMDL**. In a TMDL setting, a load reduction goal is established based on a mass balance approach. Monitoring data are used to estimate the pollutant load a waterbody can receive while complying with WQS. The pollutant load reduction goal for a watershed project becomes the difference between the current load and the TMDL which is defined by:

$$TMDL = WLA + LA + MOS$$

Where WLA is the Waste Load Allocation (the allowable point source load);

LA is the Load Allocation (the allowable nonpoint source load); and

MOS is the Margin of Safety to account for uncertainty in the other estimates.

Note that the LA term (NPS load) is often estimated by difference and is not subdivided by source type. The pollutant load reduction goal for a watershed project focused on agricultural sources, for example, will not necessarily address the full difference between current load and LA because there may be other significant nonpoint sources in the watershed such as urban and residential nonpoint sources. TMDLs are frequently based on modeling analysis, but also use available water quality data to the extent possible.

Detailed information on TMDL analysis is available through USEPA (2013). See Case Study 5 for an illustration of how water quality data can be used in the development of a watershed-scale mass balance. The accuracy of this approach, however, depends on the quality and representativeness of the data used in the analysis. In Case Study 5, for example, because internal P loading is being computed based on estimates of the other terms, underestimation of external P loading will lead to an equal overestimate of internal P loading, thus confounding interpretation of the effects of alum application. For this and other reasons, the adaptive management approach is a cornerstone of TMDL implementation. As additional data are collected, mass balances should be revisited.

- **Receiving waterbody relationships.** Numerous tools exist to evaluate the impacts of pollutant loads on waterbodies that may be helpful in estimating pollutant load reduction goals. In lakes, for example, there are many analytical procedures and modeling tools to relate phosphorus load to lake eutrophication, including the "Vollenweider models" (Vollenweider 1976, Vollenweider and Kerekes 1982) and BATHTUB (Walker 1999). Such tools may be used to "back-calculate" permissible phosphorus loads to lakes. Other receiving water models may be used for similar purposes in other types of waterbodies, e.g., QUAL2K, CONCEPTS, and WASP. All of these models can employ available monitoring data to both establish model parameter values and to conduct calibration and validation. Additional information on models useful in this kind of analysis can be found in the USEPA TMDL Modeling Toolbox. Many of these models need to be calibrated with water quality collected from the study watershed or similar watershed(s).
- **Load duration curves.** A flow or load duration curve is a cumulative frequency plot of mean daily flows or daily loads at a monitoring station (e.g., a watershed trend station or tributary outlet) over a period of record, with values plotted from their highest value to lowest without regard to chronological order (see section 7.9.3). For each flow or load value, the curve displays the corresponding percent of time (0 to 100) that the value was met or exceeded over the specified period – the flow or load duration interval. Extremely high values are rarely exceeded and have low flow duration interval values; very low values are often exceeded and have high flow duration interval values. An estimate of the pollutant reductions needed is obtained by comparing a load duration curve developed from monitored loading data against a similar curve with loads estimated as the product of monitored flows and the pollutant concentration established in a WQS. Detailed information on the application of load duration curves to pollutant load reduction estimates can be found in *An Approach for Using Load Duration Curves in the Development of TMDLs* (USEPA 2007).

CASE STUDY 5: MASS-BALANCE APPROACH USED FOR ESTIMATING PHOSPHORUS LOADS

Western Ohio

- ✓ Treated a large, shallow lake with aluminum sulfate to reduce internal phosphorus loads
- ✓ Used the mass-balance approach to estimate internal phosphorus loads pre- and post-treatment

Grand Lake St. Marys (GLSM) is located in the Grand Lake St. Marys watershed in western Ohio (Figure CS5-1). GLSM is a large (5,000 ha), man-made, shallow (mean depth: 1.6 m) lake originally constructed as a "feeder reservoir" for the Miami-Erie Canal (Hoorman et al. 2008; ODNR 2013; Tetra Tech, Inc. 2013). Over 90 percent of the watershed is in cropland with associated livestock operations. Cyanobacteria blooms in GLSM result both from external and internal phosphorus loading (Tetra Tech, Inc. 2013).

The lake was treated with aluminum sulfate (alum) in June 2011 (23.6 mg Al/L, 49.6 g/m2) and in April 2012 (21.5 mg Al/L, 45.2 g/m2) to reduce internal phosphorus loads. The combined treatments totaled approximately 70 percent of the recommended treatment for the lake (recommended treatment was 86 mg Al/L, 120 g/m2). Monitoring data from 2012 were compared against monitoring data collected between 2010 and 2011 to analyze the results of the treatments (Tetra Tech, Inc. 2013). While the assessment also included analysis of algal biomass and aluminum in the water column and sediments, this summary focuses on total phosphorus (TP).

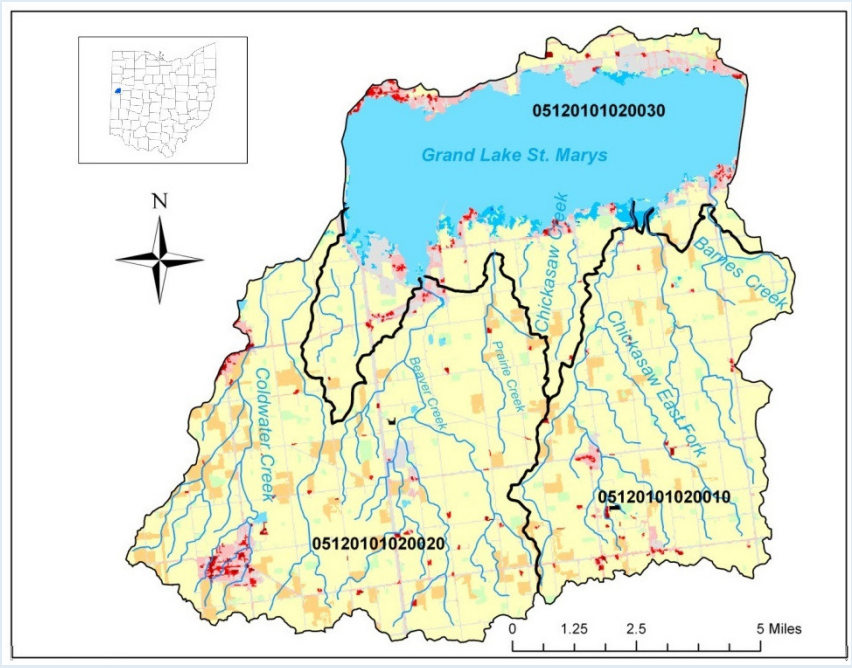


Figure CS5-1. Grand Lake St. Marys watershed

Monitoring and Sampling

Data from eleven water column monitoring sites were used in the assessment, with the five lake sites (shown in Figure CS5-2) sampled every two weeks after alum treatment. Samples at these five sites were always collected at 0.5 m from the surface, while some sampling events also included samples at the bottom of the water column. Samples were analyzed for TP, soluble reactive phosphorus, alkalinity, and chlorophyll. The Ohio Environmental Protection Agency (OEPA) also conducted routine sampling of tributaries, with sample analysis including TP (Tetra Tech, Inc. 2013).

Figure CS5-2. Tributary and lake sampling stations

Mass-Balance Approach

The mass-balance approach helped estimate internal TP loading before and after alum treatment. This approach consisted of five basic steps: (1) Estimating the water budget for GLSM; (2) Developing a basic P budget for the same time period as the water budget (May 2010 through May 2011 prior to any alum addition); (3) Predicting GLSM mean TP concentrations using a P mass balance model for which input values are based on available monitoring data for inflows and outflows; (4) Comparing estimated GLSM mean TP concentrations with measured TP concentrations; and (5) Adjusting the rates of P sedimentation and release of P into the water column (internal loading) to match predicted with measured TP concentrations in GLSM (Tetra Tech, Inc. 2013).

Water budget

A water budget for GLSM was determined at a two-week time step. Change in lake storage was determined using the following equation:

> Change in GLSM lake storage = Inflow (creek and WWTP inputs) + Precipitation – Outflow (water treatment plant withdrawal, groundwater loss, outlets) – Evaporation + Groundwater

The only tributary for which flow data were collected continuously was Chickasaw Creek where USGS has a gaging station (see Figure CS5-2). Wastewater treatment plant (WWTP) flow volumes were obtained from WWTP records and removed from the creek flow volumes so that loads from the four WWTPs in the watershed to GLSM could be calculated separately. Flow volumes from ungaged tributaries and areas draining directly to the lake were estimated by multiplying the adjusted Chickasaw Creek flow (minus WWTP) by the ratio between the other contributing drainage and Chickasaw Creek drainage areas. If creeks were observed to be dry, the flow was assumed to be zero for that period (Tetra Tech, Inc. 2013).

Precipitation records were obtained from a nearby weather station and multiplied by the surface area of the lake to get a volume of direct inflow from precipitation. Monthly mean pan evaporation rates were taken from the Hydrologic Atlas for Ohio (Harstine 1991; after Farnsworth and Thompson 1982).

Groundwater inflow was negligible and the rate for groundwater loss was assumed based on productivity of the underlying aquifer. This rate was adjusted such that there was more loss or recharge during the drier months when there was no outflow. Daily WWTP withdrawals were obtained from plant records. GLSM has two spillways, neither of which is continuously gaged. Lake level data were used to determine when losses would occur over the spillways and two instantaneous flow measurements were used to check estimated flows over the west spillway which is the major outflow. Outflow over the east spillway was assumed to be 10 percent of the west spillway outflow based on communication with local experts (Tetra Tech, Inc. 2013).

Total Phosphorus mass-balance model

A TP mass balance model was developed using the same two-week time step as used for the water budget (Perkins et al. 1997; Tetra Tech, Inc. 2013). Mass was estimated for two-week periods by multiplying the estimated flow volume and mean TP concentration. The principal use of the mass-balance model was to estimate changes in internal P loading for GLSM based on input of measured and estimated values for other terms in the model. Model calibration was based on matching predicted with measured lake TP concentration (Tetra Tech, Inc. 2013).

The following model was used to predict whole lake TP concentrations:

$$dTP/dt = W_{ext} + W_{int} – W_s – W_{out},$$

where W_{ext} is external loading, W_{int} is internal loading, W_s is loss to sediments, and W_{out} is loss through the lake outlet. Predicted whole-lake TP concentrations were compared to observed whole lake mean TP concentrations determined from monitoring at the five lake sites (Figure CS5-2).

Tributary TP concentrations were based on samples collected by OEPA during its routine monitoring. An average of all tributary TP concentrations was used for the ungaged portion of the basin. The TP concentration in direct precipitation was assumed to be 20 µg/L based on an average areal loading rate at Lake Erie from 1996 to 2002 (Dolan and McGunagle 2005). Concentration data for WWTPs were obtained from OEPA where available, and a concentration of 2 mg/L based on an OEPA analysis was assumed otherwise.

Assuming complete mixing, all but one outflow TP concentration was set equal to the whole lake average TP concentration predicted by the model. The actual measured TP concentration of the outflow, 210 µg/L, was used in the model for a single, very large storm event. Sedimentation rates (loss of TP to sediments) and sediment release rates (internal loading) of TP were adjusted in the model to reflect alum applications and to improve the relationship between predicted and measured lake TP concentrations (Tetra Tech, Inc. 2013).

Results

The phosphorus mass balance model was used to determine whole-lake mean TP concentrations based on external loading, internal loading, TP sedimentation, and TP loss through outflows. Whole-lake mean TP concentrations predicted by the 2012 model were compared to observed concentrations as collected and analyzed by OEPA. Sedimentation rates were adjusted to fit the

predicted to measured TP concentrations in the lake (Figure CS5-3). With the 2012 model thus calibrated, results were compared with those from 2010 and 2011 to determine if changes in internal TP loading had occurred as a result of alum treatments (Tetra Tech, Inc. 2013).

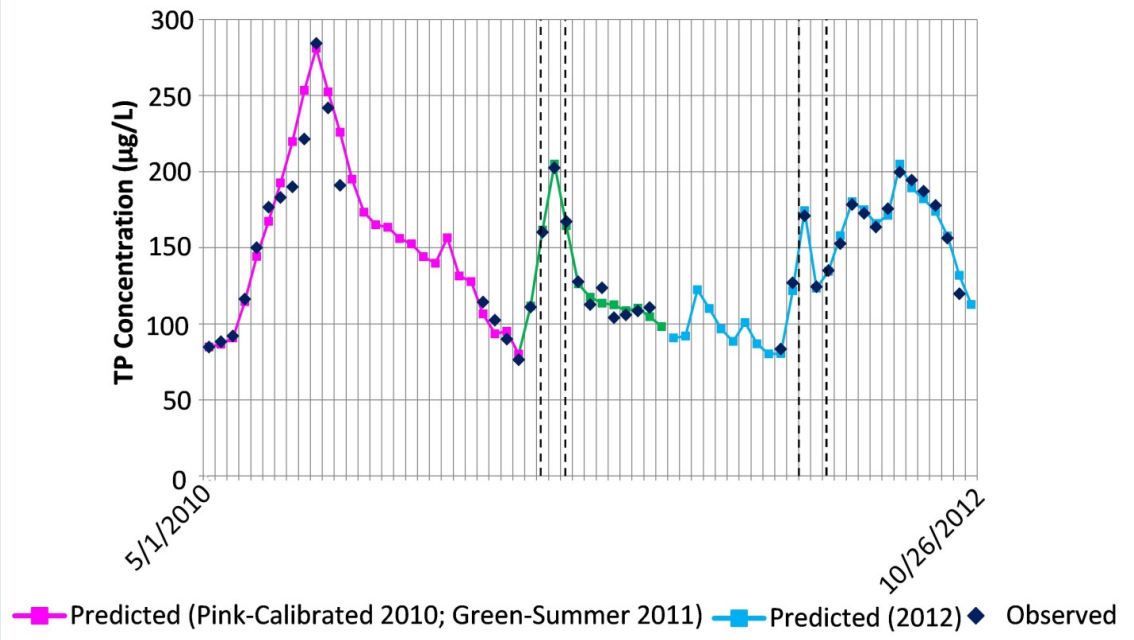


Figure CS5-3. GLSM predicted vs. observed TP concentrations from May 2010 through October 2012 (Adjustments made to internal loading estimates to match predicted November 2011 – October 2012 values to observed TP concentrations)

Table CS5-1 shows that gross summer internal TP loading to GLSM declined steadily from 2010 to 2012. The mass-balance modeling showed that average summer internal loading rate decreased from 4.0 mg/m2 per day before alum treatment to 1.8 mg/m2 per day after the two alum treatments, even though the combined 2011 and 2012 treatments totaled only 70 percent of the recommended treatment for the lake (Tetra Tech, Inc. 2013).

Table CS5-1. Comparison of internal TP loading in GLSM (2010–2012)

	2010	2011	2012
Total Gross Summer Internal TP Load (kg)	26,470	16,487	11,374
Average Summer Internal Loading Rate (SRR) (mg/m2-day)	4.0	2.4	1.8

(Tetra Tech, Inc. 2013)

References

Dolan, D.M. and K.P. McGunagle. 2005. Lake Erie total phosphorus loading analysis and update: 1996-2002. Journal of Great Lakes Research 31 (Suppl. 2):11-22. <http://www.cee.mtu.edu/~nurban/classes/ce5508/2007/Readings/dolan05.pdf>. Accessed November 2013.

Farnsworth, R.K. and E.S. Thompson. 1982. Mean Monthly, Seasonal, and Annual Pan Evaporation for the United States. NOAA Technical Report NWS 34. National Oceanic and Atmospheric Administration, National Weather Service, 82 pp. <http://www.nws.noaa.gov/oh/hdsc/PMP_related_studies/TR34.pdf> Accessed November 2013.

Harstine, L.J. 1991. Hydrologic Atlas for Ohio: Average Annual Precipitation, Temperature, Streamflow, and Water Loss for a 50-Year Period, 1931-1980. Ohio Department of Natural Resource Division of Water, Ground Water Resources Section. Water Inventory Report No. 28.

Hoorman, J., T.Hone, T.Sudman Jr., T.Dirksen, J. Iles, and K.R. Islam. 2008. Agricultural impacts on lake and stream water quality in Grand Lake St. Marys, Western Ohio. Water, Air, and Soil Pollution. 193: 309-322.

ODNR. 2013. Grand Lake St. Marys State Park. Ohio Department of Natural Resources. <http://parks.ohiodnr.gov/grandlakestmarys>. Accessed November 2013.

Perkins, W.W., E.B. Welch, J. Frodge, and T. Hubbard. 1997. A zero degree of freedom total phosphorus model: Application to Lake Sammamish, Washington, Lake and Reserv. Manage. 13:131-141.

Tetra Tech, Inc. 2013. *Preliminary Assessment of Effectiveness of the 2012 Alum Application—Grand Lake St. Marys.* <http://www.lakeimprovement.com/sites/default/files/GLSM%20Alum%20Report%20022020 13.pdf>. Accessed November 2013.

7.6.3 Estimate Land Treatment Needs

A watershed project must set land treatment goals based on estimates of pollutant reductions needed and the BMPs available to accomplish those reductions. Where aquatic habitat improvement is needed, the project's plan must also be based on an assessment of the change in habitat parameters (e.g., water temperature, cobble embeddedness, flow characteristics as well as pollutant loadings) needed to support aquatic life. Various approaches to determining land and in-stream treatment needs to restore and protect aquatic habitat have been documented (e.g., OWEB 1999, Rosgen 1997). Obviously, the BMPs selected must be those capable of addressing the pollutants and sources identified in the planning process. Setting goals for the level and extent of BMP implementation is necessary, but is an inexact science, partially because of the largely voluntary (and hence poorly predictable) nature of land treatment programs, and partially because it is difficult to predict water quality response to BMP implementation at the watershed level. See USEPA (2008) for a comprehensive discussion of watershed project planning.

Where local data on BMP performance exist (e.g., a documented 45 percent reduction in suspended sediment load through a water and sediment control structure or a 25 percent reduction in runoff phosphorus concentration from fields in conservation tillage), they can be applied to estimate pollutant reductions anticipated from different levels of implementation. Where locally-validated data do not exist, there is ample information in published literature (e.g., Simpson and Weammert 2009, USDA-NRCS 2012). Planners should use caution when applying performance data from other studies due to potential local site differences.

It should be noted that published BMP efficiencies do not generally account for interactions in multiple practice systems or address pollutant transport or delivery processes beyond the edge of field or BMP site scale. Modeling, e.g. the Soil Water Assessment Tool (SWAT), may be a better method for estimating treatment needs because some models account for routing of BMP effects through a watershed. Simple pollutant load estimation tools such as USEPA's STEPL (Spreadsheet Tool for Estimating Pollutant Load) can be used to provide general estimates of load reductions achievable via various BMP implementation options, but STEPL, for example, addresses a limited set of pollutants and simulates a limited set of BMPs.

7.6.4 Estimate Minimum Detectable Change

One critical step in watershed project planning is to use the data that have already been collected to evaluate the Minimum Detectable Change (MDC), the smallest monitored change in a pollutant concentration or load over a given period of time required to be considered statistically significant. Understanding of the MDC will assist in planning both land treatment and water quality monitoring design and will support predictions of project success. See section 3.4.2 for details.

The basic concept in the calculation of MDC is simple: variability in water quality measurements is examined to estimate the magnitude of changes in water quality needed to detect significant differences over time. The MDC is a function of pollutant variability, sampling frequency, length of monitoring time, explanatory variables or covariates (e.g., season, meteorological, and hydrologic variables) used in the analyses which 'adjust' or 'explain' some of the variability in the measured data, magnitude and structure of the autocorrelation, and statistical techniques and the significance level used to analyze the data. In general, MDC decreases with an increase in the number of samples and/or duration of sampling in a monitoring program.

The MDC for a system can be estimated from data collected within the same system during the planning or the pre-BMP project phase or from data collected in a similar system, such as an adjacent watershed.

As noted above, MDC is influenced by the statistical trend test selected. For the MDC estimate to be valid, the required assumptions must be met. Independent and identically distributed residuals are requirements for both parametric and nonparametric trend tests. Normality is an additional assumption placed on most parametric trend tests. However, parametric tests for step or linear trends are fairly robust and therefore do not require 'ideally' normal data to provide valid results.

The standard error on the trend estimate, and therefore, the MDC estimate, will be minimized if the form of the expected water quality trend is correctly represented in the statistical trend model. For example, if BMP implementation occurs in a short period of time after a pre-BMP period, a trend model using a step change would be appropriate. MDC in this case is an extension of the Least Significant Difference (LSD) concept (Snedecor and Cochran 1989). If the BMPs are implemented over a longer period of time, a linear or ramp trend would be more appropriate. Calculation of the MDC is discussed in detail in Spooner et al. (2011a) and illustrated in section 3.4.2. The reader is advised to consult that publication to calculate and apply the MDC analysis.

MDC provides an excellent feedback to whether the planned BMPs (type and location, acres served) will result in an amount of change in pollutant concentration or loads that can be statistically documented. Results of the MDC analysis can also be applied to the design of a long-term monitoring program (e.g., sampling frequency, monitoring duration). Decisions about data analysis such as the use of covariates to reduce effective variability and thereby reduce MDC can be made, or MDC calculations can be used to better understand the potential and limitations of an ongoing monitoring effort. Note that the MDC technique is applicable to water quality monitoring data collected under a range of monitoring designs including single fixed stations and paired watersheds. MDC analysis can be performed on datasets that include either pre- and post-implementation data or just limited pre-implementation data that watershed projects have in the planning phase.

7.6.5 Locate Monitoring Stations

The general location of monitoring stations is described for each monitoring design in section 2.4. Analysis of pre-project data, in conjunction with monitoring objectives, can provide insight into optimum location of monitoring stations to be used in watershed project effectiveness evaluation. Section 3.3 provides a discussion on how site characteristics, access, and logistics influence decisions on locating monitoring stations. Spatial analysis of land use and management data, including understanding of relationships between land use and management patterns and water quality (see section 7.5.2.3) can be used to inform monitoring site selection. Inferences on critical source areas (section 7.5.2.4) should also be used to guide station location. Subwatersheds showing very high and very low NO_2+NO_3-N concentrations in Figure 7-19, for example, might be selected for monitoring as treatment and control watersheds, respectively.

7.7 Data Analysis for Assessing Individual BMP Effectiveness

The availability of BMPs that perform a known water quality function is fundamental to NPS watershed projects. Many practices have a long history (e.g., buffers, conservation tillage for erosion control, grassed waterways) and their efficiency in reducing NPS pollutants is well-documented by research, although highly variable depending on site, management, and other factors. The performance of other

BMPs, such as novel practices or practices not common locally, may not be fully understood. In such cases, and in cases where specific assurance that BMPs will perform adequately in local circumstances is required, the effectiveness of individual BMPs may be assessed through monitoring.

Common monitoring designs for assessing BMP effectiveness include:

- Plot studies
- Input/output at the BMP practice scale
- Above/below at the site scale
- Paired watershed at the edge-of-field scale

Data analysis for above/below and paired-watershed BMP monitoring is essentially the same as for these designs at the watershed project level (see section 7.8). This section will focus on discussion of data analysis for plot studies and for BMP input/output studies.

7.7.1 Analysis of Plot Study Data

Controlled, replicated plot or field studies are effective for testing specific practices of undocumented effectiveness or evaluating the effectiveness of a BMP program or system at a farm or watershed scale (USEPA 1997b). To some extent, plots represent microcosms of an area where a full-scale BMP might be applied, where inputs, management, and outputs can be controlled and measured to a degree that would be extremely challenging at full scale. Most importantly, because plots are small (often less than 100 m^2), it is possible to test different levels of treatment and replicate treatments in the same experiment, thus potentially capturing enough variability to have some statistical confidence in the outcome.

As discussed in section 2.4.2.2, there are a variety of plot study designs, including factorial experiments, Latin Squares, and complete and incomplete block designs. Approaches to analyzing data from these various options differ to some degree, but most follow three basic steps:

- Test to see if there are significant differences among the treatments
- Test to find which treatments are significantly different
- Determine the magnitude of differences

Statistical approaches discussed in this section focus on one- and two-factor designs (generally Randomized Complete Block, RCB). Readers should consult statistics textbooks and other resources for information on procedures to analyze data from the more complicated designs such as Latin Squares and incomplete block designs.

Data from simple plot studies are usually analyzed using ANOVA (parametric) or the Kruskal-Wallis test (nonparametric). These procedures allow the determination of significant differences in group means for pollutant concentration or load coming from plots. When a plot study is conducted for a single precipitation/runoff event (either natural or simulated rainfall), the groups tested would be the replicate plots for each type or level of treatment, plus control plots. For a plot study conducted over a series of events, the groups tested could be data from replicate groups within individual events or mean concentration or total load over the entire series of events, depending on the study design. Note that the ANOVA and Kruskal-Wallis procedures only document that one or more group means differ significantly from the other groups. To determine which of the group means are significantly different, use a multiple

comparison test such as Tukey's or the Least Significant Difference tests (Snedecor and Cochran 1989, USEPA 1997b). Applications of the Least Significant Difference and Tukey's tests are illustrated in section 4.6.1 (pages 4-55 to 4-56) and 4.6.4 (pages 4-63 to 4-64), respectively, of the 1997 guidance (USEPA 1997b).

The ANOVA procedure can also be used where there is more than one factor or explanatory variable (e.g., plot, slope), whereas the Kruskal-Wallis test handles only one factor. The Friedman nonparametric test is recommended for more than one factor. Application of these tests is described and illustrated in section 4.6 (pages 4-52 to 4-64) of the 1997 guidance (USEPA 1997b).

One-factor comparisons using ANOVA assume random samples, independent observations, and normal distributions for each group, as well as the same variance across groups. Group sample sizes can differ, however. An illustrative example application of the Kruskal-Wallis test for one-factor comparisons is included in the 1997 guidance (USEPA 1997b), pages 4-56 to 4-58.

Two-factor comparisons using ANOVA depend on whether the factors interact. An example of an interaction is the relationship between crop yield and precipitation, both of which can independently influence soil nitrate levels; greater yields remove more nitrate from the soil profile and greater precipitation moves more nitrate through the soil profile. Yield, however, is also influenced by precipitation (e.g., drought or excessively wet soil conditions), so there is an interaction between the two factors. The plot study analysis from Vermont (see Example 7.7-1) illustrates consideration of interactions.

Both the scope of inferences that can be made and the F statistic calculation differ for fixed effect models (e.g., rainfall simulation studies in which rainfall rates are not randomly selected) versus models using randomly selected or combinations of randomly selected and fixed factors. Readers are recommended to section 4.6.2 (pages 4-58 to 4-61) of the 1997 guidance (USEPA 1997b) for an illustrative example and a discussion of these and other important considerations when applying ANOVA to two-factor comparisons. If the data are log-transformed prior to ANOVA, the treatment effects are then interpreted as multiplicative (rather than additive) in the original units. An alternative approach is to rank-transform the data prior to ANOVA, resulting in a comparison of the medians of the data in the original units (see pages 4-61 of the 1997 guidance for details).

Once a statistically significant difference has been demonstrated and the different group means have been identified, it is possible to explore the magnitude of such differences. Methods for two random samples, two paired samples, or a single sample versus a reference (e.g., criterion for a WQS) are described in section 4.5.3 (pages 4-51 to 4-52) of the 1997 guidance. It is important to take the extra step of determining confidence intervals for difference estimates.

In addition to using statistical tests to document differences among treatment groups, plot data can be evaluated by direct comparison of event mean concentration (EMC) or event load (or areal load) among treatments. For plot studies evaluating practice performance over a series of events, a cumulative export plot (where the sum of cumulative mean export from each group is plotted sequentially over the study) will illustrate the behavior of treatment groups in different events. It must be cautioned that data and quantitative inferences about practice performance from plots are usually very difficult to extrapolate to field or watershed scale because physical processes like runoff velocity are not well-represented in very small areas.

Example 7.7-1. Plot Study Analysis: Bacteria Runoff from Manure Application in Vermont

Objective

Evaluate several practical methods for controlling *E. coli* in runoff from manure application sites.

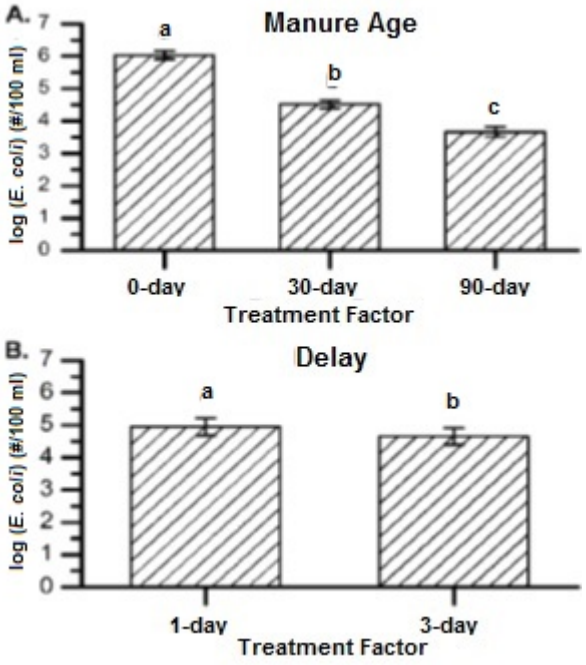
Specific objectives included: (1) determine the effect of manure storage time on *E. coli* losses in runoff from hay and corn land receiving liquid dairy manure; (2) determine the effect of manure incorporation on *E. coli* losses from corn land receiving manure; (3) determine the effect of vegetation height on *E. coli* losses in runoff from hay land; and (4) determine the effect of delay between manure application and rainfall on *E. coli* losses in runoff from hay land and corn land.

Monitoring Design

Two runoff experiments were conducted at separate hay land and corn land sites. For each experiment, 40 1.5- by 3-m plots were created, representing a factorial design of 3 replicates for each treatment combination, 3 manure ages, 2 vegetation heights (for hay) or incorporated/unincorporated (for corn), 2 delay to rain durations, resulting in 3 x 3 x 2 x 2 (36) treatments, plus three control plots (no manure applied), and one extra plot reserved as a backup. Specific treatments were assigned to plots randomly. A rainfall simulator was used to generate runoff from the test plots by continuously and uniformly applying water at an intensity resembling natural rainfall. For each experiment, the first hour or first 19 L of runoff was collected from each plot.

ANOVA table for hay land runoff experiment. Results show significant manure age, delay to rain, and two interactions

Analysis of Variance					
Source	df	Sum of Squares	Mean of Squares	*F* Ratio	*P*
Model	7	34.8385	4.9769	37.135	<0.001
Error	26	3.4846	0.1340		
Total	33	38.32311			
Effects Tests					
Source	df	Sum of Squares		*F* Ratio	*P*
Manure Age	2	31.1188		116.096	<0.001
Vegetation Height	1	0.0673		0.502	0.485
Delay to Rain	1	0.602		4.494	0.044
Manure Ag x Vegetation Height	2	0.7427		2.771	0.081
Vegetation Height x Delay to Rain	1	1.2076		9.011	0.006



Levels of *E. coli* in hay land plot runoff by two treatment factors. Error bars represent ±1 standard deviation; bars labeled with different letter(s) differ significantly ($P \leq 0.1$).

Data Analysis

Statistical analysis of *E. coli* data was conducted on $\log_{10}$ transformed data to satisfy the assumptions of normality and equal variances. All statistical tests were performed using JMP software at an α of 0.1. The effect of treatment on levels of *E. coli* in runoff was evaluated by multi-factor analysis of variance (ANOVA). After an initial pass that included all treatment factors and all possible interactions, nonsignificant ($P > 0.1$) interactions were removed from the model and a final reduced-model ANOVA was conducted. Interpretations of treatment effects were based on the reduced model.

Source: Meals and Braun 2006

7.7.2 Analysis of BMP Input/Output Data

For some BMPs, such as agricultural water and sediment control basins or stormwater treatment devices, it is possible to assess practice effectiveness by directly monitoring input and output pollutant concentration and load. In either an agricultural or an urban setting, inflow and outflow variables such as flow volume, peak flow, EMC, or pollutant loads, are measured and the effectiveness of the BMP is calculated by comparing input vs. output.

Paired input and output data can be compared by testing for significant differences in group means using the parametric paired Student's t or the nonparametric Wilcoxon Rank Sum test. Comparison of random observations from two samples (e.g., input and output from a large constructed wetland for which it is not possible to collect paired samples due to uncertain or variable flow pathways or time of travel) can also be made with a t-test if equal variance is confirmed (e.g., F test); the Mann-Whitney test is the nonparametric alternative in this case. These tests are described and illustrated in detail in chapter 4 (pages 4-34 to 4-52) of the 1997 guidance (USEPA 1997b).

Once a statistically significant difference is confirmed, BMP efficiency can be reported in a number of ways, including:

- Efficiency ratio (percent reduction in flow, EMC, or load),
- Summation of loads (percent reduction in sum of all monitored loads)
- Regression of loads (reduction efficiency is expressed as the slope of a regression line for input load vs. output load)
- Efficiency of individual storm load reductions across all monitored events
- Percent removal relative to a water quality criterion

All of these methods are described and illustrated by Geosyntec and WWE (2009). It is recommended that more than one method is used wherever possible because the results may differ. For example, results from the summation of loads and efficiency ratio (e.g., EMC) methods may not agree because of differences in how the water budgets are represented (Erickson et al. 2010b).

The EMC is the total event load divided by the total runoff volume. It should be noted that, for large practices such as some constructed wetlands, the influent EMC (EMC_I) must be adjusted to account for rain that falls directly onto the practice (Erickson et al. 2010a). Long-term performance can be determined by calculating the average EMCs (AvgEMC) for both influent (input or $AvgEMC_I$) and effluent (output or $AvgEMC_O$) and using these values to calculate the percent reduction in concentration (Erickson et al. 2010b). The simple equation becomes:

$$Long - Term\ Efficiency = 100 \times \left(\frac{AvgEMC_I - AvgEMC_O}{AvgEMC_I}\right)$$

An alternative approach that can add statistical power is to pair the input and output EMCs for each storm and calculate the average of the differences as an estimate of pollutant reduction efficiency. A paired t-test can then be used to determine both the statistical significance of and confidence interval for the reduction. See section 4.2.1 (pages 4-11 to 4-14) of the 1997 guidance (USEPA 1997b) for additional information and an illustrative example of EMC calculations.

The percent reduction in the sum of all monitored loads is calculated using the summed loads for both the input (L_I) and output (L_O):

$$Percent\ Reduction = 100 \times \frac{(L_I - L_O)}{L_I}$$

Similar to the alternative proposed for EMCs, the average differences between paired input and output loads can also be used as an estimate of pollutant reduction efficiency.

Erickson et al. (2010b) illustrate a method for determining the uncertainty of long-term performance estimates that are based on either the EMC or summation of load method they describe. Required input is the number of storm events, the standard deviation of the performance data, and a Student's t value.

Using data from Erickson et al. (2010b), Figure 7-21 illustrates regression of effluent against influent event loads. It should be noted that in this example the y-intercept was not constrained to the origin as recommended[9] by Geosyntec and WWE (2009). The slope of the line indicates that effluent concentration is 37 percent of influent concentration above the baseline level (y intercept) of 0.01 kg TP. In other words, the BMP reduces the load by 63 percent (100-37), a number that agrees well with the 57.5 percent removal rate calculated by summation of loads (Erickson et al. 2010b). Regression analysis is illustrated and described at *CADDIS Volume 4: Data Analysis*.

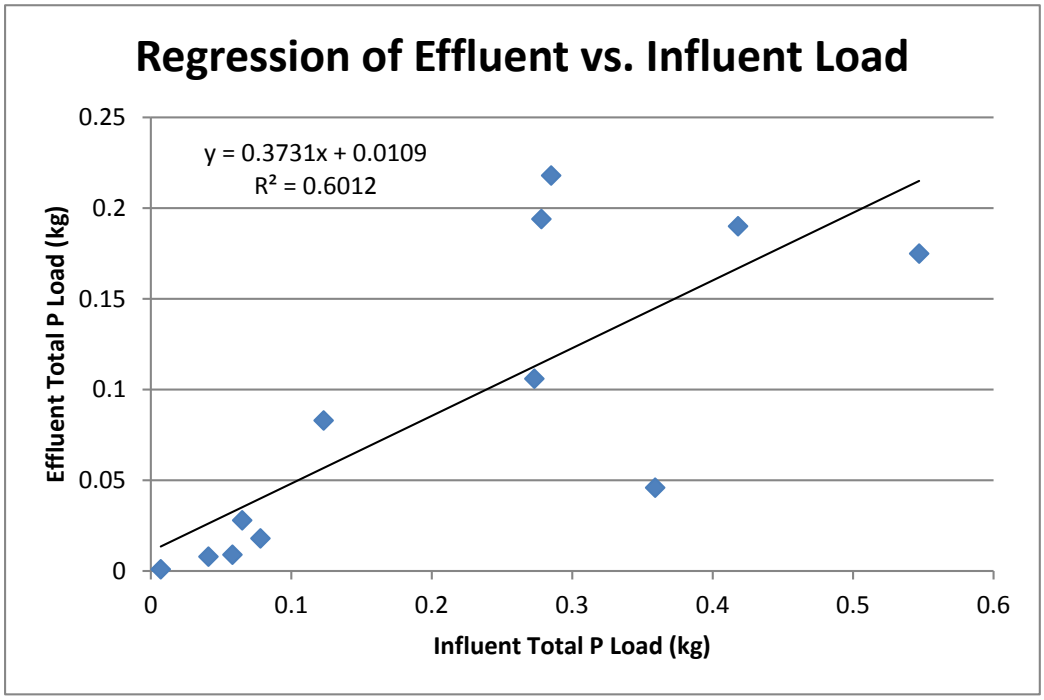


Figure 7-21. Regression of output versus input load (data from Erickson et al. 2010b)

[9] While specified in the definition of the regression of loads method, Geosyntec and WWE (2009) includes a comment suggesting that such a constraint "is questionable and in some cases could significantly misrepresent the data."

BMP efficiency evaluated by input/output monitoring is frequently reported as simply percent removal of a pollutant. In most cases, this is an inadequate basis for assessing BMP performance. Percent removal is primarily a function of input quality, and BMPs with a high apparent removal percentage may still have unacceptably high concentrations or loads in their output. Some BMPs with long retention times (e.g., constructed wetlands) show long-term performance that is not evident in comparing paired input-output samples because material from one event is not discharged until a subsequent event (i.e., the samples are not paired or matched). Finally, a simple percent removal calculation can be dominated by outliers that distort an average performance indicator.

Constructing an Effluent Probability Plot

The cumulative distribution function for the EMCs for the outflows and inflows can be created from the following steps:

- Calculate the EMC for each storm's outflows.
- Rank all EMCs for all storms from smallest to largest.
- Assign a 0 to 1 'probability' to the data based upon their ranked order. For example, if 10 storms were monitored, the ranked values would receive a 'probably ranking' value of 0.1, 0.2, … 1.0 for the lowest to highest EMC values.
- Plot the 'probability ranking' values on the Y-scale and the EMCs on the X-scale. The Y-scale should be plotted on a probability scale. Alternatively, the Y-axis could be expressed as the number of standard deviations (e.g., +/- 3). Because the EMCs are likely to follow a log-normal distribution, the X-axis should be a log scale.
- Repeat the procedure for the inflows and plot on the same graph.

For these and other reasons, USEPA and ASCE have recommended the Effluent Probability Method for evaluating input/output data from a BMP (Geosyntec and WWE 2009). In this procedure, a statistically significant difference between input and output EMC or load is verified (e.g., by Student's t Test). Then, a normal probability plot is constructed of input and output data that allows comparison of BMP performance over the full range of monitored conditions. For example, Figure 7-22 shows an effluent probability plot for chemical oxygen demand (COD) from an urban wet detention pond evaluation The plot shows that COD was poorly removed at low concentrations (<20 mg/L), but that removal increased substantially for higher concentrations.

The Effluent Probability Method is essentially a cumulative distribution function for the EMCs of the inflows and outflows. The cumulative distribution function depicts the probability of values being below a given EMC value or the EMC values that a percentage (e.g., 50 percent) of the data falls above.

The magnitude of the difference in EMC (or loads) from the inflow and outflows can be examined across the range of EMC values. The Kolmogorov–Smirnov test is based on cumulative distribution functions and can be used to determine if the two empirical distributions are significantly different (Snedecor and Cochran 1989).

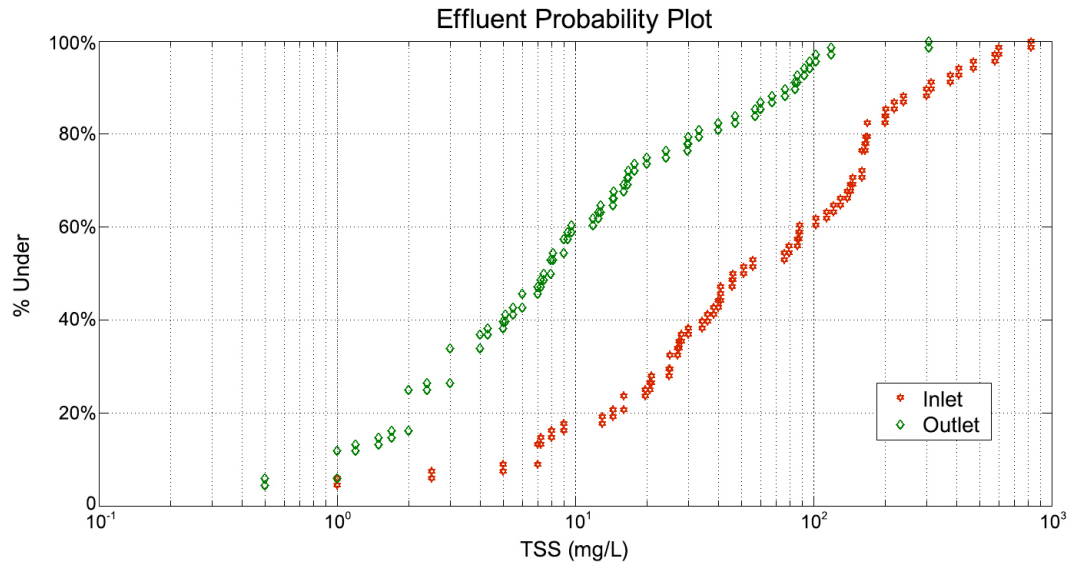


Figure 7-22. Effluent probability plot for input/output monitoring of a wet detention pond

$$Percent\ Removal\ vs\ Criterion = 100 \times \frac{(C_I - C_O)}{(C_I - C_C)}$$

Percent removal relative to a water quality criterion provides an indication of how well a BMP is performing compared to limits or expectations established for the local waterbody. Use of this method is recommended for specific event analysis, but not for a series of events (Geosyntec and WWE 2009). Calculation requires values for the criterion (C_C), input (C_I), and output (C_O), all expressed in the same units (concentration in this case):

For example, in a watershed with a target total N concentration of 0.75 mg/L, storm inlet and outlet concentrations of 3.6 mg/L N and 1.6 mg/L N, respectively, would yield a relative percent removal of 70 percent.

The reader is referred to *Urban Stormwater BMP Performance Monitoring* (Geosyntec and WWE 2009) for additional information on evaluating urban stormwater BMP performance through monitoring.

7.7.3 Analysis of BMP Above/Below Data

As noted earlier, BMP performance can be assessed using an above/below-before/after monitoring design, as long as the added area monitored by the downstream station is either entirely or predominantly influenced by the BMP. In such cases, analysis of monitoring data is done by the same approach as described in section 7.8.2.2. An example of this kind of above/below-before/after analysis of a single BMP can be found in the Otter Creek (WI) NNMP project, which assessed the effects of barnyard runoff control (see Example 7.7-2). This example illustrates application of the Hodges-Lehmann estimator described in section 4.5.3 of the 1997 guidance (USEPA 1997b).

Example 7.7-2. Above/Below-Before/After Analysis: Barnyard Runoff BMPs in Wisconsin

Monitoring Design

Sampling stations upstream and downstream of two investigated dairy barnyards were established in 1994/1995. At the upstream sampling stations, stream stage and precipitation were continuously monitored, and discrete water samples were collected automatically; at the downstream stations, only water quality samples were collected. Over the course of the study, 11 – 15 storm runoff periods were sampled at each of the sites. Continuous streamflow and instantaneous concentration data were used to estimate pollutant loads for individual storm-runoff periods.

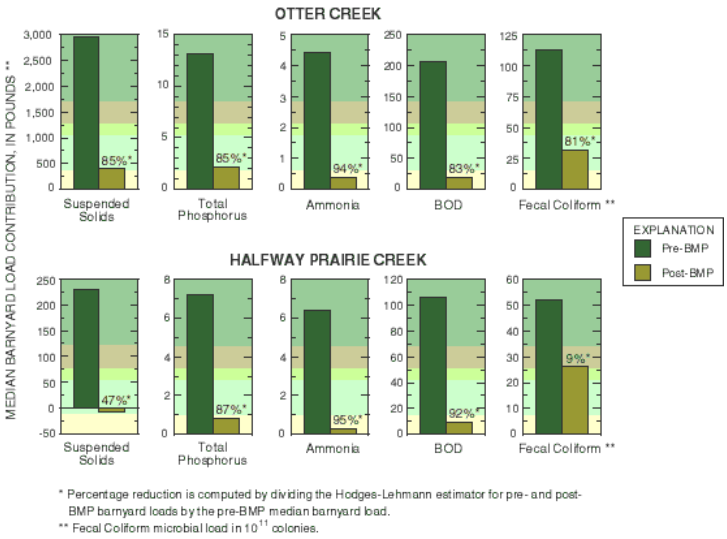
Pre-BMP Analysis

A critical aspect of obtaining useful conclusions for this study was the ability to document that downstream loads were significantly greater than upstream loads before the BMP systems were implemented. Results of t-Tests showed that, for the pre-BMP period at both creeks, downstream loads of total P, ammonia, BOD, and fecal coliform bacteria were significantly greater than upstream loads. At Otter Creek, pre-BMP downstream loads of total suspended solids also were significantly greater than those upstream. These significant differences indicated that each barnyard was an important contributor to the instream pollutant loads for the storm-runoff periods monitored.

Effects of Treatment

The difference between upstream and downstream constituent loads was computed for each pre- and post-BMP storm-runoff period. These differences were considered to be the load contributed by each barnyard. The bar graphs indicate that both barnyard BMP systems have reduced loads in the stream for each constituent. Each bar represents the median of all the differences between upstream and downstream constituent loads for both pre- and post-BMP storm-runoff periods. Although these medians could have been used to determine the percentage reduction achieved by each barnyard BMP system, it was decided that use of the Hodges-Lehmann estimator would be a more accurate approach (Helsel and Hirsch 2002). The Hodges-Lehman estimator is the median of all possible pairwise differences between pre- and post-BMP barnyard loads. This median difference was then divided by the pre-BMP median barnyard load for each constituent. The result was a percentage load reduction for each constituent.

The barnyard BMP system at Otter Creek reduced loads of total suspended solids by 85 percent, total P by 85 percent, ammonia by 94 percent, BOD by 83 percent, and microbial loads of fecal coliform bacteria by 81 percent; the respective loads at Halfway Prairie Creek have been reduced by 47, 87, 95, 92, and 9 percent.



* Percentage reduction is computed by dividing the Hodges-Lehmann estimator for pre- and post-BMP barnyard loads by the pre-BMP median barnyard load.

** Fecal Coliform microbial load in 10^{11} colonies.

Source: Stuntebeck and Bannerman 1998

7.7.4 Analysis of BMP Paired-Watershed Data

Some BMPs – especially agricultural BMPs that involve treatment of an entire field such as conservation tillage, cover crops, or nutrient management – can be evaluated using a paired-watershed design. In this case, monitoring takes place at the edge of field-sized watersheds, wherein one entire monitored field is designated to receive the BMP treatment. Automated samplers are required to collect storm event runoff. In the paired-watershed design, monitoring occurs during a calibration period in which both fields or subwatersheds have identical management. Then, after their pollutant responses to the same rainfall events are correlated, a treatment period occurs in which one of the subwatersheds receives the BMP treatment and the other remains in the 'controlled' management. Analysis of covariance (ANCOVA) is used to analyze the monitoring data from this type of study. See section 7.8.2.1 for details.

7.8 Data Analysis for Assessing Project Effectiveness

7.8.1 Recommended Watershed Monitoring Designs

Assessing the effectiveness of a watershed project where multiple BMPs are implemented in a land treatment program across a broad watershed area is a complex task with many sources of variability and uncertainty. Attributing changes in water quality documented through monitoring to land treatment, rather than to other causes such as drought or extreme weather, is another significant challenge. Monitoring designs (see chapter 2) recommended for assessing watershed project effectiveness are:

- Paired-watershed (link to section 2.4.2.3)
- Above/below-before/after (link to section 2.4.2.6)
- Nested-watershed (link to 2.4.2.3)
- Single watershed trend (link to section 2.4.2.5)

While not generally recommended because of cost and logistical constraints (see section 2.4.2.8), data analysis for multiple-watershed studies is also discussed here. These designs vary in their ability to evaluate watershed project effectiveness while controlling for sources of change other than land treatment; the designs also vary in the appropriate approach to data analysis. The paired-watershed design is generally considered to be the best design for this purpose because it strives for a controlled experiment to evaluate BMP effectiveness at a watershed scale, accounting for year-to-year variability in weather and streamflow through the use of a control watershed. Several common watershed project designs are excluded from the above list because they are not generally capable of reliably documenting water quality change and attributing the change to land treatment. Single watershed before/after and side-by-side watersheds, for example, cannot be recommended for watershed project effectiveness monitoring because they cannot be used directly to separate the effects of the BMPs from those of climate or watershed differences (e.g., soils, slope, land management) which may be the actual causes of the observed differences (see section 3.4). The single watershed before/after design can, however, be useful in comparing pollutant loads over time to determine if TMDL goals have been achieved (see section 7.9).

None of these designs will perform effectively, however, if all the requirements of the design are not met. In some cases, failure to meet a single criterion (e.g., unexpected treatment in the control watershed of a paired design, or changing analytical procedures during a long-term single-station study) may doom the effort.

Each of these designs is discussed in chapter 2; information relevant to data analysis procedures are provided in this section.

7.8.2 Recommended Statistical Approaches

The following sections recommend statistical approaches to analysis of data from recommended watershed monitoring designs. Additional details on specific statistical tests can be found in chapter 4 (Data Analysis) of the 1997 guidance (USEPA 1997b).

7.8.2.1 Paired Watershed

As described in chapter 2, the most effective practical design for evaluating watershed-level BMP effectiveness through monitoring is the paired-watershed design due to the presence of an experimental control for year-to-year hydrologic variability (Clausen and Spooner 1993). The paired-watershed design has been discussed in section 2.4.2.3. The basic design involves two watersheds (a control, where no BMPs are to be implemented, and a treatment watershed where land treatment will be applied) and two periods (a pre-treatment or calibration period, and a treatment period). Analysis of paired data (i.e., frequently collected chemical or physical data) from treatment vs. control areas should show a statistically significant correlation and result in a strong linear regression model (usually using log-transformed data) that changes from the pre-treatment to post-treatment period. In the case of biological monitoring (e.g., sampling twice per year), relationships between treatment and control watersheds should change in a more qualitative manner from pre- to post-treatment periods. For example, treatment and control watersheds may both be of "poor" quality in the pre-treatment (or pre-BMP) period, whereas the treatment watershed improves to "good" quality while the control watershed remains at "poor" quality during the post-treatment period. Additional considerations for paired-watershed designs with more than one treatment watershed are discussed at the end of this section.

Additional Information on ANCOVA

- USEPA. 1997b. *Monitoring Guidance for Determining the Effectiveness of Nonpoint Source Controls* Chapter 4;
- Clausen and Spooner. 1993. Paired Watershed Study Design. 841-F-93-009;
- Grabow et al. 1999. Detecting Water Quality Changes Before and After BMP Implementation: Use of SAS for Statistical Analysis; and
- Grabow et al. 1998. Detecting Water Quality Changes Before and After BMP Implementation: Use of a Spreadsheet for Statistical Analysis of Paired Watershed, Upstream/Downstream and Before/After Monitoring Designs.

See section 4.8 of the 1997 guidance (USEPA 1997b) for details and an example including a method for determining if enough calibration data has been collected to warrant advancing to the BMP treatment period. Failure to establish a statistically valid pre-treatment correlation will doom the evaluation design.

7.8.2.1.1 Analysis of Covariance (ANCOVA) Procedure – Paired-Watershed Analysis

The Analysis of Covariance (ANCOVA) procedure is used to analyze data from a paired-watershed study (Clausen and Spooner 1993, Wilm 1949, Clifford et al. 1986, Meals 2001). ANCOVA combines the features of ANOVA with regression (Snedecor and Cochran 1989) and is an appropriate statistical technique to use in analysis of watershed designs that compare pre- and post-BMP periods using treatment and control watershed measurements. When applied to the analysis of paired-watershed data,

ANCOVA is used both (a) to compare pre- and post-BMP regression equations between water quality measurement values (e.g., sediment concentration) for the treatment and control watersheds and (b) to test for differences in the average value (e.g., of sediment concentration) for the treatment watershed between the two time periods after adjusting measured values for covariates such as flow. Covariates are added to the analysis to decrease the residual error and give a more precise comparison between covariate-adjusted mean values.

There are three basic steps to performing ANCOVA:

1. Obtain paired observations
2. Select the proper form of linear model
3. Calculate the adjusted means (LS-means) and their confidence intervals

Paired observations could represent observations collected on the same date, the same time period for composite samples, or from the same storm event. Weekly flow-weighted composite samples taken at the outlet of both control and study watersheds would satisfy this requirement.

The second step is to select the proper form of the model. There are two basic statistical models here for paired-watershed studies:

- The change in treatment watershed concentration with change in control watershed concentration (i.e., the slope of the linear relationship between paired samples) remains constant through both the calibration and treatment periods.
- The slope of the relationship changes from calibration to treatment period.

ANCOVA for paired-watershed studies is illustrated by Figure 7-23 where pollutant concentration (or load) pairs are plotted with the treatment basin values on the Y-axis and the control basin values on the X-axis. The slopes of the pollutant concentrations plotted for both periods are tested to determine if they are significantly different (see B in Figure 7-23) or if the same slope can be assumed (see A in Figure 7-23). A change in slope and/or mean value indicates that pollutant concentrations for the treatment watershed exhibited different patterns, or magnitude, after BMPs were applied as compared to the calibration period. For example, in both A and B of Figure 7-23 the same concentration in the control watershed corresponds to a lower concentration in the treatment watershed in the post- (treatment) versus the pre-BMP (calibration) period, indicating beneficial effects from the BMPs. In the case of B, both the mean and the slope are reduced in the treatment period. The adjusted mean concentrations (LS-means) for the calibration and treatment periods are also compared for differences as described above under "ANCOVA Procedure."

The best statistical model for a particular dataset is determined with a test for homogeneity of slopes (i.e., same or different slopes) using the 'full analysis of covariance model' that allows for separate regression lines (i.e., different slopes and intercepts, Figure 7-23B) for the calibration and treatment periods (i.e., the groups) for the regression of the treatment watershed variable (Y) on the control watershed variable (X):

$$Y_{ij} = b_{0i} + \sum_{i=1}^{k} b_{1i}\left(X_{ij}\right) + e_{ij} \qquad \text{("Full statistical model" for different slopes)}$$

Where:

Y_{ij} = the j^{th} observation for Y in period i (e.g., pollutant concentration or load from treatment watershed)
b_{0i} = the intercept ($B_{0)}$ for period i
b_{1i} = the regression coefficient ($B_{1)}$ of Y on X for period i
X_{ij} = the jth observation for X in period i (e.g., pollutant concentration or load from control watershed paired with same sample time as Y_{ij})
k = number of time periods (with 'calibration' and 'treatment' periods, k=2)
e_{ij} = the residuals or experimental error for the j^{th} observation for Y in period i. Note: if the data are weekly, biweekly, or monthly, this error series is likely autocorrelated with Autoregressive, Lag 1 or AR(1) and depicted as V_{ij} or V_t. A statistical model that allows for this autocorrelated error structure should be used (e.g., PROC AUTOREG in SAS software (SAS Institute 2016d) or use a correction for the standard error on the test of LS-means (See section 7.3.6)

The F-Test for the homogeneity of slopes is used to see if the best model requires separate slopes for each period or the same (pooled) slope (Clausen and Spooner 1993). The best model will have the lowest residual sum of squares (SSE). The F-statistic for testing the homogeneity of slopes is:

$$\text{F statistic} = \left[\frac{(SSE_R - SSE_F)}{(k-1)}\right] / MSE_F$$

Where:

SSE_R = Residual sum of squares for the reduced model with a common (pooled) slope (see below)
SSE_F = Residual sum of squares for the full model which allows for separate slopes for the calibration and treatment periods
k = number of groups (calibration + treatment periods = 2 in this case)
MSE_F = Mean square error from the full model

This F-statistic is compared to an F distribution with (k-1) and (N-2k) degrees of freedom (d.f.), where k is the number of groups and N is the total sample size (i.e., the total number of paired samples used in the analysis). See Example 7.8-1 below for examples of how to test if the slopes are different using an 'interaction' term in the statistical software programs.

If there is no evidence for separate slopes, then a "reduced model" with the same slopes assumed for each group (based on pooled data) should be used (see Figure 7-23A). If the interaction term is significant, then the "full model" is the correct model and the significance of the difference between all possible pairs can be obtained (see Figure 7-23B).

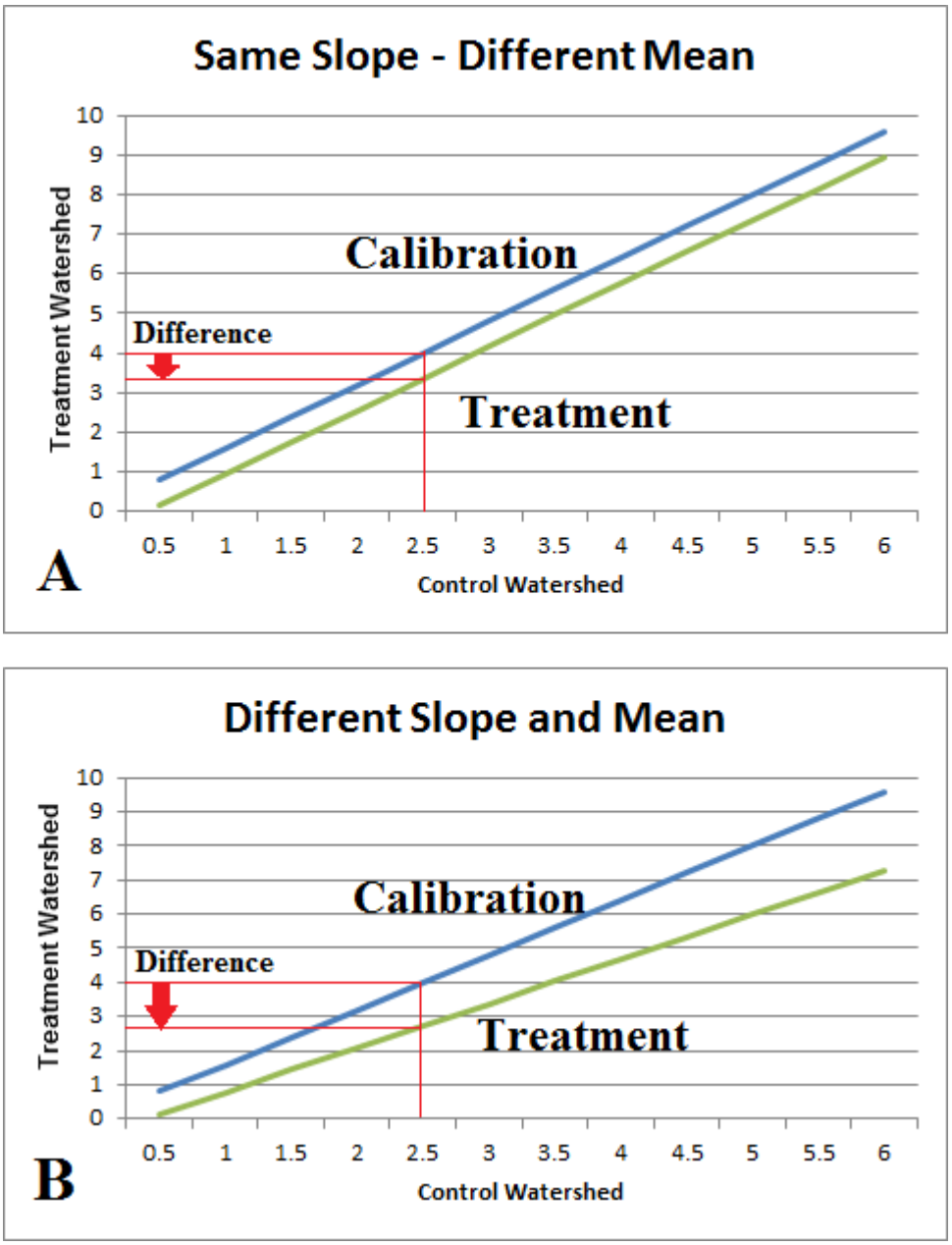


Figure 7-23. Conceptualized regression plots for paired-watershed data. The red line indicates the comparison of the treatment watershed from the calibration vs. treatment periods evaluated at the LSMEANS value of 2.5 (the mean of all sampled values in the control watershed over the entire sampling duration (both treatment and calibration period).

Example 7.8-1. Software Examples for the Statistical Analyses using Analysis of Covariance (ANCOVA) for the Paired-Watershed Study

Statistical software packages may vary in how they address ANCOVA. A few examples are given below. NOTE: We will provide a sample dataset (e.g., Walnut Creek, IA) and results for this example so users can test their own techniques and software.

A. SAS Software, assuming no autocorrelation

The SAS (SAS Institute 2010) program statements that generate a covariance model with unique slopes for each group ("full model", different slopes) are:

```
PROC GLM; CLASS PERIOD;
MODEL Y = X PERIOD PERIOD*X/ SOLUTION;
LSMEANS PERIOD /PDIFF;
```

Where the user inputs the variable names used for their project data for:

Y = Name of variable which contains the treatment watershed values (e.g., concentration/load)

X = Name of variable which contains the control watershed values (e.g., concentration/load)

PERIOD = calibration or treatment period

PERIOD*X = the "interaction" term that allows for different slopes for each PERIOD

The other terms are part of the SAS program software syntax. SOLUTION is optional but generates the regression equation for each PERIOD. The LSMEANS SAS statement generates the LS-means for each PERIOD. The PDIFF option produces significance tests to compare the LS-means for each PERIOD for statistically significant differences.

If there is no evidence for separate slopes (i.e., the PERIOD*X interaction term in the SAS output is not significant), then a "reduced model" with the same slopes assumed for each group (based on pooled data) should be used. If the interaction term is significant, then the "full model" is the correct model and the significance of the difference between all possible pairs can be obtained from the PDIFF option in the LSMEANS statement above.

The SAS program statements that generate a covariance model with common slope but unique intercepts for each period ("reduced model") are:

```
PROC GLM; CLASS PERIOD;
MODEL Y = PERIOD X/ SOLUTION SS1 SS3;
LSMEANS PERIOD /PDIFF;
```

NOTE regarding data setup:

The input data set has columns for each of the variables: Y, X, PERIOD, and DATE. Although DATE is not used in this software example, it is useful to match the values in each row for Y, X, and PERIOD to the correct sample collection date so that the Y and X values are correctly paired up. For the PROC GLM software procedure, PERIOD can be "0" and "1" or "Pre" and "Post" or any other numeric or character value desired. But, be aware that internal to SAS, "0" and "1" values will be generated based upon the alphabetical order – something to consider when interpreting the solutions for the regression line equations for each time period.

Example 7.8-1. Continued

B. SAS Software, data set with autoregressive, lag 1, AR(1) autocorrelation

The SAS (SAS Institute 2010) program statements that generate a covariance model with *unique slopes for each group ("full model", different slopes)* and accommodate an AR(1) error structure are:

```
PROC AUTOREG;
MODEL Y = X PER PER_INTER/NLAG=1 DWPROB;
```

Where the user inputs the variable names used for their project data for:

Y = Name of variable which contains the treatment watershed values (e.g., concentration/load)

X = Name of variable which contains the control watershed values (e.g., concentration/load)

PER = calibration or treatment period ("0" for pre-BMP period values; "1" for post-BMP values)

PER_INTER = the "interaction" term that allows for different slopes for each period. This is a numeric variable whose values are created by multiplying the values of X and PER for each observation

The other terms are part of the SAS program software syntax. NLAG=1 indicates a lag 1 error structure (PROC AUTOREG assumes an autoregressive error structure).

If there is no evidence for separate slopes (i.e., the PER_INTER interaction term in the SAS output is not significant), then a "reduced model" with the same slopes assumed for each group (based on pooled data) should be used. If the interaction term is significant, then the "full model" is the correct model.

The SAS program statements that generate a covariance model *with common slope but unique intercepts for each period ("reduced model")* are:

```
PROC AUTOREG;
MODEL Y = X PER /NLAG=1 DWPROB;
```

NOTE regarding data setup:

The data setup is similar to the PROC GLM software example in A above, except there is no CLASS option in PROC AUTOREG. Numeric input variables needs to be created for all input variables (e.g., 0 and 1 for pre- and post- BMP periods). Since this model includes is a time series error structure, the data must be sorted by date order and have equal spaced time intervals. PROC AUTOREG can correctly handle missing values. In such cases, a data record for the date should be included, but with missing values (indicated by a "." for the missing data input values.

When the reduced model with common slopes is used, the following equation (Snedecor and Cochran (1989) should be used to describe the linear regression for each time period, i, which would have the same slope, but be allowed to have different intercepts:

$$Y_{ij} = b_{0i} + b_1(X_{ij}) + e_{ij} \qquad \text{("Reduced model" for same slopes)}$$

Where:

Y_{ij} = the j^{th} observation for Y in period i (e.g., treatment watershed concentration or load)
b_{0i} = the intercept for period i
b_1 = the regression coefficient of Y on X pooled over all periods
X_{ij} = the j^{th} observation for X in period i (e.g., control watershed concentration or load)
e_{ij} = the residual or experimental error for the j^{th} observation for Y in period i ($V_{t\ for}$ autocorrelated error series)

Note that this version of the covariance model forces the slope of the regression of Y on X to be the same for each group, but allows the intercept to be unique (i.e., the regression lines representing each group are parallel).

Example 7.8-1. Continued

C. JMP Software, data set with no autocorrelation

Steps: Analyze => Fit Model => Select "Y' Variable, Add variables to the Model Effects ("X" and "PERIOD, highlight PERIOD and X variables in Select Colum and then select 'Cross' in Model Effects to include interaction term=>Run

NOTE regarding data setup:

The input data set has columns for each of the variables: Y, X, PERIOD, and DATE. Although DATE is not used in this software example, it is useful to match the values in each row for Y, X, and PERIOD to the correct sample collection date so that the Y and X values are correctly paired up. For the PROC GLM software procedure, PERIOD can be "0" and "1" or "Pre" and "Post" or any other numeric or character value desired. But, be aware that internal to SAS, "0" and "1" values will be generated based upon the alphabetical order – something to consider when interpreting the solutions for the regression line equations for each time period.

Note: if data has autocorrelated, autoregression, order 1 or AR(1) error series, the standard error on the differences between the LS-means can be adjusted and then the corrected significant differences can be determined by:

$$std.dev._{corrected} = std.dev._{uncorrected}\sqrt{\frac{1+\rho}{1-\rho}}$$

Where ρ = autocorrelation coefficient at lag 1
Std. dev = standard error on the differences of the LS-means

D. MiniTab Software, data set with no autocorrelation

Steps: Stat > ANOVA > General Linear Model. In the responses, model, and random factors dialogue boxes, enter "Y", "X PERIOD X*PERIOD", and "PERIOD", respectively. The user can choose whether to use adjusted or sequential sum of squares under the options button and pairwise comparisons can be chosen from the comparisons button. Pressing OK button runs the general linear model.

Reference: Minitab (2016)

Lastly, calculation of the adjusted means and their confidence intervals can be performed. After the correct model is determined ("Full" or "Reduced" model), then the adjusted LS-means[10] which correct for the bias in X between periods can be calculated. The LS-mean of each period (i.e., calibration and treatment periods in this case) is the period mean for Y adjusted to an overall common value of X. In other words, the LS-means are the calibration and treatment period regression values for the treated watershed evaluated at the mean of all the control watershed values over both time periods (e.g., mean of all the X values). Operationally, inserting the mean of all X values into the regression equations for the calibration and treatment periods will yield the LS-mean values for each period, respectively. An F-test of the adjusted LS-means then determines if there is sufficient evidence to conclude that the adjusted LS-mean for the treatment period is different from the adjusted LS-mean for the calibration period. The SAS program performs this F-test on the "Period" variable in Example 7.8-1.

[10] LS-means (least square means) are used in ANCOVA as a better comparison of average values between periods as compared to arithmetic means. LS-means are estimated values that are evaluated at the average value of the specified covariate(s) such as the control watershed values in the paired-watershed study design.

Caution must be used when interpreting the results for the comparisons of adjusted means when individual slopes are used. When the slopes are not parallel, the comparisons of adjusted means may not be the most meaningful question. One may be more interested in the behavior over the entire range of X. In this case a graphical presentation may be most appropriate.

For samples collected daily, weekly, biweekly, or monthly, autocorrelation may be significant. In these cases, autocorrelation can be addressed by using a software regression program that incorporates the autocorrelation in the error term, for example PROC AUTOREG by SAS (SAS Institute 2016d); see Example 7.8-1.

7.8.2.1.2 Multivariate ANCOVA-Paired Watershed with Explanatory Variables

Note that the above analysis employed a basic univariate ANCOVA model that included only data on the pollutant variable of interest (e.g., concentration or loads) from the control and treatment watersheds. The New York NNPSMP project demonstrated the successful use of a multivariate ANCOVA technique that included hydrologic variables (e.g., instantaneous peak flow rate, event flow volume, and average event flow rate) in the model (Bishop et al. 2005). The project found that including the flow covariates explained 80 to 90 percent of observed variability in annual and seasonal event P loads, an improvement of 16 to 50 percent versus a simpler univariate model. In addition, inclusion of covariates reduced the minimum detectable treatment effect by 11 to 53 percent versus the univariate model, a result that indicates potential cost savings through reduced sample size requirements. It is important to note that the inclusion of additional covariates (i.e., those in addition to the variable of interest in the control watershed) is prefaced upon the assumption that they are not affected by BMP implementation. In this example, testing indicated no influence of BMPs on farm runoff volume, event peak flow, or average event flow.

In the case of a paired-watershed study, explanatory variables (covariates) would be added to the statistical model. The full model which allows for different slopes for each time period and covariate is:

$$Y_{ij} = b_{0i} + \sum_{i=1}^{k} b_{1i}\left(X_{1ij}\right) + \sum_{c=2}^{d+1} b_{ci}\left(X_{cij}\right) + e_{ij}$$

Where:

Y_{ij} = the j^{th} observation for Y in period i (e.g., pollutant concentration or load from treatment watershed)
b_{0i} = the intercept (b_0) for period i
b_{1i} = the regression coefficient (b_1) of Y on X_1 for period i
b_{ci} = the regression coefficient (b_c) for covariate X_c for period i
k = number of time periods (with 'calibration' and 'treatment' periods, k=2)
X_{1ij} = the jth observation for X_1 in period i (X_1 is the pollutant concentration or load from control watershed paired with same sample time as Y_{ij})
d = number of explanatory variables in addition to the control watershed variable. For example, if only flow was used as a covariate, d=1 and the explanatory variable for flow would be X_2.
X_{cij} = the jth observation for X_c covariate in period i
e_{ij} = the residuals or experimental error for the j^{th} observation for Y in period i (V_{ij} for autocorrelated error structure)

As discussed above, a test for the homogeneity of slopes (by including interaction terms) would be performed to see if a full or reduced model is the best choice, followed by calculation of adjusted means and their confidence intervals to see if a significant difference exists between the two periods.

While the focus above has been on a basic paired-watershed study design consisting of two watersheds (control and treatment) and two periods (calibration and treatment), ANCOVA is a powerful tool that can also be applied to paired-watershed studies with multiple control and treatment watersheds and more than two periods, as well as to above-below studies that have two or more time periods.

7.8.2.1.3 Multiple Paired Watersheds

Both the Jordan Cove (CT) and Lake Champlain Basin (VT) NNMP projects included three watersheds in their paired-watershed designs. The Jordan Cove project included a previously developed drainage area as a control, and two newly developed drainage areas, one following traditional subdivision requirements and another using low-impact development BMPs (Clausen 2007). The Vermont project employed a three-way paired design including one control watershed and two treatment watersheds receiving similar BMP systems at different intensities (Meals 2001). For both studies, the two treatment watersheds were separately compared versus the control watershed using ANCOVA.

Changes versus the control watershed for the Jordan Cove project were represented by the percent change in flow, concentration, and export (Clausen 2007). These calculations were made by comparing mean predicted values (P) from the calibration regression equations to observed values (O) using the equation:

$$\%Change = \frac{(O - P)}{P} \times 100$$

Meals (2001) performed a series of analyses to examine the results of the Lake Champlain Basin study. Where full ANCOVA models were used, the calibration and treatment period regression lines intersected, suggesting, for example, that TP concentrations in one of the treatment watersheds decreased in the high range, but not in the lower range (Figure 7-24). The importance of this observation is that the higher range is where active runoff conditions occur, indicating that the BMPs may have been performing as expected.

Calculations similar to those performed for the Jordan Cove project were performed to estimate the magnitude of change (i.e., %Change), but two additional analyses were carried out to estimate this change from different perspectives:

- Breakpoint analysis for intersecting or crossed regression lines, and
- Assessment of predicted-without-treatment versus observed-with-treatment.

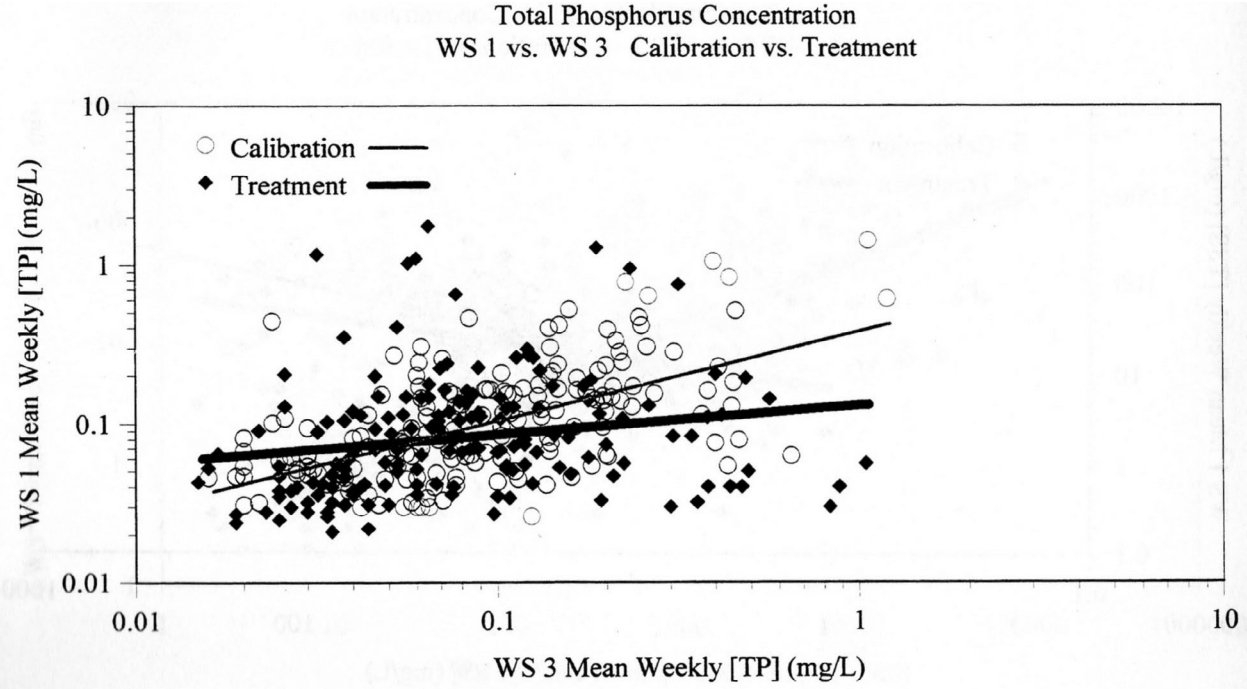


Figure 7-24. Example of intersecting regression lines (Meals 2001)

For the former analysis, the point where the regression lines crossed (the "breakpoint") was used in conjunction with the cumulative frequency of the breakpoint value in the control watershed to derive the proportion of time or conditions at which concentration or load reductions did or did not occur in the treatment watershed (Meals 2001). For example, the breakpoint in Figure 7-24 occurs at 0.055 mg/L in the control watershed (WS 3), a value for which the cumulative frequency for the entire project period was 0.32, or 32 percent. This is interpreted to mean that TP levels in the treatment watershed (WS 1) were not reduced 32 percent of the time when the concentration in the control watershed was less than 0.055 mg/L. Conversely, TP levels were reduced 68 percent of the time when control watershed concentrations exceeded 0.055 mg/L. This compares with an ANCOVA result that TP concentrations were reduced 15 percent in the treatment watershed.

The latter analysis was intended to assess the net treatment response regarding pollutant export over the full range of project conditions (Meals 2001). In this analysis, all weekly values for the treatment period in the control watershed were input to the calibration period regression for each treatment watershed to estimate what the pollutant export would have been for the hydrologic conditions of the treatment period under pre-treatment management, a what-if scenario. In other words, it is an estimate of the difference between measured loads for the treatment period and what those loads would have been if the BMPs had not been implemented.

7.8.2.1.4 Multiple Time Periods within a Paired-Watershed Study

Small watershed projects will generally have a period before BMP implementation, a period during BMP implementation, and a period after BMP implementation. The implementation and post-implementation periods are often lumped into the same period for data analysis, but this can complicate interpretation of results if the BMPs are not fully functional throughout the post-BMP period. Where feasible, it may be most appropriate to separate true implementation, and in some cases maturation of living BMPs, from post-implementation, to establish a better test of BMP or project effectiveness. There is also a very real

possibility that BMP implementation will occur in phases, creating the potential for more than two or three periods of interest. For example, in the Waukegan River NNMP project, the state Water Survey designed biotechnical and other practices to resist high velocity runoff while increasing riparian habitat for stream fisheries within the stream channel (White et al. 2011). However, as the project progressed it became clear that insufficient pool depth and the lack of pools and riffles were important impairments yet to be addressed. As a result, pool-and-riffle sequences were later added to the restoration program, creating a two-phase implementation effort. Still, however, project scientists concluded that there is a remaining need to address sewage and stormwater management problems and take steps to increase implementation of alternative conservation practices that infiltrate and treat stormwater. Were the monitoring program to be continued, these could be considered additional BMP implementation phases.

Taken to the extreme, each year could also be considered its own period or group and the groups tested for differences, but this is not recommended[11]. In some cases, BMPs may have different effects depending on the season of the year, so including a seasonal covariate(s) may be appropriate. The New York NNMP project identified four seasons that reflect seasonal variation in both source activities and hydrologic runoff processes (Bishop et al. 2005). ANCOVA was performed separately on both seasonal and full-year datasets. Despite the wide range of possibilities, time periods for the types of projects envisioned by this guidance will largely be drawn from the following set of options:

- pre-BMP or calibration,
- BMP implementation (may be subdivided by growth stage if it involves vegetative BMPs), and
- post-BMP implementation (which may include BMP implementation as well).

Where multiple phases of BMPs are to be implemented, however, there could be a separate pre-BMP implementation and post-BMP implementation for each phase. It is important to identify and plan for these phases at the beginning of the monitoring project. Adjustments may be warranted later, however, because the implementation of BMPs may be more gradual or sporadic than anticipated during the planning phases of a study, and some BMPs, like forested buffers, may take longer than expected to reach critical growth stages.

For example, in a 15-year project monitoring the effectiveness of a riparian forest buffer in an agricultural watershed, it was expected that it would take several years for the planted seedlings to have a measureable influence on water quality (Newbold et al. 2009). To account for this, the calibration period was taken to be the first five years (1992-1996) of monitoring, a period during which the seedlings became established but remained too small to affect stream nutrient concentrations. Regression analysis was used to detect gradual change and one-way ANOVA was performed on the differences between paired samples, with year treated as the main effect.

7.8.2.1.5 Other Statistical Approaches for Paired-Watershed Analyses

Paired watersheds can also be analyzed with other statistical techniques. For example, some authors have used the differences between sample pairs taken at each watershed for each sampling date (Carpenter et al. 1989; Bernstein and Zalinski 1983; MacKenzie et al. 1987; and Palmer and MacKenzie 1985) for input into t-test or intervention analysis. Hornbeck et al. (1970), Hibbert (1969), and Meals (1987) calculated a

[11] It is feasible that a 2-year study could include one year each of pre-BMP and post-BMP monitoring, but this would be highly unusual and not, in fact, recommended. A similar situation would be a 3-year study with a pre-BMP, BMP-implementation, and post-BMP year.

linear regression equation relating the observations from the two watersheds for the calibration period. Observations from the treated watershed in the treatment period were compared to predicted values from the calibration period regression. If the deviations exceeded the 95 percent confidence intervals placed about the calibration regression, the treatment was thought to be significant (Hornbeck et al. 1970).

7.8.2.2 Above/Below – Before/After

An above/below-before/after watershed design monitors a water resource (e.g., a stream) above and below the drainage area in which land treatment is applied for multiple years before and after BMP implementation (see section 2.4.2.6). Consistency of sampling regime at both stations over time is essential. Hydrologic explanatory variables (e.g., covariates) such as stream flow must also be monitored to permit correction for changes in these conditions.

7.8.2.2.1 Comparing Means and Differences between Means

Two principal approaches can be taken to statistical analysis of data from this monitoring design. Both approaches are illustrated by the projects in Examples 7.8-2–7.8-5. In the first approach, mean upstream and downstream pollutant concentrations and/or loads can be compared (e.g., with the Student's t or Wilcoxon Rank Sum tests) prior to the application of BMPs to evaluate statistically significant differences between group means. The purpose of this analysis is to confirm and quantify the pre-treatment ("before") pollutant contribution of the untreated downstream area. This analysis is then repeated for the "after" data to document the changes in pollutant contribution of the treated downstream area. Differences between upstream and downstream conditions from the before to the after condition can be evaluated simply by examining the percent reductions in concentration or load or by conducting a group means test of the differences between upstream and downstream concentrations or loads from the before to the after period. A significant decrease in this upstream/downstream difference in the "after" period, for example, would suggest a significant effect of treatment. In addition to quantitative statistical tests, it is also possible to visualize differences between above/below and before/after using comparative boxplots, bar graphs, or other graphical techniques (see section 7.3.2).

A more statistically powerful approach would be to use the paired Student's t-test to test the differences between the downstream and upstream sample values in the pre-BMP period. In the post-BMP period, a Student's t-test can be applied to the average downstream-upstream differences in the pre- vs. post-BMP periods. Other explanatory variables can be added (e.g., stream discharge) by using an ANCOVA statistical approach.

Differences between above and below stations were examined as part of the analyses performed for the Otter Creek (WI) watershed project (Stuntebeck 1995). This project also incorporated innovative sampling procedures to maximize the potential for distinguishing between upstream and downstream water quality, including programming water quality samplers to be activated by precipitation so that time-integrated samples were collected initially before stage-triggered samples were collected. This allowed sampling of barnyard runoff in the stream before stage increased, thereby isolating runoff from sources upstream. It also allowed sampling during small storms where barnyard runoff occurred in the absence of substantial upstream contributions. In addition, investigators collected concurrent samples from both the above and below sites via computer linkage to aid data interpretation. Paired Student's t-tests were used to determine that the pre-BMP average of the differences between downstream and upstream event-mean concentrations was different from zero at the 95 percent confidence level. An MDC analysis revealed that the average downstream post-BMP event-mean concentrations of TP would need to decrease by at least

50 percent for the change to be considered statistically significant at the 95 percent confidence level. In the final analysis, the Hodges-Lehmann estimator was used to determine that the barnyard BMP system at Otter Creek reduced loads of suspended solids by 85 percent, TP by 85 percent, ammonia by 94 percent, BOD by 83 percent, and microbial loads of fecal coliform bacteria by 81 percent (Stuntebeck and Bannerman 1998; See Example 7.7-2). The nonparametric Hodges-Lehmann estimator is the median of all possible pairwise differences between pre- and post-BMP barnyard loads (see section 4.5.3 of the 1997 guidance (USEPA 1997b) for a discussion of the Hodges-Lehmann estimator). This median difference was divided by the pre-BMP median load for each constituent to determine percentage load reductions.

7.8.2.2.2 ANCOVA

A second approach for analysis of the above/below-before/after design involves the application of ANCOVA. The statistical analysis approach is the same as with the paired-watershed study (see section 7.8.2.1) In this case, a significant linear regression relationship for a water quality variable (e.g., weekly mean total P concentration, weekly suspended sediment load) between the upstream and downstream stations is obtained during the "before" period. The upstream station is considered to be the "control" watershed. This regression relationship is then compared to a similar relationship during the "after" period and significant difference between the two regression models indicates the effect of treatment. Note that the analysis can include explanatory variables (e.g., covariates) like precipitation or flow in a multiple regression model that may explain more of the variability in the water quality variable than a simpler model.

Example 7.8-2. Above/Below-Before/After Design - Long Creek, NC NNPSMP

A number of successful projects have used multiple approaches to analyzing their data. For example, data from an above/below-before/after study of livestock exclusion as part of the Long Creek (NC) NNPSMP project were first log-transformed and then analyzed using t-tests, two-way ANOVA, and ANCOVA (Line et al. 2000). While the specific questions addressed by each method differ somewhat, the results all supported the conclusion that livestock exclusion and establishment of riparian vegetation reduced mean weekly loads of TSS, TKN, and TP.

Example 7.8-3. Above/Below-Before/After Design (biological data) - Waukegan River, IL NNPSMP

The Waukegan River (IL) NNPSMP project illustrates the application of the above/below design for biological monitoring. In this project, the South Branch was divided into an upstream untreated reference site designated as station S2 and a severely eroding downstream treated area designated as station S1 (Spooner et al. 2011b). At each location fish, macoinvertebrates, and habitat were sampled during the spring, summer, and fall seasons. Sampling was also conducted at stations N1 and N2 on the North Branch for reference. Qualitative analysis of biological data collected through 2006 indicated that the number of fish species and abundance in the South Branch had improved after the construction of lunkers and rock grade control structures. The IBI rose sharply from a limited aquatic resource into the moderate category after construction. Sites on both the South and North Branches where lunkers and Newbury Weirs were applied averaged higher IBI scores and fish population with more fish species than the untreated control at S2 or the N2 bank armored site from 1996 through 2006.

Example 7.8-4. Above/Below-Before/After Design with Flow as an Explanatory Variable - Pequea and Mill Creek Watershed, PA NNPSMP

A Pennsylvania study of the effects of streambank fencing on surface-water quality, near-stream ground water, and benthic macroinvertebrates employed both a paired-watershed and above/below-before/after design (Galeone et al. 2006). Data for this Section 319 NNMPMS project were collected from 1993 to 2001, with the calibration period from October 1993 through mid-July 1997. Streambank fencing was installed from May 1997 through July 1997. The above/below-before/after design featured two sites above fence installation (T-3 and T-4) and two sites located to show the effects of fencing (T-1 and T-2); T1 and T2 were paired with T3 and T4, respectively, for data analysis. Both low-flow and storm-flow samples were collected and analyzed for nutrients, suspended sediment, and fecal streptococcus (only low-flow samples). Explanatory data collected during the study included precipitation, inorganic and organic nutrient applications, and the number of cows.

Figure 7-25 illustrates the major data preparation steps and statistical procedures used by the project to analyze the chemical/physical data. Low-flow, storm-flow, pre-treatment, and post-treatment data were separated as a preliminary step. Concentrations were flow adjusted using a LOcally WEighted Scatterplot Smoothing (LOWESS) procedure (Helsel and Hirsch 2002). Statistical tests were performed on both original and flow-weighted data to determine if factoring out the variability caused by flow affected the results.

After the above steps were completed, the project applied the nonparametric rank-sum test (see Mann-Whitney test and Wilcoxon Rank Sum test on pages 4-50 of the 1997 guidance, USEPA 1997b) to determine if data for any one site significantly changed from the pre-treatment to the post-treatment period. In addition, the nonparametric Kruskal-Wallis test (see pages 4-56 of the 1997 guidance) was carried out to determine if there were significant differences between any of the sites, considering pre-treatment and post-treatment data separately. Where significant differences were found, the Tukey multiple-comparison test (see Multiple Comparisons on pages 4-63 of the 1997 guidance) was used to identify which sites were significantly different. The nonparametric signed-rank test (see Wilcoxon Signed Ranks test on pages 4-42 of the 1997 guidance) was used to determine if there were significant differences (i.e., not zero) between paired observations (e.g., matched samples from above/below sites). Finally, ANCOVA (see section 4.8 of the 1997 guidance and section 7.8.2.1 for detailed discussions of the ANCOVA procedure) was applied to determine the effects of streambank fencing using a procedure highlighted by Grabow et al. (1999). ANCOVA was performed on concentrations and loads for both low-flow and storm-flow samples. Loads were analyzed in two ways, as actual measured loads and as weighted loads adjusted with a factor determined by dividing the annual mean discharge for each water year by the mean discharge for the entire period for each station.

The procedures used by Galeone et al. (2006) demonstrated improvements relative to control or untreated sites in surface-water quality (nutrients and suspended sediment) during the post-treatment period at T-1, but T-2 showed reductions only in suspended sediment. N species at T-1 were reduced by 18 percent (dissolved nitrate) to 36 percent (dissolved ammonia); yields of total P dropped by 14 percent. Conversely, T-2 had increases in N species of 10 percent (dissolved ammonia) to 43 percent (total ammonia plus organic N), and a 51-percent increase in total P load. The average reduction in suspended-sediment load for the treated sites was about 40 percent. Two factors were evident at T-2 that helped to overshadow any positive effects of fencing on nutrient yields. One was the increased concentration of dissolved P in shallow ground water (also monitored). In addition, cattle excretions at the low-cost, in-stream cattle crossings appeared to increase concentrations of dissolved ammonia plus organic N and dissolved P. See chapter 3 Case Study #1 for a discussion of how the benthic macroinvertebrate data were analyzed.

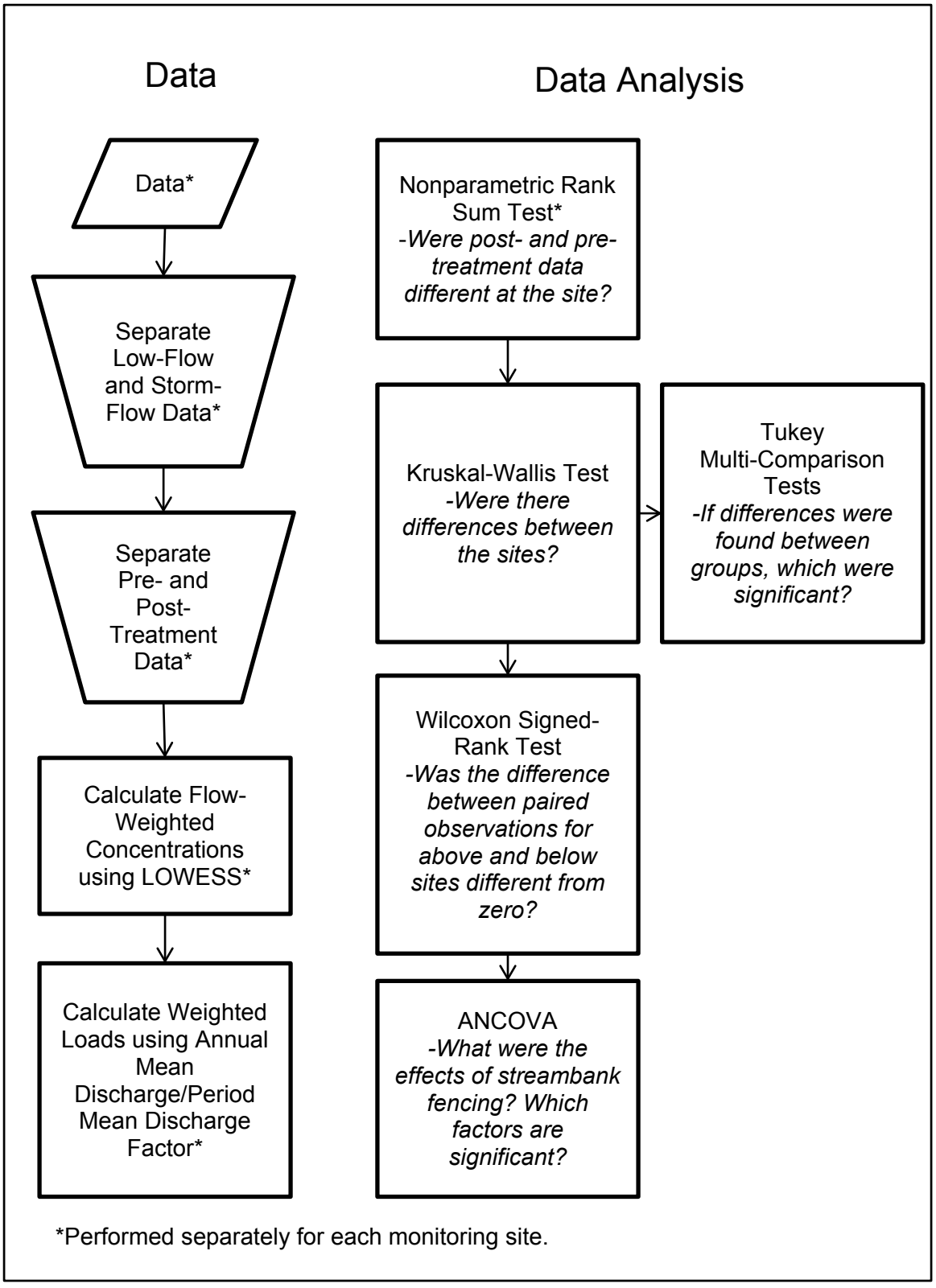


Figure 7-25. Basic data preparation and analysis procedure for above/below-before/after study in Pennsylvania (Galeone et al. 2006)

Example 7.8-5. Above/Below-Before/After Design with Upstream Concentration and Flow as Explanatory Variables - Walnut Creek, IA NNPSMP

In some cases, projects are forced to develop alternative plans for data analysis due to unforeseen circumstances they cannot control. The Walnut Creek (IA) NNMP project, for example, began as a ten-year paired-watershed study that also included an above/below-before/after design and three subwatershed single-station designs within each of the paired watersheds (Schilling and Spooner 2006). The primary purpose of the project was to evaluate the response of stream nitrate concentrations to conversion of row crops to native prairie. The normal approach of analyzing project data (both for the paired-watershed and above/below-before/after designs) using ANCOVA was compromised by two facts: prairie conversion began before the calibration period was completed, and conversion to prairie was gradual instead of rapid. Based on the guidelines and experiences of others (Spooner et al. 1987, Grabow et al. 1998 and 1999), multiple linear regression analysis on all ten monitoring sites was selected as an alternative approach to project evaluation (see Example 7.8-7 for the general form of equation used). Treatment in this case was modeled as time with covariates such as upstream concentration used to factor out hydrologic variability. For the downstream site in the treatment watershed, a model using month (for seasonality), upstream nitrate concentration, and downstream nitrate concentration in the control watershed provided the best fit to the data. For all other sites, month and the log of baseflow discharge from the same or a different site were used as covariates in the best-fit regression model. All tests resulted in detection of significant trends in nitrate concentrations, with the downstream treatment site trend indicating nitrate reductions due to conversion to prairie (the treatment). A negative coefficient on the time variable (-0.119 mg $l^{-1}yr^{-1}$) indicated a nitrate reduction of 1.2 mg l^{-1} over 10 years at this site. It was also found that in the control site, where land was unexpectedly converted from grassland to row crops, nitrate concentrations increased during the project period.

If the errors (e.g., residuals) in the statistical models are autocorrelated, a statistical software procedure should be used that incorporates the autocorrelation structure into the model. For example, PROC AUTOREG of the SAS software (SAS Institute 2010) is useful with autoregressive autocorrelation typical of weekly, biweekly, and monthly series. Alternatively, a correction of the standard deviation of the slope estimate and revised confidence intervals can be used with the correction given in section 7.3.6.

It should be cautioned that changes in pollutant concentrations or loads measured at a downstream station (either before or after land treatment) versus upstream may be difficult to detect if incoming concentrations or loads at the upstream station are high and the contribution of the additional area draining to the downstream station is small. Conversely, if the upstream contribution is very low compared to that of the treated area, a change or difference due to treatment may be difficult to attribute to BMPs because of dilution. If the upstream pollutant inputs do not respond similarly to hydraulic changes (e.g., rainfall), then the design effectively becomes a single watershed design. The Walnut Creek (IA) NNPSMP project provides an example of the former case where annual mean nitrate concentrations ranged from 10.0 to 12.7 mg/L at the upstream site and 6.8 to 9.5 mg/L at the site below the treatment area (Schilling and Spooner 2006). The treatment in this case was conversion of row crops to native prairie, and the study design (paired watersheds and above/below-before/after) was compromised by the fact that land conversion began before pre-treatment conditions could be established. See Example 7.8-5 for a discussion of how data from this project were analyzed using multiple linear regression, a technique typically applied to single watershed trend designs.

7.8.2.3 Nested Watershed

As described in section 2.4.2.3, it is preferred that the nested subwatershed is used as the control watershed[12] and is located above the remainder of the watershed where treatment occurs (Hewlett and Pienaar 1973). However, a valid nested design can also entail the treatment watershed in a small headwater subbasin; the control being the much larger watershed outlet. This design requires calibration (before) and treatment (after) periods similar to the paired-watershed design.

Analysis of data from a nested watershed design can be done using the same ANCOVA procedure described in section 7.8.2.1 for the paired-watershed design. In the case of nested watersheds, the paired data represent observations collected on the same date, time period, or storm at both the nested and main watershed stations. As noted above, data from the nested watershed should represent the control watershed, while data from the main watershed outlet represent the treatment watershed.

7.8.2.4 Single Watershed Trend Station

As noted in section 2.4.2.5, monitoring at a single watershed outlet is not a strong design for documenting the effectiveness of watershed land treatment on water quality. Without the ability to control for the effects of varying weather and hydrology, it is difficult to attribute any observed changes in water quality to the land treatment program. However, because the coupling of budget limitations and accountability requirements often leads to single-station designs, the unfortunate fact that some paired-watershed and other superior designs fail due to unforeseen circumstances, and the simple reality that some NPS watershed programs must rely on watershed outlet monitoring conducted by another party (e.g., a state long-term surveillance program or a USGS network station), it is useful to discuss how best to analyze data from such stations to assess the effects of a watershed project. In addition, experience has shown that projects with failed paired-watershed or above/below-before/after designs may resort to trend analysis as the best option for analyzing project data (see Example 7.8-6).

Long-term water quality data may show a *monotonic* trend (a continuous change, consistent in direction, either increasing or decreasing) or a *step* trend (an abrupt shift up or down). Trend analysis may be the best — or perhaps only — approach to documenting response to treatment in situations where water quality data are collected only at a single watershed outlet station or where land treatment was widespread, gradual, and inadequately documented. Data from long-term, fixed-station monitoring programs where gradual responses such as those due to incremental BMP implementation or continual urbanization are likely to occur are more aptly evaluated with monotonic trend analyses that correlate the response variable (i.e., pollutant concentration or load) with time or other independent variables. These types of analyses are useful in situations where vegetative BMPs like the riparian buffers implemented in the Stroud Preserve NNPSMP project (Newbold et al. 2008) must mature, resulting in gradual effects expressed over time. Analysis of step trends, on the other hand, is most appropriate when the change in response to BMP implementation is rapid and abrupt (e.g., when a municipal stormwater management regulation is enforced) and the timing of that change is known and well-documented. Biological data can also be evaluated with either monotonic or step-trend tests. A potential limitation is that most biological programs will only sample once a year and the time to acquire sufficient samples to detect a meaningful trend might be longer than what is practical.

[12] A reverse situation, where the downstream subwatershed area is the control is possible in theory, but all effort would need to be made to ensure that upstream contributions to constituents measured at the downstream control area are minimized.

Example 7.8-6. Single Trend Watershed with Covariates - Sycamore Creek, MI NNPSMP

This project planned a paired-watershed study with two treatment watersheds (Willow Creek and Marshall Drain) and one control watershed (Haines Drain), but implementation of no-till and continuous cover in the control watershed compromised the study (Suppnick 1999). Each watershed was then analyzed independently, with regression analysis ultimately successful in linking reductions in TSS (95 percent confidence level) and TP (90 percent confidence level) loads to the percentage of land in no-till in the Willow Creek watershed (Grabow 1999, Suppnick 1999). Following is a summary of the steps taken to establish the TSS relationship for Willow Creek (Grabow 1999):

1. Regression analysis on sediment yield versus storm discharge and/or peak flow to reduce the analysis to water quality change over time independent of hydrologic variability. All variables were log-transformed.
2. Two methods were then used to answer the question of whether there was a water quality trend over time.
    a. Regression equation incorporating elapsed time and explanatory variables. This addresses the question of whether there has been a change in water quality over time while simultaneously accounting for hydrologic variability.
    b. Regression of residuals[1] from regression on the water quality variable and explanatory variables versus elapsed time. This addresses the question of whether there has been a water quality change over time after adjusting for hydrologic variability.
3. Correlation of land use change to water quality change via multiple linear regression analysis. Terms incorporated in the regression model were percent of land in no-till, percent of land in continuous cover, storm discharge, and peak flow.

Step 1 yielded correlation between TSS load (kg/storm) and both storm discharge (mm) and peak flow (liters/second). Discharge and peak flow were tested for collinearity which was found to be not an issue (see Box 7.8-1).

Step 2 analyses indicated statistically significant trends in TSS and TP in Willow Creek watershed. Method “a” used the following basic equation:

$$log[TSS] = \beta_0 + \beta_1 log[Q] + \beta_2 log\left[Q_p\right] + \beta_3 t$$

Where TSS is the TSS storm load (kg), Q is the total storm discharge, Q_p is the peak stream discharge, t is elapsed time in days, and the β terms are regression parameter estimates. A significant negative value for β_3 indicated a reduction in TSS load over time. Insertion of average log values of total storm discharge and peak discharge, and setting the beginning and ending days (1 and 2,629 for t_{begin} and t_{end} in this case) would then yield the average change in loadings from the first to last data of data collection.

[1]Residuals are the differences between actual and predicted values: Actual-Predicted.

Example 7.8-6. (continued)

Sycamore Creek, MI NNPSMP

Method "b" of Step 2 used the following equation:

$$TSSres = \beta_0 + \beta_1 t$$

Where TSSres is the residuals (log kg/storm) from the regression in Step 1 and t is again elapsed time. In this approach, a statistically significant value for β_1 would indicate a change in the relationship between TSS and the explanatory variables (total and peak discharge), suggesting an impact due to land use change. The value $\beta_1 \times t_{end}$ would then estimate the change in loading (in log units) over the data collection period. The average change in loading is determined by then plugging the average values for log [Q] and log[Q_p] into the regression equation used in Step 1.

In this case, method "a" indicated a 60 percent reduction in TSS load, whereas method "b" estimated a 59 percent reduction.

With a statistically significant reduction in TSS load now documented, Step 3 explored the linkage between that reduction and land use change by adding the percentage of land in no-till (NoTill) and the percentage of land in continuous cover (ContCov) as additional terms in the multiple linear regression used for method "a" in Step 2. Statistically significant regression parameters β_3 and/or β_4 in the following equation would indicate correlation between log[TSS] and the percentage of land in no-till and/or continuous cover.

$$log[TSS] = \beta_0 + \beta_1 log[Q] + \beta_2 log[Q_p] + \beta_3 NoTill + \beta_4 ContCov + \beta_5 t$$

A statistically significant value of -0.01969 was found for β_3, but β_4 was insignificant, suggesting that for every percent increase in the percentage of land under no-till, the TSS load (as log kg) would decrease by 0.01969 log units. Regression estimates based on average storm discharge and peak flow were then used in conjunction with first-year and last-year values of no-till percentages to estimate a TSS load reduction of 52 percent, with a 95 percent confidence interval of 18-72 percent. This agreed well with the estimates of 59 and 60 percent reduction from Step 2.

Combining the results from the above analyses by Grabow (1999) with additional project information, it was concluded that it is very likely that streambank stabilization also contributed to the reduction in TSS observed in Willow Creek (Suppnick 1999).

Box 7.8-1. Collinearity

What is Collinearity?

Collinearity in multiple regression analysis occurs when there is a linear relationship between two explanatory (x) variables. Although this does not impact the reliability of the overall model, it does create great uncertainty regarding the model coefficients. There are ways to address collinearity, including recognizing the ambiguity in the interpretation of regression coefficients (USF n.d.) or simply removing one of the variables from the regression model (Martz 2013).

Various statistics programs have tests for collinearity (or multicollinearity), including the Variance Inflation Factor (VIF), Tolerance (1/VIF), and the Condition Index (SAS 2016a and 2016c, USF n.d.). Guidelines vary, but VIF values greater than 5 to 10, Tolerance values close to 0, and Condition Index values greater than 15 to 30 indicate problems with collinearity. See Belsley et al. (1980) for additional details.

Several statistical trend analysis techniques will be mentioned in this section; the topic of trend analysis is covered in more detail in *Tech Notes 6: Statistical Analysis for Monotonic Trends* (Meals et al. 2011). Before proceeding, it is important to recognize some limitations of trend analysis. First, trend analysis is most effective with long periods of record; general guidelines are ≥5 years of monthly data for monotonic trends and ≥2 years of monthly data before and after a step trend (Hirsch 1988). Short monitoring periods and small sample sizes make documentation of trends difficult, and it must be recognized that - especially over the short term - some increasing or decreasing patterns in water quality are not trends. A snapshot of water quality data over a few months may suggest a trend, but examination of a full year may show this "trend" to be part of a regular cycle associated with temperature, precipitation, or cultural practices. Autocorrelation may also be mistaken for a trend, especially over a short time period. Changes in sampling schedules, field methods, or laboratory practices can cause shifts in data that could be erroneously interpreted as step trends.

Perhaps most importantly, statistical trend analysis can help to identify trends and estimate the rate of change, but will not provide much insight into attributing a trend to a particular cause (e.g., land treatment). Interpreting the cause of a trend requires knowledge of the watershed and a deliberate study design (see section 7.8.1).

Before proceeding to numerical analysis, it is useful to examine time series plots for visual evidence of a trend. Visualization of trends in noisy data can be clarified by various data smoothing techniques. Plotting moving averages or medians, for example, instead of raw data points, reduces apparent variation and may reveal general tendencies. Spreadsheets can display a moving-average trend line in time-series scatterplots with adjustable averaging periods. A more complex smoothing algorithm, such as *LOWESS* (*LO*cally *We*ighted *S*catterplot *S*moothing), can reveal patterns in very large datasets that would be difficult to resolve by eye (see Helsel and Hirsch 2002). Most pollutant concentrations and loads in surface waters show strong seasonal patterns. Seasonal variations in precipitation and flow are often main drivers of these patterns, but seasonal changes in land management and use may also play a role. See section 4.3 of the 1997 guidance (USEPA 1997b) for additional information on seasonality.

Some techniques to address seasonality beyond controlling for the effects of flow covariates are often necessary for water quality trend analysis. For example, the relationship between concentration and discharge may not be consistent over time, perhaps due to seasonal variations in BMP implementation. The relationship (or slope) can be allowed to change between time periods by the use of interaction terms between the time periods and discharge in an analysis of covariance (ANCOVA) statistical model. An alternative that might develop more traction with experiences is to consider a weighted regressions on time, discharge and season (WRTDS) proposed by Hirsch et al. (2010) (see section 7.9.2 for more information on WRTDS).

When multiple explanatory variables are included in the trend models, it is common that these variables will be related to each other (collinearity) and/or a few data points may have a lot of 'influence' over the regression results (Belsley et al. 1980). Regression analysis performed with various software programs will provide leverage plots as part of the output to help identify these data features.

7.8.2.4.1 Monotonic Trends

Table 7-8 lists some monotonic trend tests available for different circumstances, including adjustments for a covariate and the presence of seasonality. The tests are further divided into parametric, nonparametric, and mixed types. Regression tests require that the expected value of the dependent variable is a linear function of each independent variable, the effects of the independent variables are additive, the errors in

the model are independent (e.g., no correlation between consecutive errors in the case of time series data), and the errors exhibit both normality and constant variance. Nonparametric tests require only constant variance and independence. Parametric trend tests (see Examples 7.8-7 and 7.8-8) are considered more powerful and/or sensitive to detect significant trends than are nonparametric tests (see Example 7.8-9), especially with a small sample number. However, unless the assumptions for parametric statistics are met, it is generally advisable to use a nonparametric test (Lettenmaier 1976, Hirsch et al. 1991, Thas et al. 1998).

Table 7-8. Classification of tests for monotonic (nonparametric) or linear (parametric) trend (adapted from Helsel and Hirsch 2002)

	Type of Test	Not Adjusted for covariate (X)	Adjusted for covariate (X)
No Seasonality	Parametric	Linear regression of Y on t	Multiple linear regression of Y on X and t
	Mixed	-	Mann-Kendall on residuals from regression of Y on X
	Nonparametric	Mann-Kendall	Mann-Kendall on residuals from LOWESS of Y on X
Seasonality	Parametric	Linear regression of Y on t and periodic functions or indicator X's for months	Multiple linear regression of Y on X, t, and periodic functions or indicator X's for months
	Mixed	Regression of deseasonalized Y on t	Seasonal Kendall on residuals from regression of Y on X
	Nonparametric	Seasonal Kendall on Y	Seasonal Kendall on residuals from LOWESS of Y on X
Other Explanatory variables or covariates (e.g., stream discharge)	Parametric	Linear regression of Y on t and covariates (X)	Multiple linear regression of Y on t, X covariates
	Mixed	Regression of deseasonalized Y on X	Seasonal Kendall on residuals from regression of Y on X
	Nonparametric	Seasonal Kendall on Y	Seasonal Kendall on residuals from LOWESS of Y on X

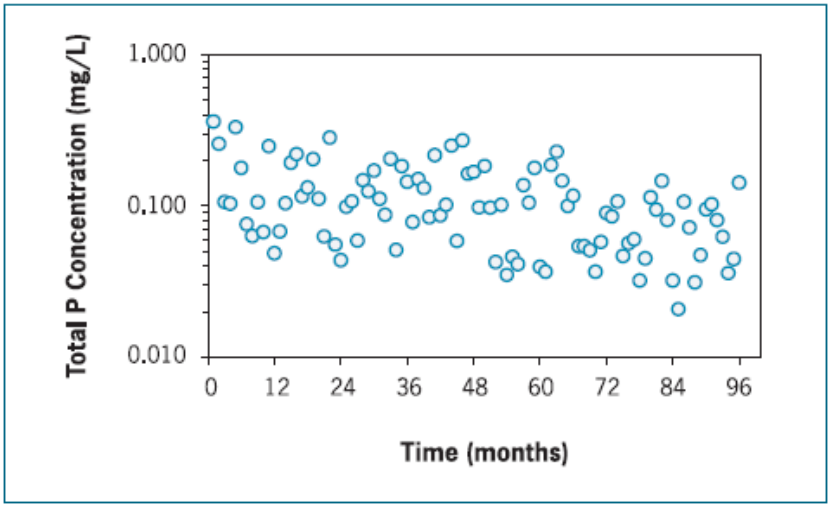
Y = dependent variable of interest; X = covariate; t = time

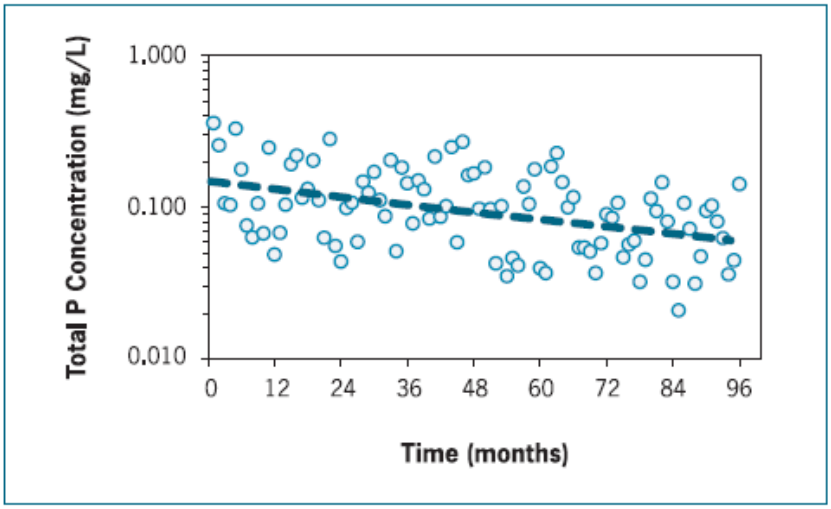
Refer to *Tech Notes 6: Statistical Analysis for Monotonic Trends* (Meals et al. 2011) for details on the tests listed in Table 7.8-1. Chapter 4 (pages 4-86 through 4-89) of the 1997 guidance (USEPA 1997b) also discusses the computation of Mann-Kendall and Seasonal Kendall statistics.

If the trend model has autocorrelated errors, a statistical model that incorporates the autoregessive errors should be employed. Alternatively, a correction of the standard error of the slope that is given in section 7.3.6 can be used to calculate the correct confidence interval of the slope on t (time, date) to determine if it is significantly different from zero (e.g., evidence of a trend over time) in the pollutant concentration or load.

Example 7.8-7. Simple Linear Regression - Samsonville Brook in Vermont

- Eight years of monthly TP concentration data from Samsonville Brook in Vermont
- Data satisfy assumptions for regression after log transformation: normal distribution, constant variance, independence (low autocorrelation)





Simple linear regression (using Excel® or any basic statistical package)

Log[TP] = -0.8285-0.00414(Time)
r^2 = 0.18, F = 21.268 P ≤ 0.001

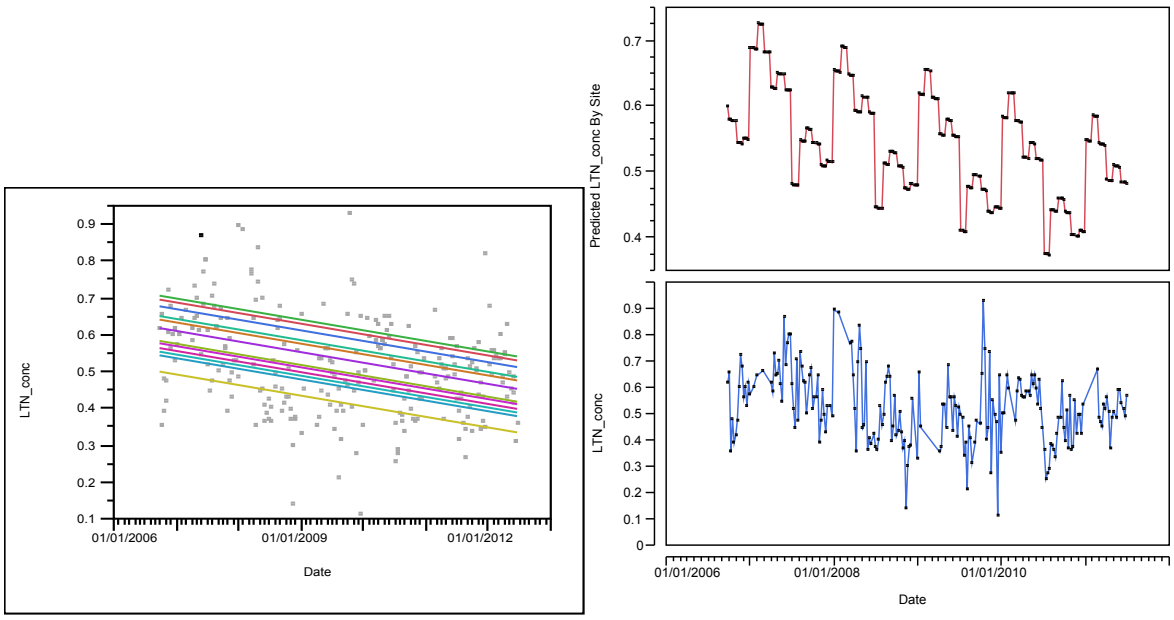
Rate of change:
Slope of log-transformed date = -0.00414
$(10^{-0.00414} - 1) \times 100$ = -0.95%/month or about -11%/year

This result suggests that TP concentrations have decreased significantly over the period at a rate of approximately 11 percent per year.

Note: Data used in this example are taken from the Vermont NNMP project, *Lake Champlain Basin agricultural watersheds section 319 national monitoring program project, 1993 – 2001* (Meals 2001).

Example 7.8-8. Linear Regression with Monthly Seasons as a Covariate - Corsica River, MD NNPSMP

A significant trend was detected in a small watershed within the Corsica River Basin, Maryland, using times series analysis that adjusted for autocorrelation as well as monthly (seasonal) differences for log transformed, flow-weighted total nitrogen (TN) concentrations. In this example, monthly indicator variables were used to adjust for seasonality in an ANOVA regression model. See section 7.3.6 for details on adjustments for autocorrelation and seasonality.



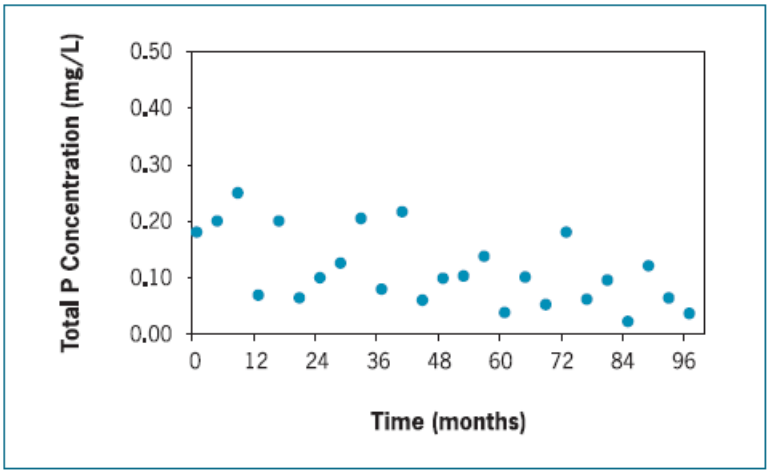
By addressing seasonality in the regression model with monthly indicator variables, most of the regression degrees of freedom were preserved, a more powerful approach than if each month was evaluated separately. Each line in the plot on the left represents the trend line (log transformed, ffflow-weighted TN concentration) for a single month (i.e., January, February … December). The trend slopes for each month were assumed to be the same, but the intercept was allowed to vary, enabling the differences in concentration due to season to be removed from the test for trends and therefore making it easier to isolate and detect trends due to other factors (e.g., BMPs).

The bottom right graph shows the raw data. The noise due to seasonal differences and other factors makes it difficult to pick out any trends. The top right graph shows the predicted value from the seasonal regression model with the indicator variables. A downward trend is apparent and it is also clear from this graph that the highest TN concentration is found in February, followed by January, March, May, April, June, Sept, August, October, November, December, and July (lowest).

Example 7.8-9. Mann-Kendall Procedure – Single Trend Watershed - Samsonville Brook in Vermont.

The data from Samsonville Brook in Vermont:

- Eight years of quarterly mean TP concentration data
- Data satisfy assumptions for constant variance and independence, but are not normally distributed without transformation



Month (n=25)	[TP] mg/L
1	**0.180**
5	**0.200**
9	**0.250**
13	**0.068**
17	**0.201**
21	**0.063**
25	**0.099**
29	**0.125**
33	**0.205**
37	**0.078**
41	**0.216**
45	**0.059**
49	**0.098**
53	**0.102**
57	**0.137**
61	**0.037**
65	**0.100**
69	**0.051**
73	**0.180**
77	**0.060**
81	**0.095**
85	**0.021**
89	**0.120**
93	**0.063**
97	**0.035**

The Mann-Kendall trend test for this example may be evaluated in two ways. First, in a manual calculation, use the formulas below. The value of S (sum of the signs of differences between all combinations of observations) can be determined either manually or by using a spreadsheet to compare combinations, create dummy variables (-1, 0, and +1), and sum for S.

Mann-Kendall $S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign\left(y_j - y_i\right) = -106$

$$\tau = \frac{S}{n(n-1)/2} = \frac{-106}{300} = -0.353$$ **(decreasing trend)**

Calculating Z_s as (S±1)/σ_s where

$$\sigma_s = \sqrt{(\frac{n}{18}) \times (n-1) \times (2n+5)} = 42.817$$

$$Z = \frac{-105}{42.817} = -2.454$$ **(USEPA 1997a)**

This Z statistic is significant at P=0.014, indicating a significant trend, i.e., we are 98.6 percent confident there is a decreasing trend in TP. See USEPA (1997a) for the calculation of σ_s when there are ties among the data.

To estimate the rate of change, use the Sen slope estimator:

$\beta_1 = median\left(\frac{y_j - y_i}{x_j - x_i}\right)$ **211 individual slopes -000945 to +0.00766**

Median slope = -0.0011 mg/L/month = -0.013 mg/L/yr

This result suggests that TP concentration decreased significantly over the period at a rate of about 0.013 mg/L/yr.

Note: Data used in this example are taken from the Vermont NNMP project, Lake Champlain Basin agricultural watersheds section 319 national monitoring program project, 1993 – 2001 (Meals 2001).

7.8.2.4.2 Step Trends

Monotonic trend analysis may not be appropriate for all situations. Other statistical tests for discrete changes (step trends) should be applied where a known discrete event (like BMP implementation over a short period) has occurred. Testing for differences between the "before" and "after" conditions is done using two-sample procedures such as the Student's t test and ANCOVA (parametric techniques with and without covariates) and nonparametric alternatives such as the rank-sum test, Mann-Whitney test, and the Hodges-Lehmann estimator of step trend magnitude (Helsel and Hirsch 2002, Walker 1994). Application of the Mann-Whitney/Wilcoxon's rank sum test and the Hodges-Lehmann estimator are illustrated in sections 4.5.2 and 4.5.3, respectively, of the 1997 guidance (USEPA 1997b). A key principle in step trend analysis is that the hypothesized timing of the step change must be selected in advance (i.e., define the pre- and post- periods before conducting statistical tests). Knowledge of watershed management activities and examination of data plots will be helpful in identifying a potential step in time.

For example, the Mann-Whitney test was used to associate changes in P management practices with a decrease in annual median soluble reactive P concentration from a 9-ha grassland catchment in Northern Ireland (Smith et al. 2003). Weekly samples were collected from 1989 through 2000, with the change in P management instituted in 1998. A comparison of data from 1997 with data from 2000 indicated that the change from whole-farm to site-specific P management reduced SRP concentrations significantly.

If the trend model has autocorrelated errors, a statistical model that incorporates the autoregessive errors should be employed. Alternatively, a correction of the standard error of the slope that is given in section 7.3.6 can be used to calculate the correct confidence interval of the step change (difference) between time periods to determine if it is significantly different from zero (e.g., evidence of a step change) in the pollutant concentration or load.

7.8.2.5 Multiple Watersheds

In the simplest case of a multiple watershed design, where monitored watersheds fall into two groups, treated and untreated, data may be analyzed by Student's t test or the non-parametric Wilcoxon Rank-Sum test. Such an analysis would test the (null) hypothesis that there was no significant difference in mean pollutant concentration or load between the treated and untreated watershed groups. Where monitored watersheds occur in more than two groups (e.g., untreated, treatment A, treatment B, etc.), significant differences in group means can be evaluated using ANOVA or the Kruskal-Wallis test. For example, Clausen and Brooks (1983) assessed mining impacts on MN peat lands using a multiple watershed design. Results – analyzed by ANOVA for normally distributed variables and otherwise by nonparametric Kruskal-Wallis and Chi-Square tests – documented significant impacts of peat mining on water quality. Lewis (2006) describes application of fixed-effect and mixed-effect (i.e., includes random effects) regression models to multiple-watershed studies involving logging. A 13-watershed study involving 3 controls, 5 clear-cuts, and 5 partial cuts was carried out over sixteen years with monitoring of storm volumes during four years before cutting, three years of logging, and nine years[13] of post-logging. The best fit was obtained when the proportion harvested, antecedent wetness, regrowth, and spatial autocorrelation were all incorporated into the model. This study design and analytic approach allows the prediction of streamflow response to harvesting in other watersheds considered part of the same population of watersheds included in the study.

[13] Three years of post-cut monitoring at seven stations and nine years at six stations.

7.8.3 Linking Water Quality Trends to Land Treatment

A central objective of many NPS watershed projects is to determine not only if water quality changes can be documented but also if water quality changes can be associated with changes in land treatment. Such documentation is necessary to help build an information base to support continued improvement in preventing and solving water quality problems. It is also needed in many cases to justify expenditure on clean-up efforts.

For a range of reasons, including budgets and programmatic constraints, watershed project monitoring efforts are almost never designed to satisfy the rigorous criteria for establishing true cause and effect relationships (see Box 7.8-2). Rather, project effectiveness monitoring designs are generally intended to measure improvements in water quality and, hopefully, relate that improvement to activities undertaken to influence water quality. A plausible argument that what was done on the ground improved water quality is often the best that can be hoped for and that is usually not a simple task at the watershed level. The ability to control for factors other than land treatment (e.g., weather, hydrology, land use change) is a key ingredient in making such a plausible argument.

Box 7.8-2. Cause-effect requirements (Mosteller and Tukey 1977).

A cause-effect relationship must satisfy the following criteria:

- *Consistency* - the direction and degree of the relationship between the measured variables (such as TP loads and acres treated with nutrient management) holds in each data set.
- *Responsiveness* - as one variable changes in a known manner, the other variable changes similarly. For example, as the amount of land treatment increases, further reduction of pollutant delivery to the water resource is documented.
- *Mechanistic* - the observed water quality change is that which is expected based on the known or hypothesized physical processes involved in the installed BMPs.

Control refers to eliminating or accounting for all factors that may affect the response to the treatment so that the treatment effect can be isolated. In a laboratory experiment, control is usually obtained by subjecting the entire system to the same conditions, varying only the treatment variable and selecting replicates at random to assure that unmeasured sources of variability do not affect the interpretation. Such an approach is rarely if ever possible for monitoring projects in watersheds dominated by nonpoint sources. Instead, we hope to show an association between change in water quality and change in land use or management by selecting a project design that includes monitoring for important explanatory variables (covariates) and applying appropriate statistical tools to include and adjust for these covariates in the analysis. By factoring explanatory variables into trend analyses, we remove some of the noise in the data to uncover water quality trends that are closer to those that would have been measured had no changes in climatic or other explanatory variables occurred over time. When performing statistical analyses with both water quality and land treatment data, it is important to note that it is not necessary to summarize the water quality data on the same (less frequent) time scale as the land treatment data. Rather, land treatment data can be incorporated within a trend analysis, for example, as repeating explanatory variables. That is, the values of land treatment and land use are treated as X variables in a statistical trend model. Because land management data are usually taken less frequently than water quality data, the land management information for a given X variable can be repeated for the time range of water quality samples that is represented by the land management value.

Although association by itself is not sufficient to infer causal relationships, it can contribute to a plausible argument that pollution control activities have resulted in environmental improvement. Thus, knowledge

of land management and land treatment in the watershed is essential to demonstrate an association between changes on the land and changes in water quality. For example, section 7.8.2.2 described how the Sycamore Creek (MI) NNMP project used multiple linear regression to link log[TSS] load to the percentage of land under no-till cropping (Grabow 1999). Additional explanatory variables included the logs of total storm discharge and peak stream discharge.

Data on both the temporal progress and spatial extent of land treatment and other watershed land use/management activities should be used to build an association between land treatment and observed water quality. For example, on a temporal scale, land treatment and management data can be analyzed and linked to water quality in these ways:

Define monitoring periods: Documentation of BMP implementation can be used to define critical project periods, like pre- and post-treatment periods in before/after and paired-watershed designs or to establish a hypothesis on the timing of a step trend.

Explain observed water quality: Knowledge of not only BMP implementation history but also dates of tillage, manure or agrichemical applications, street sweeping, and other watershed management activities can be extremely useful in qualitatively explaining observed water quality patterns, especially extreme or unusual values.

Quantify the level of treatment: Quantitative expressions of land treatment can become the independent variable in an analysis of correlation between land management and water quality. Analyze land treatment data collected in the watershed monitoring program to form such variables as:

- Number or percent of watershed animal units under animal waste management
- Acres or percent of cropland in cover crops
- Acres or percent of cropland under conservation tillage
- Annual manure or fertilizer application rate and extent
- Extent and capacity of stormwater infiltration practices

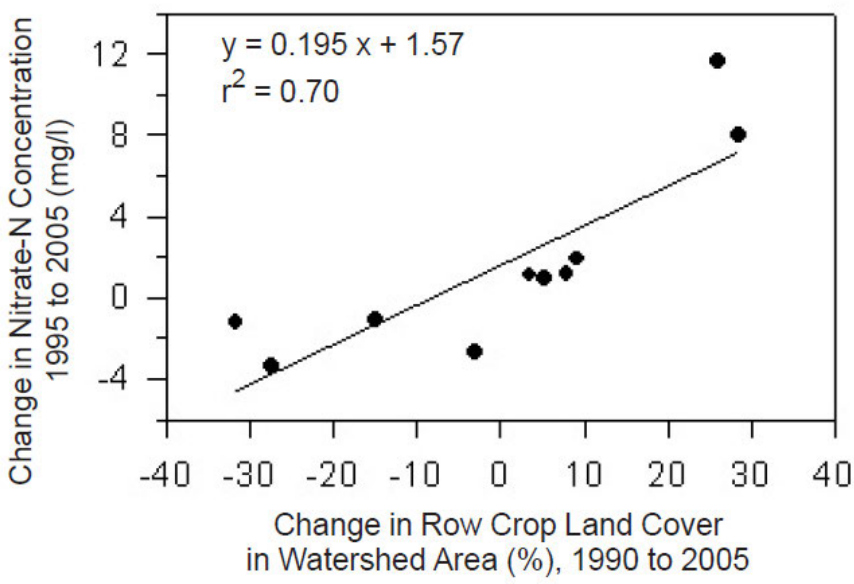
Such variables can be tested for correlation with mean total P concentration, annual suspended sediment load, or other annual water quality variables.

Document areas receiving BMPs: Use knowledge of locations of land treatment to:

- Select appropriate watersheds for analysis in a multiple watershed design
- Confirm conditions in above/below and nested watershed designs
- Document the integrity of the control and treatment watersheds in a paired-watershed design

Relate land treatment to critical source areas: A comparison of critical pollutant sources to locations that received treatment can assist in evaluating effectiveness of land treatment efforts and establish expectations for how much of the NPS problem the land treatment program potentially addresses.

The Walnut Creek (IA) NNPSMP project, for example, monitored stream NO_3-N concentrations and tracked conversion of row crop land to restored prairie vegetation (Schilling and Spooner 2006). By linking the two monitored variables, the project was able to suggest a clear association between restoring native prairie and reducing stream nitrate levels (see Figure 7-26).



Relationship between change in percentage of land cover in row crops and change in stream NO_3-N concentrations in Walnut Creek, IA

Figure 7-26. Linking stream nitrate concentration to land cover (Schilling and Spooner 2006)

7.9 Load Estimation

Determination of pollutant load is a key objective for many NPS monitoring projects. The mass of nutrients delivered to a lake or estuary drives the productivity of the waterbody. The annual suspended sediment load transported by a river is usually a more meaningful indicator of soil loss in the watershed than is a suspended sediment concentration. The foundation of water resource management embodied in the TMDL concept lies in assessment of the maximum pollutant load a waterbody can accept before becoming impaired and in the measurement of changes in pollutant loads in response to implementation of management measures.

Estimation of pollutant load through monitoring is a complex task that requires accurate measurement of both pollutant concentration and water flow and careful calculation, often based on a statistical approach. It is imperative that an NPS monitoring program be designed for good load estimation at the start. This section addresses important considerations and procedures for developing good pollutant load estimates in NPS monitoring projects. Much of the material is taken from an extensive monograph written by Dr. R. Peter Richards, of Heidelberg University, *Estimation of Pollutant Loads in Rivers and Streams: A Guidance Document for NPS Programs*. The reader is encouraged to consult that document and its

associated tools for additional information on load estimation. Much of this information is also summarized in Meals et al. (2013).

7.9.1 General Considerations

7.9.1.1 Definitions

Load may be defined as the mass of a substance that passes a particular point of a river (such as a monitoring station on a watershed outlet) in a specified amount of time (e.g., daily, annually). Mathematically, load is essentially the product of water discharge and the concentration of a substance in the water. Flux is a term that describes the loading rate, i.e., the instantaneous rate at which the load passes a point in the river. Water discharge is defined as the volume of water that passes a cross-section of a river in a specified amount of time, while flow refers to the discharge rate, the instantaneous rate at which water passes a point. Refer to Meals and Dressing (2008) for guidance on appropriate ways to estimate or measure surface water flow for purposes associated with NPS watershed projects.

Basic Terms

Flux – instantaneous loading rate (e.g., kg/sec)

Flow rate – instantaneous rate of water passage (e.g., L/sec)

Discharge – quantity of water passing a specified point (e.g., m^3)

Load – mass of substance passing a specified point (e.g., metric tons)

If we could directly and continuously measure the flux of a pollutant, the results might look like the plot in Figure 7-27. The load transported over the entire period of time in the graph would simply be equal to the shaded area under the curve.

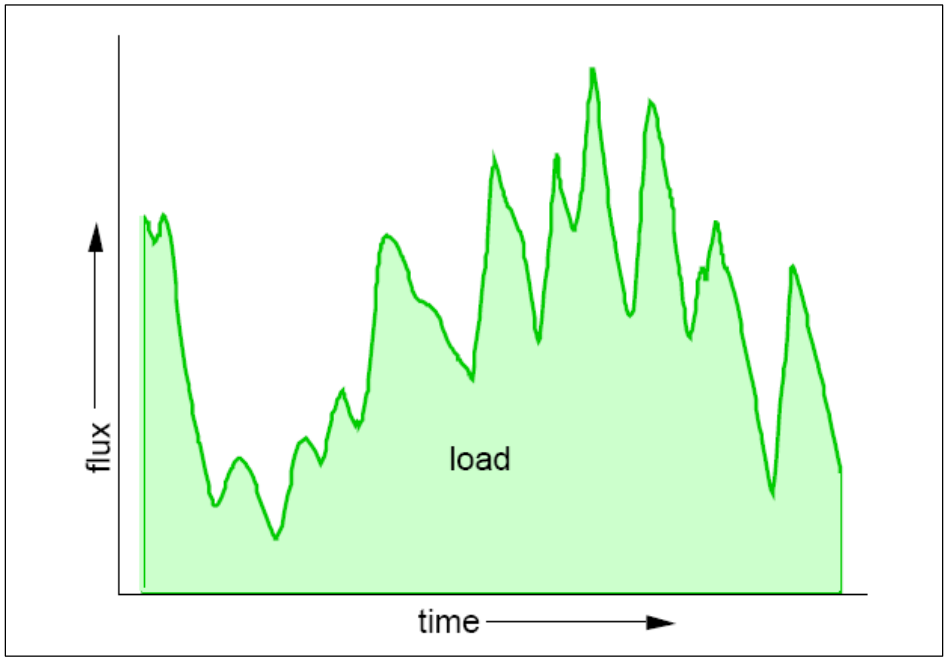


Figure 7-27. Imaginary plot of pollutant flux over time at a monitoring station (Richards 1998)

However, we cannot measure flux directly, so we calculate it as product of instantaneous concentration and instantaneous flow:

$$Load = k \int_{t} c(t)q(t)dt$$

where c is concentration and q is flow, both a function of time (t), and k is a unit conversion factor. Because we must take a series of discrete samples to measure concentration, the load estimate becomes the sum of a set of *n* products of concentration (c), flow (q), and the time interval (Δt) over which the concentration and flow measurements apply:

$$Load = k \sum_{i=1}^{n} c_i q_i \Delta t$$

The main monitoring challenge becomes how best to take the discrete samples to give the most accurate estimate of load. Note that the total load is the load over the timeframe of interest (e.g., one year) determined by summing a series of unit loads (individual calculations of load as the product of concentration, flow, and time over smaller, more homogeneous time spans). The central problem is to obtain good measures of concentration and flow during each time interval; calculation of total load by summing unit loads is simple arithmetic.

7.9.1.2 Issues of Variability

Both flow and concentration vary considerably over time, especially in NPS situations. Accurate load estimation becomes an exercise in both how many samples to take and when to take them to account for this variability.

Sampling frequency has a major influence on the accuracy of load estimation, as shown in Figure 7-28. The top panel shows daily suspended solids load (calculated as the products of daily total suspended solids (TSS) concentration and mean daily discharge measured at a continuously recording USGS station) for the Sandusky River in Ohio. The middle panel represents load calculated using weekly TSS samples and mean weekly discharge; the lower panel shows load calculated from monthly TSS samples and mean monthly discharge data. Clearly, very different pictures of suspended solids load emerge from different sampling frequencies, as decreasing sampling frequencies tend to miss more and more short-term but important events with high flow or high TSS concentrations.

Because in NPS situations most flux occurs during periods of high discharge (e.g., ~80 – 90 percent of annual load may be delivered in ~10 – 20 percent of time), choosing *when* to sample can be as important as how often to sample. The top panel in Figure 7-29 shows a plot of daily suspended solids load derived from weekly sampling superimposed on daily flux data; the bottom panel shows daily loads derived from monthly and quarterly sampling on top of the same daily flux data. Weekly samples give a reasonably good visual fit over the daily flux pattern. The monthly series gives only a very crude representation of the daily flux, but it is somewhat better than expected, because it happens to include the peaks of two of the four major storms for the year. A monthly series based on dates about 10 days later than these would have included practically no storm observations, and would have seriously underestimated the suspended solids load. Quarterly samples result in a poor fit on the actual daily flux pattern.

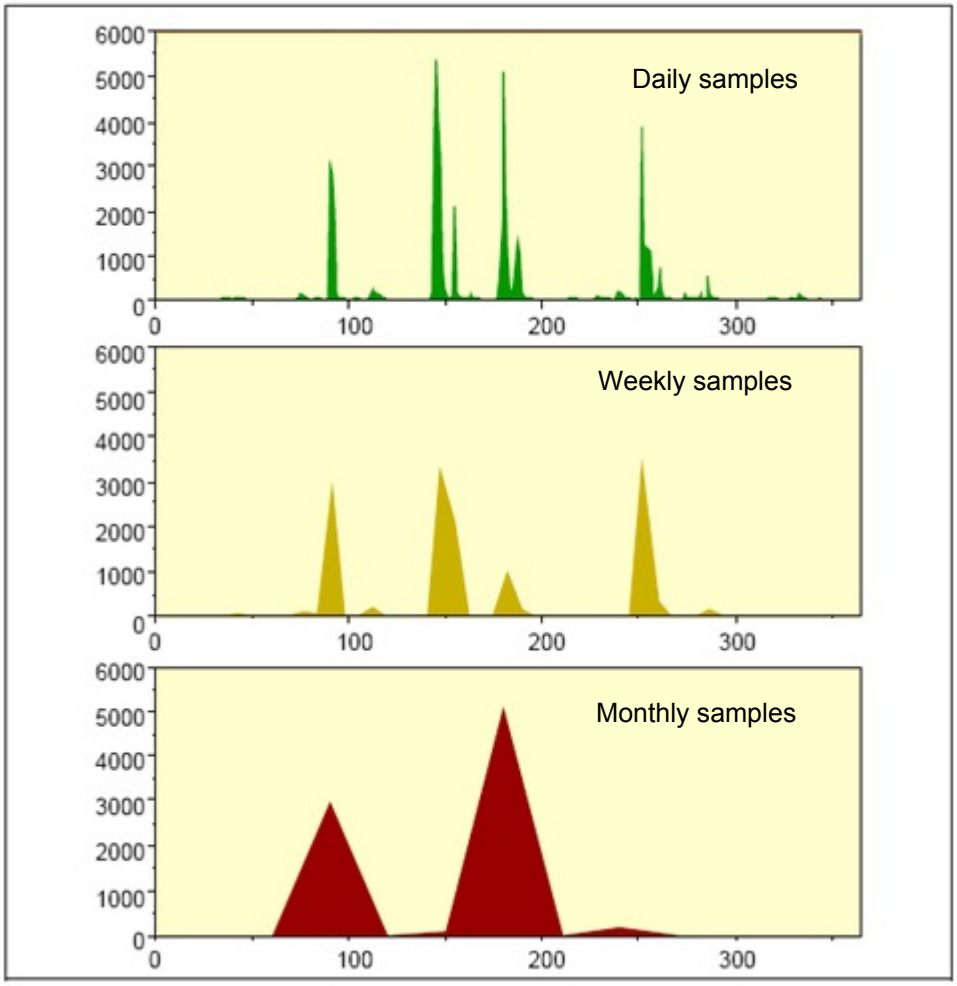


Figure 7-28. Plot of suspended solids loads for the Sandusky River, water year 1985 (Richards 1998). *Top,* daily TSS samples; *Middle*, weekly samples; *Bottom,* monthly samples. Weekly and monthly sample values were drawn from actual daily sample data series. Flux is on y-axis, time is on x-axis, and area under curve is load estimate.

The key point here is that many samples are typically needed to accurately and reliably capture the true load pattern. Quarterly observations are generally inadequate, monthly observations will probably not yield reliable load estimates, and even weekly observations may not be satisfactory, especially if very accurate load estimates are required to achieve project objectives.

7.9.1.3 Practical Load Estimation

Ideally, the most accurate approach to estimating pollutant load would be to sample very frequently and capture all the variability. Flow is relatively straightforward to measure continuously (see Meals and Dressing 2008), but concentration is expensive to measure and in most cases impossible to measure continuously. It is therefore critically important to choose a sampling interval that will yield a suitable characterization of concentration.

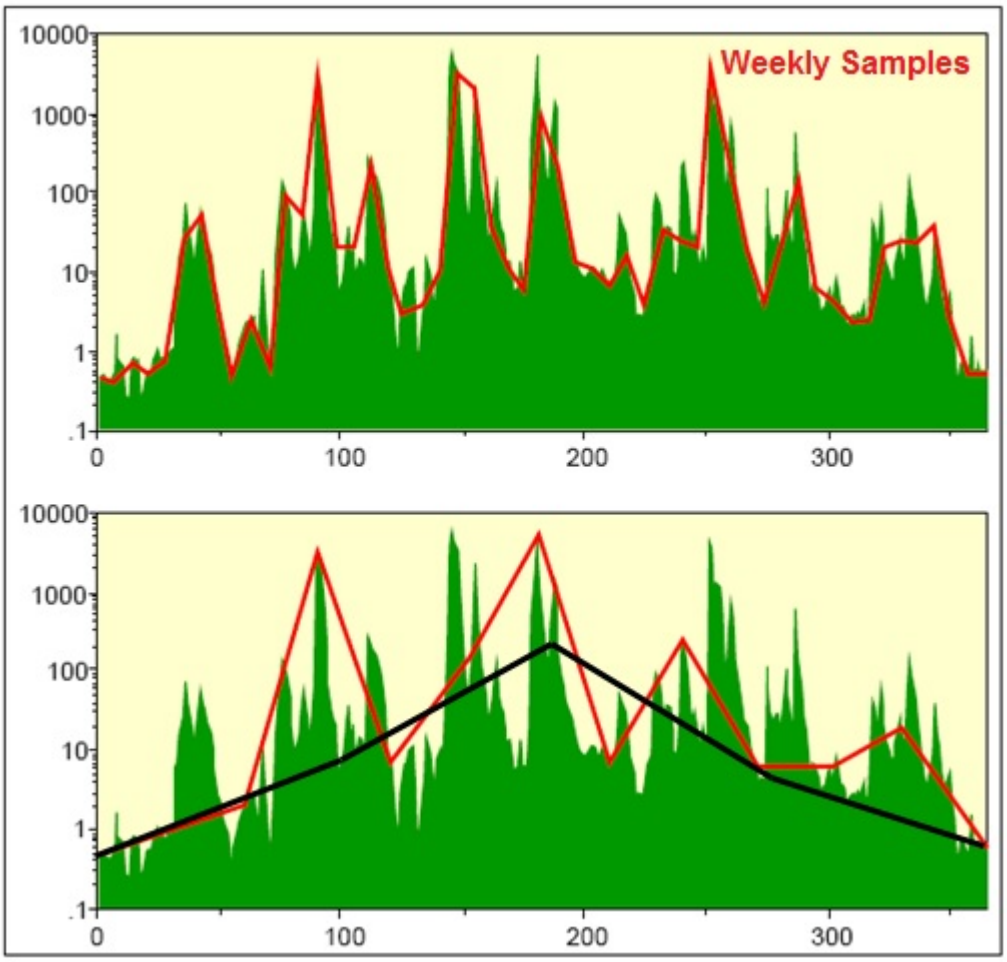


Figure 7-29. Weekly (red line in top panel), monthly (red line), and quarterly (black line in bottom panel) suspended solids load time series superimposed on a daily load time series (Richards 1998). Log of flux is on y-axis, time is on x-axis, and area under curve is load estimate.

There are three important considerations involved in sampling for good load estimation: sample type, sampling frequency, and sample distribution in time. Grab samples represent a concentration only at a single point in time and the selection of grab sampling interval must be made in consideration of the issues of variability discussed above. Integrated samples (composite samples made up of many individual grab samples) are frequently used in NPS monitoring. Time-integrated or time-proportional samples are either taken at a constant rate over the time period or are composed of subsamples taken at a fixed frequency. Time-integrated samples are poorly suited for load estimation because they are taken without regard to changes in flow (and concentration) that may occur during the integration period and are usually biased toward the low flows that occur most often. Flow-proportional samples (where a sample is collected for every n units of flow that pass the station), on the other hand, are ideally suited for load estimation, and in principle should provide a precise and accurate load estimate if the entire time interval is properly sampled. However, collecting flow-proportional samples is technically challenging and may not be suitable for all purposes. Also, even though a flow-proportional sample over a time span (e.g., a week) is a good summation of the variability of that week, ability to see what happened within that week (e.g., a transient spike in concentration) is lost. Flow-proportional sampling is also not compatible with some monitoring demands, such as monitoring for ambient concentrations that are highest at low flow or for documenting exceedance of critical values (e.g., a water quality standard).

Sampling frequency determines the number of unit load estimates that can be computed and summed for an estimate of total load. Using more unit loads increases the probability of capturing variability across the year and not missing an important event (see Figure 7-29); in general, the accuracy and precision of a load estimate increases as sampling frequency increases. Over a sufficiently short interval between samples, a sampling program will probably not miss a sudden peak in flux. If, for example, unit loads are calculated by multiplying the average concentration for the time unit by the discharge over the same time unit, the annual load that is the sum of four quarterly unit loads will be considerably less accurate than the annual load that is the sum of twelve monthly loads. Note that this example does not mean that an annual load calculated from 12 monthly loads is sufficiently accurate for all purposes.

There is a practical limit to the benefits of increasing sampling frequency, however, due to the fact that water quality data tend to be autocorrelated (see section 7.3.6). The concentration or flux at a certain point today is related to the concentration or flux at the same point yesterday and, perhaps to a lesser extent, to the concentration or flux at that spot last week. Because of this autocorrelation, beyond some point, increasing sampling frequency will accomplish little in the way of generating new information. This is usually not a problem for monitoring programs, but can be a concern, however, when electronic sensors are used to collect data nearly continuously.

Consideration of the basic sampling frequency – n samples per year – does not address the more complex issue of timing. The choice of *when* to collect concentration samples is critical. Most NPS water quality data have a strong seasonal component as well as a strong association with other variable factors such as precipitation, streamflow, or watershed management activities such as tillage or fertilizer application. Selecting when to collect samples for concentration determination is essentially equivalent to selecting when the unit loads that go into an annual load estimate are determined. That choice must consider the fundamental characteristics of the system being monitored. In northern climates, spring snowmelt is often the dominant export event of the year; sampling during that period may need to be more intensive than during midsummer in order to capture the most important peak flows and concentrations. In southern regions, intensive summer storms often generate the majority of annual pollutant load; intensive summer monitoring may be required to obtain good load estimates. For many agricultural pesticides, sampling may need to be focused on the brief period immediately after application when most losses tend to occur. Issues of random sampling, stratified random sampling, and other sampling regimes should be considered. Simple random sampling may be inappropriate for accurate load estimation if, as is likely, the resulting schedule is biased toward low flow conditions. Stratified random sampling – division of the sampling effort or the sample set into two or more parts which are different from each other but relatively homogeneous within – could be a better strategy. In cases where there is a conflict between the number of observations a program can afford and the number needed to obtain an accurate and reliable load estimate, it may be possible to use flow as the basis for selecting the interval between concentration observations. For example, planning to collect samples every x thousand ft^3 of discharge would automatically emphasize high flux conditions while economizing on sampling during baseflow conditions. Sampling levels following this strategy could be based on an annual average flow, recognizing that the number of samples per year would vary.

7.9.1.4 Planning for Load Estimation

Both discharge and concentration data are needed to calculate pollutant loads, but monitoring programs designed for load estimation will usually generate more flow than concentration data. This leaves three basic choices for practical load estimation:

1. Find a way to estimate un-measured concentrations to go with the flows observed at times when chemical samples were not taken;
2. Throw out most of the flow data and calculate the load using the concentration data and just those flows observed at the same time the samples were taken; and
3. Do something in between - find some way to use the more detailed knowledge of flow to adjust the load estimated from matched pairs of concentration and flow.

The second approach is usually unsatisfactory because the frequency of chemical observations is likely to be inadequate to give a reliable load estimate when simple summation is used. Thus almost all effective load estimation approaches are variants of approaches 1 or 3.

Unfortunately, the decision to calculate loads is sometimes made after the data are collected, often using data collected for other purposes. At that point, little can be done to compensate for a data set that contains too few observations of concentration, discharge, or both, collected using an inappropriate sampling design. Many programs choose monthly or quarterly sampling with no better rationale than convenience and tradition. A simulation study for some Great Lakes tributaries revealed that data from a monthly sampling program, combined with a simple load estimation procedure, gave load estimates which were biased low by 35 percent or more half of the time (Richards and Holloway 1987).

To avoid such problems, the sampling regime needed for load estimation must be established in the initial monitoring design, based on quantitative statements of the precision required for the load estimate. The resources necessary to carry out the sampling program must be known and budgeted for from the beginning.

The following steps are recommended to plan a monitoring effort for load estimation:

- Determine whether the project goals require knowledge of load, or if goals can be met using concentration data alone. In many cases, especially when trend detection is the goal, concentration data may be easier to work with and be more accurate than crudely estimated load data.
- If load estimates are required, determine the accuracy and precision needed based on the uses to which they will be put. This is especially critical when the purpose of monitoring is to look for a change in load. It is foolish to attempt to document a 25 percent load reduction from a watershed program with a monitoring design that gives load estimates ±50 percent of the true load (see Spooner et al. 2011a).
- Decide which approach will be used to calculate the loads based on known or expected attributes of the data.
- Use the precision goals to calculate the sampling requirements for the monitoring program. Sampling requirements include both the total number of samples and, possibly, the distribution of the samples with respect to some auxiliary variable such as flow or season.
- Calculate the loads based on the samples obtained after the first full year of monitoring, and compare the precision estimates (of both flow measurement and the sampling program) with the initial goals of the program. Adjust the sampling program if the estimated precision deviates substantially from the goals.

It is possible that funding or other limitations may prevent a monitoring program from collecting the data required for acceptable load estimation. In such a case, the question must be asked: is a biased, highly uncertain load estimate preferable to no load estimate at all? Sometimes the correct answer will be no.

7.9.2 Approaches to Load Estimation

Several distinct technical approaches to load estimation are discussed below. The reader is encouraged to consult Richards (1998) for details and examples of these calculations. Do not estimate annual loads based on simple multiplication of an annual average concentration and average discharge as load estimates will be biased low for positively correlated parameters such as suspended sediment and total phosphorus.

7.9.2.1 Numeric Integration

The simplest approach is numeric integration, where the total load is given by

$$Load = \sum_{i=1}^{n} c_i q_i t_i$$

where c_i is the concentration in the ith sample, q_i is the corresponding flow, and t_i is the time interval represented by the i^{th} sample, calculated as:

$$\frac{1}{2}(t_{i+1} - t_{i-1})$$

It is not required that t_i be the same for each sample.

The question becomes how fine to slice the pie – few slices will miss much variability, many slices will capture variability but at a higher cost and monitoring effort. Numeric integration is only satisfactory if the sampling frequency is high - often on the order of 100 samples per year or more, and samples must be distributed so that all major runoff events are captured. Selection of sampling frequency and distribution over the year is critical – sampling must focus on times when highest fluxes occur, i.e., periods of high discharge.

As noted above, flow-proportional sampling is a special case of mechanical rather than mathematical integration that assumes that one or more samples can be obtained that cover the entire period of interest, each representing a known discharge and each with a concentration that is in proportion to the load that passed the sampling point during the sample's accumulation. If this assumption is met, the load for each sample is easily calculated as the discharge times the concentration, and the total load for the year is derived by summation. In principle, this is a very efficient and cost-effective method of obtaining a total load.

7.9.2.2 Regression

When, as is often the case with NPS-dominated systems, a strong relationship exists between flow and concentration, using regression to estimate load from continuous flow and intermittent concentration data can be highly effective. In this approach, a regression relationship is developed between concentration and flow based on the days for which concentration data exist. Usually, these data are based on grab samples for concentration and mean daily flow for the sampling day (see Example 7.9-1). This

relationship may involve simple or multiple regression analysis using covariates like precipitation. In most applications, both concentration and flow are typically log-transformed to create a dataset suited for regression analysis (see section 7.3.2 and Meals and Dressing 2005) for basic information on data transformations). The regression relationship may be based entirely on the current year's samples, or it may be based on samples gathered in previous years, or both. This method requires that there be a strong linear association between flow and concentration that does not change appreciably over the period of interest. If BMP implementation is expected to affect the relationship between flow and concentration, such relationships must be tracked carefully - if BMPs change the relationship, the concentration estimation procedure must be corrected.

Once the regression relationship is established, the regression equation is used to estimate concentrations for each day on which a sample was not taken, based on the mean daily flow for the day. The total load is calculated as the sum of the daily loads that are obtained by multiplying the measured or estimated daily concentration by the total daily discharge.

The goal of chemical sampling under this approach is to accurately characterize the relationship between flow and concentration. The monitoring program should be designed to obtain samples over the entire range of expected flow rates. If seasonal differences in the flow/concentration relationship are likely, the entire range of flows should be sampled in each season. In some cases, separate seasonal flow-concentration regressions may need to be developed and used to estimate seasonal loads. Examples of such flow-concentration regressions are shown in Figure 7-30 and example 7.9-1.

This approach is especially applicable to situations where continuous flow data already exist, e.g., from an ongoing USGS hydrologic station. Grab samples can be collected as needed and then associated with the appropriate flow observations. Economy is another significant advantage of this approach. After an initial intensive sampling period to develop the regression, it may be possible to maintain the regression model with ~20 samples a year for concentration, focusing on high-flow or critical season events. Software exists to calculate and manage this approach, e.g. Flux32 (Walker 1990, Soballe 2014). Flux32 is an interactive program designed for use in estimating the loadings of nutrients or other water quality components passing a tributary sampling station over a given period of time. Data requirements include (a) grab-sample nutrient concentrations, typically measured at a weekly to monthly frequency for a period of at least 1 year, (b) corresponding flow measurements (instantaneous or daily mean values), and (c) a complete flow record (mean daily flows) for the period of interest. Using six calculation techniques, Flux32 maps the flow/concentration relationship developed from the sample record onto the entire flow record to calculate total mass discharge and associated error statistics. An option to stratify the data into groups based upon flow, date, and/or season is also included. The USGS program LOADEST is also available and is widely used to estimate loads together with an estimate of precision using the regression approach. LOADEST includes an adjusted maximum likelihood estimation method that can be used for censored data sets and a least absolute deviation method to use when the regression residuals are not normally distributed. A web-based version of LOADEST program is available at https://engineering.purdue.edu/~ldc/LOADEST/. Another USGS load estimation calculation tool – FLUXMASTER – has been used in the SPARROW (SPAtially Referenced Regressions On Watershed attributes) watershed modeling technique to compute unbiased detrended estimates of long-term mean flux, and to provide an estimate of the associated standard error (Schwarz et al. 2006). These models include seasonal and temporal terms in their formulation that can improve the estimate of load; however, care is needed to ensure the model form is correct by reviewing the diagnostic plots.

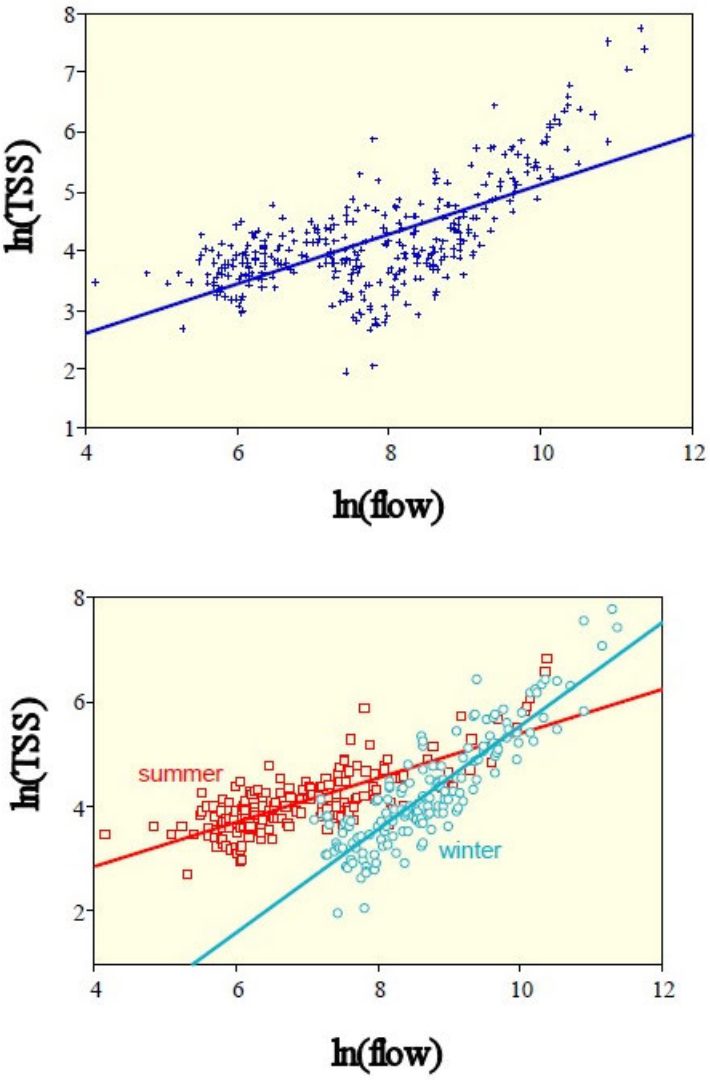


Figure 7-30. Flow-concentration regressions from the Maumee River, Ohio (Richards 1998). *Top panel,* regression relationship between log of total suspended solids concentration and log of flow for the 1991 water year dataset; *Bottom panel*, plot of same data divided into two groups based on time of year. Within each season, the regression model is stronger, has lower error, and provides a more accurate load estimate.

Example 7.9-1. Mill Creek Watershed, PA NNPSMP

In this project, loads per unit area of nutrients and suspended sediment were estimated by combining the non-storm (i.e., low flow) and storm-flow loads (Galeone et al. 2006). Low-flow and storm-flow loads were computed using a multiple regression technique that included explanatory variables such as discharge, season, and time to estimate concentrations (and subsequently loads). Regressions were developed separately for low-flow and storm-flow periods, and for both low flow and storm flow, separate models were generated for the pre- and post-treatment periods for each site. Models were selected on the basis of the highest adjusted R^2 and residuals plots to detect trends, and all F-values had to exceed the value for the F distribution for the appropriate degrees of freedom and an alpha equal to 0.05.

Continuous discharge data for the four sites was first separated into low-flow and storm-flow periods using site-specific criteria defining a storm event. Sampled storms were reviewed to determine the typical rate of stage-height increase that initiated storm sampling. The recession and subsequent completion of storm sampling was also reviewed to determine the typical endpoint of storm sampling at each of the four sites. This information was used with 5- or 15-minute stage data to manually separate storm-flow discharge data from low-flow data.

For low-flow periods, a subset of the grab-sample data was used to develop the relation between constituent concentrations and explanatory variables. Prior to using the grab-sample data, the cumulative frequency distribution for each site was determined using the continuous discharge data for the entire period of record. Grab samples collected at flows above the 97th percentile were deleted prior to load analysis. With these higher flows deleted, the relation between constituent concentrations and explanatory variables was developed. The low-flow constituent concentrations were estimated on a daily basis using the daily-mean discharge data for low-flow periods. The estimated concentrations were multiplied by the daily-mean discharge to estimate daily loads.

Storm-flow loads for nutrients and suspended sediment were estimated by use of the mean discharge and mean constituent concentration for sampled storms. The mean discharge-concentration relation developed for sampled storms using regression analysis was used to predict the concentrations for unsampled storms. The mean discharge was calculated for unsampled storms using the 5- or 15-minute continuous-stage data for the sites. This mean discharge was applied to the predicted concentration to estimate constituent loads for unsampled storms. Increases in stage caused by snowmelt events were analyzed separately by subsetting the storm events sampled during snowmelts and using these regression relations to estimate loads for non-sampled snowmelt events. The percentage of the storms sampled at each site was somewhat dependent on the location of the surface-water site, ranging from about 50-60 percent at outlet sites and 35-45 percent at upstream sites where flashiness was greater and defined storms more frequent.

Constituent loads for each continuous surface-water site were estimated by summing the low-flow and storm-flow loads. The annual load data for the constituents were divided by the basin drainage areas to determine constituent yields. The percentage of the total yield in storm-flow was determined by summing the sampled and unsampled storm yields and dividing by the total yield. The remaining yield was attributed to low-flow periods. Data also were separated into pre- and post-treatment periods.

There are a few potential disadvantages to this approach. First, as mentioned earlier, potential changes or trends in the concentration-flow relationship – sometimes a goal of watershed projects – must be tracked. If the relationship changes a new regression model must be constructed. Second, the monitoring program must be managed to effectively capture the entire range of flows/conditions that occur; the use of data from fixed-interval time-based sampling is not appropriate for this purpose because of bias toward low flow conditions.

Hirsch et al. (2010) propose a weighted regression on time, discharge, and season (WRTDS) method that addresses some of these shortcomings. Principally, the WRTDS method relies on the same function regression structure as LOADEST; however, the fitted coefficients are allowed to vary with time. For example, the amplitude of the seasonal cycle could be relatively large in some periods of the record and then dampen to smaller cycles in other portions of the record. This is achieved through using a weighted regression that "windows in" on a portion of the record in time, flow and season. It is noteworthy that the researchers recommend that this method is primarily developed for data sets with more than 200 samples collected over 20 years. Like other flow adjustment tools there is a requirement of flow stationarity, that is, there isn't a basis for expecting a change in flow over time such as a new reservoir whether that change is observed over the entire year or just during a portion of the year. Extended dry or wet periods are simply an expected part of the long term record. WRTDS is generally intended for gradual changes that might be expected with NPS projects or sites that represent the cumulative effect of multiple point sources, and less for abrupt changes. WRTDS has been built into Exploration and Graphics for RivEr Trends (EGRET): An R-package for the analysis of long-term changes in water quality and streamflow. User guidance is available at https://github.com/USGS-R/EGRET/wiki although more current releases are available through R (R Core Team 2013). The WRTDS method was applied to eight monitoring sites on the Mississippi River investigating nitrate (Sprague et al. 2011) and compared to the more traditionally recommended ESTIMATOR by Moyer et al. (2012) in an evaluation using data from the Chesapeake Bay.

7.9.2.3 Ratio Estimators

The concept of ratio estimators is a powerful statistical tool for estimating pollutant load from continuous flow data and intermittent concentration data. Ratio estimators assume that there is a positive linear relationship between load and flow that passes through the origin. On days when chemistry samples are taken, the daily load is calculated as the product of grab-sampled concentration and mean daily flow, and the mean of these loads over the year is also calculated. The mean daily load is then adjusted by multiplying it by a flow ratio, which is derived by dividing the average flow for the year as a whole by the average flow for the days on which chemical samples were taken. A bias correction factor is included in the calculation, to compensate for the effects of correlation between discharge and load. The adjusted mean daily load is multiplied by 365 to obtain the annual load.

When used in a stratified mode (e.g., for distinct seasons), the same process is applied within each stratum, and the stratum load is calculated by multiplying the mean daily load for the stratum by the number of days in the stratum. The stratum loads are then summed to obtain the total annual load. The Beale Ratio Estimator is one technique, with an example provided by Richards (1998). Several formulas are available to calculate the number of samples (random or within strata) required to obtain a load estimate of acceptable accuracy based on known variance of the system. Stratification may improve the precision and accuracy of the load estimate by allocating more of the sampling effort to the aspects which are of greatest interest or which are most difficult to characterize because of great variability such as high flow seasons.

7.9.2.4 Comparison of Load Estimation Approaches

Although strongly driven by available resources, the monitoring program design (that should have included consideration of load estimation issues from the beginning), and the natural system itself, the choice of load estimation approach can make an enormous difference in the resulting load estimate.

In an analysis of total suspended solids data from the Maumee River in water year 1991, Richards (1998) demonstrated that different methods of load estimation applied to different datasets can result in substantially different estimates of pollutant load. Richards (1998) found that loads were often underestimated with the Beale Ratio Estimator and regression techniques, attributing this finding to missed high flow/TSS events and/or the estimation methods being biased toward low flow conditions. Notably, the Beale Ratio Estimator gave a load estimate closer to the true load (estimated through numeric integration) than did the regression method. For the full daily dataset, the single flow-concentration regression over the entire year appeared to seriously underestimate suspended solids load; while separating the data into summer and winter seasons improved the fit and the accuracy of the load estimate. In a summary of findings, Harmel et al. (2006) reported that the USGS regression method could result in annual constituent loads to within 10 percent of true loads in larger watersheds but no less than 30 percent for smaller watersheds.

Harmel and King (2005) and Harmel et al. (2006) concluded that flow-proportional, composite sampling was the most effective method to obtain high quality data for estimating loads from small agricultural watersheds. They concluded that composite sampling extended the sampler capacity with little effect on error, noting that intensive sampling strategies could achieve errors less than 10 percent. In their study, smaller sampling intervals should be used for constituents such as sediment which varies more during the course of a rainfall event in comparison to other constituents which vary less during a rainfall event.

Dolan et al. (1981) evaluated total phosphorus loadings to Lake Michigan from Grand River in 1976-77. They found that the Beale ratio estimator performed better than regression or other simplified calculations. Quilbé et al. (2006) evaluated a 1989-1995 nutrient and sediment data set from the Beaurivage River (Québec, Canada). They chose to estimate loadings with a Beale ratio estimator because they found that the correlation between flow and various water quality parameters was too weak to develop regression equations while noting that regression techniques would have been preferred if good correlations were found. Marsh and Waters (2009) also found few cases with strong correlations in their evaluation of 31 storm events in Queensland. They concluded that there was no clear best technique, but noted that the ratio methods were more robust and regression techniques worked well when there was a "tight" correlation. Using hourly model output, Zamyadi et al. (2007) found that the Beale ratio did not perform well in comparison to averaging and interpolation procedures.

Taking the above literature into account, this guidance recommends that numeric integration be used when the full time series of water quality and flow data are available as in the case of flow-proportional composited samples. Regression approaches are appropriate for incomplete water quality records if good correlations between water quality and flow exist, with the Beale ratio recommended otherwise. It is important to take into account stratification by flow regime, season, and other covariates for both regression and the Beale ratio.

7.9.3 Load Duration Curves

A particularly useful diagnostic tool for load estimation data is the load duration curve. Simply stated, a duration curve is a graph representing the percentage of time during which the value of a given parameter (e.g., flow, concentration, or load) is equaled or exceeded. A load duration curve is therefore a cumulative frequency plot of mean daily flows, concentrations, or daily loads over a period of record, with values plotted from their highest value to lowest without regard to chronological order. For each flow, concentration, or load value, the curve displays the corresponding percent of time (0 to 100) that the value was met or exceeded over the specified time – the flow, concentration, or load duration interval.

Extremely high values are rarely exceeded and have low duration interval values; very low values are often exceeded and have high duration interval values.

The process of using load duration curves generally begins with the development of a flow duration curve, using existing historical flow data (e.g., from a USGS gage), typically using mean daily discharge values. A basic flow duration curve runs from high to low along the x-axis, as illustrated in Figure 7-31. The x-axis represents the duration or percent of time, as in a cumulative frequency distribution. The y-axis represents the flow value (e.g., ft^3/sec (cfs)) associated with that percent of time. Figure 7-31 illustrates that the highest observed flow for the period of record was about 5,400 cfs, while the lowest flow was 6 cfs. The median flow – the flow exceeded 50 percent of the time – was about 200 cfs.

In the next step, a load duration curve is created from the flow duration curve by multiplying each of the flow values by the applicable numeric water quality target (usually a water quality criterion) and a unit conversion factor, then plotting the results as for the flow duration curve. The x-axis remains as the flow duration interval, and the y-axis depicts the load rather than the flow. This curve represents the allowable load (e.g., the TMDL) at each flow condition over the full range of observed flow. An example is shown in Figure 7-32 for the same site as shown in the flow duration curve, using a target of 0.05 mg/L total P. Then, observed P load observations associated with the flow intervals are plotted along the same axes. Points located above the curve represent times when the actual loading is exceeding the target load, while those plotting below the curve represent times when the actual loading is less than the target load.

A key feature of load duration curve analysis is that the pattern of loads – and impairment – can be easily visualized over the full range of flow conditions. Because flow variations usually correspond to seasonal patterns, this feature can address the requirement that TMDLs account for seasonal variations. The pattern of observed loads exceeding target loads can be examined to see if impairments occur only at high flows, only during low flows, or across the entire range of flow conditions. A common way to look at a load duration curve is by dividing it into zones representing, for example: high flows (0-10 percent flow duration interval), moist conditions (10-40 percent), mid-range flows (40-60 percent), dry conditions (60-90 percent), and low flows (90-100 percent). Data may also be grouped by season (e.g., spring runoff versus summer base flow). Sometimes, analysis of the load duration curve can provide insight on the source of pollutant loads. Measured loads that plot above the curve during flow duration intervals above 80 percent (low flow conditions), for example, may suggest point sources that discharge continuously during dry weather. Conversely, measured loads that plot above the curve during flow duration intervals of about 10 to 70 percent tend to reflect wet weather contributions by NPS such as erosion, washoff, and streambank erosion. Figure 7-32 illustrates that allowable total P loads in the Sevier River were exceeded during all flow intervals, and that P concentrations were independent of flow.

It should be noted that an individual load duration curve applies only to the point in the stream where the data were collected. A load duration curve developed at a watershed outlet station (e.g., for a TMDL) applies only to loads observed at that point. If significant pollution sources exist upstream, a single load duration analysis at the watershed outlet can underestimate the extent of impairment in upstream segments. For this reason, it is usually wise to develop multiple load duration curves throughout the watershed to address the spatial distribution of impairments. Such an exercise can also be useful in targeting land treatment to critical watershed source areas.

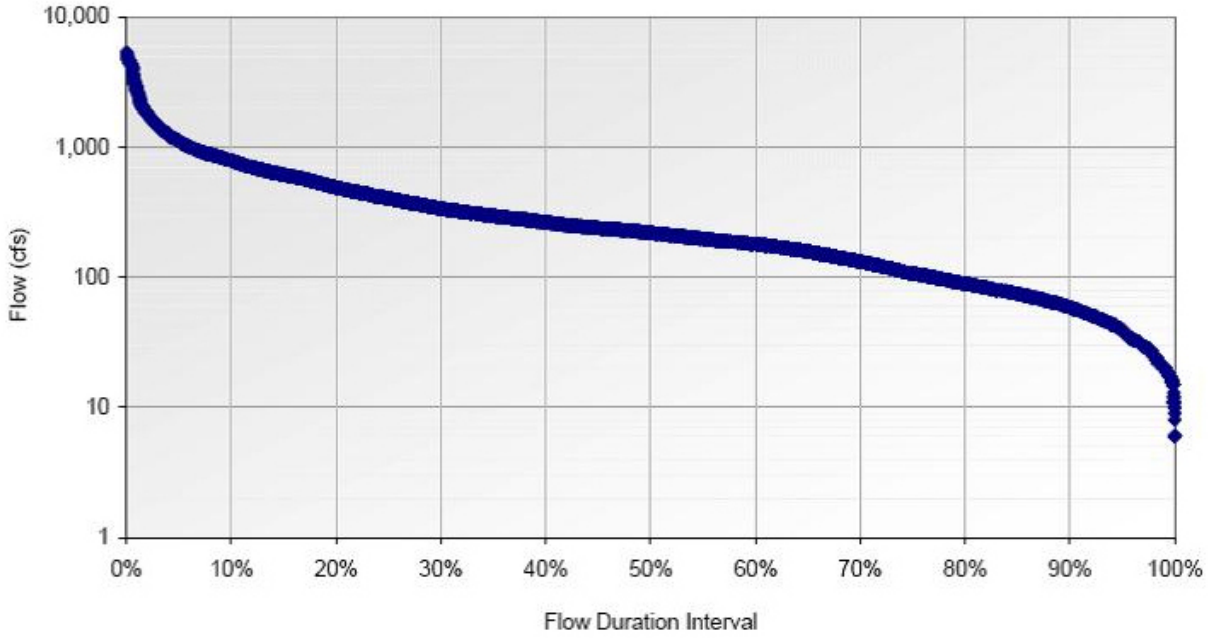


Figure 7-31. Flow duration curve for the Sevier River near Gunnison, UT, covering the period January 1977 through September 2002

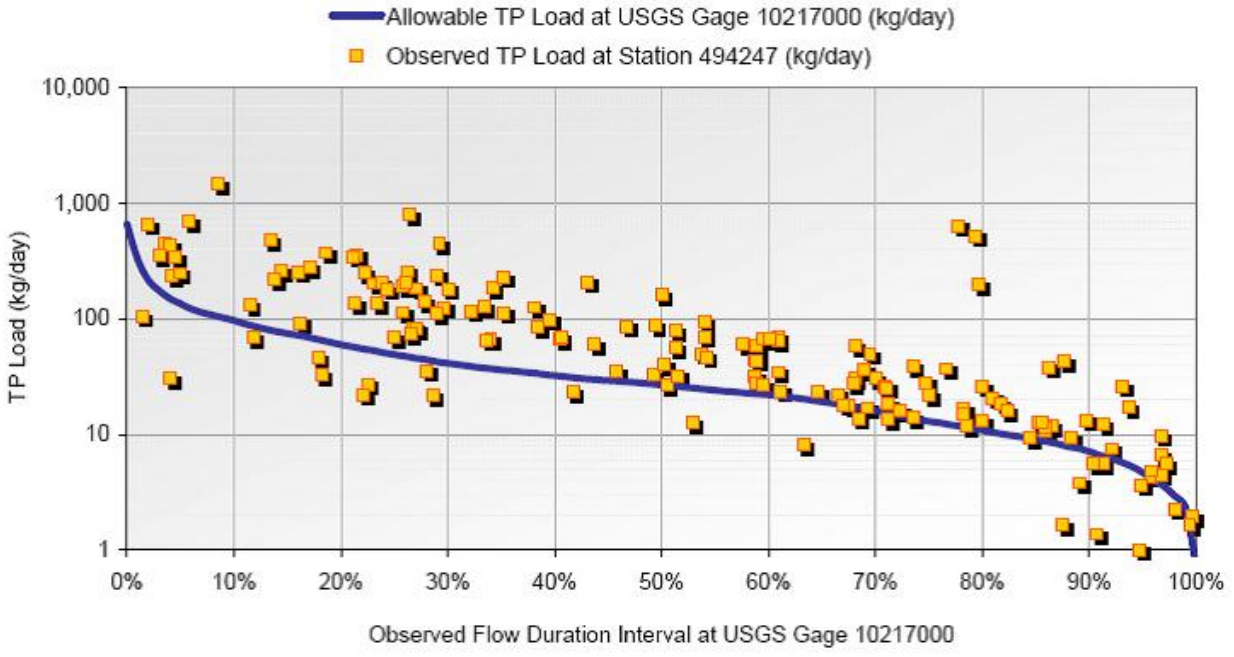


Figure 7-32. Load duration curve for the Sevier River near Gunnison, UT, January 1977 through September 2002. Blue line represents allowable total P load calculated as the product of each observed flow duration interval and the target total P concentration of 0.05 mg/L. Yellow points represent observed total P loads at the same flow duration intervals.

For more detailed discussion of load duration curves, particularly their application to the TMDL process, refer to:

- USEPA. 2007. *An Approach for Using Load Duration Curves in the Development of TMDLs*

7.9.4 Assessing Load Reductions

The same statistical tools recommended for flow and concentration data in section 7.8.2 and elsewhere in this chapter can be used to analyze program effectiveness with regard to load reductions. For example, loads might be estimated on a weekly basis using numeric integration and flow-proportional, composite sample data. Under a paired-watershed approach, the weekly-paired loads would be grouped as pre- and post-treatment and analyzed using ANCOVA.

For comparisons of annual loadings, the analyst will have limited data to perform analyses (i.e., one annual loading value per site-year) and will be generally limited to reporting simple change in loading and drawing anecdotal comparisons to the control watershed. Normalizing the loadings based on watershed size, annual rainfall, and other covariates might prove helpful.

Depending on the watershed and the types of installed BMPs, it is also appropriate to compare storm loadings from individual storms before and after BMP implementation in a single watershed. The particular challenge here is to control for other covariates and select/analyze storms of a certain size (e.g., rainfall between 2.5-5.0 cm) and occurring at key times during the year (e.g., within 6 weeks of spring planting). This type of analysis might also be limited to drawing simple comparisons due to sample size.

7.10 Statistical Software

Modern computers and software packages make it simple to perform the statistical analyses described in this chapter. Most standard spreadsheet programs include basic statistical functions and graphing capabilities, but more sophisticated and powerful statistical software packages might be needed for advanced analyses such as ANCOVA or cluster analysis. An extensive list and comparison of statistical software packages is available at Wikipedia. Practical Statistics, a web site maintained by Dennis Helsel, provides a more environmental-centric review of low-cost software tools. Table 7-9 lists some examples and websites to visit for more information about the many statistical packages available.

Table 7-9. Sampling of available statistics software packages

Package Name	Web Site URL
Analyse-It (add in for MS Excel)	http://www.analyse-it.com
DataDesk	http://www.datadesk.com
JMP	http://www.jmp.com/en_gb/software.html
Mathematica	http://www.wolfram.com/mathematica/
MATLAB	http://www.mathworks.com/products/matlab/
MINITAB	https://www.minitab.com/en-us/
R	https://www.r-project.org/
SAS/Stat, SAS/Insight	http://www.sas.com/technologies/analytics/statistics/index.html
SPSS	http://www.spss.com/spss/
SYSTAT	http://www.systat.com/products/Systat/
WINKS	http://www.texasoft.com/

7.11 References

Belsley, D.A., E. Kuh, and R.E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons, New York.

Bernstein, B.B. and J. Zalinski. 1983. An optimum sampling design and power tests for environmental biologists. *Journal of Environmental Management* 16(1):35-43.

Bishop, P.L., W.D. Hively, J.R. Stedinger, M.R. Rafferty, J.L. Lojpersberger, and J.A. Bloomfield. 2005. Multivariate analysis of paired watershed data to evaluate agricultural best management practice effects on stream water phosphorus. *Journal of Environmental Quality* 34:1087–1101.

Box, G.E.P. and D.R. Cox. 1964. An analysis of transformations - series B (methodological). *Journal of the Royal Statistical Society* 26(2):211-252.

Box, G.E.P. and G.M. Jenkins. 1976. *Time Series Analysis: Forecasting And Control*. Revised Edition. Holden-Day, Oakland, CA.

Carpenter, S.R., T.M. Frost, D. Heisey, and T.K. Kratz. 1989. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. *Ecology* 70(4):1142-1152.

Chambers, J.M., W.S. Cleveland, B. Kleiner, P.A. Tukey. 1983. *Graphical Methods for Data Analysis.* Duxbury Press, Boston.

Clausen, J.C. 2007. *Jordan Cove Watershed Project Final Report*. University of Connecticut, College of Agriculture and Natural Resources, Department of Natural Resources Management and Engineering. Accessed January 8, 2016. http://jordancove.uconn.edu/jordan_cove/publications/final_report.pdf.

Clausen, J.C. and K.N. Brooks. 1983. Quality of runoff from Minnesota peatlands: II. a method for assessing mining impacts. *Water Resources Bulletin* 19(5):769-772.

Clausen, J.C. and J. Spooner. 1993. *Paired Watershed Study Design*. 841-F-93-009. Prepared for S. Dressing, U.S. Environmental Protection Agency, Office of Water, Washington, DC. Accessed February 12, 2016.

Cleveland, W.S. 1993. *Visualizing Data.* AT&T Bell Laboratories/Hobart Press, Murray Hill, NJ/Summit, NJ.

Clifford, R., Jr., J.W. Wilkinson, and N.L. Clesceri. 1986. Statistical Assessment of a Limnological Data Set. In *Statistical Aspects of Water Quality Monitoring, Proceedings of the Workshop Held at the Canada Centre for Inland Waters*, October 7-10, 1985, Volume 27 of Developments in Water Science series, ed. A.H. El-Shaarawi and R.E. Kwiatkowski. Elsevier Publishers, New York. pp. 363-380.

Dolan, D.M., A.K. Yui and R.D. Geist. 1981. Evaluation of river load estimation methods for total phosphorus. *Journal of Great Lakes Research* 7(3):207-214.

Dolan, D.M. and K.P. McGunagle. 2005. Lake Erie total phosphorus loading analysis and update: 1996-2002. *Journal of Great Lakes Research* 31(Supplement 2):11-22. Accessed March 24, 2016. http://www.cee.mtu.edu/~nurban/classes/ce5508/2007/Readings/dolan05.pdf.

Drake, D. 1999. *Multivariate Analysis of Fish and Environmental Factors in the Grande Ronde Basin of Northeastern Oregon*. Oregon Department of Environmental Quality, Biomonitoring Section, Laboratory Division, Portland. Accessed March 24, 2016. http://www.deq.state.or.us/lab/techrpts/docs/Bio012.pdf.

Elliott, A.C. 2012. *Descriptive Statistics Using Microsoft Excel*. Texa Soft Mission Technologies, Cedar Hill, TX. Accessed March 24, 2016. http://www.stattutorials.com/EXCEL/EXCEL-DESCRIPTIVE-STATISTICS.html.

Erickson, A.J., P.T. Weiss, J.S. Gulliver, R.M. Hozalski. 2010a. Analysis of Individual Storm Events, In *Stormwater Treatment: Assessment and Maintenance* ed. J.S. Gulliver and A.J. Erickson, University of Minnesota, St. Anthony Falls Laboratory. Minneapolis, MN. Accessed March 24, 2016. http://stormwaterbook.safl.umn.edu/

Erickson, A.J., P.T. Weiss, J.S. Gulliver, and R.M. Hozalski. 2010b. Analysis of Long-Term Performance. In *Stormwater Treatment: Assessment and Maintenance*. ed. J.S. Gulliver, A.J. Erickson, and P.T. Weiss. University of Minnesota, St. Anthony Falls Laboratory. Minneapolis, MN. Accessed March 24, 2016. http://stormwaterbook.safl.umn.edu/content/analysis-long-term-performance.

Farnsworth, R.K. and E.S. Thompson. 1982. *Mean Monthly, Seasonal, and Annual Pan Evaporation for the United States*. Technical Report NWS 34. National Oceanic and Atmospheric Administration, National Weather Service. Accessed March 24, 2016. http://www.nws.noaa.gov/oh/hdsc/PMP_related_studies/TR34.pdf.

Fuller, W.A. 1976. *Introduction to Statistical Time Series*. John Wiley & Sons, Inc. New York.

Galeone, D.G., R.A. Brightbill, D.J. Low, and D.L. O'Brien. 2006. *Effects of Streambank Fencing of Pastureland on Benthic Macroinvertebrates and the Quality ofSurface Water and Shallow Ground Water in the Big Spring Run Basin of Mill Creek Watershed, Lancaster County, Pennsylvania, 1993-2001*. Scientific Investigations Report 2006-5141. U. S. Geological Survey, Reston, VA. Accessed March 24, 2016. http://pubs.usgs.gov/sir/2006/5141/.

Geosyntec and WWE (Wright Water Engineers, Inc.). 2009. *Urban Stormwater BMP Performance Monitoring*. Prepared for U.S. Environmental Protection Agency, Water Environment Research Foundation, Federal Highway Administration, and Environmental and Water Resources Institute of the American Society of Civil Engineers, by Geosyntec Consultants and Wright Water Engineers, Inc., Washington, DC. Accessed March 24, 2016. http://www.bmpdatabase.org/Docs/2009%20Stormwater%20BMP%20Monitoring%20Manual.pdf.

Gilliom, R.J., R.M. Hirsch, and E.J. Gilroy. 1984. Effect of censoring trace-level water-quality data on trend-detection. *Environmental Science and Technology* 18(7):530-535.

Grabow. G.L. 1999. *Summary of Analyses Performed for Sycamore Creek Section 319 NNMP Project*. North Carolina State University, NCSU Water Quality Group, Raleigh, NC. Accessed April 29, 2016. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/monitoring-and-evaluating-nonpoint-source-watershed.

Grabow, G.L., J. Spooner, L.A. Lombardo, and D.E. Line. 1998. Detecting water quality changes before and after BMP implementation: use of a spreadsheet for statistical analysis. *NWQEP Notes* 92:1-

9. North Carolina State University Cooperative Extension, Raleigh. Accessed March 24, 2016. http://www.bae.ncsu.edu/programs/extension/wqg/issues/92.pdf.

Grabow, G.L., J. Spooner, L.A. Lombardo, and D.E. Line. 1999. Detecting water quality changes before and after BMP implementation: use of SAS for statistical analysis. *NWQEP Notes* 93:1-11. North Carolina State University Cooperative Extension, Raleigh. Accessed March 24, 2016. http://www.bae.ncsu.edu/programs/extension/wqg/issues/93.pdf.

Harmel, R.D. and K.W. King. 2005. Uncertainty in measured sediment and nutrient flux in runoff from small agricultural watersheds. *Transactions of the American Society of Agricultural and Biological Engineers* 48(5): 1713-1721.

Harmel, R.D., K.W. King, B.E. Haggard, D.G. Wren, and J.M. Sheridan. 2006. Practical guidance for discharge and water quality collection in small watersheds. *Transactions of the American Society of Agricultural and Biological Engineers* 49(4):937-948.

Harstine, L.J. 1991. *Hydrologic Atlas for Ohio: Average Annual Precipitation, Temperature, Streamflow, and Water Loss for a 50-Year Period, 1931-1980*. Water Inventory Report No. 28. Ohio Department of Natural Resource, Division of Water, Ground Water Resources Section.

Helsel, D.R. 2012. *Statistics for Censored Environmental Data Using Minitab and R*. 2nd ed. Wiley and Sons, New York.

Helsel, D.R. and T.A. Cohn. 1988. Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research* 24(12):1997-2004.

Helsel, D.R., and R.M. Hirsch. 2002. Statistical Methods in Water Resources. Book 4, Chapter A3 in *Techniques of Water-Resources Investigations*. U.S. Geological Survey, Reston, VA. Accessed February 10, 2016. http://pubs.usgs.gov/twri/twri4a3/.

Hewlett, J.D. and L. Pienaar. 1973. Design and Analysis of the Catchment Experiment, In *Proceedings of a Symposium on Use of Small Watersheds in Determining Effects of Forest Land Use on Water Quality*, ed. E. H. White, University of Kentucky, Lexington, KY, May 22-23, 1973. Accessed January 26, 2016. http://coweeta.uga.edu/publications/hewlett_73_catachment.pdf.

Hibbert, A.R. 1969. Water yield changes after converting a forested catchment to grass. *Water Resources Research* 5(3):634-640.

Hirsch, R.M. 1988. Statistical methods and sampling design for estimating step trends in surface water quality. *Water Resources Research* 24:493-503.

Hirsch, R.M. and J.R. Slack. 1984. A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research* 20(6):727-732.

Hirsch, R.M., R.B. Alexander, and R.A. Smith.1991. Selection of methods for the detection and estimation of trends in water quality. *Water Resources Research* 27:803-813.

Hirsch, R.M. D.L. Moyer, and S.A. Archfield. 2010. Weighted regressions on time, discharge and season (WRTDS), with an application to Chesapeake Bay river inputs. *Journal of the American Water Resources Association* 46(5):857-880.

Hoorman, J., T.Hone, T.Sudman Jr., T.Dirksen, J. Iles, and K.R. Islam. 2008. Agricultural impacts on lake and stream water quality in Grand Lake St. Marys, Western Ohio. *Water, Air, and Soil Pollution* 193: 309-322.

Hornbeck, J.W., R.S. Pierce, and C.A. Federer. 1970. Streamflow changes after forest clearing in New England. *Water Resources Research* 6(4):1124-1132.

Jambu, M. 1991. *Exploratory and Multivariate Data Analysis.* Academic Press, Inc., Boston.

Kilpatrick, F.A. and J.F. Wilson, Jr. 1989. Measurement of Time of Travel in Streams by Dye Tracing. Book 3, Chapter A9 in *Techniques of Water-Resources Investigations*. U.S. Geological Survey, Reston, VA. Accessed March 24, 2016. http://pubs.usgs.gov/twri/twri3-a9/pdf/twri_3-A9.pdf.

Lettenmaier. D.P. 1976. Detection of trends in water quality data from records with dependent observations. *Water Resources Research* 12:1037-1046.

Lettenmaier, D.P. 1978. Design considerations for ambient stream quality monitoring. *Water Resources Bulletin* 14(4):884-902.

Lewis, J. 2006. *Fixed and Mixed-Effects Models for Multi-Watershed Experiments*. U.S. Department of Agriculture, Forest Service, Pacific Southwest Research Station, Arcata, CA. Accessed March 24, 2016. http://www.fs.fed.us/psw/publications/4351/Lewis06.pdf.

Line, D.E., W.A. Harman, G.D. Jennings, E.J. Thompson, and D.L. Osmond. 2000. Nonpoint-source pollutant load reductions associated with livestock exclusion. *Journal of Environmental Quality* 29: 1882–1890.

Loftis, J.C. and R.C. Ward. 1980a. Sampling frequency selection for regulatory water quality monitoring. *Water Resources Bulletin* 16:501-507.

Loftis, J.C. and R.C.Ward, 1980b. Water quality monitoring – some practical sampling frequency considerations. *Environmental Management* 4:521-526.

MacKenzie, M.C., R.N. Palmer, and S.P. Millard. 1987. Analysis of statistical monitoring network design. *Journal of Water Resources Planning and Management* 113(5):599-615.

Marsh, N. and D. Waters. 2009. Comparison of Load Estimation Methods and Their Associated Error. In *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*, ed. R.S. Anderssen, R.D. Braddock and L.T.H. Newham, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, pp. 3322-3328. Accessed March 24, 2016. http://mssanz.org.au/modsim09/I4/marsh_I4.pdf.

Martz, E. 2013. *Enough Is Enough! Handling Multicollinearity in Regression Analysis*. The Minitab Blog, Minitab Inc. Accessed March 24, 2016. http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis.

Matalas, N.C. and W.B. Langbein. 1962. Information content of the mean. *Journal of Geophysical Research* 67(9):3441-3448.

Meals, D.W. 1987. Detecting changes in water quality in the LaPlatte River watershed following implementation of BMPs. *Lake and Reservoir Management* 3(1):185-194.

Meals, D.W. 2001. *Lake Champlain Basin Agricultural Watersheds Section 319 National Monitoring Program Project, Final Project Report: May, 1994-September, 2000*. Vermont Department of Environmental Conservation, Waterbury, VT.

Meals, D.W. and S.A. Dressing. 2005. *Monitoring Data – Exploring Your Data, the First Step*, Tech Notes #1, July 2005. Prepared for U.S. Environmental Protection Agency, by Tetra Tech, Inc., Fairfax. Accessed March 24, 2016. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoring-technical-notes.

Meals, D.W. and D.C. Braun. 2006. Demonstration of methods to reduce E.coli runoff from dairy manure application sites. *Journal of Environmental Quality* 35:1088-1100.

Meals, D.W. and S.A. Dressing. 2008. *Surface Water Flow Measurement for Water Quality Monitoring Projects*, Tech Notes #3, March 2008. Prepared for U.S. Environmental Protection Agency, by Tetra Tech, Inc., Fairfax, VA. Accessed March 24, 2016. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoring-technical-notes.

Meals, D.W., J. Spooner, S.A. Dressing, and J.B. Harcum. 2011. *Statistical Analysis for Monotonic Trends*, Tech Notes #6, September 2011. Prepared for U.S. Environmental Protection Agency, by Tetra Tech, Inc., Fairfax, VA. Accessed March 24, 2016. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoring-technical-notes.

Meals, D.W., R.P. Richards, and S.A. Dressing. 2013. *Pollutant Load Estimation forWater Quality Monitoring Projects.* Tech Notes #8. Prepared for U.S. Environmental Protection Agency, by Tetra Tech, Inc., Fairfax, VA. Accessed March 24, 2016. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoring-technical-notes.

Minitab. 2016. Minitab 17. Minitab Inc., State College, PA. Accessed January 22, 2016. http://www.minitab.com/en-US/products/minitab/.

Mosteller, F. and J.W. Tukey. 1977. *Data Analysis and Regression: Second Course in Statistics*. Addison-Wesley Pub. Co., Reading, MA.

Moyer, D.L., R.M. Hirsch, and K.E. Hyer. 2012. *Comparison of Two Regression-Based Approaches for Determining Nutrient and Sediment Fluxes and Trends in the Chesapeake Bay Watershed*. Scientific Investigations Report 2012-5244. U.S. Geological Survey, Reston, VA. Accessed March 24, 2016. http://pubs.usgs.gov/sir/2012/5244/.

NCDENR (North Carolina Department of Environment and Natural Resources). 2016. *Sanitary Survey*. North Carolina Department of Environment and Natural Resources, Division of Marine Fisheries, Raleigh, NC. Accessed March 24, 2016. http://portal.ncdenr.org/web/mf/sanitary-survey.

Newbold, J.D., S. Herbert, B.W. Sweeney, and P. Kiry. 2009. Water quality functions of a 15-year-old riparian forest buffer system. *NWQEP Notes* 130:1-9. North Carolina State University Cooperative Extension, Raleigh. Accessed March 15, 2016. http://www.bae.ncsu.edu/programs/extension/wqg/issues/notes130.pdf.

Newbold, J.D., S. Herbert, and B.W. Sweeney. 2009. *Mitigation of Nonpoint Pollution by a Riparian Forest Buffer in an Agricultural Watershed of the Mid-Atlantic Piedmont: Stroud Preserve Watersheds National Monitoring Project Final Report*. Stroud Water Research Center, Avondale PA. Accessed March 24, 2016. http://www.stroudcenter.org/research/projects/StroudPreserve/StroudNMPFinalReport2009.pdf.

ODNR (Ohio Department of Natural Resources). 2013. *Grand Lake St. Marys State Park*. Ohio Department of Natural Resources. Accessed March 24, 2016. http://parks.ohiodnr.gov/grandlakestmarys.

OWEB (Oregon Watershed Enhancement Board). 1999. *Oregon Aquatic Habitat Restoration and Enhancement Guide*. Oregon Watershed Enhancement Board, Salem, OR. Accessed April 25, 2016. http://www.oregon.gov/OWEB/docs/pubs/habguide99-complete.pdf.

Palmer, R.N. and M.C. MacKenzie. 1985. Optimization of water quality monitoring networks. *Journal of Water Resources Planning and Management* 111(4):478-493.

Perkins, W.W., E.B. Welch, J. Frodge, and T. Hubbard. 1997. A zero degree of freedom total phosphorus model: application to Lake Sammamish, Washington. *Lake and Reservoir Management* 13:131-141.

Porter, P.S., R.C. Ward, and H.F. Bell. 1988. The detection limit, water quality monitoring data are plagued with levels of chemicals that are too low to be measured precisely. *Environmental Science and Technology* 22:856-861.

Primrose, N.L. 2003. *Report on Nutrient Synoptic Surveys in the Corsica River Watershed, Queen Annes County, Maryland, April 2003*. Maryland Department of Natural Resources, Watershed Services, Annapolis, MD.

Quilbé, R., A.N. Rousseau, M. Duchemin, A. Poulin, G. Gangbazo, J. Villeneuve. 2006. Selecting a calculating method to estimate sediment and nutrient loads in streams: application to the Beaurivage River (Québec, Canada). *Journal of Hydrology* 326(2006):295-310.

R Core Team. 2013. *The R Project for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Accessed April 25, 2016. http://www.R-project.org/.

Richards, R.P. 1998. *Estimation of Pollutant Loads in Rivers and Streams: A Guidance Document for NPS Programs*. Prepared for U.S. EPA Region VIII, by Heidelberg University, Water Quality Laboratory, Tiffin, OH. Accessed February 5, 2016. http://141.139.110.110/sites/default/files/jfuller/images/Load_Est1.pdf.

Richards, R.P. and J. Holloway. 1987. Monte Carlo studies of sampling strategies for estimating tributary loads. *Water Resources Research* 23:1939-1948.

Roseboom, D., T. Hill, J. Rodsater, J. Beardsley, and L. Duong. 1999. *Evaluation of Sediment Delivery to Lake Pittsfield After Best Management Practice Implementation-National Watershed Monitoring Project*. Illinois Environmental Protection Agency, Springfield, IL.

Rosgen, D.L. 1997. A Geomorphological Approach to Restoration of Incised Rivers. In *Proceedings of the Conference on Management of Landscapes Disturbed by Channel Incision*, ed. S.S.Y. Wang, E.J. Langendoen and F.D. Shields, Jr. 1997. Accessed March 24, 2016.

http://www.wildlandhydrology.com/assets/A_Geomorphological_Approach_to_Restoration_of_Incised_Rivers.pdf.

SAS Institute. 1985. *SAS® User's Guide; Statistics. Version 5 Edition*. SAS Institute Inc., Cary, North Carolina 27513.

SAS Institute. 2010. *SAS® Version 9.2, SAS/ETS*. SAS Institute Inc., Cary, NC.

SAS Institute. 2016a. *Collinearity Diagnostics, SAS 9.2 Documentation, SAS/STAT(R) 9.22 User's Guide*. SAS Institute, Inc., Cary, NC. Accessed March 24, 2016. http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_reg_sect038.htm.

SAS Institute. 2016b. *JMP® Version 9.0.2, 2010*. SAS Institute Inc., Cary, NC. Accessed March 24, 2016. http://www.jmp.com/software/.

SAS Institute. 2016c. *Multiple Linear Regression, SAS 9.2 Documentation.* SAS Institute, Inc., Cary, NC. Accessed March 24, 2016. http://support.sas.com/documentation/cdl/en/anlystug/58352/HTML/default/viewer.htm#chap11_sect3.htm.

SAS Institute. 2016d. *The Autoreg Procedure, SAS/ETS(R) 9.2 User's Guide*. SAS Institute, Inc., Cary, NC. Accessed March 31, 2016. http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/viewer.htm#autoreg_toc.htm.

Schilling, K.E., and J. Spooner. 2006. Effects of watershed-scale land use change on stream nitrate concentrations. *Journal of Environmental Quality* 35:2132–2145.

Schwarz, G.E., A.B. Hoos, R.B. Alexander, and R.A. Smith. 2006. *The SPARROW Surface Water-Quality Model: Theory, Application and User Documentation*. Techniques and Methods 6-B3. U.S. Geological Survey, Reston, VA. Accessed March 24, 2016. http://pubs.usgs.gov/tm/2006/tm6b3/.

Simpson, T. and S. Weammert. 2009. *Developing Best Management Practice Definitions and Effectiveness Estimates for Nitrogen, Phosphorus, and Sediment in the Chesapeake Bay Watershed*. Final Report. University of Maryland Mid-Atlantic Water Program. Accessed March 24, 2016. http://archive.chesapeakebay.net/pubs/BMP_ASSESSMENT_REPORT.pdf.

Smith, R.V., S.D. Lennox, and J.S. Bailey 2003. Halting the upward trend in soluble phosphorus transported from a grassland catchment. *Journal of Environmental Quality* 32:2334–2340.

Snedecor, G.W. and W.G. Cochran. 1989. *Statistical Methods*. 8th Edition. Iowa State University Press, Ames, IA.

Soballe, D.M. 2014. Flux32. Developed in conjunction with Minnesota Pollution Control Agency, by U.S. Army Corps of Engineers Waterways Experiment Station, Vicksburg, MS. Accessed March 24, 2016.

Spooner, J., C.J. Jamieson, R.P. Maas, and M.D. Smolen. 1987. Determining statistically significant changes in water pollutant concentrations. *Lake Reservoir Management* 3:195–201.

Spooner, J., S.A. Dressing, and D.W. Meals. 2011a. *Minimum detectable change analysis.* Tech Notes #7. Prepared for U.S. Environmental Protection Agency, by Tetra Tech, Inc., Fairfax, VA. Accessed March 24, 2016. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoring-technical-notes.

Spooner, J., L.A. Szpir, D.E. Line, D. Meals, G.L Grabow, D.L. Osmond and C. Smith. 2011b. *2011 Summary Report: Section 319 National Monitoring Program Projects, National Nonpoint Source Watershed Project Studies*. North Carolina State University, Biological and Agricultural Engineering Department, NCSU Water Quality Group, Raleigh, NC. Accessed March 15, 2016. http://www.bae.ncsu.edu/programs/extension/wqg/319monitoring/toc.html.

Sprague, L.A., R.M. Hirsch, and B.T. Aulenbach. 2011. Nitrate in the Mississippi River and its tributaries, 1980 to 2008: are we making progress? *Environmental Science and Technology* 45 (17):7209–7216.

Statistics Solutions. 2016. *Correlation (Pearson, Kendall, Spearman)*. Statistics Solutions, Clearwater, FL. Accessed March 24, 2016. http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/.

Stuntebeck, T.D. 1995. *Evaluating Barnyard Best Management Practices in Wisconsin Using Upstream-Downstream Monitoring*. Fact Sheet FS-221-95. U.S. Geological Survey, Madison, WI. Accessed February 5, 2016. http://pubs.usgs.gov/fs/1995/fs221-95/.

Stuntebeck, T.D. and R.T. Bannerman. 1998. *Effectiveness of Barnyard Best Management Practices in Wisconsin*. Fact Sheet FS-051-98. U.S. Geological Survey, Madison, WI. Accessed March 24, 2016. http://pubs.er.usgs.gov/publication/fs05198.

Suppnick, J. 1999. *Water Chemistry Trend Monitoring in Sycamore Creek and Haines Drain, Ingham County, Michigan 1990-1997*. Staff report MI/DEQ/SWQD-99-085. Michigan Department of Environmental Quality, Surface Water Quality Division. Accessed April 29, 2016. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/monitoring-and-evaluating-nonpoint-source-watershed.

Tetra Tech. 2013. *Preliminary Assessment of Effectiveness of the 2012 Alum Application—Grand Lake St. Marys*. Tetra Tech, Inc., Fairfax, VA. Accessed March 24, 2016. http://www.lakeimprovement.com/sites/default/files/GLSM%20Alum%20Report%2002202013.pdf.

Thas O., L. Van Vooren, and J.P. Ottoy. 1998. Nonparametric test performance for trends in water quality with sampling design applications. *Journal of the American Water Resources Association* 34(2):347-357.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Co., Reading, MA.

USDA-NRCS (U.S. Department of Agriculture-Natural Resources Conservation Service). 2012. *Assessment of the Effects of Conservation Practices on Cultivated Cropland in the Upper Mississippi River Basin*. U.S. Department of Agriculture, Natural Resources Conservation Service. Accessed March 24, 2016. http://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/technical/nra/ceap/na/?&cid=nrcs143_014161.

USEPA (U.S. Environmental Protection Agency). 1997a. *Linear Regression for Nonpoint Source Pollution Analysis*. EPA-841-B-97-007. U.S. Environmental Protection Agency, Office of Water, Washington, DC. Accessed March 24, 2016.

USEPA (U.S. Environmental Protection Agency). 1997b. *Monitoring Guidance for Determining the Effectiveness of Nonpoint Source Controls*. EPA 841-B-96-004. U.S. Environmental Protection Agency, Office of Water, Washington, DC. Accessed March 24, 2016. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/monitoring-guidance-determining-effectiveness-nonpoint.

USEPA (U.S. Environmental Protection Agency). 2007. *An Approach for Using Load Duration Curves in the Development of TMDLs*. EPA 841-B-07-006. U.S. Environmental Protection Agency, Office of Wetlands, Oceans and Watersheds, Watershed Branch, Washington, DC. Accessed March 24, 2016. https://www.epa.gov/tmdl/approach-using-load-duration-curves-development-tmdls.

USEPA (U.S. Environmental Protection Agency). 2008. *Handbook for Developing Watershed Plans to Restore and Protect Our Waters.* EPA 841-B-08-002. U.S. Environmental Protection Agency, Office of Water, Washington, DC. Accessed March 24, 2016. http://www.epa.gov/polluted-runoff-nonpoint-source-pollution/handbook-developing-watershed-plans-restore-and-protect.

USEPA (U.S. Environmental Protection Agency). 2010. *Causal Analysis/Diagnosis Decision Information System (CADDIS)*. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC. Accessed March 24, 2016. http://www.epa.gov/caddis.

USEPA (U.S. Environmental Protection Agency). 2013. Information on impaired waters and total maximum daily loads. U.S. Environmental Protection Agency, Office of Water, Washington, DC. Accessed March 24, 2016. http://water.epa.gov/lawsregs/lawsguidance/cwa/tmdl/index.cfm.

USF (University of South Florida). n.d. *Collinearity*. University of South Florida , College of Arts and Sciences, Tampa, FL. Accessed March 24, 2016. http://faculty.cas.usf.edu/mbrannick/regression/Collinearity.html.

Vollenweider, R.A. 1976. Advances in defining critical loading levels for phosphorus in lake eutrophication. *Memorie dell'Istituto italiano di idrobiologia dott. Marco De Marchi* 33:53-83.

Vollenweider, R.A., and J. Kerekes. 1982. *Eutrophication of Waters: Monitoring, Assessment and Control*. Organization for Economic Co-Operation and Development, Paris.

Walker, J.F. 1994. Statistical techniques for assessing water quality effects of BMPs. *Journal of Irrigation and Drainage Engineering* 120(2):334-347.

Walker, W.W. 1990. *Flux Stream Load Computations. DOS Version 4.4*. Prepared for U.S. Army Corps of Engineers Waterways Experiment Station, Vicksburg, MS.

Walker, W.W. 1999. *Simplified Procedures for Eutrophication Assessment and Prediction: User Manual*. Prepared for U.S. Army Corps of Engineers, Water Operations Technical Support Program, Instruction Report W-96-2, September 1996 (Updated April 1999). Accessed March 24, 2016. http://www.wwwalker.net/bathtub/Flux_Profile_Bathtub_DOS_1999.pdf.

White, W., J. Beardsley, and S. Tomkins. 2011. *Waukegan River Illinois National Nonpoint Source Monitoring Program Project*. Contract Report 2011-01. , University of Illinois at Urbana-

Champaign, Institute of Natural Resource Sustainability, Illinois State Water Survey, Champaign, Illinois. Accessed March 24, 2016. http://www.isws.illinois.edu/pubdoc/CR/ISWSCR2011-01.pdf.

Whitfield, P.H. 1983. Evaluation of water quality sampling locations on the Yukon River. *Water Resources Bulletin* 19(1):115-121.

Whitfield, P.H. and P.F. Woods. 1984. Intervention analysis of water quality records. *Water Resources Bulletin* 20(5):657-668.

Wilm, H.G. 1949. How long should experimental watersheds be calibrated? *Transactions of the American Geophysical Union* 30(2):272-278.

Zamyadi, A., J. Gallichand, and M. Duchemin. 2007. Comparison of methods for estimating sediment and nitrogen loads from a small agricultural watershed. *Canadian Biosystems Engineering* 49:1.27-1.36.