University of Tennessee, Knoxville
Trace: Tennessee Research and Creative Exchange

Doctoral Dissertations

Graduate School

8-2003

Assessment and Redesign of the Synoptic Water Quality Monitoring Network in the Great Smoky Mountains National Park

Kenneth Ray Odom
University of Tennessee - Knoxville

Recommended Citation

Odom, Kenneth Ray, "Assessment and Redesign of the Synoptic Water Quality Monitoring Network in the Great Smoky Mountains National Park. " PhD diss., University of Tennessee, 2003.
https://trace.tennessee.edu/utk_graddiss/2193

To the Graduate Council:

I am submitting herewith a dissertation written by Kenneth Ray Odom entitled "Assessment and Redesign of the Synoptic Water Quality Monitoring Network in the Great Smoky Mountains National Park." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Civil Engineering.

R. Bruce Robinson, Major Professor

We have read this dissertation and recommend its acceptance:

Chris D. Cox, William Seaver, Bruce A. Tschantz

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council

I am submitting herewith a dissertation written by Kenneth Ray Odom entitled "Assessment and Redesign of the Synoptic Water Quality Monitoring Network in the Great Smoky Mountains National Park." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Civil Engineering.

R. Bruce Robinson
Major Professor

We have read this dissertation
and recommend its acceptance:

Chris D. Cox

William Seaver

Bruce A. Tschantz

Accepted for the Council:

Anne Mayhew
Vice Provost and Dean of
Graduate Studies

(Original signatures are on file with official student records.)

ASSESSMENT AND REDESIGN OF THE SYNOPTIC WATER QUALITY MONITORING NETWORK IN THE GREAT SMOKY MOUNTAINS NATIONAL PARK

**A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

Kenneth Ray Odom

August 2003

Dedication

This dissertation is presented to my wife, Shanda, and my boys, Nolan and Garrett, who graciously supported me and gave me the time to chase a dream, and to my parents, Bud and Brenda Odom, for raising me, supporting me, and always telling me that I could do it, and to my late grandfathers James T. Cook and Olen Odom who both served our country in WWII to preserve freedom and grant me this opportunity.

Acknowledgments

I wish to thank Dr. Bruce Robinson for his guidance and support through the development and writing of this dissertation. He spent countless hours with me through this process and was very patient and understanding.

I also want to thank Steve Moore, fisheries biologist, with the National Park Service in the Great Smoky Mountains National Park for serving as an ex officio member of my committee and helping to make this dissertation a tool that can be used to monitor and protect our national park lands.

Many thanks are also extended to my committee members Dr. Bruce A. Tschantz, Dr. Chris D. Cox, and Dr. William Seaver for their time spent with me on this dissertation and their classroom instruction.

A special thanks goes to Roger Campbell, Assistant City Manager for the City of Maryville, Tennessee, who provided stedfast encouragement and support from the first day that I entered graduate school.

I also wish to thank Gary Hensley, Greg McClain and my fellow employees of the City of Maryville, Tennessee for support and time off during the work day to attend class and work on this dissertation.

Many thanks to Charlie and Roberta Walker for being my weekly encouragers.

Finally, thanks to my wife Shanda Odom and Pam Arnett for proofreading this dissertation and helping me to pull it together.

Abstract

The purpose of this study was to assess and redesign an existing 83-site synoptic water quality monitoring network in the Great Smoky Mountains National Park. The study involved a spatial analysis of water quality data (pH, ANC, conductivity, chloride, nitrate, sulfate, sodium, and potassium), watershed characteristics (geology, morphology, and vegetation), and collocated site information to determine which sites were redundant and a temporal analysis to determine the effectiveness of the current sampling frequency to detect long-term trends. The spatial analysis employed a simulated annealing algorithm using the variable costs of the network and the results of multivariate data techniques to identify an optimized subset of the existing sampling sites based on a maximization of benefits. A second simulated annealing algorithm was created to identify optimum user-defined monitoring networks of *n* sites and to validate the results of the first simulated annealing program. The first simulated annealing program identified an optimized network consisting of 67 of the existing 83 sampling sites. The second simulated annealing algorithm bracketed the same 67 sites and also provided a basis for an ordered discontinuation of sampling sites by identifying the best ten-site monitoring network through the best 70-site monitoring network.

The temporal analysis employed the “effective” sample method, Sen's slope estimator, Mann-Kendall test for trend, and a boxplot analysis to determine the effectiveness and the power of the current sampling frequency to detect long-term trends. The results showed that the current sampling frequency of four samples per year presents a low statistical power for short historical records. However, increasing the sampling frequency to more than 12 samples per year creates serial dependence between samples.

By combining the results of the spatial and temporal analyses a new network is proposed by dividing the network into primary, secondary, and tertiary sites with sampling frequencies of six and 12 samples per year. Seventeen new sites are also proposed to collect additional data above 3000 feet MSL because the existing number of sampling sites is not proportional to park area in certain elevation ranges.

TABLE OF CONTENTS

CHAPTER **PAGE**

LIST OF TABLES

TABLE **PAGE**

TABLE **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

FIGURE **PAGE**

Executive Summary

Introduction

This study was performed to assess and redesign the synoptic water quality monitoring network in the Great Smoky Mountains National Park. The process used for conducting this study focuses on improving and preserving the long-term trend monitoring capabilities so that the future of natural resources can be effectively managed. The data used for the spatial analyses were the means of water quality variables (pH, ANC, conductivity, chloride, nitrate, sulfate, sodium, and potassium), watershed characteristics (geology, morphology, and vegetation), and collocated sampling site information. The sampling frequency analyses were conducted using high frequency sampling data from the Noland Divide, southwestern stream.

Spatial Analyses

The current network includes 83 stream sampling sites and seven high elevation springs for a total of 90 sites that costs approximately $139,575 per year to operate. The total cost of the network can be divided into approximate fixed and variable costs of $70,410 and $69,165, respectively. The variable cost of $69,165 can be broken down further to $50,000 for laboratory costs and database management, and $19,200 for collecting the samples. The variable costs are based on the current sampling frequency of four samples per year for each site. Benefits are assigned to the network based on a 20 percent return of the variable costs (multiplier of 1.2 applied to the variable costs), or $60,000 for laboratory costs and database management, and $23,040 for the collection of samples.

The 83 stream sampling sites were the focus of this study; however, recommendations were also made as to the future of the high elevation springs. The existing stream sampling sites were assessed using the multivariate statistical techniques of principal components analysis (PCA), cluster analysis (CA), and discriminant analysis (DA) to determine the degree of data-redundancy among the sampling sites. Some redundancy in the data is needed; however, too much may require costs that result in only a very small increase in information. The money from excessive data-redundant sites might be better spent on increasing the sampling frequency or adding new sampling sites to the existing network. The collocated site information and the centroid distances from the cluster analyses of the water quality variables and the watershed characteristics were used for apportioning the variable cost of $83,040 to each sampling site. After benefit and cost assignments have been made the network is optimized using simulated annealing by maximizing the net benefits (benefits – costs).

The spatial analyses resulted in a priority ranking of sites into primary, secondary, and tertiary groups. The primary group consists of the following sampling sites: 4, 13, 14, 20, 23, 24, 30, 34, 47, 49, 66, 71, 73, 74, 114, 137, 142, 143, 144, 147, 148, 149, 173, 174, 191, 193, 194, 233, 234, 237, 251, 252, 253, 266, 268, 291, 293, 488, 489, 492, and 493 . The primary group is made up mainly of the 40 best sites from the simulated annealing analysis. Representation from each water quality cluster, elevation class, and watershed characteristic cluster is maintained in the primary group of sampling sites.

The secondary group consists of sampling sites: 115, 156, 184, 192, 221, 310, 311, 337, 472, 473, 479, 480, 481, 482, 483, 484, and 485 . These sites are needed to guarantee that at least three sampling sites from each cluster and elevation class are represented in the redesigned network (if each cluster or class contained three or more sites originally).

The tertiary group consists of sampling sites: 50, 52, 150, 186, 190, 209, 210, 213, 214, 215, 474, and 475 . The addition of these sites would complete the optimal network formed by the initial simulated annealing that produces the maximized net benefit. At this point these sites are recommended for discontinuation.

It is recommended that 17 additional new sites be added to the network. The elevation range from 3000 feet MSL and above is not well represented in the monitoring network in terms of the proportion of sampling sites to park area in this elevation range. The addition of these sites would correct the proportionality.

Sites definitely recommended for discontinuation because of their ranking in the simulated annealing include: 1, 3, 43, 45, 46, 103, 104, 106, 107, 127, 138, 200, and 336. However, the NPS will need to make the final decision on which sites are to be discontinued. A detailed view of all these sites are shown in Figures 40 and 41.

Frequency Analyses

The study also assessed the sampling frequency using the “effective” sampling method, Sen's slope estimation method, Mann-Kendall test for trend, and a boxplot analysis of the Noland Divide, southwestern stream data, to determine the effectiveness of the current sampling frequency to detect long-term trends. The "effective" sample method was used to determine the maximum number of independent samples that could be collected each year. The remaining methods were used to determine the adequacy of larger sampling frequencies to identify trends that were found in the original weekly sampling data.

An increase in the sampling frequency may increase the effectiveness of trend detection to a point, but a sampling frequency that results in the collection of dependent samples produces no additional information. An increase in the sampling frequency up to the collection of the maximum number of independent samples in a a given year will nevertheless decrease the amount of time required to reliably detect trends with a specified statistical confidence and power.

The "effective" sample method revealed that the maximum number of independent samples per year ranged from a low of eight for potassium to a high of 23 for sulfate. Disregarding potassium, since potassium is not regarded as a robust measure of stream health, the range is between ten and 23 samples per year with the most of the variables being between 10 and 13 samples per year. This would seem to indicate that the frequency of sampling should be at least six samples per year but not more than 12 samples per year. For a sampling record of 15 years with a specified statistical confidence and power of 95 and 90 percent, respectively, the detectable trend level would be approximately 1.4 for a sampling frequency of four samples per year. If the sampling frequency were increased to 12 samples per year the detectable trend level would decrease to approximately 0.8. In other words an increase in sampling frequency from four to 12 samples per year would allow the detection of a trend slope, with the same confidence and power, that is 43 percent less.

Sen's slope estimation method, Mann-Kendall test for trend, and the boxplot analyses showed that on average a sampling interval of bimonthly to monthly can detect the trends that are evident in the weekly sampling data at Noland Divide. There are some exceptions to this but remember that a sampling frequency greater than 12 samples per year will result in samples that are serially dependent.

Recommendations

The following recommendations and costs are presented based on the results of the spatial and frequency analyses:

- Sample the primary sites [4, 13, 14, 20, 23, 24, 30, 34, 49, 66, 71, 73, 74, 144, 147, 148, 149, 173, 174, 193, 194, 233, 237, 251, 252, 253, 266, 268, 291, 293, 488, and 489] using a sampling frequency of 12 samples per year. (+ $73,728 variable cost)
- Sample the primary sites [47, 114, 137, 142, 143, 191, 234, 492, and 493] using a sampling frequency of 6 samples per year. (+ $10,368 variable cost)
- Sample the secondary sites [115, 156, 184, 192, 221, 310, 311, 337, 472, 473, 479, 480, 481, 482, 483, 484, and 485] using a sampling frequency of six samples per year. (+ $19,584 variable costs)
- Add 17 new sites above elevation 3000 feet MSL with a sampling frequency of six samples per year. (+ $19,584 variable cost)

- Tertiary sites [50, 52, 150, 186, 190, 209, 210, 213, 214, 215, 474, and 475] could be considered on a site-by-site basis with a sampling frequency of six samples per year. (+ $1154 per site per year if included) (+ $13,848 variable cost if all are included)
- Discontinue sites: 1, 3, 43, 45, 46, 103, 104, 106, 107, 127, 138, 200, and 336 . (- $9,984 savings from existing network if all are discontinued and based on current sampling frequency of four samples per year)
- Discontinue the sampling of high-elevation springs. (- $5,384 savings from the existing network if all are discontinued and based on current sampling frequency of four samples per year)
- Cost summary based on the recommendations above:

Existing monitoring network (fixed and variable)		$139,575
Redesigned monitoring network		
Primary sites (variable costs)	$84,096	
Secondary sites (variable costs)	$19,584	
New sites (variable costs)	$19,584	
Fixed costs	$70,410	
TOTAL redesigned monitoring network		$193,674
Net increase from existing to redesigned monitoring network		$54,099

Chapter 1 INTRODUCTION

1.1 Purpose of this Study

The purpose of this study is to statistically assess the existing synoptic stream sampling network of the Great Smoky Mountains National Park (GRSM), make recommendations on which of these sites could be discontinued without a significant loss of data, and to determine if the sampling frequency is adequate to detect long-term trends in water quality variables. Since 1988, the number of sampling sites has been reduced from 300 to 90; eighty-three samples are taken from streams and seven are taken from high-elevation springs. The National Park Service (NPS) is now faced with further reductions in sampling locations because funding has not kept pace with inflation and existing funds. To assist the NPS in this decision-making process, multivariate statistical procedures will be applied to the current network to determine the redundancy of data that is being produced between sampling sites across the monitoring network. Time series techniques were applied to the weekly sampling data from Noland Divide to determine an optimal sampling frequency. These measures will ensure the effectiveness of the network while maximizing the benefits and minimizing the costs.

1.2 Objectives of this Study

Six objectives were integral to the purpose of this study and include both the spatial and temporal analyses of the data. These objectives are as follows:

- Determine the degree of data-redundancy of water quality variables between the sampling sites using Principal Components Analysis (PCA), Clustering Analysis (CA), and Discriminant Analysis (DA).

- Assess the similarities of the sampling sites based on the morphology, vegetation, and geology of the watersheds in which the sampling sites are located using PCA, CA, and DA.
- Assign benefits to collocated sampling sites where auxiliary information on fish and benthic organisms are collected.
- Develop a simulated annealing (SA) optimization algorithm that will integrate the results of the above analyses to maximize the benefits and minimize the costs of the network by identifying sites that could be discontinued without a significant loss of information.
- Evaluate the current sampling frequency using Sen's slope method, Mann-Kendall test for trend, "effective" sample method, and box-plot comparison.
- Compile the spatial and temporal findings and make final recommendations to the NPS.

1.3 Justification for this Study

The history of this network has been a reduction in the number of sampling sites and this study is needed to assess its current state before further changes are made. If a sampling network is facing a reduction in the number of sampling sites, analyses should be performed to ensure that the effectiveness of the network is not placed in jeopardy. Historical records also show that sampling for this network has been conducted at irregular intervals in the past, especially in the early years. An assessment of the current monitoring network will assist the NPS in making confident decisions on which sampling sites to discontinue and if the current sampling frequency is adequate to detect trend. Also under consideration will be the cost effectiveness of collecting data and the auxiliary needs at each sampling site.

Chapter 2 REVIEW OF LITERATURE

The literature review for this study has focused on five key areas:

- Design issues of a new monitoring network
- Considerations for redesign of an existing monitoring network
- Historical perspectives of network design
- Review of proposed statistical methods
- Network optimization using heuristic methods

The literature review for this project began by searching the following electronic databases:

- Applied Science and Technology Index
- Chemical Abstracts
- GeoRef
- Web of Science

Keyword phrases used in the search included: "principal component analysis," "water quality network," "sampling network," "network design," "water sampling," "hydrologic network," "water quality monitoring," "network optimization," "stream sampling," and "redundancy." There were other keyword phrases used but these, or a combination of these, led to most of the literature used in this study. Approximately 80 percent of the references are from refereed journals while the remaining 20 percent are from books, publications, and technical reports.

2.1 Introduction

This study involves the redesign of an existing monitoring network. However, knowledge of the requirements for building a new monitoring network remains very important because of the goals or objectives that are formed during the design process. In the case of the network considered in this study, the primary goal is to be able to detect long-term trend. This is not to ignore the fact that this sampling network must also provide an understanding of the drivers of water quality, the suitability of water quality for aquatic life, and assistance to the NPS for managing resources. The following literature review presents a synopsis of the major works used to support the concepts of this study. It also examines some of the methods used in the past to accomplish some of these same objectives.

2.2 Design Issues of a New Monitoring Network

Monitoring of environmental variables provides a strong basis for being able to make decisions about natural resources and conservation, and for monitoring changes of these variables (Moss, Lettenmaier, et al. 1978). It is perhaps the only way that these issues can be addressed directly. The main purpose of the monitoring network must be realized early in the planning process so that goals or objectives can be determined properly. However, the importance of the main purpose and clearly defined objectives is often overlooked during the planning process (Palmer and Mackenzie 1985). There is no single approach to designing a monitoring network (Palmer and Mackenzie 1985) and for good reason. The physical systems are very diverse and produce innumerable responses from various inputs (Lettenmaier 1979). However, in recent years more attention has been focused on network design (Reinelt, Horner, et al. 1988). Sampling networks have been established to monitor water quality for a number of purposes. Moss (1979) and Whitfield (1988) list some goals of monitoring networks. These are:

- Determination of trends
- Compliance with regulations or stated objectives
- Estimation of mass transport
- Assessment of environmental impacts
- General surveillance

It is easy to see that each one of these goals requires a different sampling scheme. For example, determination of trends would probably require a lower sampling frequency but over a much longer period of time than a network where compliance with regulations is the main goal. This may require sampling at a high frequency over a short period of time. The spatial pattern of sampling is also likely to be very different. Long-term trend sampling may require a broad network over a large area while compliance sampling may require a high density network over a small area. These issues and the primary goal of the network are a prelude to determining the number and location of sampling sites, the type of data needed, the length of recordation required, and the sampling frequency (Moss, Lettenmaier, et al. 1978; Reinelt, Horner, et al. 1988; Ward 1989). Furthermore, quantitative answers to these issues guided by the goal of the network will help to guarantee the success of the network by providing information that is understandable and usable (Moss, Lettenmaier, et al. 1978; Ward 1989).

The process of designing a monitoring network is very complicated and requires many resources. All these issues point to a planning process that requires knowledge from different fields of expertise including chemistry, biology, hydrology, and statistics. It is also very important for the decision makers to be included and to fully understand the goals of the network, its limits, and the information that it will produce early in the design process (Messer 1989).

Another issue that needs to be known early in the design process is the available funding. The cost of the monitoring will be affected by a number of issues including:

- Number and location sampling sites
- Sampling frequency
- Water quality variables measured

Lettenmaier (1978) suggested that sites should be located to minimize travel and collection time. It should be noted, however, that he puts this statement at the end of a list of issues to stress its lack of importance when compared to other issues. Some of the considerations for station selection stated by Smith and McBride (1990) are:

- Accessibility
- Transport time from the field to the lab
- Ease and safety of sampling

In the absence of data to use for spatial analysis, much of the literature infers that station location for a new network is somewhat subjective and should be guided by the goals of the network. If no prior water quality data are available at the location of the proposed network, the preliminary design locations may be chosen using a method proposed by Sharp (1970; 1971). He suggested that stream order could be viewed as a degree of uncertainty and, therefore, can be used to locate sampling sites. Once data are collected, the monitoring sites can then be analyzed for redundancy (Langbein 1979; Whitfield 1988) to determine if sites can be discontinued. If the data between sites are not redundant, this may indicate that there is a shortage of sampling sites. Lettenmaier (1978) suggests a number of factors to consider for locating long-term trend detection sampling sites. Of those factors, the most important for this project are:

- Sites should be located to monitor as much of the watershed as possible
- Sites should be selected based on the length and quality of data already recorded at that particular site (if historical data exists)
- Sites should be located so that travel and collection time is minimized

The first factor is to make certain that as much of the watershed as possible is monitored. An objective of this study is to ensure that selected watersheds are sampled from the upper elevations to the lower elevations resulting in a stream profile of the measurement variables. Liebetrau (1979) suggests that a stratified sampling scheme with respect to physiographic or topographic attributes be used. Harwell's (2001) findings of different trends at different elevation classes provides a good basis for this scheme. The second factor pertains to length and quality of data. The record lengths for the sites in this study are the same for the most part and the quality of the data is currently very good. However, a cursory review of record lengths and data quality from the early years is still recommended and may play a role in whether or not a sampling site is discontinued. The third factor is cost. Funding cutbacks and/or redirection of existing funds are resulting in downsizing of existing networks. Methods are needed to guarantee that the goals of the sampling network are not compromised.

Sampling frequency should be long enough to minimize costs yet short enough to capture the natural variability in the data (Whitfield 1988; Biswas 1997). Whitfield suggests that initially a high sampling frequency should be used. After a "reasonable" amount of data is gathered, an analysis can be performed to determine if adjustments can be made to the sampling frequency. In the case of a new network where no historical data are available, it may be possible to find a nearby site with similar morphological characteristics that is being sampled. These data can then be used to estimate an initial sampling frequency. This method was used by Lachance, Bobée, et al. (1989) to determine the sampling frequency on a previously unsampled lake in Québec. Lettenmaier (1978) recommends a method based on the serial correlations of an AR(1) model to calculate the "effective" independent sample size. This method will be discussed in greater detail in Section 4.4.

As data are collected, the network can be assessed to determine if the number of sampling sites or the sampling frequency need to be adjusted (Langbein 1979; Whitfield 1988). As mentioned earlier, both of these factors are very important, especially from an economic standpoint.   From an operational standpoint, it may be that some sites are redundant and can be discontinued. On the other hand, there may be a need for additional sites. Sampling frequency may also need to be adjusted because of changes in trend magnitude, seasonality, or variability.

2.3 Considerations for Assessment of an Existing Monitoring Network

Assessment of an existing monitoring network is usually easier than designing a new network because of the data available. Recalling some of the issues raised previously regarding sampling locations and frequency, one can see how collected data can make these determinations less subjective and mainly quantitative. Langbein (1979) uses the term "audit" to define redesigning or assessing an existing monitoring network. He suggests the following outline for auditing a monitoring network:

- Describe the purpose of the network, all associated costs, and funding resources
- Define the objectives of the network and how they may have evolved over time
- Identify sources of error (measurement, model, recordation, sampling, etc.)
- Analyze results, efficiency, and redundancy
- Study the implications of the results of the network

Some of these issues, especially those that are operational in nature, should probably be evaluated after one year of operation while others may require a large amount of data from the field. This would allow for a full understanding of their behavior and ability to make sound judgments. It should be noted that such an assessment with limited data (only one year) should not include a determination of whether or not the number or location of sites should be adjusted. In addition, the sampling frequency should not be changed unless there is strong evidence to do so. Smith and McBride (1990) assessed the nationwide monitoring network in New Zealand one year after operation began. This network consisted of 77 river sampling sites and 30 lake sampling sites sampled monthly and bimonthly, respectively. Their work focuses on the state of staff training, accuracy, laboratory and field operations, QA/QC procedures, and data management. These all involve the reduction and minimization of error in the laboratory and field, which is listed as the third outline point above by Langbein.

The issues of costs and funding in the first outline point above are part of the main focus of this study. The second point involves assessing the stated objectives and determining if they have evolved as the sampling has progressed. The main objective of this project is clear and has remained so. However, there have been redirections of some of the secondary objectives from time to time because of other water quality needs such as collocated studies. The fifth and final point is being addressed in ongoing work at the University of Tennessee.

Costs of operating monitoring networks are constantly under the scrutiny of decision makers who control the funding; therefore, it is of utmost importance that the information gained from the network be efficient and cost-effective. Cost-effectiveness has not been a major focus of many monitoring network designs (Sanders and Adrian 1978; Lettenmaier, Anderson, et al. 1984; Mackenzie, Palmer, et al. 1987). Ward (1996) recognizes that the decision-makers, as well as the informed taxpayer want to know what they are getting for their money. This statement should not cause one to immediately think that monitoring must be held to a bare minimum. What it should imply is that a monitoring system must be accountable for producing high quality information in a cost-effective manner. The definition of high quality in this case would be that the goals of the system are continuously being satisfied. Periodic assessment of the monitoring system is needed to stay on-track with the system goals. It must be noted that the system should not be assessed so often that there is a possibility of data fragmentation (Lettenmaier, Conquest, et al. 1982). A good example of an assessment and consolidation of an existing network is where Lettenmaier, Anderson, et al. (1984) worked with the City of Seattle to reduce their sampling sites from 81 to 47. He combined a scoring system that included information about salmon habitat, recreation, fecal coliform counts, years of record, and basin information with an optimization routine that maximizes scores for sampling sites divided into primary basins that discharge into Puget Sound. This ensured that all primary basins would be assigned at least one sampling site. The total cost savings was approximately $87,000 per year, which allowed the City of Seattle to conduct needed short-term, intensive monitoring at other locations.

It is clear that probably the most significant and direct impact to the cost of a monitoring network is the location and number of sampling sites and the sampling frequency. The locations of the sites are normally controlled by those constraints mentioned in the previous section. The number of sampling sites and the sampling frequency, however, can now be explored further using the available data. It is worthwhile to once again mention assessment of the operational aspects of the network at the end of the first year to make certain that the data used to assess the number of sampling sites and the sampling frequency at some point will be reliable information.

2.3.1 Statistical Methods

Many types of statistical tests have been utilized over the years in monitoring programs to study water quality data and to determine if trends exist. Monitoring programs range in size from small municipal projects, seeking to meet the requirements of the EPA's NPDES Phase I and II storm water mandates, to nationwide programs such as the U.S. Geological Survey's NAWQA (National Water-Quality Assessment) network. Many of the smaller programs are just beginning because of the more recent EPA requirements, while the NAWQA program began in 1991 with 137 surface water networks, and a planned addition of 21 new networks by 2009 (Mueller, Lapham, et al. 2002). The marked increase in monitoring programs demonstrates the concern placed on the nation's water quality. Reliable techniques are needed to insure that the most cost-effective data are obtained in these programs.

Statistical methods have been used to treat a variety of data from univariate to multivariate cases. However, it seems that the majority of the focus found in the historical literature has been placed on univariate and bivariate data. Ideally, users of statistical methods find the greatest ease when the water quality data are linear, normal, independent and identically distributed with no outliers. However, this is almost never the case. Over the past several decades, many water quality researchers have developed new approaches for dealing with problems such as non-normal data, non-linearity, seasonality, outliers, dependence, irregular spaced intervals, missing data, and serial correlation (Lettenmaier 1976; Lettenmaier 1978; Lettenmaier 1979; Liebetrau 1979; Moss 1979; Hirsch, Slack, et al. 1982; McLeod, Hipel, et al. 1983; Whitfield 1983; Hirsch and Slack 1984; Ward and Loftis 1986; Berryman, Bobée ,et al. 1988; Hirsch 1988; Whitfield 1988; Hirsch, Alexander, et al. 1991; Somerville and Evans 1995; Thas, Van Vooren, et al. 1998; Urquhart, Paulsen, et al. 1998). One important facet of almost every new approach is the need to determine the statistical power of the test performed, and the sensitivity of the statistic to changes in the variable of interest. Many of the previously mentioned authors have addressed this concern. The methods by the aforementioned authors are mainly evolutions of earlier statistical techniques and by now have become quite commonplace. Therefore, to focus more on the main objectives of this study, a detailed review will not be presented here but the reader is urged to consult the prior credited sources if more information is desired.

2.3.2 Long-term Trend Monitoring

Since long-term trend detection is the primary focus of the monitoring network of the Great Smoky Mountains National Park, it is necessary to examine some of the research that focuses on this aspect. Trends of water quality variables are often determined using a linear approach or a step approach. The linear approach, of course, almost always involves a regression analysis. The step approach involves determining the change in the mean of a water quality variable from one time period to another. Time periods can range from seasonal to multi-year periods. Lettenmaier (1978) indicates that determining a step change in mean levels from one year to the next may provide more insight than determining a linear trend because biochemical processes tend to seek an equilibrium. The known point of equilibrium is, however, affected by exogenous variables. The desired knowledge of a trend, whether linear, step, or both, would seem to depend on the timescale of interest. In either case, a more important aspect would be the reliability or the statistical power of the trend detection. Increased statistical power is obtained by increasing the length of record and by having a relatively higher trend magnitude (Lettenmaier 1978; Somerville and Evans 1995;

Urquhart, Paulsen, et al. 1998). The latter is, of course, uncontrollable. The former is ensured by installing a sampling frequency that obtains 80 to 90 percent of the maximum number of independent samples for a given sampling site (Lettenmaier 1978; Lachance, Bobée, et al. 1989). Sampling beyond this number yields little added value and statistical adjustments must be made to account for serial dependency (Lettenmaier 1976; Lettenmaier 1978; Sanders, Ward, et al. 1983; Hirsch and Slack 1984; Lachance, Bobée, et al. 1989). Seasonality in the data is often another concern in time series analysis. Many traditional non-parametric tests, such as the Mann-Kendall test for trend, have been adapted to deal with seasonal trends. This is usually performed by dividing the data into *n* seasons and calculating *n* trends (Helsel and Hirsch 1992). If seasonal variation in trends is not as important as simply understanding the long-term trend, then decomposition, smoothing, or seasonal differencing could be used to remove seasonality from the data before performing trend tests.

2.4 Historical Perspectives of Network Design

This section will explore some of the historical methods used for network design, redesign, and trend detection. A review of past network design techniques is necessary for knowing the pitfalls and successes that have occurred and may also lie ahead. Many techniques were developed with no particular method that is applicable to all situations. This is understandable because of the seemingly infinite number of conditions that are possible in the field. As mentioned in the previous section, the number and location of sampling sites and the frequency at which those sites are sampled more directly influences the costs associated with a monitoring network. Each of these aspects is designed according to the goals of the network. In the GRSM monitoring network the primary goal is long-term trend detection. The cursory review of past techniques to follow will focus on these same aspects as they relate to long-term trend detection. A more in-depth review of these plus some of the lesser-known techniques will follow in the dissertation.

In the past, approaches to network design have included both statistical and non-statistical methods (Lettenmaier 1978; Hughes and Lettenmaier 1981; Ward and Loftis 1986; Husain 1989; Lachance, Bobée, et al. 1989; Harmancioglu and Alpaslan 1992). These methods focus on either sampling number and location or frequency, while others focus on both.

One statistical approach where prior data are needed is the entropy-based method. Harmancioglu, Fistikoglu, et al. (1999) defines entropy as 'a measure of the degree of uncertainty of random hydrological processes.' Using this method, the amount of transinformation between sampling sites is determined based on the degree of uncertainty (Husain 1989; Harmancioglu and Alpaslan 1992; Harmancioglu, Fistikoglu, et al. 1999). Two or more sampling sites that are thought to produce serially independent data may actually produce data that are dependent. Serial dependence between sampling sites results in reduced entropy, or uncertainty, between the sampling sites (uncertainty and entropy can be used interchangeably). If over time the serial dependence is consistent, one or more of the sampling sites may be discontinued with a minimal loss of information. In simple terms, one sampling site is producing the same information as another sampling site. So the question must be asked – why are both sites needed? Although not specifically a parametric approach, the knowledge of the type of probability distribution is needed for a univariate case. The data must be normal or lognormal for the multivariate case. The entropy-based approach does not handle other skewed distributions with multivariate data very well (Harmancioglu and Alpaslan 1992; Yang and Burn 1994). The following is a brief description of how the entropy technique is applied by Harmancioglu and Alpaslan (1992) and Harmancioglu, Fistikoglu, et al. (1999). For a multivariate case and assuming a multivariate normal probability in the data, the joint entropy of **X** is defined by

$$H(X)=\left(\frac{M}{2}\right)\ln 2\pi+\left(\frac{1}{2}\right)\ln|C|+\frac{M}{2}-M\ln(\Delta x) \tag{2.1}$$

where **X =** vector of *M*-variables
|C| = determinant of the covariance matrix
Δx = class interval size for the M-variables

This equation results in a single value that expresses the joint entropy over the whole network for M variables. Marginal entropy for each sampling site is computed using the same equation but letting M = 1, adding the subscript *m* to **X** to represent each sampling station, and substituting the variance ($\boldsymbol{\sigma}^2$) for |C|. Calculations must now be performed one variable at a time. The station with the largest degree of uncertainty yields the largest marginal entropy. The information from this unique site is then compared with every other site, one by one, and the conditional entropy is calculated until the pair with the greatest conditional entropy is found. The pair are then joined with the remaining sites, one by one, until the set of three with the greatest conditional entropy is found. This process continues until the last site is joined. The procedure ultimately provides a ranking of the sites from highest to lowest degree of entropy.

After repeating the entire process for each variable the sites can be evaluated for discontinuance starting at those stations with the lowest degrees of entropy.

The entropy method can also be used to assess sampling frequencies (Harmancioglu and Alpaslan 1992; Harmancioglu, Fistikoglu, et al. 1999). The same equation shown above can be used to calculate the marginal entropies in the temporal dimension using different time lags. The change in entropy from the initial time to the *k*th lag is calculated. A point where the change in entropy becomes negligible indicates a time window where the sampling frequency may be adjusted. The entropy method has been applied to multivariate data but only one variate is analyzed at a time. True handling of multivariate data should be conducted simultaneously rather than sequentially.

Husain (1989) used the entropy concept to identify the most crucial rain gaging stations in the Sleeper River Research Watershed in Beltsville, Vermont. Use of this method would prevent decision-makers from unknowingly removing the most important gaging stations should a reduction in network size become necessary. It also identifies areas where additional gages may need to be installed because the information between adjacent sampling stations had a high degree of uncertainty. This is a good example of where economic decisions and information quality can be coordinated to preserve the goals of the network. Harmancioglu, Fistikoglu, et al. (1992) used the entropy concept to assess the station locations and the sampling frequency of an existing network using historical monthly records for conductivity, dissolved oxygen and chloride on the Porsuk River in Turkey. They found that conductivity requires more frequent sampling than dissolved oxygen or chloride but failed to state exactly what the recommended frequency should be. Their spatial study revealed that the six stations in operation were slightly more than needed to explain 90 and 95 percent of the uncertainty for chloride and dissolved oxygen, respectively; however, the six stations explained only 35 percent of the uncertainty for conductivity meaning that additional sampling sites are needed.

Another approach that has seen a considerable amount of use in this and other fields is kriging, which is a technique used to interpolate spatial information (Hughes and Lettenmaier 1981). This technique is an optimization (non-statistical) approach and has been more widely used in the fields of groundwater and mining; however, many researchers (Hughes and Lettenmaier 1981; Eynon 1988; Venkatram 1988; Jager, Sale, et al. 1990; Ben-Jemaa, Marino, et al. 1995; Christensen, Phoomiphakdeephan, et al. 1997; Huang and Yang 1998) have adapted its use to surface water quality, precipitation networks, precipitation chemistry, and stream flow. In a very basic sense, the point estimate at one location is determined using a linear combination of all other observations, expressed as

$$\hat{Y}_0 = \sum_{j=1}^{N} \lambda_j Y_j \qquad (2.2)$$

and

$$\sum_{j=1}^{N} \lambda_j = 1 \qquad (2.3)$$

where $\hat{Y}_0$ = point estimate of Y_0
$\boldsymbol{\lambda}_j$ = weighting factor for the *j*th location
Yj = known observation at the *j*th location
N = total number of sampling locations

The weighting factors must be obtained using a constrained optimization technique that utilizes Lagrange multipliers to minimize the error variance ($\hat{Y}_0$ - Y_0) while working within the specified constraints, one of which is shown above. The accuracy of these equations is highly dependent on the generalized covariances (Hughes and Lettenmaier 1981). This ultimately results in a system of *N* equations that must be solved using matrix algebra to determine the weighting factors. If locations where the variables are known can be predicted with a high degree of accuracy using the other sites, then the intermediate site may be discontinued. If intermediate locations cannot be predicted, it may be an indication that additional sampling sites are needed. Kriging has evolved into cokriging and universal cokriging, which can handle multiple variables and time trends (Isaaks and Srivastava 1989). The method is data-dependent because it has no advantage over other techniques when the number of observations is small (Hughes and Lettenmaier 1981). This method is classified as a redesign tool since prior data is needed to determine if a sampling location can be successfully predicted using the data from the remaining sites. For the case of trend detection, Lettenmaier (1979) indicates that kriging may be well suited for network design where changes in trend are expected to occur slowly.

Sanders and Adrian (1978) developed the following equation for determining the sampling frequency at a single sampling site where multiple water quality variables are measured. This technique could be useful at a site such as Noland Divide. Equation 2.4 provides “a sampling frequency such that the average 95 percent confidence

$$n = \left[(2)(2)(1.96)\frac{\sum_{i=1}^{k}\sigma_i}{\sum_{i=1}^{k}\mu_i} \right]^2 \tag{2.4}$$

where n = number of samples required per year
k = number of water quality variables considered
σ = sample standard deviation
μ = sample mean

interval width about the means will be equal to one-half of the average of the means." In short, this equation is calculating an average sampling frequency using all variables, and is weighted toward the variable with the greatest variability. One issue with this simple approach may be a scaling problem depending on the magnitudes of the variances and measurement scales of the data. However, if the issue can be resolved, the equation would perform correctly by assigning more importance to the variable with the greatest variability. Another problem with this equation is that it assumes that all observations are serially independent. Closely spaced water quality data rarely adhere to such an assumption (Lettenmaier 1976). When serial correlation is a concern, Sanders and Adrian (1978) offer another technique that was developed by Lettenmaier (1976). Using this method, sampling frequencies using the Noland Divide data will be analyzed. A more comprehensive review of this particular technique will be presented in the Section 4.3. Reviews of other methods used in the analyses of this study will also be included.

2.5 Multivariate Statistical Methods

Multivariate statistical analysis, as the name implies, involves the study of more than one variable in a concerted effort. This section of the literature review focuses on the multivariate methods that have seen the most use in studies similar to the one presented here and provides some of the background necessary for their understanding.

Multivariate statistics have had a wide-range of applications including: trend tests for water quality variables (Lettenmaier 1988); trend tests for pharmaceutical trials (Dietz and Killeen 1981); grouping of homogenous precipitation stations (Morin, Fortin, et al. 1979); identification of outliers and variability in air quality and meteorological data (Smeyers-Verbeke, Denhartog, et al. 1984; Smeyers-Verbeke, Denhartog, et al. 1984); the study of shell measurement changes over time of bivalve mollusks (Symons and Ringele 1976); industrial process monitoring at Tennessee Eastman for six process variables of chemical reactors (Chen and Liu 1999); separation of long-term trends and periodic variation for water quality variables using a functional principal components analysis (Champley and Doledec 1997); identification of correlation patterns between physical and chemical variables in urban and agricultural soils (Salman and Abu Ruka'h 1999); the interaction between pore water in peat moss with groundwater and selected chemical constituents (Reeve, Siegel, et al. 1996); correlations between trace metals and their origins from three different geologic types (Weissberg and Singers 1982); correlations between physical and chemical properties of carbonate-rock aquifers in Pennsylvania (Rauch and White 1970; Brown 1998); correlation between topography and extreme precipitation events on the island of Tahiti (Wotling, Bouvier, et al. 2000); correlations between streamflow in the western United States with El Nińo patterns (Piechota, Dracup, et al. 1997); shape and magnitude classification of hydrographs resulting from glacial runoff (Hannah, Smith, et al. 2000); evaluation of regional water quality data patterns for the state of Nebraska (Crisp 1989);

evaluation of a water quality monitoring network in Queensland, Australia (McNeil, McNeil, et al. 1989); correlations between chemical parameters and volcanic lakes (Varekamp, Pasternack, et al. 2000). As mentioned, the examples listed cover a wide-range of applications but many of them have one important concept in common with this study: principal components analysis and cluster analysis are used together.

The following sections will address in more detail the use of principal components analysis (PCA), cluster analysis (CA), and discriminant analysis (DA) and their application in this study.

2.5.1 Principal Components Analysis (PCA)

Harold Hotelling introduced PCA in 1933. Since that time the procedure has seen many uses but the main purposes include data interpretation, pattern recognition, dimensional analysis, and multicollinearity detection. PCA is a data technique rather than a statistical technique meaning that many of the commonly applied statistical inference tests are not directly applicable (Hintze 2001). PCA transforms a set of correlated variables into a smaller set of uncorrelated variables called principal components (Flury 1988; Flury and Riedwyl 1988; Dunteman 1989; Everitt and Dunn 1991; Jackson 1991; Jobson 1992; Johnson 1998; Jolliffe 2000). If the variables of interest do not have significant correlations then nothing is gained by the use of PCA because each variable has a significant portion of the total variance that only it can explain. When variables are correlated, PCA can be used to find linear combinations of the correlated variables that explain a significant amount of the total variability. The user must ultimately decide how much of the variability is needed in their analysis using the results of the eigenanalysis on the variance-covariance matrix (∑) or the correlation matrix (**R**) of the data. Kaiser (1960) suggested that principal components with eigenvalues less than one should be ignored. Jolliffe (1972; 2000) felt that this could cause the removal of still important principal components and suggested that all principal components with an eigenvalue greater than 0.7 be retained when performing the analysis on the correlation matrix. A more subjective approach to the eigenanalysis was suggested by Cattell (1966) and Cattell and Jaspers (1967). Their idea was to decide which principal components to retain for further analysis based on a scree plot of the percent of variability explained. Eigenvalues for each of the principal components are represented by bars denoting percent of variability explained. Eigenvalues are retained to a point where the bars level off and variability explained by additional principal components is minimal.

A PCA is performed using either ∑ or **R** from the original data. For the purpose of this study, the PCA is performed using the **R** matrix because the scales of measurement are different for the variables and the variances are quite different. Because a correlation matrix is always symmetric and nonnegative, the eigenvalues (λ_p) will be positive real numbers, where p is the number of variables (or the number of principal components). It is important to note that

$$\sum_{n=1}^{p} \lambda_n = p \tag{2.5}$$

where

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$$

because the variability explained by each principal component is computed by

$$\sigma_{11} = \frac{\lambda_1}{p} \tag{2.6}$$

where λ_1 is the variance explained by the first principal component. The first principal component always explains the greatest amount of variability in the data followed by second principal component and so on. Maximum variance in the first principal component is gained by calculating its linear combination based on the largest normalized eigenvector vector through the ellipsoid of concentration of the data. The next linear combination of the variables is chosen based on the second largest normalized eigenvector, which is orthogonal to the first eigenvector. Principal component scores (y_{rj}) can then be calculated from the eigenvectors of the principal components by the equation

$$y_{rj} = \hat{a}_j{}'(x_r - \hat{u}) \tag{2.7}$$

where $j = 1$ to p, $r = 1$ to N (number of observations), $\hat{a}$ is the eigenvector for the jth principal component, x is the observed value for the rth observation, $\hat{u}$ is the sample mean. The scores represent distances of the observations from the jth principal component axis and are the first step in identifying sites which are similar based on multivariate distances.

Multivariate variable selection and multiple regression in the NCSS software package are sometimes used to determine the principal original variables (of all possible variables) that formulate the principal components. In some cases a reduced set of the original variables may produce better clusters than all the variables combined. The principal variables are the source of most of the variability in the principal components and can often be analyzed as a subset without losing any significant information. NCSS uses McHenry's algorithm (McHenry 1978) for multivariate variable selection by minimizing Wilks' lambda to find the best combination of variables (Hintze 2001). Wilks' lambda is a goodness-of-fit measure similar to a correlation measure except that it is used for the multivariate case.

Multiple regression can be used by regressing the original variables from the multivariate variable selection against the principal component scores. If the residuals of the regression are normally distributed the significance of each independent original variable as a predictor can be determined. If the residuals are not normally distributed, the significance cannot be determined accurately. However, the goodness-of-fit measures (R^2 and *press*-R^2) resulting from the regression can be used to determine how well the subset of original variables predicts the dependent variable (principal component scores). To calculate *press*-R^2, observations are removed one at a time and a new regression model is created without that observation. The new regression model is then used to predict the removed observation. The procedure is performed *n* times, once for each observation removed. *Press*-R^2 is then the summation of the squared differences of the predicted value from the *N* regression equations and the value predicted using the full model. Even if the residuals are not normally distributed, the goodness-of-fit measures alone give an indication of how well the subset of principal variables predicts the principal component scores.

2.5.2 Cluster Analysis (CA)

Cluster analysis (CA) is an exploratory data technique used to group similar observations into clusters based on distances where the within-cluster variance is minimized and the between-cluster variance is maximized (Peck, Fisher, et al. 1989; Jobson 1992; Johnson 1998). In earlier years, the validity of clustering techniques were questioned because of the lack of inferential tests offered by many other statistical techniques (Baker and Hubert 1975; Wong 1982). However, additional tests such as estimation of the bootstrap confidence intervals (Peck, Fisher, et al. 1989), performing a discriminant analysis on the final clusters (Jobson 1992), or combining hierarchical and non-hierarchical clustering techniques (Jobson 1992) can provide validation for the chosen clusters. Additionally, statistical software such as SAS and NCSS has the ability to approximate statistical inference tests within clustering algorithms.

Observations can be clustered using one of several techniques that generally fall into one of three categories: (1) hierarchical, (2) non-hierarchical, and (3) fuzzy. Hierarchical methods most commonly use agglomerative techniques where each observation starts in a cluster by itself (*n* observations = *n* clusters). As the algorithm progresses, observations are joined based on a proximity measure until there is only one cluster composed of all the observations (*n* clusters = 1). Different variations of proximity measures are used for each technique. One such proximity measure, for example, is Equation 2.8 for Euclidean distance.

$$d_{rs} = \left[\left(\boldsymbol{x}_r - \boldsymbol{x}_s \right)' \left(\boldsymbol{x}_r - \boldsymbol{x}_s \right) \right]^{\frac{1}{2}} \tag{2.8}$$

A benefit of using these hierarchical methods is the generation of the dendrogram, which can be very useful in locating outliers and visualizing “good” cluster partition points based on dissimilarity measures. Agglomerative hierarchical methods do not provide a single solution (Jobson 1992) and for this reason are somewhat subjective. However, there are tools to assist the user in choosing the best clustering method and an optimum number of clusters. The most common methods of hierarchical clustering are:

- Single-linkage
- Complete-linkage
- Simple-average
- Group-average
- Median
- Centroid
- Ward’s minimum variance

For each of these methods a cophenetic correlation measure (Hintze 2001) and a *delta* (Mather 1976) can be calculated. The cophenetic correlation is simply a correlation measure between the original distances between the observations and the final distances after clustering. Hintze (2001) recommends that the clustering techniques with cophenetic correlation measures of 0.75 and greater would indicate that the technique is acceptable. The *delta* that Mather (1976) suggests is based on the degree of distortion and is given by

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} \left| d_{jk} - d_{jk}^{*} \right|^{\frac{1}{A}}}{\sum_{j<k} \left(d_{jk}^{*} \right)^{1/A}} \right]^{A} \qquad (2.9)$$

where *d* and *d** are the original distance and the distance after clustering, respectively. *A* is assigned 0.5 or 1. Δ_A values closest to zero are desired. Once the best method of clustering is chosen, the question of the number of clusters must be addressed. Jobson (1992) suggests graphing the proximity measures for each number of clusters and picking the number of clusters prior to a noticeable increase in the proximity measure. This point identifies a large increase in proximity distance that is needed to join additional observations.

Most of the hierarchical methods listed above are very susceptible to the effects of outliers and extreme observations; some are more susceptible than others. Another drawback of using the hierarchical methods alone is that once an observation is assigned to a cluster, it cannot be reassigned to another cluster (Jobson 1992). The non-hierarchical methods do allow an observation to be reassigned.

The non-parametric density estimation technique in SAS (called *MODECLUS METHOD = 6*) was developed by Koontz and Fukunaga (1972a; 1972b) and is the type of hierarchical clustering mainly used in this study (SAS Institute 1999). The nonparametric density estimation technique is inherently hierarchical and agglomerative because of the process used to form the clusters. The *nearest-neighbor* form of the non-parametric density estimation technique constructs an optimum solution of *n* clusters by iteratively joining the total number of observations until the estimated density of *n* cluster seeds is not less than any of the neighbors in the cluster. This particular technique is more robust to outliers and scale differences than other types of hierarchical methods.

As mentioned earlier, a two-step process using hierarchical and non-hierarchical methods can be used to assist in validating the clusters. Most of the analyses in this study will be conducted using this procedure. Non-hierarchical methods, such as *k*-

means, require initial seeds to be generated automatically by the software, or to be input by the user. The initial seeds may be obtained from a previous clustering method (Jobson 1992). If seeds are generated by the software the final clusters formed can be influenced by the order in which the data are read into the program (Johnson 1998). Seeding from a hierarchical method provides the non-hierarchical method with an initial partition but does not force the non-hierarchical method to strictly match the hierarchical results. Seeding merely provides a rational starting point that increases confidence in the final results and decreases the time that the non-hierarchical method takes to arrive at an optimum solution.

The non-hierarchical clustering methods used are the *k*-means (called *FASTCLUS* in SAS) and the medoid partitioning method. *K*-means clustering is a partitioning method based on the Euclidean distance where the optimum cluster centroids are obtained in an iterative fashion by minimizing the sum of squared distances between the cluster means and the cluster members. Although convergence is not required for an optimum solution, cluster means and cluster centroids are equal when convergence is achieved. The FASTCLUS algorithm in SAS uses adaptive training during cluster formulation by allowing the centroids to be updated each time a reassignment is made (SAS Institute 1999). This method was developed by Anderberg (1973) based on previous work by MacQueen (1967) and Hartigan (1975). The *k*-means clustering algorithm uses the initial seeds and clusters from the non-hierarchical results and reassigns observations until all observations are located in the cluster with the nearest centroid. *FASTCLUS* is also an effective method for locating outlying observations by placing outliers in clusters of single memberships.

Medoid clustering is similar to *k*-means in the sense that it is a non-hierarchical partitioning method. NCSS software uses two methods developed by Spath (1985) and (Kaufman and Rousseeuw 1990) to estimate the cluster memberships. Medoid partitioning finds a representative observation (called the medoid) for each cluster where the dissimilarity measure of each observation in that cluster is minimized with respect to the representative observation (Hintze 2001). Spath's (1985) algorithm reaches an optimal solution by using multiple random starting points so as to minimize the objective function where the objective function is the summation of the minimized Euclidean distances between all of the observations and their representative observations. The method by Spath overcomes local optima by using numerous random starting configurations. The algorithm developed by Kaufman and Rousseeuw (1990) first finds a representative set of observations then proceeds iteratively through the set of unselected observations in an overall effort to minimize the distance objective function.

Fuzzy clustering is similar to *k*-means and medoid partitioning with one major difference. In fuzzy clustering all observations are allowed to be a member of more than one cluster (Hintze 2001). Each observation has a membership value of one and can have partial memberships as long as the sum of the membership value is one. The membership value is actually the probability that the observation belongs to a certain cluster. Fuzzy clustering was developed in the NCSS software based on the algorithm by Kaufman and Rousseeuw (1990). Fuzzy clustering can be used initially to assist in determining the optimum number of clusters for each set of variables being analyzed. This is helpful to validate the results of other clustering methods. An objective function measure similar to the medoid partitioning method is used in fuzzy clustering to determine membership probabilities. The disadvantage of fuzzy clustering is that there is more information to interpret in the final results (Hintze 2001).

2.5.3 Discriminant Analysis (DA)

Discriminant analysis (DA) is a multivariate statistical method that is somewhat similar to CA except for one major difference. In DA the groups are already known and the user is testing the ability of a set of variables to discriminate between the various groups. DA has been developed for parametric (Rao 1973) and nonparametric (Rosenblatt 1956; Parzen 1962) cases. The nonparametric form is also known as the *k-nearest-neighbor* form and can be executed in the SAS software under the *DISCRIM* procedure. This procedure is appropriate when the data are not multivariate normal. DA can be used as a validation measure to test the discriminating ability of the clusters from a previous CA. The *nearest-neighbor* DA is based on the Mahalanobis distance measure

$$d_i=(x-\hat{\mu}_i)'\Sigma^{-1}(x-\hat{\mu}_i) \tag{2.10}$$

where an observation is classified into a cluster based on the Mahalanobis distance from a cluster mean. The cluster means are recalculated in the DA based on the cluster assignments from the CA. The DA in SAS uses two discriminating procedures: resubstitution and cross-validation. In resubstitution, DA develops a discriminate rule based on the Mahalanobis distances using the full set of cluster assignments from the CA, and then tests whether each observation can be assigned to the same cluster as in the CA. The method's one drawback is that it tends to overestimate the number of correct classifications (Johnson 1998). Cross-validation, on the other hand, removes one observation at a time and each time a new discriminant rule is developed. The observation removed is then tested against the new discriminant rule to determine whether it can be classified in the same cluster as the CA. This method is more robust than the resubstitution method because the observation is being tested against a discriminant rule that is constructed without the information of the removed observation. A high misclassification rate usually means that the CA has problems, and therefore focus is shifted back to the CA. DA can be performed twice--the first time using the principal component scores and the second using the raw data. By using the raw data, a "full-circle" validation is provided because the raw data are the source of the principal components.

2.6 Network Optimization using Heuristic Methods

Heuristic methods are often used to solve problems where an exact solution is almost impossible to obtain because of the seemingly infinite number of possible solutions. Heuristic methods employ techniques that generally use trial and error or systematic elimination approaches to find optimal solutions without having to perform an exhaustive search of all possible solutions. Some of the most common heuristic methods used are local search, gradient methods, neural networks, linear programming, greedy algorithms, dynamic programming, branch and bound, genetic algorithms, fuzzy algorithms, and hill-climbing techniques. All of these techniques usually have three basic components in common that form the algorithm: (1) definition of the objective, (2) an evaluation function, and (3) a scheme for searching possible solutions. The listed components can range from very simple to highly complex depending on what the constraints of the model are.

The type of optimization sought in this study is one of finding a combination of water quality sampling sites that forms an optimum monitoring network from an existing sampling network of 83 sampling sites. The definition of an “optimum” solution for this problem is described in detail in later sections. For now it is just important to remember that the focus is to find an optimized subset of the 83 sampling sites. Problems of this type are commonly called “combinatorial optimization” problems where an optimum vector of sites is searched for instead of an optimum scalar quantity.

Combinatorial optimization problems can be solved using most of the heuristic algorithms previously mentioned above, but some are more efficient and simpler than others. At each end of the spectrum, from very complex to relatively simple, are genetic algorithms and simulated annealing. Both of these methods were given early consideration as possible solutions to the monitoring network optimization in this study. Sections 2.6.1 and 2.6.2 will review genetic algorithms and simulated annealing. Section 2.6.3 will present some examples of simulated annealing.

2.6.1 Genetic Algorithms (GA)

One of the most popular techniques of combinatorial optimization currently used is genetic algorithms (GA). Mitchell (2001) lists the following applications of GA: (1) combinatorial optimization, (2) automatic programming, (3) machine learning, (4) economics, (5) immune systems, (6) ecology, (7) population genetics, (8) evolution and learning, and (9) social systems. This list reveals that GA have almost unlimited applications. The technique can be quite complicated and for this reason are often a good choice when the constraints of the model are very complex. GA require a sophisticated system of encoding the problem and then decoding for a final solution (Holland 2001; Mitchell 2001). Therefore, the algorithm formulation is more susceptible to human error than simpler methods.

GA are based on the theory of evolution and natural selection (Goldberg 1989) and uses many of the biological terms such as chromosomes, alleles, reproduction, mutation, and crossover to describe the operations of the GA (Holland 1995; Holland 2001). The simplest GA model contains the elements of populations of chromosomes, a fitness function, and a crossover and mutation operator to generate new offspring (new solutions) (Mitchell 2001). In general terms each chromosome represents a point in the set of candidate solutions. The fitness function is the equation that is being optimized (minimized or maximized). It is used to decide if a new solution should replace a current solution. Riola (1992) described a simple example of a fitness function as maximizing

$$f\left(y\right) = y + \left|\sin\left(32y\right)\right|, \quad 0 \le y < \pi \tag{2.11}$$

where candidate solutions are denoted by *y.* The values for *x* in this problem are the binary equivalent of the *y*-values. Candidate solutions (*y*) are encoded to binary values (*x*), then after reproduction, the binary values are decoded to real values to be evaluated by the fitness function (Equation 2.11). Crossover and mutation combine to make up the reproduction scheme where new solutions are generated and then evaluated by the fitness function. One potential problem that must be overcome in any optimization is escaping a local optima solution. GA does this by using the mutation operator where random solutions are introduced to the solution set. This presents the algorithm with a solution that might produce a better fitness function evaluation than what would have been discovered through the normal scheme. Finally, the algorithm terminates when a better solution cannot be found. Actual termination procedures for a GA can be very simple to very complex as well.

The use of the GA has covered a wide range of disciplines over the last few decades but probably none more than in computer science. Knuth (1973) documented the efforts of well-known computer scientists from 1962 to 1973 that focused on the development of sorting algorithms to sort elements into specified lists, groups, etc. In the years following their works, there was much criticism about how efficient the sorting algorithms actually were. Hillis (1990; 1992) decided to approach this problem using a GA. His first results were disappointing because the GA could not perform as well as the earlier non-GA sorting algorithms. Only after increasing the level of sophistication and by adding a "predator-prey" co-evolution constraint to the GA, did he finally get results that were better and more efficient than the earlier sorters. The high level of complexity to which GA can be applied is obvious from this example. In this type of problem, the population of solutions is infinite because the sorter must be able to operate efficiently on any list when called. The question now is the outcome when the possible set of solutions, albeit very large, is well-defined. If the population of all possible solutions is well-defined, GA can be outperformed using simpler techniques (Mitchell 2001). This would naturally indicate that the complexity of the problem should drive the selection criteria of the method.

The complexity of the GA shows its ability to handle very simple to highly sophisticated problems. However, it was mentioned by one author that other methods may be more robust when the population is very well-defined. In this study where the population consists of 83 sampling sites and with five categories defining each sampling site there was ample argument for use of a different technique other than GA. The literature mentioned time and time again the usefulness and robustness of the simulated annealing method in combinatorial optimization problems. After searching, it was believed that simulated annealing would be an effective solution because of its relative simplicity and its ability to escape local optima using a "cooling schedule" (Michalewicz and Fogel 2002).

2.6.2 Simulated Annealing (SA)

Simulated annealing (SA) comes from a family of combinatorial optimization approaches that searches for a globally optimum solution from a seemingly infinite number of permutations. Annealing is most commonly known as a metallurgical process where metal is treated to a high temperature and then gradually decreased so that the atoms can arrange themselves in a minimal energy state or ground state (Aarts and Korst 1989). If the temperature is started sufficiently high and reduced slowly enough, the ground state can be achieved. In terms of the SA algorithm, this is called the annealing schedule, or "statistical cooling" (Vidal 1993), and is analogous to allowing a system to obtain a globally optimum solution. If the initial temperature is not high enough or if the rate of temperature decrease is too great, ground state cannot be achieved and a less desirable, locally optimum solution is achieved. This procedure is governed by the Boltzmann probability equation and the Metropolis (1953) algorithm, and is described in greater detail in the following section.

Simulated annealing (SA) is mainly based on the use of the Boltzmann probability equation. The Boltzmann probability equation, developed by J. Willard Gibbs (1839-1903) and Ludwig Boltzmann (1844-1906), is part of their contributions to statistical mechanics (Halliday and Resnick 1981). The Boltzmann distribution can be written as

$$\Pr\left\{\mathbf{E}=E\right\}=\frac{1}{Z\left(T\right)}\times \mathrm{e}^{\left(-\frac{E}{k_B T}\right)} \tag{2.12}$$

where E = energy state of the system
$Z(T)$ = normalization factor
T = temperature
k_b = Boltzmann constant

The Boltzmann distribution in Equation 2.12 is used to calculate the probability that a system has reached an energy state such that the system is at a point of thermal equilibrium for a given temperature **T** (Kirkpatrick, Gelatt, et al. 1983; Laarhoven and Aarts 1987). In SA, Equation 2.12 is often written without the normalization factor as shown in Equation 2.13. Here, the energy is a metaphor for some variable describing

$$Pr(\Delta E)=e^{(\frac{-\Delta E}{k_b T})} \tag{2.13}$$

where ΔE = change in energy between two permutations
k_b = Boltzmann constant
T = temperature

the system such as benefit or cost. Equation 2.13 is known as the Metropolis criterion and is used as the perturbation mechanism to introduce new solutions into the system that might not be introduced on the basis of energy alone. The main tenet in SA is to hold the system at a constant temperature long enough to allow the system to reach a state of thermal equilibrium. At this point, the molecules arrange themselves in a state of minimum energy. Metropolis (1953) showed that the generation of numerous instances of a state in a system would follow the Boltzmann probability distribution (Rodrigues and Anjo 1993) and ultimately reach a state of thermal equilibrium, or in the case of SA, a global optimum solution. He further developed what is called the Metropolis algorithm which consists of two mechanisms: the first being the perturbation mechanism in Equation 2.13 and the second being an acceptance mechanism based on the energy of the system (Aarts and Korst 1989). The perturbation mechanism of Equation 2.13 allows the algorithm to accept almost any change in energy at very high temperatures and reject change when the temperature is low. Thus, at high temperatures, random perturbations are introduced into the system with progressively fewer perturbations as the temperature cools so that a global optimum solution can be discovered. The second mechanism compares the new energy in the system to the current energy in the system. Depending on whether the objective function is to be minimized or maximized, the new energy, or solution, will be accepted or rejected. The program flow in a typical SA problem usually involves testing the new solution against the acceptance criterion first. If the acceptance criterion rejects the new solution then it is tested for acceptance again using the perturbation mechanism. In the SA algorithm presented here acceptance of new permutations based on the net benefit of sampling sites are determined using the Metropolis algorithm.

2.6.3 Examples of Using SA

The SA algorithm has been used to solve the traditional Traveling Salesman Problem (TSP). The TSP is considered a traditional problem because it is often used as a benchmark test to compare the results of different combinatorial algorithms, and because it is simple to understand (Kirkpatrick, Gelatt, et al. 1983). The TSP problem involves finding a route where the traveling salesman would depart from a beginning city and travel to each city in the dataset once while minimizing the total distance that is traveled, and finally return to the city of origin. The objective function in the TSP is a simple minimization of total distance traveled. In the past, the TSP has been adjusted to make it more complicated by using additional constraints. Press, Teukolsky, et al. (1992) added the placement of a river through the dataset of cities in the TSP. A new constraint was added by specifying that the traveler was afraid to cross the river so the number of times crossed while visiting all of the cities had to be minimized also. The consideration of the constraint became part of the objective function. Both total distance and number of times the river was crossed had to be minimized. The above example is indicative of a traditional problem where the constraints can be very simple to increasingly complex if the user desires. The objective function of the problem presented in this study is similar in the sense that the distance is somewhat minimized. However, the main priority is to maximize the net benefit of the monitoring network while taking distance into consideration.

In the field of hydrology, SA has recently been used to optimize a rainfall gaging network (Pardo-Iguzquiza 1998) and a collection of river sampling sites (Dixon, Smyth, et al. 1999). Pardo-Iguzquiza's (1998) main goal was to design a rainfall gaging network in such a way that the best estimation of mean areal precipitation amounts could be determined while minimizing the estimation variance. The objective function for minimizing the estimation variance was a function of the number of sites, locations, and cost. The variance was calculated using a variogram for subareas in which gages were required. SA estimated the optimal location within each subarea for a rain gage. He then constructed a graphical curve of variance versus number of gages. The graph shows that as the number of gaging stations increase past seven or eight, the decrease in variance is minimal. Another graph was presented showing how cost increased with smaller variances. Together these graphs and information could be used to optimize a rainfall gaging network that minimizes the variance and maintains a reasonable cost.

Dixon, Smythe, et al. (1999) used SA for the evaluation of existing flow and water quality measuring sites on Logan River and Albert River in Queensland, Australia. Their work was based on previous studies by Sharp (1970; 1971) where stream order was the basis for selection of potential sampling sites on stream reaches. The authors identified eight of 12 existing sampling sites that would minimize cost yet provide an optimal detection network for pollutant sources.

Skaggs, Mays, et al. (2001) used an enhanced version of SA with directional search and memory capabilities to design optimal groundwater remediation techniques. Their example involved determining optimal pumping rates so that desired contaminant levels could be achieved within selected timeframes at a minimal cost. The objective function of total cost was a function of extraction, injection, installation cost, and a penalty function for not being able to achieve the targeted contaminant level in a given timeframe.

Chapter 3 ASSESSMENT OF SPATIAL VARIABILITY AMONG SAMPLING SITES

3.1 Synoptic Sampling Data

The data from the synoptic sampling network for determining data-redundancy are quarterly samples from January 1996 to November 2001 and include the following eight measurements: pH, ANC (acid neutralizing capability), conductivity, nitrate, sulfate, chloride, sodium and potassium. Water quality data within this period of time were chosen because of the consistency in sampling frequency and locations, whereas the earlier years seemed to lack consistency in sampling frequency and location. These samples were collected by volunteers and by staff and students from the University of Tennessee. Collected samples are returned to the Science and Engineering Research Building at the University of Tennessee where they undergo a collection of water quality tests including the eight measurements listed previously. Dr. Bruce Robinson of the Civil and Environmental Engineering Department oversees these tests including a rigorous quality assurance program.

The mean values of each chemical constituent were calculated for individual sampling sites for the period from 1996 through 2001. Computations were performed using macros and filters in the Microsoft Access database where the data are stored. Mean values were calculated over the full six-year period rather than 3 two-year periods (as was originally proposed) because the number of observations was low. Using the full six-year period, the results are based on 24 observations; whereas, a two-year period is based on eight observations. Means calculated using only eight observations would have high standard deviations and wide confidence intervals. The water quality means data for the chemical constituents of each site are listed in Appendix B. At this point it should be mentioned that the 83 sites sampled from streams are considered in this study. The high-elevation springs are not considered here because of the natural differences in chemical characteristics compared to surface waters.

A series of descriptive tests are presented on the water quality data. Such tests include univariate and multivariate normality, identification of outliers, correlations, scatterplots, and boxplots. The descriptive statistics were performed using the NCSS software and are necessary for two main reasons. First, the descriptive tests establish fundamental knowledge of the data that are needed to make good decisions about which methods can be used in the cluster and discriminant analysis. Some of the methods should only be used for multivariate normal data while others can be used for non-normal data. Second, knowledge of which sampling sites are outliers and which sampling sites cause non-normality can aid in the interpretation of the final results.

3.2 Geology, Morphology, Vegetation and Collocated Studies of the Monitoring Network

The similarities between the sampling sites are also analyzed with respect to the watershed characteristics that include geology, morphology, and vegetation. Watershed characteristics are analyzed using multivariate statistical techniques and are necessary to ensure that uniqueness, based on these characteristics, is not overlooked. Most of the information for the watershed characteristics were compiled by Harwell (2001). The terminology "watershed characteristics" will be used throughout the remainder of this study and whenever encountered is meant to represent geology, morphology, and vegetation data. Table 1 shown below shows the components of each of the watershed characteristics that are included in the cluster analysis.

Many of the water quality sampling sites are collocated with fish and benthic ecological studies that are performed by the NPS. It is necessary to identify such collocated sampling sites and attribute a higher benefit score into the optimization routine so that the collocated sampling sites will receive an overall higher benefit than those sampling sites that are not collocated with other studies. Since much of the information gathered in the monitoring network is used to assist in the interpretation of the fish and benthic studies, extra consideration should be given before discontinuing collocated sampling sites.

3.3 Costs and Benefits of Sampling

Two primary considerations in this study are the costs of collecting samples and determining the benefit of sampling at each water quality sampling site in the monitoring network. Costs and benefits are reflected on a monetary scale so that all sampling sites are based on an equivalent measure. Costs are calculated based on access distances, average sampling time at each site, laboratory analysis, data interpretation, project administration, and overhead. The monetary benefits are a combined function of the estimated total current benefits of sampling the entire network and the results of the multivariate analyses of the water quality and watershed characteristics data.

Most of the information for determining the costs of sampling was obtained

Table 1. Watershed characteristics analyzed.

Geology	**Morphology**	**Vegetation**
Thunderhead sandstone	Site elevation	Spruce-fir
Limestone	Mean basin elevation	Northern hardwood
Cades Cove sandstone	Stream order	Cove hardwood
Anakeesta	Maximum channel length	Mesic Oak
Elkmont sandstone	Basin length	Mixed-mesic oak
Basement group	Basin area	Tulip poplar
Great Smoky group	Stream density	Pine
	Mean basin slope	Heath bald
	Channel slope	Xeric oak
	Basin width	Pine-oak

through discussions with Dr. Bruce Robinson and with staff and graduate students that work very closely with the project. Distances between sampling sites were measured from topographic maps of the GRSM using Wildflower Productions' TOPO! software for North Georgia, Great Smoky Mountains, and Atlanta (1999). Using the trail and road distance information from the aforementioned source and average speeds for various modes of transportation, an equivalent man-hour (MH) calculation was developed for accessing each sampling site in the network. The MH equation standardizes the distances so that all sampling sites can be compared on an equal basis. The man-hours (MH) were calculated using Equation 3.1 below.

$$MH = \frac{M_1}{45mph} + \frac{M_2}{1.5mph} + \frac{M_3}{3.5mph} + SST \tag{3.1}$$

where MH = man-hours determined by average hiking speed
M_1 = automobile miles (45mph average automobile speed)
M_2 = hiking miles (1.5mph average hiking speed)
M_3 = all-terrain vehicle miles (3.5mph average all-terrain speed)
SST = time spent onsite to collect sample (hours)

Average time spent conducting sampling at each sampling site (*SST*) is assumed to be 10 minutes.

Determination of the access and onsite collection costs required a separation of the variable cost from the total costs for sampling the present 90 site network (83 stream sites and seven high-elevation springs). It was decided in this study to only include the variable costs in the optimization because these costs are directly associated with the number of sampling sites in the network. The total variable annual cost of approximately $69,165 was determined from the balance sheet presented in Appendix

A. There it can be seen that the total cost of the monitoring project administered by the University of Tennessee is approximately $199,600. Approximately 70 percent, or $139,575, of this amount is spent on the synoptic monitoring network focused on in this study. Of the $139,575, approximately $69,165 are variable costs and $70,410 are fixed costs. The fixed costs of $70,410 are incurred whether the number of sampling sites are 10 or 100 and include project administration, data analysis, reporting, presentations, equipment replacement, etc. Based on access calculations mentioned earlier in this section and assuming that two people are present for all sample collections, it takes approximately 640 man-hours each year to access and sample the 83-site stream sampling network. At a nominal cost of $30 per man-hour this calculates to $19,200 each year for the access and collection costs for 83 sampling sites. The prorated costs for sampling all 90 sites, including the high elevation springs, would equal $20,820 and the total costs of these 90 sites for lab analysis and database management would equal $48,345 or $538 per site each year. Based on the battery of tests conducted, this amount seems quite reasonable. For the purpose of the analysis in this dissertation the cost used in the optimization associated with access and sampling time was $19,200 and the cost associated with laboratory analysis and database management was $50,000 for 83 stream sampling sites. This is a total of $69,200 which is approximately equal to the variable costs of the synoptic monitoring network. Note that 55 percent of the costs are cost-shared of are in-kind contributions.

In order to assign an overall benefit dollar value to the current sampling network, a multiplier of 1.2 was assumed and applied to the total of the variable costs for the network, i.e. a total benefit value of 1.2 times $69,200 equals $83,040. The multiplier is a reasonable assumption based on the notion that the information and benefit gained by sampling outweighs the associated costs. Additionally, a 20 percent return on an investment is a modest expectation. The total network benefit value can then be apportioned over the sampling network based on the total benefit score for each sampling site, where the individual sampling site benefit score is calculated using the collocation identification and clustering scores of the water quality data and watershed characteristics (geology, morphology, and vegetation). The cost and benefit equations used in the optimization are described in detail in sections 3.5.1 and 3.5.2.

3.4 Multivariate Statistical Methods and Composite Benefits

Water quality variables and watershed characteristics for the 83 site sampling network are analyzed using the following three multivariate statistical methods that were explored in the literature review:

- Principal component analysis (PCA)
- Hierarchical/non-hierarchical and fuzzy cluster analysis (CA)
- Discriminant analysis (DA)

The flowchart shown in Figure 1 provides a conceptual view of how these three methods are linked together to analyze the data and validate the results. The process is routinely followed in most cases. SAS software and NCSS software are used for these analyses.

The process in Figure 1 begins by taking the raw data and performing descriptive statistics. Information obtained in this phase will aid in the interpretation of subsequent analyses. If significant correlations exist between variables then a PCA is

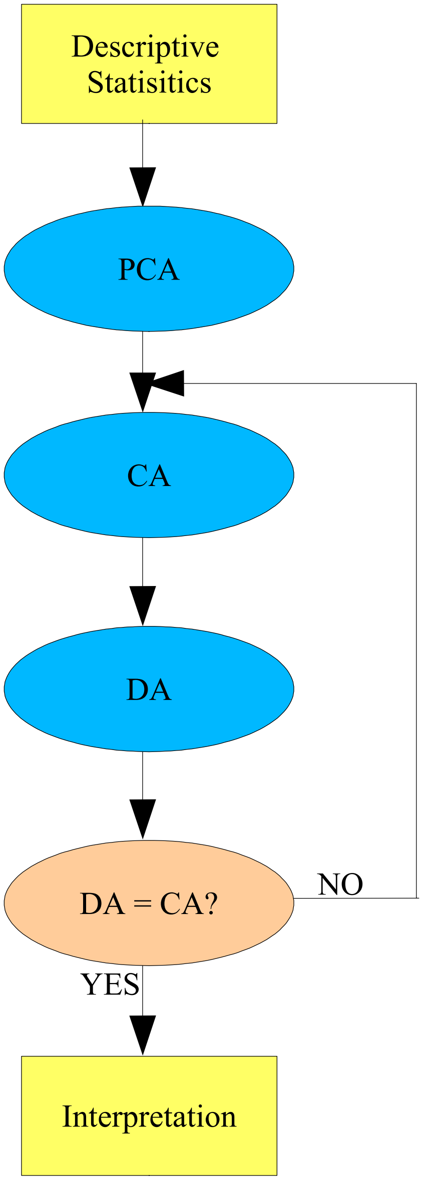

Figure 1. Conceptual flowchart of analyses for examination of water quality data.

performed to reduce the dimensionality and to identify the underlying structure of the dataset for further analyses. Principal component scores, according to the eigenanalysis of the PCA, are then selected for the cluster analysis. A two-step CA is performed using hierarchical and non-hierarchical methods which will produce the number of clusters, cluster memberships, and the distance of each member from its respective cluster centroid. Sampling sites with the greatest distance from their cluster centroid are indicative of sampling sites within that cluster that explain a greater amount of variability, therefore, being of the greatest benefit in terms of information gained. Finally, DA will be used to test the discriminating ability of the clusters formed in the previous step using the principal component scores and the original data. By using the original data, a "full-circle" validation is provided because the raw data are the source of the principal components. If the DA produces poor discriminating results, the methods used for the CA will be revisited.

The culmination of the above multivariate analyses will result in the formulation of a composite benefit score for each sampling site to be used in the network optimization algorithm later in this study. The composite benefit score is the factor by which the dollar amount of benefit is apportioned to each sampling site, and is the foundation of choosing which sites could be discontinued and which sites should be retained. The equation used to calculate the composite benefit score is the summation of the water quality clustering results, the watershed characteristics clustering results, and the collocation component. This equation has the form

$$\Psi_i = \omega_1 W_i + \omega_2 M_i + \omega_3 G_i + \omega_4 V_i + \omega_5 C_i \tag{3.2}$$

where Ψ_i = composite benefit score for the *i*-th sampling site
W_i = clustering results from the water quality data component
M_i = clustering results from watershed morphology component
G_i = clustering results from the watershed geology component
V_i = clustering results from watershed vegetation component
C_i = collocation component
ω = weight assigned to each component score

Weights may be applied to each component benefit score to accentuate the importance of a certain factor. Weights could also be applied to equalize the importance of a subset of variables or all of the variables.

Using Equation 3.2, the clustering results for water quality and watershed characteristics are converted into a score. Each member (sampling site) of a cluster has a calculated distance from the centroid of the cluster. The site with the greatest distance from the centroid explains more of the variability within the cluster than a site that is located near the centroid of the cluster. Therefore, a site with the greatest distance from its respective cluster centroid is more beneficial to sample than a site that is near the centroid. In the water quality data or watershed characteristics there are a number of clusters for the 83 site network. For example, in the water quality data (W_i) clustering, suppose that the 83 site network divides into eight clusters and the largest of these clusters contains 20 of the sampling sites. Likewise, the geology data for each sampling site may divide the network into 10 clusters with the largest cluster membership of 30 sampling sites. Therefore, the water quality data and geology data each define a unique set of clusters.

Consider again the hypothetical clustering case mentioned above for the water quality data where the largest cluster contained a membership of 20 sampling sites. Instead of using the cluster distances directly as a measure of benefit, the sites were ranked from one to 20 with one and 20 representing the sites with the closest and the farthest distances from the centroid, respectively. Therefore, the site with the highest ranking (farthest from the centroid) receives the highest benefit. The same scale (one to 20) was used to rank the other seven clusters of the water quality clusters. Based on this scale, a cluster that had only three members would have its sites ranked as 20, 19, and 18. The process accomplishes two goals. First, it places all clusters on an equal scale. Second, it helps to insure that smaller clusters are not removed during optimization. The original number of clusters should survive the optimization so that each cluster will still be represented in the future monitoring network. In clusters with a very small number of members, or sampling sites, it may be necessary to not remove any of the sites since some redundancy is needed. The same approach is used for the geology, morphology, and vegetation clusters. In cases where there are ties, the rank of the tied sampling sites is calculated by averaging the range of ranks over the sampling sites that are tied.

The composite benefit scores should not used solely to indicate whether a site should or should not be discontinued because of other factors relating to cost. As mentioned earlier, the scores are used to determine a dollar amount of benefit for a single site. A very low composite benefit score would indicate at least one and probably more of the following:

- Explanation of variability within its respective cluster is relatively low
- A sampling site is very similar to other sites in terms of watershed characteristics
- A sampling site is very costly to sample because of distance

- A sampling site is not collocated with another NPS study

Conversely, a high score would indicate a very unique sampling site that should not be discontinued.

3.5 Monitoring Network Optimization using Simulated Annealing

The SA algorithm formulated in this study uses two different approaches and is written and run in the Matlab programming environment. The first approach takes an initial set of sampling sites, selected by the user, and then progresses until an optimum solution is reached. In the second approach, the user specifies the number of sites (*n*) desired in the final network and the program searches to find an optimum solution of the *n* best sites. The initial selection of sampling sites in the second approach is generated randomly by the computer. Discussions thus far have identified SA as a minimization technique; however, in the problem presented here SA will be used to maximize the monetary benefits of the sampling network. Maximization is achieved by changing the sign of ΔE in Equation 2.13 (Michalewicz and Fogel 2002). Confirmation of this approach can be verified graphically by plotting the objective function and this will be presented in Section 3.6.9.

The SA algorithm begins with the initial set of sampling sites that is chosen by the user or generated by the computer depending on which approach is used. The term “energy” can now be replaced by “monetary benefits.” The initial set becomes what is commonly known in SA as the “current solution”, and then the monetary benefit is evaluated for the current solution. The monetary benefit for a subsequent (new) set of sampling sites, generated randomly by the computer, is calculated and compared to the current solution. The latter set of sampling sites is commonly referred to as the “new solution.” If the new solution has a higher monetary benefit than the current solution, the new solution replaces the current solution. If the monetary benefit of the new solution is lower than the current solution, it may still replace the current solution by the Boltzmann probability *P(ΔE)*, i.e. if *P(ΔE)* is greater than a random uniformly distributed number [0,1). This is known as the Metropolis algorithm or in statistics the Monte-Carlo method and was discussed in Section 2.6.2 of the Literature Review. As the temperature decreases the probability that a new solution will replace the current solution based solely on the Boltzmann probability becomes less and less. It can be seen that if the initial temperature in Equation 2.13 is very high, the probability of replacement will be close to one and even a poor solution can replace a better solution. Perturbations are introduced to the system when a better solution is replaced by a poor solution thereby escaping local maxima. The annealing schedule (initial temperature and rate of decrease) must therefore be chosen carefully based on the range of monetary benefits from the data used in the algorithm (Aarts and Korst 1989). It is suggested that the initial temperature be no greater than the temperature corresponding to the largest difference in monetary benefits of two permutations of sampling sites (Press, Teukolsky, et al. 1992). Initial temperatures higher than this result in no improvement of accuracy and a significant increase in computational time. The initial temperature must ultimately be estimated through trial-and-error. The decay of the temperature must be slow enough so that a globally optimum solution can be found. Suggestions for determining the rate of decrease are given by Press, Teukolsky, et al. (1992), and will be discussed further in Section 3.5.4. A termination condition is reached after the decrease in energy is negligible indicating that the maximum monetary benefit has been identified, or when a user-specified maximum number of iterations have been exceeded.

3.5.1 Calculating the Cost of Sampling a Site in Matlab

The goal of this section is to describe how the cost of sampling a network of sites is determined by the Matlab algorithm. Section 3.3 described how the costs and benefits were calculated. Remember that costs are based on the components of distance, sample collection time, and laboratory and associated costs. Cost breakdown is similar to the one used by MacKenzie, Palmer, et al. (1987) where they were also minimizing the cost of a monitoring network while maximizing the statistical power of trend detection. The equation in Matlab terms used in the SA algorithm to represent this is shown in Equation 3.3.

$$COST_p = \sum_p LABCOST + \sum_p ACCESS \qquad (3.3)$$

where $COST_p$ = total cost for p sampling sites
$LABCOST_p$ = laboratory and associated costs for p sampling sites
$ACCESS_p$ = costs for accessing p sampling site

ACCESS is determined from by the number of man-hours required to reach *p* sampling sites where *p* is the permutation of sampling sites being considered. *ACCESS* is calculated in the objective function subroutine. Figure 2 shows the tree diagram of how all the sites were assembled. The illustration is based on the most likely route of accessing the sites following roads and trails that connect the array of sites. Site 30 (Sugarlands Visitor Center) is assumed to be the starting point of collection. From here the tree branches into three major directions of US Highway 321 towards Cosby, State

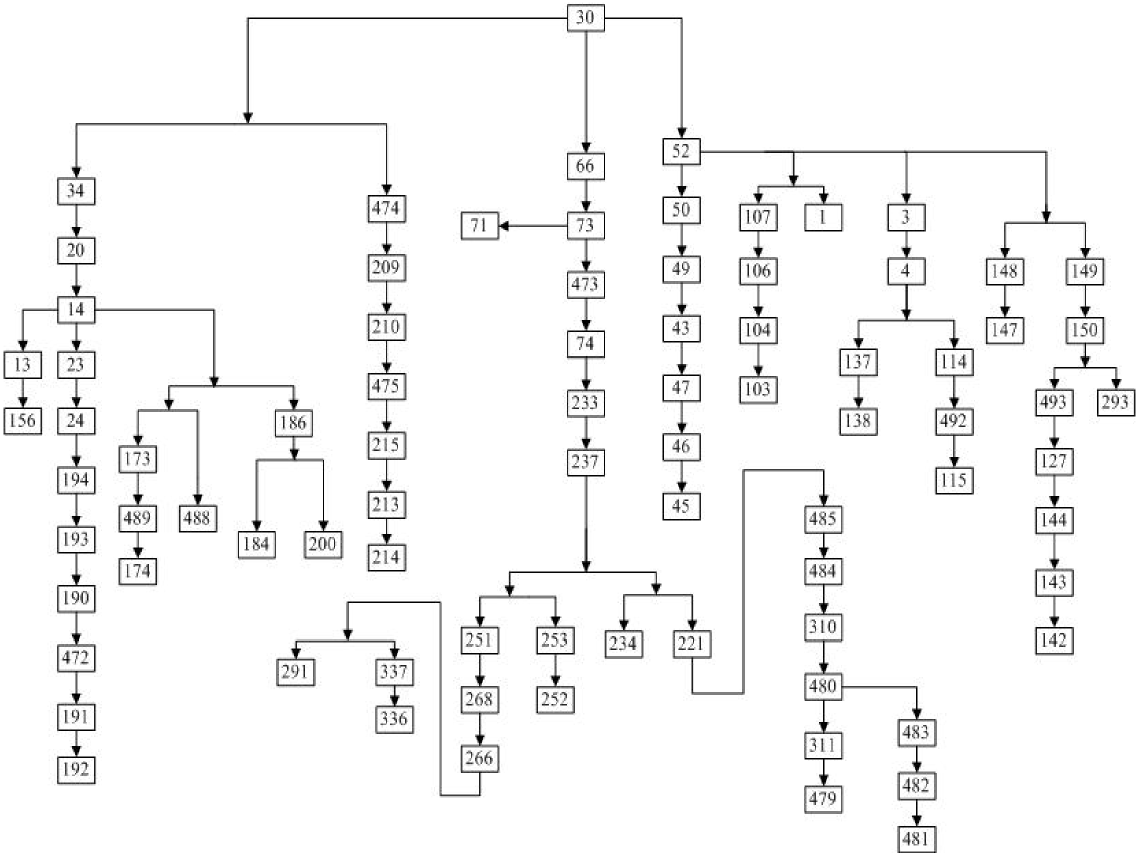

Figure 2. Tree diagram of the 83-site sampling network.

Highway 73 towards Cades Cove, and US Highway 441 to Cherokee. When a permutation of sites is chosen, a linkage from the furthest sites back to site 30 is assembled to calculate the cost using the man-hours needed to access all sampling sites. A separate function in the subroutine is used to calculate the man-hours through the intersections to insure that man-hours through branches are not counted twice. Finally, *LABCOST* calculates the costs of collecting the sample once arriving at the site and performing the laboratory analyses and all other associated costs based on the number of sites in the current permutation.

The incremental cost of sampling a particular site can change through the iterations of the algorithm as sites go in and out of network. It is at this point that the algorithm is most useful. For example, in Figure 3, assume that the current permutation of a sampling network includes site 30 and site 3. A new permutation of sampling sites includes not only sites 30 and 3, but also adds site 2. Because the current permutation would already include the access cost from site 30 to site 3, there would be no additional access cost to sample site 2. The costs of laboratory and associated tasks for site 2 would be added to the total costs. If site 3 was removed and site 2 was added, a new access cost with a shorter distance to the end of the branch at site 2 would need to be calculated.

3.5.2 Calculating the Benefit of Sampling a Site in Matlab

The following section describes how the benefit of sampling a network of sites is determined by the Matlab algorithm. Equation 3.2 determines the composite benefit score for each sampling site based on the multivariate statistics and the collocation

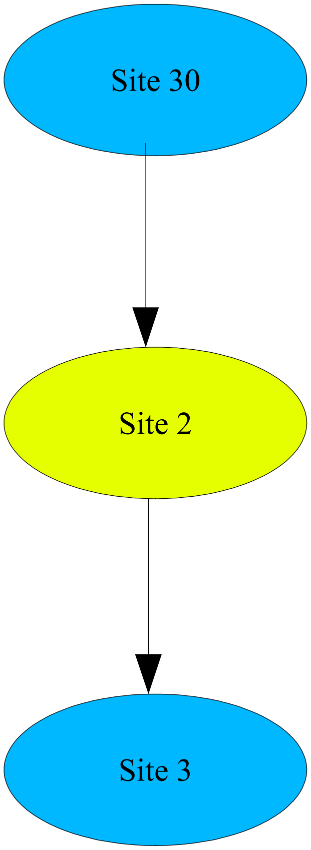

Figure 3. Diagram example of determining site access cost.

information. Additionally, Section 3.3 describes the monetary value of the benefits. The results are combined in Matlab to form the benefit portion of the objective function. Equation 3.4 shows how the benefit for a sampling site is calculated based on access costs, laboratory costs, overhead, interpretation, etc. It can be seen in this equation that the benefit is a

$$BENEFIT_i = \frac{\Psi_i}{\Psi_{total}}(1.2 * LABCOST_n) + (1.2 * ACCESS_i) \tag{3.4}$$

where $BENEFIT_i$ = benefit in terms of dollars for the *i*th site
Ψ_i = composite benefit score for the *i*th site (described in Section 3.3)
Ψ_{total} = total of composite benefit scores for all 83 sites
$LABCOST_n$ = total laboratory and associated costs for 83 sites
$ACCESS_i$ = cost of accessing and sampling the *i*th site

ratio of the site composite score to the sum of all composite scores plus the benefit of accessing the *i*th site. Use of this equation allows the site with the highest clustering scores and collocation scores to receive the greater benefits. The $BENEFIT_i$ for each site is calculated in an Excel spreadsheet beforehand and exported to a Matlab file. The benefits remain constant because they are a product of the multivariate analysis performed initially on the current network and are retrieved by the objective function subroutine of the SA algorithm when the particular site is included in the current permutation.

There are also benefits that are a function of site access and sampling time. These benefits are variable and are based on the access route and number of sampling sites in a permutation of sites being tested by the SA algorithm. The benefits for access and sampling time are simply 1.2 times the access cost for a permutation of sites being tested and are calculated during execution of the SA algorithm.

3.5.3 Objective Function of the SA Algorithm

The aim of the algorithm is to maximize the objective function which is the sum of the benefits (+) and costs (-) for a permutation of sampling sites. The equation for the objective function is shown below in Equation 3.5.

$$NETBENEFIT_p = \sum_p BENEFIT - \sum_p COST \qquad (3.5)$$

where $NETBENEFIT_p =$ total benefit for permutation of p sites
$\Sigma BENEFIT =$ sum of the monetary benefits for the permutation of p sites
$\Sigma COST =$ sum of the monetary costs for the permutation of p sites

3.5.4 Annealing Schedule

The annealing schedule is perhaps the most important part of the SA algorithm because it controls the loops in the program and the rate at which the temperature is changed (Michalewicz and Fogel 2002). Press, Teukolsky, et al. (1992) suggests that finding the best annealing schedule is often a trial-and-error process. The process followed in this study was the one used by Press, Teukolsky, et al. (1992) in a traveling salesman problem and is as follows:

1. At various high temperature values T, generate random permutations of sampling networks and evaluate the object function for each permutation. This is achieved by simply writing the objective function evaluation to a file and then requesting the maximum and minimum evaluation. At each value of T, the difference between the maximum and minimum evaluation is the largest ΔE for that T.

2. A starting value of T corresponding to the largest ΔE is chosen for the initial T in the SA algorithm.
3. T should be decreased at steps no greater than 10 percent. T should not be changed until 100N (where $N = 83$ sites) permutations are made or until 10N permutations are accepted by the objective function criteria, whichever is first.
4. The algorithm is terminated at the end of the annealing schedule or when ΔE cannot be improved upon after a specified number of successive iterations.

After several trials through this procedure, optimal operating values were found. However, because this problem is different from a traveling salesman problem, the annealing schedule proved to be somewhat simpler. The initial temperature T was set at 200 based on the procedure in Step 1. Step 2 generally required the temperature T to be decreased at a rate no greater than 10 percent. After numerous trials it was found that 3 to 5 percent decreases were better. Steps greater than 5 percent resulted in unstable effects by generating significantly different solutions for different runs. Step 3 specified 100N as an initial number of permutation trials for a given T, but it was found that 10N was sufficient and the computation time was reduced from approximately 3.5 hours to approximately 30 minutes.

3.5.5 Optimum Solution Search for the Full Network

This version of the algorithm assesses all of the sites and searches until an optimum solution is found. A user-specified set of sites is initially marked to be in the network and this becomes the current solution. The main program then calls a subroutine function program to calculate the objective function for the current solution.

The initial configuration can simply be a random selection of sites and can be as large or small as wanted. In order to overcome a possible division-by-zero error in the first step, at least two sites must be selected initially. Other than this, there are no requirements for the initial set of sites.

There are two main controls in the Matlab SA program. The first control or "outer" loop is the annealing schedule that controls the number of temperature steps allowed by the program and the rate at which the temperature is decreased. The "outer" loop terminates when the temperature step counter reaches the number of temperature steps allowed by the program. The second control loop is interior to the first control or the "outer" loop. The second control loop contains the Metropolis algorithm and creates the random permutations of sampling sites. The second control also sums the number of successful acceptances of sampling network permutations. The second control or "inner" loop is terminated when the number of successful permutations reaches 10 times the number of sites considered, i.e. 830 successes. When program execution reaches the "inner" loop, the program randomly selects a site from the 83 site network. If that site is already in the current network, the program tests that site for removal from the current network using the objective function and the Metropolis algorithm. Conversely, if the site is not in the current network the site is tested using the objective function and the Metropolis algorithm to determine if it should be added to the current network. In either case, the test is that a new objective function is calculated and compared to the objective function of the current network. If the objective function of the new network is greater than that of the current network, then the new network configuration replaces the current network configuration. The process is known as the local or "*nearest-neighbor*" search. If the objective function is less it can still be accepted using the Boltzmann probability rule, allowing the optimum solution to be searched for outside of the *nearest-neighbor* area, thereby permitting the program to find a globally optimum solution rather than a local solution. If this solution is not accepted with either rule, the new configuration is discarded and the current configuration continues with the program beginning a new iteration.

3.5.6 Optimum Solution Search for a User-specified Number of Sites

The user-specific method of the SA algorithm is mechanically very similar to the method above with a few exceptions. At the beginning of the program execution, the user specifies *n,* the number of sites desired in the final network. The program then generates a random selection of *n* sampling sites for the current network. The remaining sampling sites (83-*n*) are placed into a second set of sites that are not in the current network. For example, the user specifies a desired network size of 10 sampling sites. The current network is formed by random selection from the pool of 83 sites to form the current network of 10 sampling sites. The remaining 73 sampling sites form a set of sites that are not in the current network. The objective function is then called and the current network is evaluated. The new network is generated by randomly switching a single site from the current network with a single site from the second set of 73 sites not in the network. The objective function is then calculated for the new configuration and compared to the current network. The rules for acceptance of the new network as the current network are the same as those in the previous section. If the new configuration cannot be accepted as the current configuration, the sites are switched back to their original position and the process starts over. The annealing schedule is also the same as mentioned in the previous section.

3.6 Results of the Analyses

The results of the analyses are divided into eight sections that present the outcomes of the methods used to optimize the water quality monitoring network and provide recommendations for discontinuance of selected sampling sites. The sections are as follows:

1. Data Screening of the water quality data
2. Analysis of water quality data
3. Analysis of geology
4. Analysis of morphology
5. Analysis of vegetation
6. Identification of sites in overlapping regions of clusters
7. Collocated sampling sites
8. Compilation of data for network optimization
9. Network optimization using simulated annealing (SA)

It should be mentioned beforehand that data screening of the watershed characteristics was not performed to the extent of the water quality data. Much of the watershed characteristics database has entries of zero because they are based on percentages of a certain characteristic in the watershed; analysis of these data in the data screening section would lend minimal interpretation to the final results. The watershed characteristics are also considered as constants because changes occur over relatively large timescales, especially when compared to the highly variable nature of the water quality data.

A variety of multivariate statistical methods have been used in the analysis of the various types of data presented. There is no single stepwise recipe for the application of the statistical methods presented. Rather, logical steps from one method to another are used in searching for the one method or combination of methods that produce the best results in the DA.

3.6.1 Data Screening of the Water Quality Data

Basic univariate and multivariate data screening results are presented in this section for the water quality data. The data screening results are an integral part of the interpretation of the final results and recommendations presented. The water quality variables analyzed here include hydrogen ion, ANC, conductivity, chloride, nitrate, sulfate, sodium, and potassium. Obvious abbreviations for these variables may be used in some of the graphs and tables that follow to meet margin restrictions. Appendix B contains the means of the water quality variables and Appendix C contains the results of data screening that are mentioned but not shown in this section. References to the information presented in Appendix C will be made in the remainder of this section. Boxplots were constructed for each water quality variable listed in the preceding paragraph and are shown in Appendix C. Table 2 presents the outliers identified by the boxplots for the water quality variables. The most notable feature in these boxplots is the identification of mild and severe outliers for the univariate case. Sampling sites 156, 174, 237, 251, 252, 253, and 489 appear as outliers numerous times for different

Table 2. Mild and severe outliers of the water quality data.

Water Quality Variable	**Sampling Sites identified as outliers**
Hydrogen ion	NONE (103, 156, 174, 237, 489 as pH)
ANC	156, 173, 174, 489
Conductivity	174, 256, 251, 252, 489
Chloride	156, 174, 251, 252, 253, 489
Nitrate	NONE
Sulfate	15, 74, 233, 251, 252
Sodium	148
Potassium	106

water quality variables. At this point it is important to remember the observations that are identified as outliers as this will be very worthwhile later in the study. It also worth noting that nitrate contained no outliers, while sampling site 148 was an outlier only in the sodium variable and sampling site 106 was an outlier only in the potassium variable. Univariate normality tests were performed using the D́Agostino Omnibus method. The method combines tests for skewness and kurtosis which are typically 0 and 3, respectively, for normal data. The D́Agostino Omnibus method calculates whether the actual skewness and kurtosis of the water quality data are significantly different from 0 and 3. If skewness and kurtosis are significantly different, then the data are believed to be non-normal. According to the D́Agostino Omnibus method using an α-level of 0.05, the data are not normally distributed with the exception of potassium. The Shapiro-Wilk method was also used to test normality. The Shapiro-Wilk test agreed with the D́Agostino Omnibus test in every case except one. The Shapiro-Wilk test suggested acceptance of normality for the sodium data while D́Agostino Omnibus rejected normality. The p values for the normality tests of the water quality variables are shown below in Table 3. The overwhelming amount of non-normal data is a powerful indicator that the water quality variables will not be multivariate normal.

A multivariate outlier test for the water quality variables was also performed. The test is run using the NCSS software and is based on the Mahalanobis distance from the variable means and the relationship between the t^2-distribution and the F-distribution. The outlier test was conducted using an α-level of 0.10. Sampling sites identified as outliers were: 106, 148, 156, 174, 234, 237, 251, 252, 253, and 489. It is interesting to recall the outlying observations that were identified by the univariate tests

Table 3. p values of the normality tests for the water quality variables.

	Normality Test Name	
WQ Variable	**D́Agostino Omnibus**	**Shapiro-Wilk**
Hydrogen ion	0.0023	0.0001
ANC	0.0000	0.0000
Conductivity	0.0000	0.0000
Chloride	0.0000	0.0000
Nitrate	0.0000	0.0000
Sulfate	0.0000	0.0000
Sodium	0.0156	0.0546
Potassium	0.5794	0.2406

above. Many of the same sampling sites identified as outliers in the multivariate case are also univariate outliers. The test results may help to pinpoint the variables that are producing most of the multivariate outliers.

Multivariate normality of the water quality data was checked by visual observation of a multivariate normality plot. The multivariate normality plot was generated using the SAS software and is shown in Figure 4. The plots are essentially interpreted the same way as a univariate normality plot. Normal scores are plotted along a 45-degree line from the origin. If there is significant departure from the line or if the normal scores plot to only one side of the line or the other, the data are assumed non-normal. At this point multivariate non-normality is almost guaranteed because the majority of the water quality data are univariate non-normal. The 45-degree line is represented by the line of X's from the origin and the normal scores by the O's. The differences in the scales of the axes make the line appear to be less than 45-degrees. Outliers are those points at the upper-right part of the plot that are major departures from the 45-degree line. Although data can be inherently non-normally distributed, outlying points alone can cause non-normality in otherwise normally distributed data. The points that show major departures at the upper tail do coincide with those that were identified as univariate and multivariate outliers in the previous tests.

One essential data requirement for applying principal components analysis (PCA) is that correlation exists between some of the variables. Correlation is the next analysis applied to the water quality data to verify that PCA is a viable choice for analyzing the water quality data. Table 4 below shows a high number of correlations with *p* values less than 0.05 based on Pearson's correlation. The information reinforces

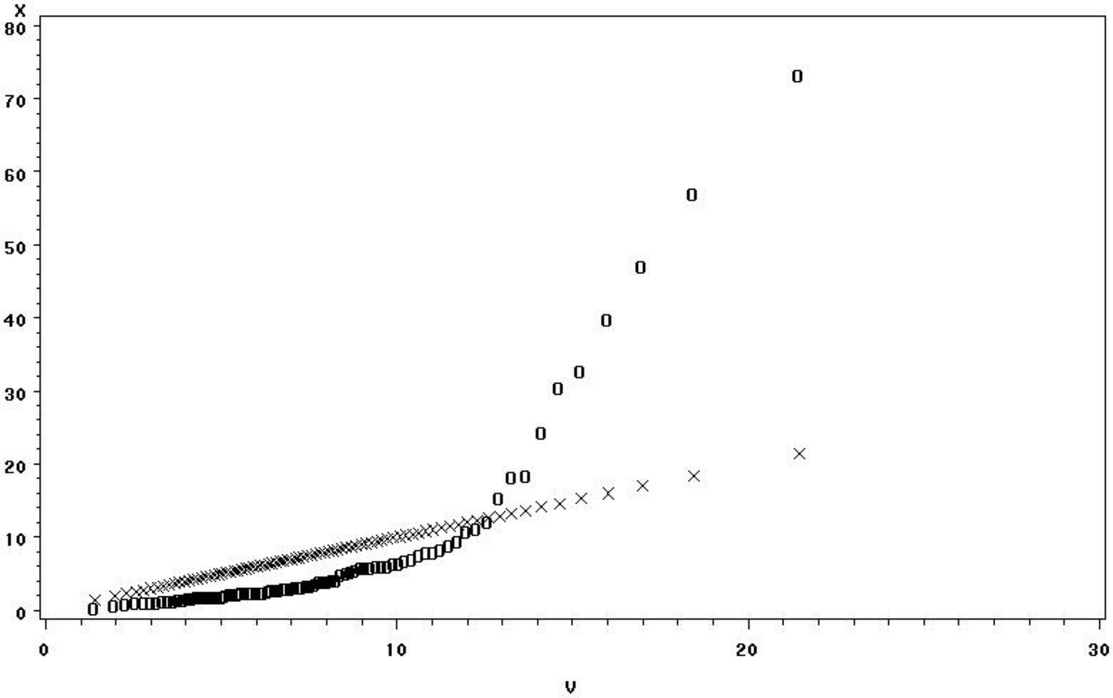

Figure 4. Multivariate normality plot of the water quality data.

Table 4. Pearson's correlations for the water quality data.

	Hydrogen ion	**ANC**	**Con**	**Chl**	**Nit**	**Sul**	**Sod**
Hydrogen ion	1.000						
p-value							
ANC	-0.192	1.000					
p-value	0.0827						
Conduct.	-0.032	0.933	1.000				
p-value	0.773	<.0001					
Chloride	-0.101	0.572	0.694	1.000			
p-value	0.366	<.0001	<.0001				
Nitrate	-0.373	-0.241	0.052	0.302	1.000		
p-value	0.010	0.028	0.641	0.006			
Sulfate	0.280	-0.043	0.297	0.355	0.566	1.000	
p-value	0.001	0.700	0.006	0.001	<.0001		
Sodium	-0.417	0.463	0.336	0.255	-0.434	-0.177	1.000
p-value	<.0001	<.0001	0.002	0.020	<.0001	0.109	
Potassium	-0.383	0.443	0.304	0.155	-0.380	-0.272	0.700
p-value	<0.001	<.0001	0.005	0.161	0.000	0.013	<.0001

the validity of using PCA in the analyses to follow. The correlation tables for the watershed characteristics are shown in Appendix C and also exhibit correlations worthy of being analyzed using PCA. A scatterplot matrix for the water quality variables is shown in Appendix C to complement the results of the correlation analysis. The plots are a useful tool for visually observing the bivariate relationships and the correlations between the variables. Pearson's correlation measures the linear association between two variables. It is noteworthy to mention that based on the scatterplots of the water quality data, there does appear to be some relationships that might be more suited to a correlation measure for monotone nonlinear associations such as Spearman's correlations. Although in most of these cases the apparent nonlinear associations seem to be reinforced by outlying observations as can be seen in the relationship between hydrogen ion and ANC. Four observations appear to be highly influential to this association and mask the correlation of the majority of the data for hydrogen ion and ANC. When these four observations (sampling sites 156, 174, 237, and 489) are removed, the Pearson correlation increases from -0.192 to -0.699. Another relationship that is hard to overlook in the scatterplots is the one between conductivity and ANC. The relationship is a good example of where correlation and multicollinearity is induced because of three influential observations (sampling sites 156, 174, and 489). The Pearson correlation for conductivity and ANC is 0.935, the highest of all the correlations. The Pearson correlation is insignificant at 0.011 when these three observations are removed. The strong influence of sampling sites 156, 174, 237, and 489 will need to be dealt with in the analyses to follow.

Because an essential requirement of performing PCA is that correlations exist in the data, correlations for geology, morphology, and vegetation are also calculated. The correlation matrices for geology, morphology, and vegetation are shown in Appendix C. It can be seen in these tables that significant correlations do exist meaning that PCA can be used to analyze their data.

3.6.2 Analysis of the Water Quality Data

The results of the PCA, CA, and DA are presented below in a summary list and in a step-by-step format for the sampling sites and their water quality data (hydrogen ion, ANC, conductivity, chloride, nitrate, sulfate, and sodium+potassium). It should be mentioned that sodium and potassium have been added together for this analysis. These two variables are generally not revered as being robust measures of stream health. Rather, potassium and sodium may act as surrogate variables for calcium and magnesium or other water quality measures, which of course are not included in this study. Since these variables are positively correlated, addition should not induce a canceling effect on their contribution. The step-by-step format provides more of the detailed information that is omitted in the summary. These analyses were performed using the NCSS software and the SAS software package.

SUMMARY

- Robust PCA was performed in NCSS to confirm the results of the data screening that sampling sites 156, 174, 237, and 489 should be classified as outliers. After discovering low down-weights (0.06, 0.01, 0.07, and 0.02, respectively) in the robust PCA, it was decided to remove these four sampling sites from further analyses. Down-weights are applied to outlying observations by the robust PCA to reduce their effect on the total analysis. PCA, CA, and DA with these sites included produced undesirable positive classification rates.

- PCA was performed on the remaining 79 sites. The first three principal components, explaining 86.4 percent of the variability in the water quality variables, were selected for the cluster analysis.

- Fuzzy clustering results showed that the optimum number of clusters for the 79 sampling sites was between eight and 10.
- MODELCLUS and FASTCLUS clustering were performed in SAS. This resulted in the formation of nine clusters using the first three principal components derived from all of the water quality variables. A total of 11 clusters actually exist with sampling site 237 being in cluster 10 alone and sampling sites 156, 174, and 489 being in cluster 11.
- The DA positively classified 90 percent and 95 percent of the sampling sites into the correct cluster using the principal component scores and the original water quality variables, respectively.

STEP 1

The initial concern was the effect of sampling sites 156, 174, 237, and 489 (physical locations are listed in Table 5 below) on the analyses. The sampling sites were analyzed to determine if they should be removed from further analyses because of their influence. Data screening showed that these sampling sites had a profound effect on correlations between certain water quality variables that does not exist between the remaining sampling sites. A robust PCA showed that sampling sites 156, 174, 237, and 489 were assigned small down-weights of 0.06, 0.01, 0.07, and 0.02, respectively. Robust PCA assigns down-weights to outliers to mitigate their influence on the analysis. Initial CA placed site 237 in a cluster alone and sites 147, 156, and 489 were placed in a cluster together. Multiple DA trials usually agreed with the clusters identified by the CA, but the effect of the outlying sampling sites caused the DA classification results for the remaining sampling sites to be very poor. The positive classification results in the DA were between 69 percent and 85 percent when the

Table 5. Physical location of sampling sites 156, 174, 237, and 489.

Sampling Site	Physical Location
156	Abrams Creek above Abrams Creek Ranger Station
174	Abrams Creek below Cades Cove
237	Walker Camp Prong on US 441 west of Newfound Gap
489	Abrams Creek 300m below trailhead bridge in Cades Cove

outlying sampling sites were included. The resulting conclusion was to remove these sampling sites from further analyses.

STEP 2

Step 2 describes the process of selecting the principal components to be used in the cluster analysis. PCA was performed using the remaining 79 sampling sites with sites with all water quality variables. The first three principal components were selected to explain 86.4 percent of the variability in the water quality variables. The eigenvalues of the principal components and the variability explained by each are shown in Table 6 below. The eigenvectors for the principal components are shown in Table 7. Some authors recommend that an eigenvalue cut-off be set at one because a principal component with an eigenvalue below one explains less variability than is contained in a single variable. However, Jolliffe (1972; 2000) believed that this was too restrictive and recommended that a cut-off be set at 0.7. It was decided to add the third principal component because of the increase in the variability explained and the fact that the eigenvalue is near 0.7. The eigenvectors can often be used to provide some interpretation as to the makeup of the principal components based on the original variables. Table 7 shows that the first principal component is mainly a factor of all the water quality variables since the eigenvectors are relatively equal. The difference in signs is important as noticed in the first principal component where the relationship between pH and the variables easily influenced by pH (ANC, nitrate, and sulfate) are contrasted. The second principal component is still influenced by most of the water quality variables with the exception of hydrogen ion and nitrate. Principal component three exhibits a strong influence by chloride with an eigenvector of -0.793.

Table 6. Eigenanalysis in STEP 2 for water quality variables.

Principal Component	Eigenvalue	Proportion of Variability Explained	Cumulative Variability Explained
1	3.536	50.51	50.51
2	1.847	26.39	76.90
3	0.668	9.54	86.44
4	0.498	7.12	93.56
5	0.239	3.41	96.96
6	0.207	2.95	99.92
7	0.006	0.08	100.00

Table 7. Eigenvectors in STEP 2 for water quality variables.

Water Quality Variables	1	2	3
Hydrogen ion	0.413	-0.238	0.230
ANC	-0.382	0.455	0.109
Conductivity	0.339	0.521	0.315
Chloride	0.281	0.380	-0.793
Sulfate	0.476	-0.007	-0.208
Nitrate	0.424	0.307	0.404
Sodium+Potassium	-0.291	0.475	0.035

STEP 3

The principal component scores of the first two principal components for 79 sampling sites in STEP 5 were passed to the MODECLUS procedure in SAS. The MODECLUS procedure is a hierarchical, non-parametric clustering algorithm that uses the *nearest-neighbor* method. MODECLUS provides an initial starting point for subsequent non-hierarchical clustering algorithms that increases cluster accuracy and reduces computation time. A *k* value of three, representing the three *nearest-neighbor*s, was used. *Nearest-neighbor k* values can range from two to *N*. Trial and error using different *k* values proved that a *k* value of three yielded DA results with the highest positive classification ratios. (It can be mentioned at this point that a *k* value of three also produced the best results for the remaining analyses.) When clusters are being formed, the membership of an observation will be determined by the membership of the three *nearest-neighbor*s in the hyper-dimensionality of the data. The result was 12 initial clusters.

STEP 4

The principal component means of the 12 clusters from STEP 6 were passed to the FASTCLUS *k*-means clustering algorithm in SAS. One advantage to the FASTCLUS procedure is that it allows memberships to be re-assigned during the process. In MODECLUS, once a membership has been assigned it cannot be changed. The *maxclusters* parameter was set at 20 initially to give the FASTCLUS procedure the latitude to increase, as well as decrease, the number of final clusters. Subsequent trial and error runs of FASTCLUS using different numbers of *maxclusters* were applied to find the best results based on the DA positive classification results. NCSS fuzzy clustering was used to gain some idea of how many clusters existed for the 79 water quality sampling sites. Fuzzy clustering showed that the maximized and minimized Dunn and Kaufman partition coefficients would support there being between eight and 10 clusters. Table 8 presents the FASTCLUS results of nine clusters that are best in terms of the DA, which will be explained in the following step. Table 9 presents the means of the water quality variables for each cluster. A plot of the principal component scores by cluster is shown in Figure 5. Several of the clusters do exhibit overlapping; this will be discussed in the next section.

STEP 5

The final part of the analysis of the water quality variables was performed to test the discriminating ability of the clusters that were formed in the FASTCLUS analysis. The nine clusters and the principal component scores for the first three principal components were passed to the DISCRIM procedure in SAS. A non-parametric *nearest-neighbor* form of the discriminant analysis is used to cross-validate the cluster memberships. Cross-validation involves removing one observation at a time and forming discriminant rules using *n*-1 sites. The rules are then applied to the observation that was removed to determine if that observation can be classified into the cluster that it was originally assigned (Johnson 1988). A resubstitution method of validation is also available but this method usually overestimates the correctness of the cluster memberships (Johnson 1988). The cross-validation procedure using the first three principal components from STEP 2 misclassified only eight of the 79 sampling sites for a positive classification ratio of 90 percent. The clusters were also cross-validated using the original means of all the water quality variables. The test more rigorously

Table 8. Cluster memberships of sampling sites for water quality variables.

Cluster Number								
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
30	24	23	3	114	43	1	251	47
74	148	34	4	184	45	49	252	191
253	173	52	13	200	46	50		192
	293	127	14	291	66	71		213
		142	20	337	73	115		
		144	147		103	143		
		186	149		104	190		
		193	150		106	194		
		215	266		107	209		
		268	310		137	210		
		311	479		138	214		
		336	480		233	221		
		475	481		234	472		
		484	482		473	474		
		488	483		492			
			485					
			493					

Table 9. Means of the water quality variables by cluster membership.

Cluster	**Hydrogen ion**	**ANC**	**Con**	**Chl**	**Nit**	**Sul**	**Sod+Pot**
1	3.88	70.94	23.07	22.18	37.89	65.23	47.43
2	2.46	128.99	18.89	18.63	7.71	28.53	70.88
3	4.01	57.88	12.67	18.57	11.28	25.62	50.14
4	3.23	81.82	14.30	16.24	6.09	26.23	58.20
5	6.77	33.43	13.57	20.63	30.38	27.66	48.93
6	15.19	18.11	17.86	17.95	37.87	63.64	37.53
7	5.80	41.03	13.16	17.05	16.87	35.17	44.56
8	14.35	21.08	33.88	23.30	46.71	199.56	59.23
9	9.91	20.02	11.47	16.40	19.73	33.87	36.07

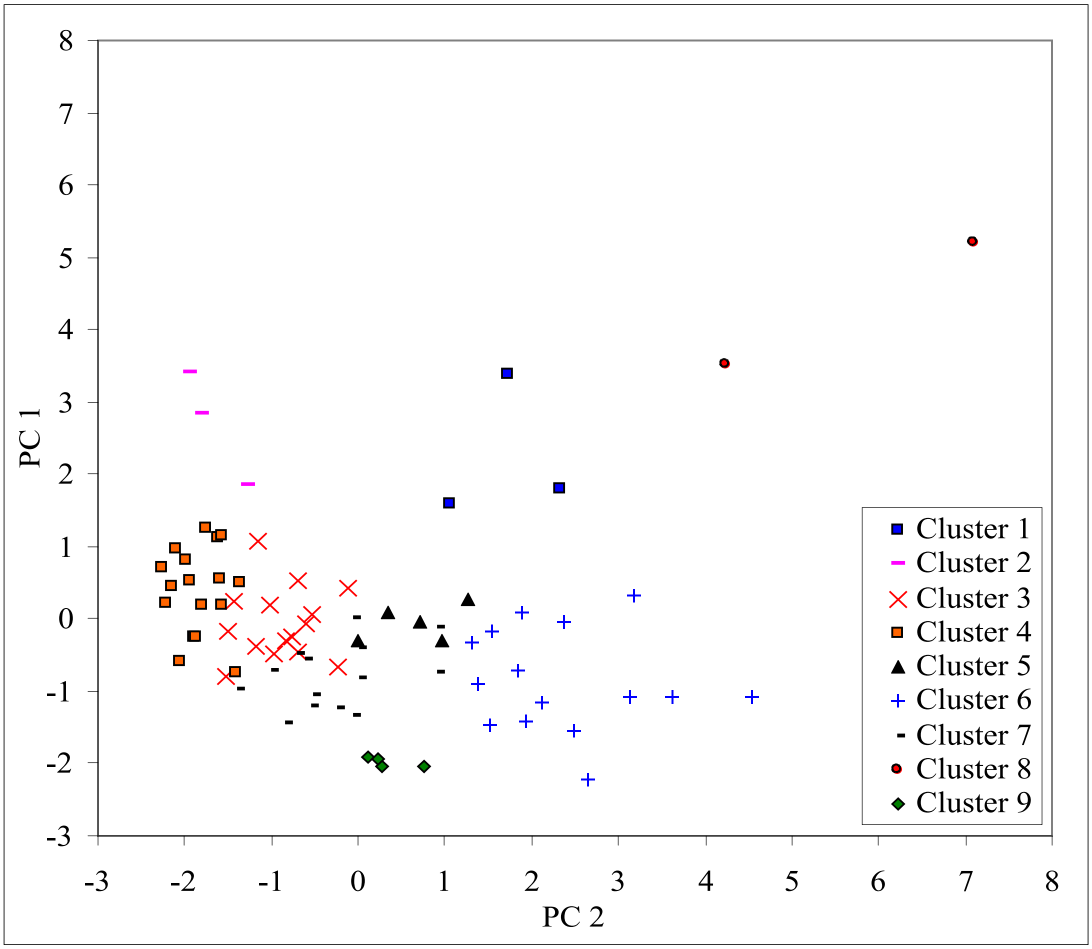

Figure 5. Principal component score plot of the water quality clusters.

determined the discriminating ability of the clusters because the clusters were formed from the first three principal components, which, as an analogy, could be considered a "distilled" version of the original water quality variables. The positive classification ratio using the original variables was 95 percent with five observations misclassified. Table 10 and Table 11 show how the misclassifications were distributed using the principal components and the water quality variables, respectively. Table 12 lists the misclassifications by sampling site ID. Figure 6 presents a plot of the final clusters within the NPS boundary.

Although the DA results were very high, there was some concern about how several of the clusters exhibited overlapping in Figure 5. Overlapping is most prevalent in clusters three, four, five, and seven. These clusters are the source of most of the misclassifications in the DA. Individual misclassified sampling sites were subjectively compared to the means of their respective cluster and the means of the cluster that they were assigned to by the DA. There are no compelling reasons why any of the cluster memberships should be manually changed. Supplementary FASTCLUS procedures were performed by constraining the number of formed clusters to eight, seven, six, and five to forcibly reduce the number of clusters. A new DA was performed at each level to determine if forcible joining would produce better clusters. The results showed that in no case could the reduced number of clusters out perform the nine clusters originally formed.

Table 10. Cross-validation misclassifications using principal components.

Cluster	Frequency	Number Misclassified	Classified to Cluster
3	15	3	4, 5, 7
6	15	1	7
7	14	3	6, 9
8	2	1	1

Table 11. Cross-validation misclassifications using original water quality variables.

Cluster	Frequency	Number Misclassified	Classified to Cluster
2	4	1	4
3	15	1	7
6	15	1	9
7	14	2	5, 9

Table 12. Misclassifications by site ID for the water quality variables.

Site ID	CA Assignment	DA Assignment*
43	6	9
71	7	6
115	7	6, 5
221	7	9 **
251	8	1
268	3	4
293	2	4
475	3	5
484	3	7 **
492	6	7

* *If two clusters are listed, the first is the misclassification in the DA of the principal components and the second is the misclassification in the DA of the original variables*

** *Classifications and misclassifications were the same for the principal components and the original variables*

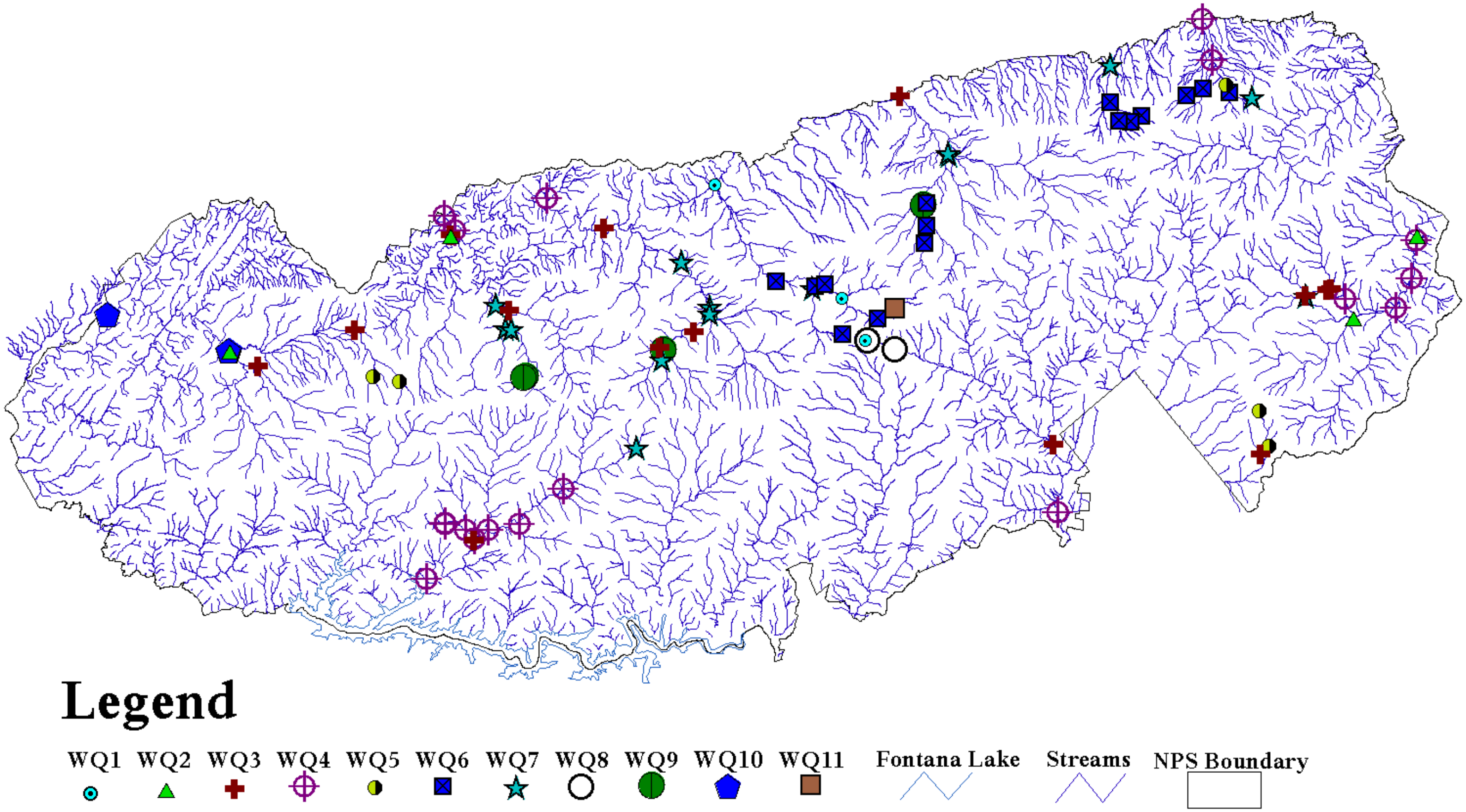

Figure 6. Plot of the water quality clusters within the NPS boundary.

3.6.3 Analysis of the Geology

The analysis of the geology data for the sampling site watersheds are presented in this section using the same format as the water quality variables. The geology data are measured by the percentage of each formation in the watershed upstream of the sampling site and includes Thunderhead sandstone, limestone, Cades Cove sandstone, Anakeesta formation, Elkmont sandstone, Basement complex, and Great Smoky group. This information was compiled by Harwell (2001) from previous work by King, Neuman, et al. (1968) and is shown in tabular form in Appendix D. As in the previous section, a stepwise format will be presented.

<u>SUMMARY</u>

- The initial PCA using all geology variables resulted in five significant principal components with eigenvalues greater than 0.7.
- Fuzzy clustering in NCSS identified the optimum number of clusters being between eight and 10.
- Because there were five significant principal components multivariate variable selection and multiple regression were used to further explore the data in hopes of reducing the dimensionality. Multivariate variable selection resulted in Great Smoky group being removed from the analysis to prevent a singularity problem. With five significant principal components in the initial PCA it was decided to proceed to the multiple regression using all of the vegetation variables except for Great Smoky group.
- The multiple regression results produced perfect R^2 and almost perfect *press*-R^2 values using Thunderhead sandstone, limestone, Cades Cove sandstone, Anakeesta formation, Elkmont sandstone, and Basement complex. Multicollinearity problems were apparent when Great Smoky group was introduced to the analysis.
- A second PCA was performed using Thunderhead sandstone, limestone, Cades Cove sandstone, Anakeesta formation, Elkmont sandstone, and Basement complex because of the significant correlations that remained between these variables. This resulted in four principal components with eigenvalues greater than 0.7.

- The scores of principal components one through four were passed to the MODECLUS procedure in SAS where 12 initial clusters were formed. The seeds of these clusters were passed to FASTCLUS in SAS.

- The FASTCLUS procedure also produced 12 clusters. *Maxclusters* was also set at 11, 10, nine and eight because of the results of the fuzzy analysis saying that the optimum number of clusters should be between eight and 10. Ten clusters ultimately provided the best results in the DA.

- DA was performed on the 10 clusters formed by FASTCLUS. This led to only one misclassification out of 83 sampling sites for a positive classification ratio of 98.8 percent using the principal components and all of the geology variables.

STEP 1

An initial PCA was performed in SAS using the percentages of each formation in the watersheds. The results are shown in Table 13. Based on the eigenanalysis and a cutoff eigenvalue of 0.7, the first five principal components would be needed to analyze the geology data. The eigenvectors of the first five principal components are shown in Table 14. Based on the eigenvectors, principal component one is a measure of the sandstone formation and limestone because of their higher eigenvectors. Principal component two is comprised mainly by Thunderhead sandstone and somewhat by Anakeesta and Great Smoky group. Principal component three is a contrast between Anakeesta and Great Smoky group. Principal component four is strongly influenced by Basement group and principal component five is mainly explained by Elkmont sandstone. Although not shown, the seventh principal component was zero meaning that multicollinearity exists between some of the variables.

STEP 2

Fuzzy clustering in NCSS was performed on the geology variables for all the sampling sites to approximate the optimum number of clusters. NCSS reported that the

Table 13. Eigenanalysis in STEP 1 for the geology variables.

Principal Component	Eigenvalue	Proportion of Variability Explained	Cumulative Variability Explained
1	2.201	31.44	31.44
2	1.420	20.28	51.73
3	1.352	19.32	71.04
4	0.913	13.04	84.08
5	0.765	10.93	95.02
6	0.349	4.98	100.00
7	0.000	0.00	100.00

Table 14. Eigenvectors in STEP 1 for the geology variables.

Geology Variables	1	2	3	4	5
Thunderhead sandstone	-0.413	0.629	0.017	-0.194	0.189
Limestone	0.508	0.227	-0.089	0.179	0.474
Cades Cove sandstone	0.544	0.232	-0.099	0.129	0.255
Anakeesta	-0.129	-0.441	-0.647	0.357	0.069
Elkmont sandstone	0.458	0.166	-0.122	-0.160	-0.765
Basement group	-0.127	0.237	0.328	0.870	-0.251
Great Smoky group	0.180	-0.469	0.664	-0.064	0.147

optimum number of clusters was between eight and 10. The Dunn partition coefficients ranged from 0.9904 to 0.9925 while the Kaufmann partition coefficients ranged from 0.0003 to 0.0054.

STEP 3

Multivariate variable selection was performed in NCSS using principal component scores one through five as the dependent variables and the geology variables as the independent variables to determine if the dimensions of the data could be reduced before clustering. The procedure helped to identify the variable causing the multicollinearity problem discovered in the initial PCA by notification of a singularity problem in the independent variables. Through trial-and-error Great Smoky group had to be removed from the process in order to finalize the multivariate variable selection results. All of the geology variables, with the exception of Great Smoky group, were found to be needed to minimize Wilks' lambda to near zero.

STEP 4

Multiple regression was performed in NCSS to determine the R^2 and *press*-R^2 values between each of the principal components and the set of geology variables. Initially, all geology variables were included in the analysis and this, of course, resulted in perfect correlation between the independent variable (principal component) and the geology variables. However, this exercise did identify Great Smoky group as the source of the multicollinearity problem by producing high condition numbers and variance inflation factors. When Great Smoky group was removed the correlations were still perfect, i.e. equal to one. Forthcoming analysis will conclude that even fewer geology variables are actually needed to explain the geologic variability among the watersheds. However, because five principal components had eigenvalues greater than 0.7 for only seven variables, it was decided to leave all variables in the analysis with the exception of Great Smoky group.

STEP 5

A second PCA was performed using Thunderhead sandstone, limestone, Cades Cove sandstone, Anakeesta formation, Elkmont sandstone, and Basement complex. The results of the eigenanalysis are shown in Table 15. The first four principal components should be retained based on an eigenvalue cutoff of 0.7. The eigenvectors of the first four principal components are shown in Table 16. The eigenvectors of the first principal component show the large influence of Cades Cove sandstone, limestone, and Elkmont sandstone. The second principal component is displays the contrast between Anakeesta and Thunderhead sandstone. The third principal component is strongly influenced by Basement complex and the fourth principal component is a contrast between Elkmont sandstone and limestone.

STEP 6

The first four principal components were passed to MODECLUS in SAS. The MODECLUS analysis resulted in the initial formation of 12 clusters. The cluster means of the principal components were passed to FASTCLUS for *k*-means clustering.

STEP 7

FASTCLUS *k*-means clustering was performed next using SAS. With *maxclusters* set equal to 20 there were 12 clusters formed. Because the NCSS fuzzy clustering results suggested that the actual number of clusters was between eight and 11, *maxclusters* was set to 11, 10, nine, and eight on successive attempts. Ten clusters

Table 15. Eigenanalysis in STEP 5 for the geology variables.

Principal Component	Eigenvalue	Proportion of Variability Explained	Cumulative Variability Explained
1	2.168	36.13	36.13
2	1.398	23.29	59.42
3	0.917	15.28	74.70
4	0.791	13.18	87.89
5	0.379	6.32	94.21
6	0.347	5.79	100.00

Table 16. Eigenvectors in STEP 5 for the geology variables.

Geology Variables	1	2	3	4
Thunderhead sandstone	-0.368	0.542	-0.264	0.317
Elkmont sandstone	0.484	0.062	-0.142	-0.676
Limestone	0.528	0.135	0.131	0.508
Cades Cove sandstone	0.567	0.133	0.091	0.305
Anakeesta	-0.101	-0.724	0.296	0.229
Basement group	-0.134	0.376	0.893	-0.195

produced the lowest misclassification rates in the DA, which will be presented in the step that follows this section. The cluster memberships for the sampling sites based on the geology data are shown in Table 17 and the cluster means are shown in Table 18. Figure 7 shows the principal components plot by cluster for principal components one and two. Although additional principal components were used in the clustering, principal components one and two account for almost 60 percent of the variability in the geology data and the plot shown is presented for displaying the segregation of the clusters.

STEP 8

DA was performed on the cluster results from the FASTCLUS procedure. The positive classification ratios were extremely high at 98.8 percent. In both DA cases using the principal components and the original geology data for all 83 sampling sites only one misclassification occurred for the cross-validation tests. The single misclassification occurred in cluster two where one member was incorrectly classified into cluster five. Figure 8 shows a map of the geologic clusters within the NPS boundary.

3.6.4 Analysis of the Morphology

The morphology of the watersheds was analyzed using the same process as the previous cases. This approach provided the best results; however, the exact number of clusters was more difficult to determine compared with previous analyses. The variables included in the morphology analysis were: stream elevation, mean basin elevation, stream order, maximum channel length, basin length, basin area, stream density, mean

Table 17. Cluster memberships of sampling sites for geology variables.

Cluster Number									
1	2	3	4	5	6	7	8	9	10
1	13	45	142	3	46	30	147	156	24
4	23	74	144	14	73	43	149	174	173
47	49	215	266	20	192	66	150	489	184
103	148	233	268	34	252	190	493		186
104	191	234		50	253	193			200
106	221	237		52	473	194			488
107	310			71		251			
114	311			209		472			
115	479			210					
127	480			213					
137	481			293					
138	482			474					
143	483			475					
214	484								
291	485								
336									
337									
492									

Table 18. Percentage means of geology data by cluster (rounded to nearest tenth).

Cluster	Thunderhead sandstone	Limestone	Cades Cove sandstone	Anakeesta	Elkmont sandstone	Basement complex	Great Smoky group
1	97.6	0	0	0.2	0	0	2.2
2	10.6	0	0.2	5.7	2.6	0	80.8
3	2.0	0	0	96.6	0	0	1.4
4	69.8	0	0	9.3	0	4.4	16.4
5	69.1	0	0.1	16.1	4.1	0	10.6
6	19.6	0	0	75.1	0	0	5.4
7	44.5	0	0	44.6	0.7	0	10.2
8	66.1	0	0	0	0	1.3	32.7
9	0.4	13.2	20.1	0	42.5	0	23.6
10	7.0	0.5	4.3	0	75.8	0	12.4

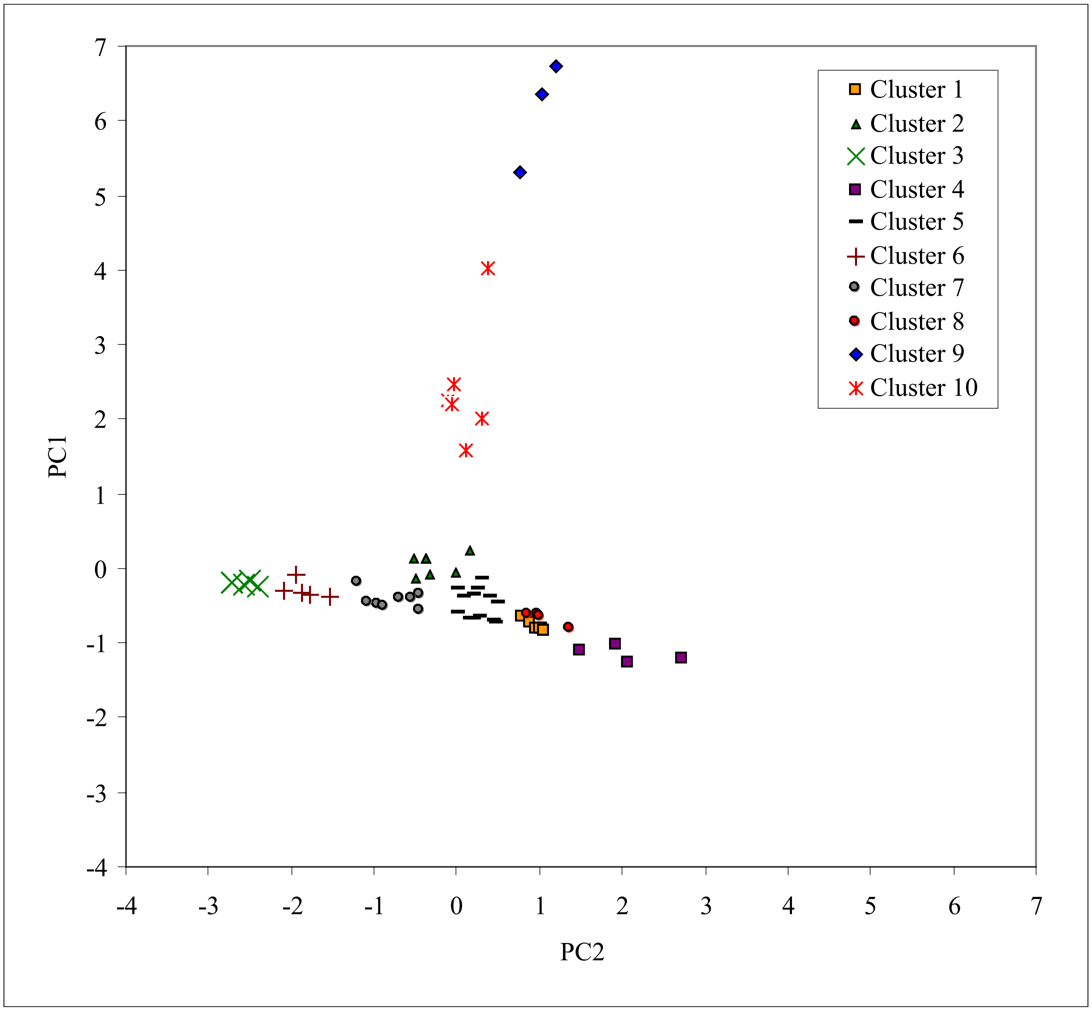

Figure 7. Principal component score plot of the geology clusters.

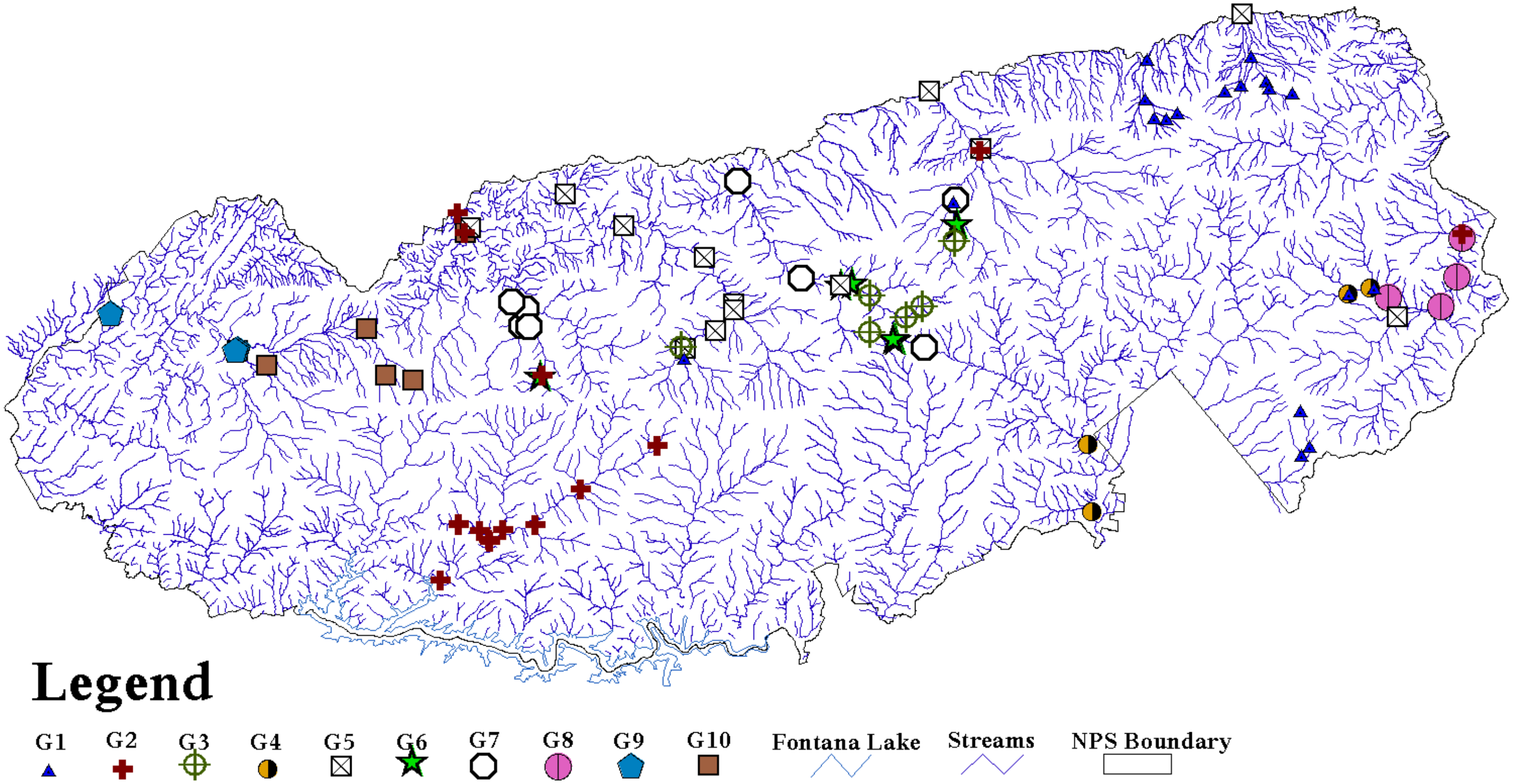

Figure 8. Plot of the geology clusters within the NPS boundary.

basin slope, channel slope, and basin width. The units of measurement are metric except for stream order which is an ordinal variable. The data were also compiled by Harwell (2001). The morphology data are included in Appendix D. The process for clustering the morphology data are recorded in a summary and step-by-step format similar to previous analyses.

SUMMARY

- An initial PCA was performed using all of the morphology variables. The first three principal components with eigenvalues greater than 0.7 explained 84.7 percent of the variability in the morphology variables.

- Multivariate variable selection was performed using the first three principal components and the morphology variables. Stream elevation, stream order, basin area, stream density, mean basin slope, and channel slope were identified as the main components of the first three principal components. The remaining variables caused singularity problems in the analysis.

- Multiple regression was performed using the first three principal components as dependent variables and stream elevation, stream order, basin area, stream density, mean basin slope, and channel slope as independent variables. R^2 and *press*-R^2 values ranged from 0.90 to 0.99 meaning that these variables were good predictors of the principal components. Multicollinearity was a problem when additional variables were introduced to the analysis.

- Fuzzy clustering was performed using all of the morphology variables. The results were rather vague with the number of possible clusters ranging from three to 15.

- A second PCA was performed in SAS using stream elevation, stream order, basin area, stream density, mean basin slope, and channel slope because of significant correlations that remained between these variables. The first three principal components had eigenvalues greater than 0.7 and explained 85.3 percent of the variability in the morphology variables.

- MODECLUS analysis was performed using the first three principal components from the second PCA. Eleven initial clusters were formed by this procedure.

- A FASTCLUS analysis was performed using the cluster seeds from the MODECLUS analysis. Nine clusters were initially formed. Smaller numbers of clusters were tried by iteratively changing *maxclusters* from nine to three and observing the DA results. Five clusters ultimately provided the best positive classification results in the DA.

- DA was performed using the cluster results from the FASTCLUS procedure. The positive classification result using the principal components was 98 percent while the positive classification results using the original variables was 90.

STEP 1

PCA was performed on the morphology data using all of the sampling sites. The eigenvalues are shown in Table 19 with the proportional and cumulative variability explained. The first three principal components, with eigenvalues greater than 0.7, explain 84.7 percent of the variability in the data. The eigenvectors of the first three principal components are shown in Table 20. Based on the eigenvectors in Table 20, the first principal component is a contrast between elevation, slope and the remaining variables. In general, as slope and elevation increase, values for the remaining variables decrease. The first principal component is a general measure of all the morphology variables because the majority of the eigenvectors are fairly equal in magnitude. However, the magnitudes of the eigenvectors for principal components two and three are different. Principal component two is mainly explained by the contributions of elevation and stream density. Once again the difference in signs gives additional interpretation that stream density decreases as elevation increases. Principal component three is strongly influenced by mean basin slope.

STEP 2

Multivariate variable selection was performed in NCSS to determine which of the morphology variables are most important for the explanation of the first three

Table 19. Eigenanalysis in STEP 1 for the morphology variables.

Principal Component	Eigenvalue	Proportion of Variability Explained	Cumulative Variability Explained
1	6.153	61.53	61.53
2	1.258	12.58	74.11
3	1.058	10.58	84.69
4	0.570	5.70	90.40
5	0.493	4.93	95.33
6	0.235	2.35	97.68
7	0.144	1.44	99.12
8	0.042	0.42	99.55
9	0.031	0.31	99.86
10	0.014	0.14	100.00

Table 20. Eigenvectors in STEP 1 for the morphology variables.

Morphology Variables	1	2	3
Stream elevation	0.335	0.405	-0.030
Mean basin elevation	0.277	0.428	0.326
Stream order	-0.358	-0.022	0.174
Max. channel length	-0.373	0.172	0.099
Basin length	-0.385	0.124	0.094
Basin area	-0.363	0.229	0.138
Stream density	-0.175	-0.602	-0.157
Mean basin slope	0.096	-0.315	0.856
Channel slope	0.296	-0.225	0.195
Basin width	-0.367	0.204	0.170

principal components. Using the principal components as the dependent variables and the morphology variables as the independent variables, stream elevation, stream order, basin area, stream density, mean basin slope, and channel slope were identified to minimize Wilks' lambda to zero.

STEP 3

Multiple regression was performed using NCSS. The principal components were once again used for the dependent variables and the morphology variables identified in STEP 2 were used for the independent variables. Stream order is an ordinal variable that ranges from one to five. In multiple regression ordinal variables must be converted to a set of n-1 dummy variables where $n = 5$ for the number of stream orders. The stream order of three was used as the baseline order and four dummy variables were created for stream orders one, two, four, and five. The R^2 and *press*-R^2 results showed that stream elevation, stream order, basin area, stream density, mean basin slope, and channel slope are in fact all good predictors of the first three principal components. R^2 and *press*-R^2 values ranged from 0.90 to 0.99. Variables omitted were added to the regression procedure only to verify that multicollinearity existed when they were included, which turned out to be the case.

STEP 4

Fuzzy clustering was performed in NCSS using the morphology variables. The results were not as definite as in the previous analyses. The optimum number of clusters fell in a wide range from three to 15, all with high Dunn partition coefficients and low Kaufman partition coefficients.

STEP 5

A new PCA was performed in SAS using stream elevation, stream order, basin area, stream density, mean basin slope, and channel slope. Significant correlations exist between some of these variables meaning that PCA remains to be a valid tool for analysis. The first three principal components had eigenvalues greater than 0.7. The eigenvalues for the first three principal components are shown in Table 21 and the corresponding eigenvectors are shown in Table 22. The first three principal components explain 85.3 percent of the variability in the morphology variables listed. Principal component one can be interpreted in the same way as the PCA in STEP 1. Principal components two and three are mainly explained by mean basin slope and stream density, respectively.

STEP 6

The principal component scores for the first three principal components were passed to the MODECLUS procedure in SAS. Eleven clusters were formed by MODECLUS. The principal component cluster seeds were passed to the FASTCLUS procedure.

STEP 7

The FASTCLUS *k*-means clustering procedure was performed using the seeds from the MODECLUS procedure. Initially, nine clusters were formed with the *maxclusters* parameter set to 20. The cross-validation DA produced a low positive classification rate at 72 percent. The *maxclusters* parameter was then iteratively decreased from nine to three while the DA results were observed. A cluster number of

Table 21. Eigenanalysis in STEP 5 for the morphology variables.

Principal Component	**Eigenvalue**	**Percentage of Variability Explained**	**Cumulative Variability Explained**
1	3.160	52.67	52.67
2	1.144	19.07	71.74
3	0.811	13.52	85.25
4	0.416	6.94	92.19
5	0.298	4.97	97.17
6	0.170	2.83	100.00

Table 22. Eigenvectors in STEP 5 for the morphology variables.

Morphology Variable	**1**	**2**	**3**
Stream elevation	-0.475	-0.291	0.024
Stream order	0.512	0.093	0.239
Basin area	0.468	-0.107	0.262
Stream density	0.293	0.395	-0.793
Mean basin slope	-0.135	0.792	0.473
Channel slope	-0.434	0.334	-0.144

five produced the best positive classification results, which will be explained in the following step. The final clusters memberships for the sampling sites are shown in Table 23 and the morphology cluster means are shown in Table 24.

STEP 8

DA was performed on the FASTCLUS cluster results in SAS using the principal components and all of the original morphology variables. The positive classification rate using the principal components was 98 percent. One observation from cluster one was classified into cluster three and one observation from cluster four was classified into cluster one. The positive classification rate using the original morphology variables was 90 percent. The misclassifications using the morphology variables are shown in Table 25. The principal components plot for principal components one and two by cluster is shown in Figure 9. A plot of the clusters within the boundary of the NPS is shown in Figure 10.

3.6.5 Analysis of the Vegetation

The vegetation variables are analyzed in this section of the study. The vegetation data for the GRSM were compiled by MacKenzie (1993) and divided into the percentages of each type of vegetation in the sampling site watersheds by Harwell (2001). The vegetation types used in this analysis include the following: northern hardwood, spruce-fir, cove hardwood, mesic oak, mixed mesic oak, tulip poplar, pine, heath bald, xeric oak, and pine-oak. Other vegetation types were identified by MacKenzie (1993) but their contribution on a percentage basis was very small. According to his findings, the vegetation types listed above account for approximately

Table 23. Cluster memberships of sampling sites for morphology variables.

Cluster Number				
1	2	3	4	5
1, 43, 49, 66, 73, 127, 144, 148, 186, 190, 193, 210, 293, 310, 472, 473, 475, 480, 484, 488, 493	103, 115, 234, 252, 253, 291	3, 13, 14, 20, 23, 24, 30, 34, 50, 52, 147, 149, 150, 156, 173, 174, 194, 266, 268, 311, 474, 479, 489	71, 74, 142, 143, 184, 191, 200, 213, 214, 215, 221, 233, 237, 251, 336, 337, 482, 483, 485, 492	4, 45, 46, 47, 104, 106, 107, 114, 137, 138, 192, 481

Table 24. Means of the morphology variables by cluster membership.

Morphology Variable	1	2	3	4	5
Stream elevation (m)	762	1416	535	1056	914
Mean basin elevation (m)	1250	1566	1070	1345	1332
Stream order	3	1	4	2	2
Max. channel length (km)	8.9	0.9	20.0	4.2	3.4
Basin length (km)	6.7	0.7	12.8	3.2	2.9
Basin area (ha)	2429	43	10403	459	369
Stream density (km/km^2)	1.5	0.4	1.7	1.4	1.8
Mean basin slope (%)	48	45	43	44	55
Channel slope (%)	9.3	27.3	4.8	12.5	23.4
Basin width (km)	3.5	0.5	7.9	1.3	1.2

Table 25. Cross-validation misclassifications using the morphology variables.

Cluster	Frequency	Number Misclassified	Classified to Cluster
1	21	1	4
3	23	2	1
4	21	2	1
5	12	3	4

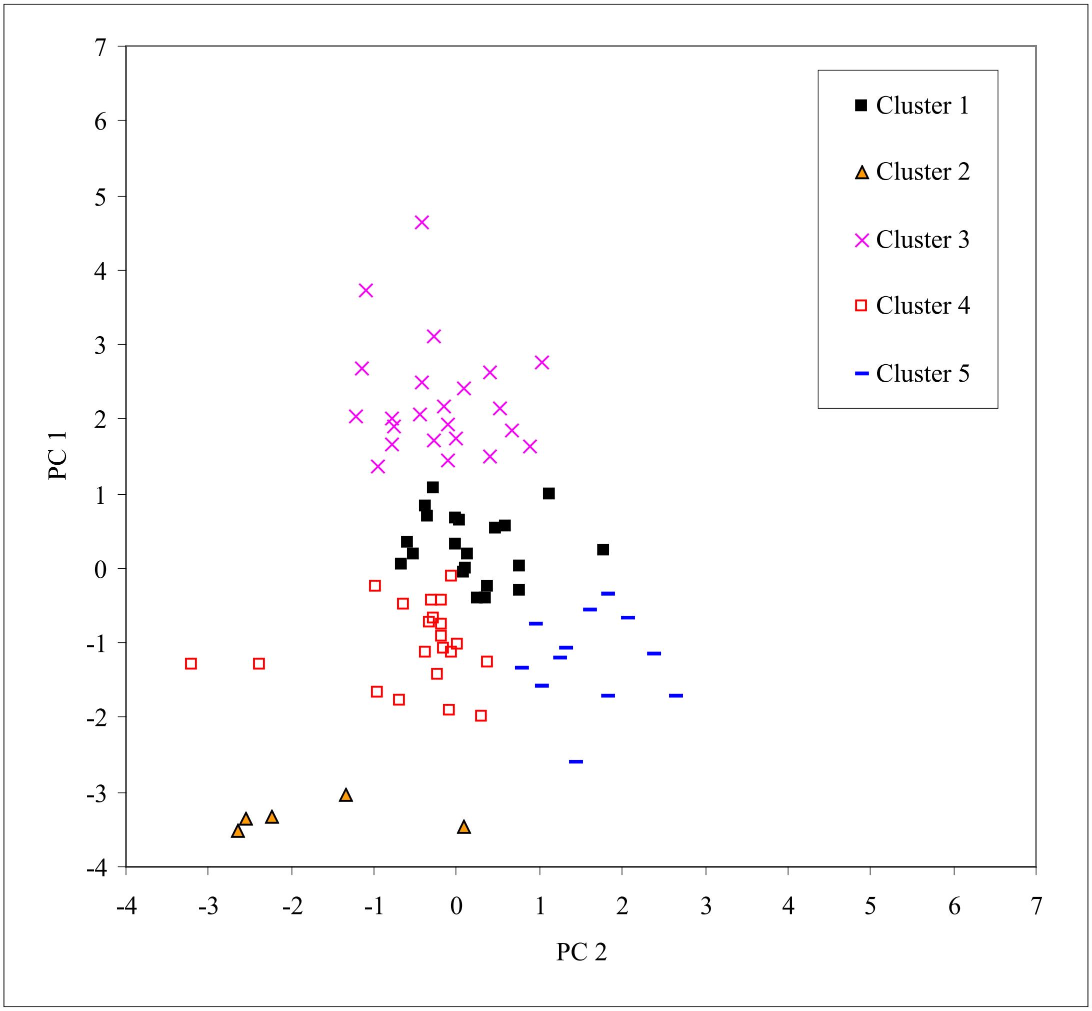

Figure 9. Principal component score plot of the morphology clusters.

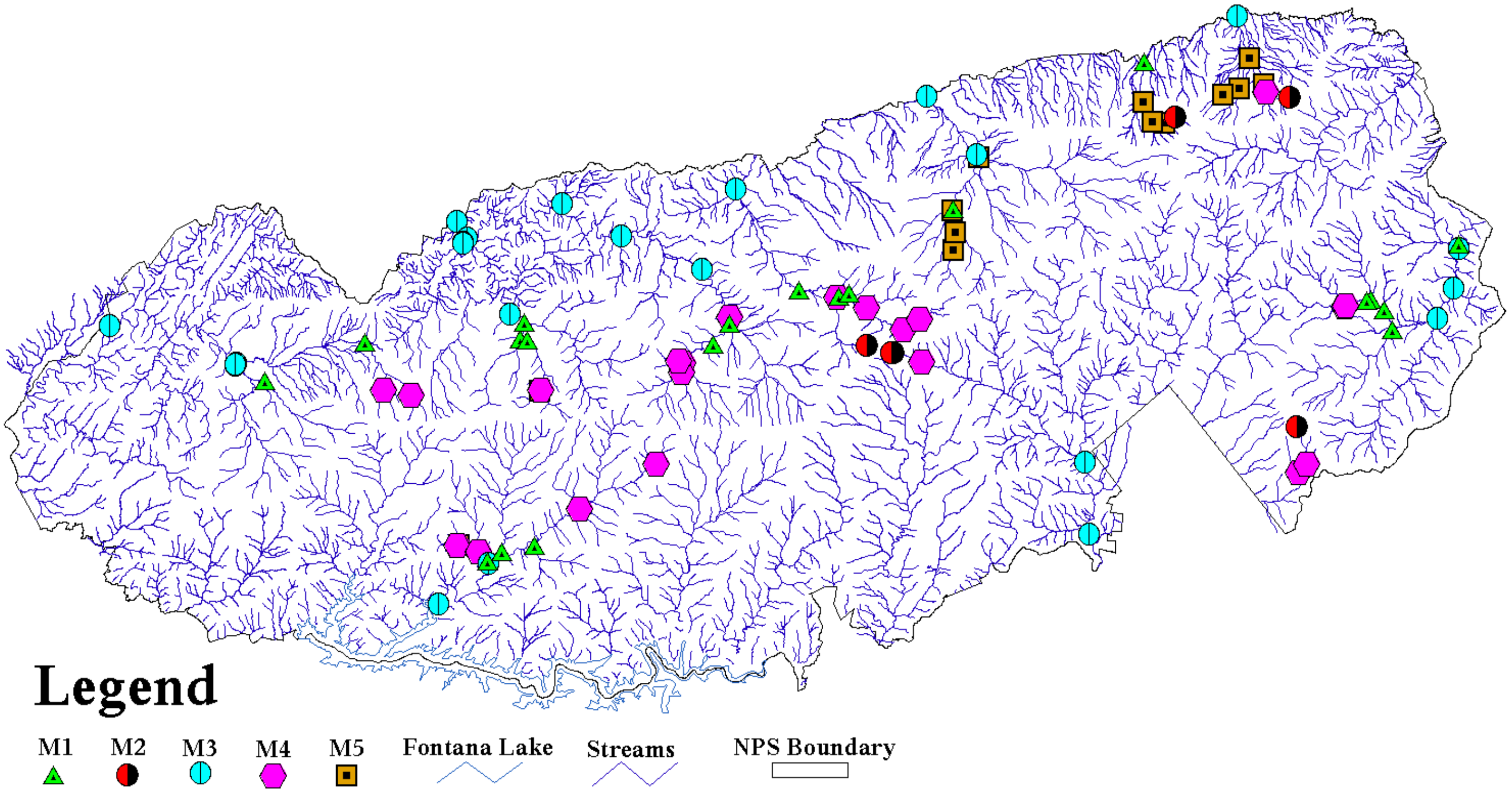

Figure 10. Plot of the morphology clusters within the NPS boundary.

98 percent of the vegetation in the GRSM. The vegetation variables are similar to the geology variables in the sense that the measurements are based on the percentages of each type of vegetation in the watersheds of the sampling sites. Summary and step-by-step formats of the clustering procedure are presented below.

The methods used in this section of the study are somewhat different from the previous sections. In short, the results of the clustering using the PCA components did not form distinct clusters. Although the clusters did perform well in the DA, the principal components plot showed that four of nine clusters have severe overlapping problems. However, PCA was used to help determine if a smaller number of variables can be used for clustering.

SUMMARY

- An initial PCA was performed on all of the vegetation variables. This resulted in the identification of four principal components with eigenvalues greater than 0.7 and 82.5 percent of the variability explained.

- Multivariate variable selection was performed using the first four principal components and all of the vegetation variables. This resulted in a four-variable model composed of spruce-fir, northern hardwood, mesic oak, and heath bald, and a five-variable model composed of spruce-fir, northern hardwood, mesic oak, pine, and heath bald.

- Multiple regression was performed using the four and five-variable models from the multivariate variable selection step. Principal components one through four were used as the dependent variables. Both models performed very well with R^2 and *press*-R^2 values ranging from 0.91 to 0.99. Multicollinearity problems were identified when mixed-mesic oak was introduced. At this point neither model seemed superior to the other.

- Fuzzy clustering failed to produce any concrete results as to the number of clusters in the vegetation variables. The group-average hierarchical method and *k*-means non-hierarchical method were used in NCSS to gain some foreknowledge as to the number of clusters in the vegetation variables. Using joining distances in the group-average hierarchical method nine clusters were identified as optimal. The *k*-means method identified either nine or 10 clusters as optimum based on the sum of the minimized within cluster sum of squares.

- *K*-means clustering was then applied to the four and five-variable models identified in the multivariate variables selection process. The five-variable model of spruce-fir, northern hardwood, mesic oak, pine, and heath bald for nine clusters outperformed the four-variable model in the DA.

- DA was performed in SAS using the clusters of the *k*-means analysis in NCSS. Using all of the original variables, the positive classification rate was 95 percent.

STEP 1

An initial PCA was performed using all of the vegetation variables. The eigenvalues of the PCA are shown in Table 26. Four principal components can be identified with eigenvalues greater than 0.7 and with 82.5 percent of the variability in the vegetation variables explained. The eigenvectors of the first four principal components are listed in Table 27. It can be seen that principal component one does not have a clear interpretation in terms of one or two of the vegetation variables because most of the eigenvectors are fairly equal. However, xeric oak and pine-oak do have somewhat higher eigenvectors meaning that these vegetation variables contribute most to the explanation of principal component one. Principal component two has relatively higher eigenvectors for mesic oak, cove hardwood, and spruce-fir. Principal component three is mainly explained by the contributions of tulip poplar, mesic oak, and heath bald. The main contributing vegetation variables for principal component four are spruce-fir, northern hardwood, and heath bald.

STEP 2

Multivariate variable selection was performed with the first four principal components and the vegetation variables. Wilks' lambda was minimized to zero to form

Table 26. Eigenanalysis in STEP 1 for the vegetation variables.

Principal Component	**Eigenvalue**	**Percentage of Variability Explained**	**Cumulative Variability Explained**
1	3.981	39.81	39.81
2	1.914	19.14	58.95
3	1.615	16.15	75.09
4	0.737	7.37	82.46
5	0.660	6.60	89.06
6	0.417	4.17	93.22
7	0.324	3.24	96.46
8	0.242	2.42	98.88
9	0.107	1.07	99.95
10	0.005	0.05	100.00

Table 27. Eigenvectors in STEP 1 for the vegetation variables.

Vegetation Variable	**1**	**2**	**3**	**4**
Spruce-fir	0.282	-0.490	0.058	-0.563
Northern hardwood	0.322	-0.344	-0.138	0.423
Cove Hardwood	0.272	0.508	0.188	0.139
Mesic oak	0.074	0.498	-0.444	-0.040
Mixed mesic hardwood	0.392	0.112	-0.276	-0.234
Tulip poplar	0.184	0.105	0.634	0.151
Pine	0.388	-0.162	0.251	0.319
Heath bald	-0.160	0.267	0.419	-0.513
Xeric oak	0.459	-0.079	0.131	-0.000
Pine-oak	0.440	-0.089	-0.106	-0.202

a five-variable model comprised of spruce-fir, northern hardwood, mesic oak, pine, and heath bald.

STEP 3

Multiple regression was performed using the five-variable model identified in STEP 2. The five-variable performed very well with high R^2 and *press*-R^2 values that ranged from 0.91 to 0.99 for all four principal components as dependent variables. Other variables were added to the regression model to determine their predictive capabilities. The models did not provide a marked increase in predictive ability compared to the five-variable model identified by the multivariate variable selection technique. The models did verify a multicollinearity problem when mixed mesic hardwood was introduced as an independent variable by generating high condition numbers and variance inflation factors.

STEP 4

Fuzzy clustering was performed to determine an optimum number of clusters for the vegetation variables. The method failed to provide optimized Dunn and Kaufman partition coefficients for cluster sizes from two through 15. Two alternative methods were used to gain some understanding as to the number of clusters that should be expected from the vegetation variables.

The first method was the group-average hierarchical method of clustering. Remember that hierarchical clustering begins with all observations in different clusters and eventually joins the observations to form one cluster using distance measures. Observations are joined to form a smaller number of clusters by selecting observations with the smallest distance to a particular cluster. There usually comes a point where the addition of other observations to a cluster requires a large increase in distance to allow the next closest observation to join. The number of clusters prior to the largest increase in joining distance has been identified as a possible solution for the optimum number of clusters (Jobson 1992). The iterative joining report generated by NCSS allows the identification of this point. The largest incremental increase in cluster joining distance for the vegetation variables occurred when decreasing from nine clusters to eight clusters. Two goodness-of-fit measures are used to assess the results. The first is the cophenetic correlation measure, which was 0.865 for this analysis. The cophenetic correlation measures the correlation between the original distances and the distances created by the clustering. The second goodness-of-fit measure in NCSS is the deltas. Delta (0.5) and delta (0.1) were 0.171 and 0.211, respectively. The deltas measure the degree of distortion in the formed clusters (Hintze 2001) and ideally should be very close to zero. Hence, the optimum number of clusters identified by the group-average hierarchical method in NCSS is nine.

The second method for determining the number of clusters was performed using a *k*-means cluster analysis for a range of clusters. The range of the number of clusters for the vegetation variables was set from two to 15. The minimum iteration section of the *k*-means cluster analysis gives the summation of the within-cluster sum of squares for all the clusters in a solution. The number of clusters where the summation of the within-cluster sum of squares reaches a point of diminishing return is often noted as the optimum number of clusters (Hintze 2001). This point occurred at nine and 10 clusters. Based on the *k*-means cluster analysis, the optimum number of clusters is thought to be between nine and 10.

STEP 5

K-means clustering was performed using FASTCLUS in SAS for the five-variable model from STEP 2. Nine clusters were formed to give the best results in terms of positive classification percentage in the DA. The positive classification ratio for the five-variable model of spruce-fir, northern hardwood, mesic oak, pine, and heath bald was 95 percent (four misclassifications out of 83 sampling sites). Other cluster sizes were tested but none could outperform the nine cluster model in the DA. The DA was tested against all 10 of the vegetation variables, which means that the reduced model outperforms the cluster results of the full 10 variable model. The cluster memberships are shown in Table 28 and the cluster means are shown in Table 29. A plot of the clusters within the NPS boundary is shown in Figure 11.

3.6.6 Identification of Sites in Overlapping Regions of Clusters

Sampling sites in the overlapped regions of clusters have questionable memberships because some of their properties may be characterized by the elements of two or more clusters. These sites are unique and should probably be retained in redesigned network because their watershed characteristics are unusual when compared to the majority of the sites. The sites in the overlapped regions of clusters can easily be identified in the DA by noting those sites that are misclassified in the cross-validation. Sites in overlapping clusters were identified on the basis of combined watershed characteristics and the water quality variables. PCA, CA, and DA were performed on the sampling sites after combining the geology, morphology, and vegetation variables into one data file. This three step method is very similar to the one that has been used

Table 28. Cluster memberships of sampling sites for the vegetation variables.

Cluster Number								
1	2	3	4	5	6	7	8	9
47	148	3	20	184	233	66	24	1, 4, 115,
114	480	13	30	221	234	71	156	127, 137,
190	481	14	34	336	237	73	173	138, 142,
191	482	23	43		291	74	174	143, 144,
192	483	186	45			103	489	147, 149,
193		488	46			104		150, 200,
213			49			106		251, 266,
214			50			107		268, 293,
215			52			252		310, 311,
475			194			253		337, 479,
492			209			473		484, 485,
			210					493
			472					
			474					

Table 29. Means of the vegetation variables by cluster membership.

Vegetation variable	1	2	3	4	5	6	7	8	9
Spruce-Fir	1.9	0.02	0.8	8.5	0.8	46.8	22.4	0	2.3
Northern hardwood	22.5	0.3	11.9	11.2	61.4	30.0	30.8	4.9	17.8
Cove hardwood	51.9	20.3	38.6	54.0	26.2	21.1	41.4	18.3	51.2
Mesic oak	10.5	18.3	5.3	5.0	5.3	0.6	0.7	2.9	12.9
Mixed mesic hardwood	4.0	41.0	15.8	7.9	5.3	0.1	1.0	17.1	11.7
Tulip Poplar	2.0	0.7	6.0	5.5	0.6	0.2	0.5	4.5	0.5
Pine	0.2	3.0	11.2	2.0	0.2	0.1	0.1	23.1	1.0
Heath Bald	4.7	0	1.1	2.2	0	0.6	1.4	0.1	0.4
Xeric oak	2.2	13.7	8.0	3.3	0	0	0.03	20.6	1.4
Pine-oak	0.03	2.7	1.0	0.2	0	0	0.05	2.0	0.3

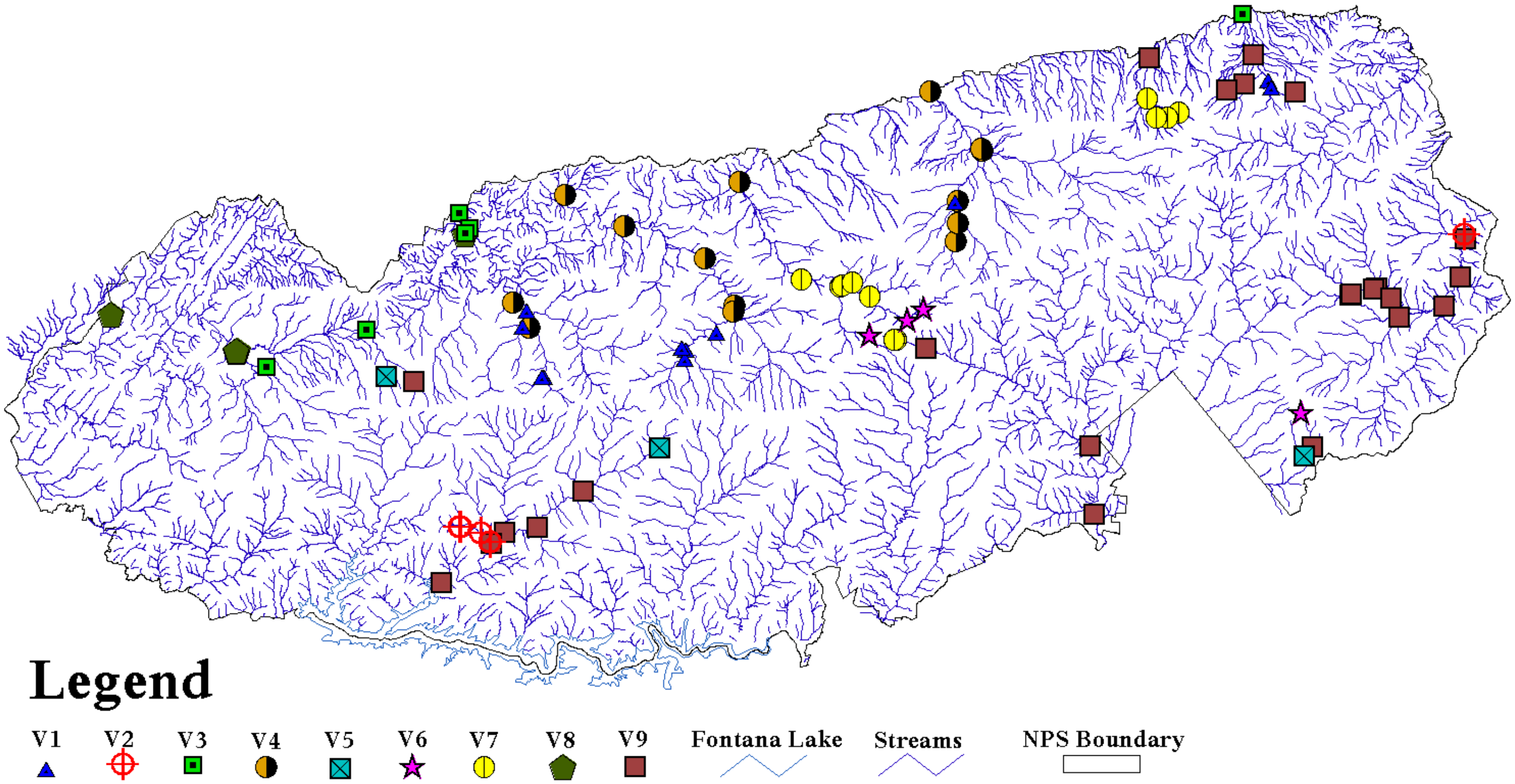

Figure 11. Plot of the vegetation clusters within the NPS boundary.

thus far. Therefore, only a brief summary of the results will be presented. The sites in the overlapped regions of the water quality clusters were determined from the analysis in Section 3.6.2 and are also listed in this section.

The PCA was performed using the combined total of 28 variables in the watershed characteristics. Nine principal components with eigenvalues greater than 0.7 explained 87.6 percent of the total variability in the 28 variables. The CA resulted in the formulation of 10 clusters. The DA resulted in positive classification rates of 88 percent and 89 percent using the principal components and the original variables, respectively. The misclassified sites were: 49, 50, 52, 71, 74, 127, 147, 149, 150, 191, 192, 210, 221, and 474. The sites in the overlapped regions of the water quality clusters of Section 3.6.2 were 43, 71, 115, 221, 251, 268, 293, 475, 484, and 492. Combining these two sets of sites gives a total of 22 sites that are considered to be in overlapping regions of clusters.

3.6.7 Collocated Sampling Sites

The information for the collocated sampling sites was compiled using ESRI ArcView GIS software. The collocated site information was obtained directly from the GRSM NPS. (In order to differentiate between water quality sampling sites and the sites obtained from the NPS, the water quality sampling sites will simply be referred to as “sampling sites” and the sites obtained from the NPS will be referred to as “ecological sites” in this section) The coordinates of the ecological sites were downloaded as a database table to an existing ArcView project of the GRSM that contained the sampling sites. A query was performed in ArcView to find the sampling sites which had ecological sites located within a radius of 100 meters. The number of ecological sites inside this radius for each sampling site was also counted. A sampling site with one ecological site within 100 meters received a collocation benefit score of 20 and a sampling site with two ecological sites received a collocation benefit score of 30. These collocation benefit scores were chosen because they were similar in magnitude to the clustering benefit scores. Sampling sites with no ecological sites located within the 100 meter radius received a collocation benefit score of zero, even if ecological sites were located upstream or downstream outside of the 100 meter radius. The assigned collocation benefit scores are listed in Appendix E.

3.6.8 Compilation of Data for Optimization

The collocation information and the clustering results of the water quality data and the watershed characteristics were compiled to generate a total benefit score for each sampling site to be used in the SA algorithm. The information is shown in tabular form in Appendix E. The total benefit score was calculated according to Equation 3.7 and all weights were assigned a value of one except for water quality which was assigned a weight of two. Additional analyses will be conducted in the sensitivity analysis to explore the use of weights assigned to each category of water quality, geology, morphology, vegetation, and collocation. Appendix E presents a table of the one-way distances between sites, sampling time spent at each site, and the total dollars spent each year based on four samples per year at $30 per man-hour. The distances are based on the scheme presented in Figure 2.

3.6.9 Sampling Site Selection Optimization using SA

This section presents the results of performing SA using the benefits and the costs of the GRSM water quality monitoring network using the clustering results, collocation information, and the costs of the present sampling scheme. The computational runs are performed using the current sampling scheme of four samples per sampling site location per year. However, recommendations may be submitted at the end of this study that change the location and frequency of sampling. The combination of possible sampling frequency changes and selection of which sites can be discontinued will be very important in saving money and more importantly, ensuring that a monitoring network is in place that will have the ability to detect long-term trends on a site-by-site basis and on an overall network basis.

Recall from earlier discussions that the SA algorithm was written to perform the retaining or discarding of sampling sites for two different problems. The first was where the user would input an arbitrary selection of sites and the algorithm would evaluate the network as a whole and maximize the objective function for a global solution that would recommend sampling sites to be retained and sampling sites to be discarded. The second form of the algorithm begins with a user-specified number of sites, *n*, (the initial network of sites are chosen randomly by the algorithm) for the final monitoring network to consist of, and then seek to find the best subset of *n* sampling sites with a maximum objective function for the final monitoring network. The results of both of these analyses will be presented here and some inference will be given as to how their relationship complements each other. Results of sensitivity analyses will also be presented to assess the algorithm's reaction to minor changes in certain aspects of the program that may seem somewhat arbitrary to the reader. Summary and scenario sections are presented for each of the two SA program types. From this point forward, SA1 will be used to reference the first form of the SA program where the full network is optimized, and SA2 will be used to reference the second form of the SA program where the best subset of *n* sampling sites are selected. The Matlab code for SA1 and SA2 is listed in Appendix F.

It was mentioned earlier in the study that weights could be assigned to the water quality clusters, watershed characteristics clusters (geology, morphology, and vegetation), or the collocation data. For the purpose of the initial analysis, the water quality clusters have been assigned a weight of two and the watershed characteristics clusters have been assigned a weight of one. The sampling sites that are collocated with other NPS projects have been assigned a value of 20 if they have one NPS site located within a 100 meter radius and 30 if they have two NPS sites located within the same radius. All other sites were assigned a zero collocation benefit. Values were arbitrarily chosen because they were similar in magnitude to the water quality cluster benefits and the watershed characteristics benefits.

<u>Summary of Scenario 1 – SA1 analysis for full network optimization</u>

- The objective function was maximized at a value of 15,280 with 124,500 iterations and a runtime of 47.6 minutes.

- A total of 67 sampling sites were retained and 12 sampling sites were discarded at the maximized objective function.

- Results of a sensitivity analysis on the benefit values of water quality, geology, morphology, vegetation, and collocation show that vegetation and collocation are more sensitive than the other variables when the benefit multiplier is increased to 2.0.

Scenario 1 – SA1 analysis for full network optimization

The first SA1 program run was initialized with a user-selection of 39 sampling sites in the initial monitoring network: 1, 4, 14, 24, 34, 45, 47, 50, 52, 71, 73, 114, 115, 127, 137, 138, 143, 148, 150, 174, 200, 210, 215, 221, 237, 253, 266, 268, 293, 311, 336, 337, 473, 475, 480, 482, 483, 488, and 489. The runtime for the algorithm was 47.6 minutes and the maximum objective function (maximized net benefit) achieved was 15,280. The algorithm produced 124,500 monitoring network permutations with 9,403 of those permutations being accepted by either the objective function rule or by the Boltzmann probability distribution, together known as the Metropolis algorithm. The retained and the discarded sampling sites are shown in Table 30 below.

Figure 12 shows a sampling site diagram with the sites that were discarded identified by the gray highlight. Figure 13 presents the objective function improvement of the SA1 algorithm by iterations and by temperature step decreases. The top graph in Figure 13 shows how the algorithm begins with a low objective function (poor solution) and gradually finds a better solution iteration by iteration. At a point, there is a noticeable "ceiling" value at which the objective function cannot be improved upon. At a later point, around 100,000 iterations, the algorithm flat lines because a better solution cannot be found. The lower graph shows vertical bars at each temperature step that represent the range from the lowest to the highest objective function evaluation. Notice that the bars generally become shorter and closer together as the temperature decreases toward the termination point. Again notice that the maximum objective function value is identified by a "ceiling" value. At a point near the termination point, the bar height is infinitesimally small indicating once again that the objective function cannot be

Table 30. SA1 results.

Sampling Sites Retained	Sampling Sites Discarded
4, 13, 14, 20, 23, 24, 30, 34, 47, 49, 50, 52, 66, 71, 73, 74, 114, 137, 142, 143, 144, 147, 148, 149, 150, 156, 173, 174, 186, 190, 191, 192, 193, 194, 209, 210, 213, 214, 215, 221, 233, 234, 237, 251, 252, 253, 266, 268, 291, 293, 310, 311, 472, 473, 474, 475, 479, 480, 481, 482, 483, 484, 485, 488, 489, 492, 493	1, 3, 43, 45, 46, 103, 104, 106, 107, 115, 127, 138, 184, 200, 336, 337

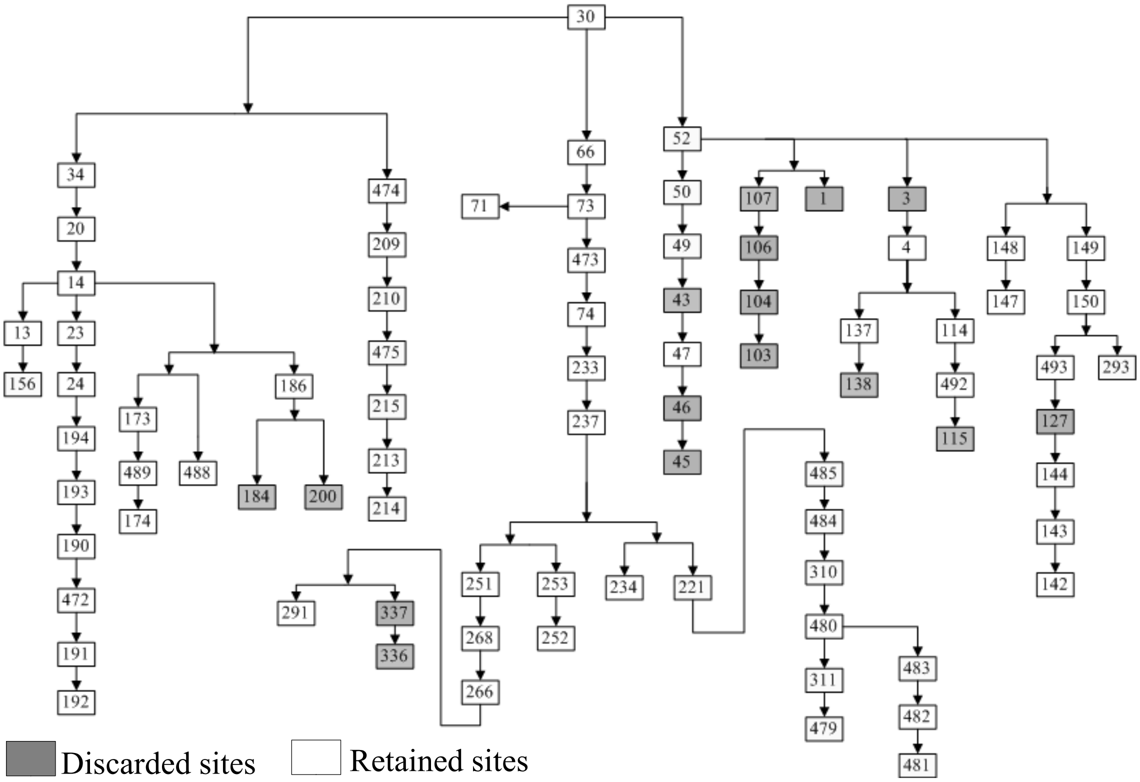

Figure 12. Sampling site diagram of the SA1 solution.

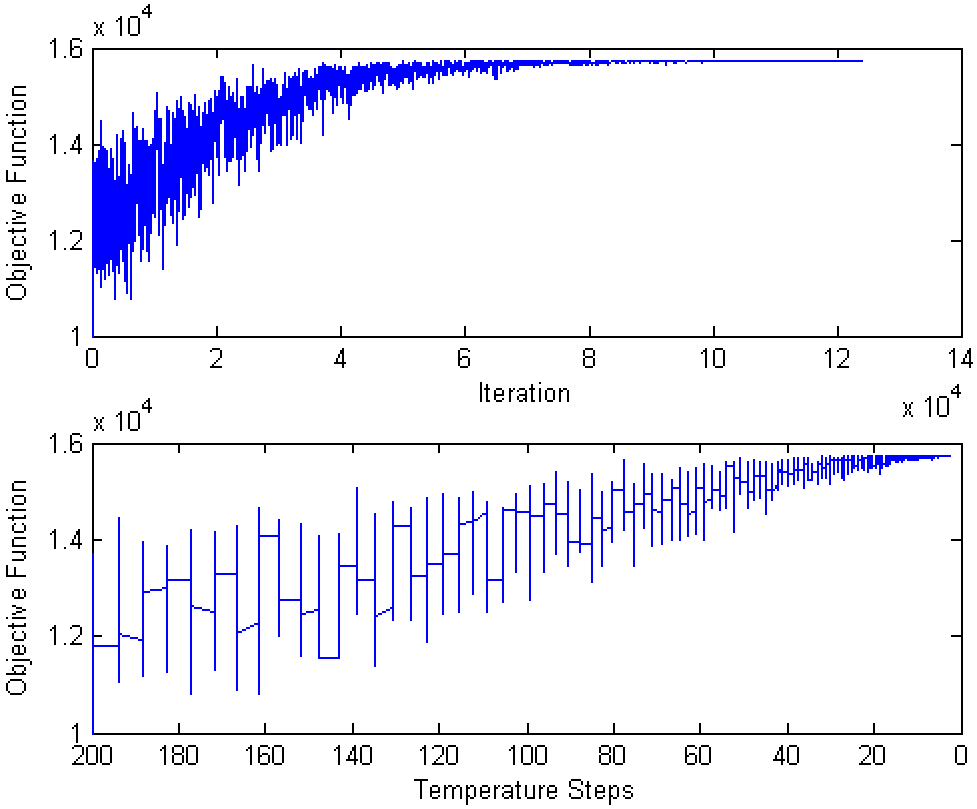

Figure 13. Objective function tracking through the SA1 solution.

improved upon. These graphs were created in Matlab as part of the SA algorithm and are standard output each time a SA program run is executed.

Because the SA1 algorithm requires that the user specify an initial selection of sampling sites in the network, a second analysis was performed by reducing the number of sites in the initial monitoring network. The previous paragraph describes the situation in which 22 sampling sites out of 83 sampling sites were initially selected for discontinuation. The sites were not chosen for any specific reason; they were arbitrarily marked to be included in the initial monitoring network. One additional program run was performed with only sampling sites 480 and 489 in the initial configuration. The outcome was the same as the previous case in which there were 22 sampling sites selected for discontinuation in the initial network.

A sensitivity analysis was performed on the SA1 algorithm by multiplying the water quality, geology, morphology, vegetation, and collocation benefit values by factors of 0.8, 1.0, 1.5, and 2.0. The analysis was performed on one set of benefit values at a time while holding the other benefits at their original values. The SA1 program sensitivity analysis results are shown in Tables 31 through 35. Each table presents the results for a set of sensitivity analyses corresponding to each data group (water quality, geology, morphology, vegetation, and collocated sites). The first column of each table shows the multipliers that were applied to the individual benefit values of the data group. The second column gives the maximized net benefit obtained in the SA1 program. The third column presents the changes in the retained sampling sites for each analysis. The second row of each table, with a multiplier of 1.0, gives the sensitivity analysis results of the data with their original values, i.e. all benefits are

Table 31. Sensitivity analysis results of water quality benefits.

Multiplier	Net Benefit	Retained sampling site changes from base
0.8	15915	No change
1.0	15752	BASE
1.5	15470	-43, +52, +156
2	15282	-43, +52, -107, +336, +337

Table 32. Sensitivity analysis results of geology benefits.

Multiplier	Net Benefit	Retained sampling site changes from base
0.8	17172	-43, -221, -310, -311, -479, -480, -481, -482, -483, -484, -485
1.0	15752	BASE
1.5	15713	+3, +45, +46, +127, +156
2.0	15840	+3, +45, +46, -107, +127, +156,

Table 33. Sensitivity analysis results of morphology benefits.

Multiplier	Net Benefit	Retained sampling site changes from base
0.8	15811	156
1.0	15752	BASE
1.5	16736	-43, -221, -310, -311, -479, -480, -481, -482, -483, -484, -485
2.0	15792	-43, +45, +46, +103, +104, +106

Table 34. Sensitivity analysis results of vegetation benefits.

Multiplier	Net Benefit	Retained sampling site changes from base
0.8	15765	-43
1.0	15752	BASE
1.5	15785	-52, -156
2.0	15826	-52, -156

Table 35. Sensitivity analysis results of collocation benefits.

Multiplier	Net Benefit	Retained sampling site changes from base
0.8	15496	+45, +46, +156
1.0	15752	BASE
1.5	16421	No change
2.0	17036	-209

multiplied by a factor of 1.0. This is referred to as the “base” in the third column. For example, row four of Table 31 for the water quality variables is based on a multiplier of 2.0. The multiplier was applied to all the water quality benefit values, and then the SA1 program was executed. The third column lists the sampling sites that were added (+) or removed (-) from the retained sampling sites list relative to the base row with multiplier of 1.0. It is apparent from Tables 31 through 35 that the SA program is reasonably stable considering the magnitude of the multipliers applied to the benefit values. It can be inferred that in most cases the program is more dependent on the results of the cluster analyses than it is on the magnitude of the benefit values. In other words, if the clusters for the data are reasonable the results of the SA should be very accurate. It can be noticed that geology and morphology are more sensitive to multiplier changes than water quality, vegetation, or collocation. Water quality, vegetation, and collocation exhibit only a small number of changes compared to the base. It is interesting to note at this time that none of the sensitive sites listed in Tables 31 through 35 will be included in the retained networks of *n* = 10 through *n* = 40 of SA2 presented later in this section. The sensitive sites should be termed as “pivotal” because their benefits are neither very high or very low.

A sensitivity analysis was also performed by changing the benefit multiplier that was used to apportion the dollar amount of benefit to each sampling site. Recall that a factor of 1.2 was multiplied by the costs of laboratory expenses, overhead, interpretation, etc. The costs remain fixed and equal for each site and do not include the variable cost of accessing each sampling site. Costs are multiplied by the factor 1.2, then summed up for the entire network, and apportioned out according to the benefit values for each sampling site. The factor of 1.2 is the focus of this part of the sensitivity analysis. Table 36 shows the results of using multipliers 1.0, 1.2, 1.5, and 2.0. This table uses the same format as the previous sensitivity analyses. The changes in this table are much more dramatic than those of the previous sensitivity analyses. The SA1 algorithm is very sensitive to the changes in the cost-based multiplier. The dramatic changes do, however, make logical sense. When the benefit dollar amount is decreased for a sampling site the access and lab costs will exceed the benefit, making that site less beneficial on a cost basis. Conversely, when the benefits are increased to a sampling site the benefit cannot be exceeded by the access and lab costs.

One final sensitivity analysis was performed by removing the collocation benefit from the analysis since the collocation benefit was an assumed value of the same magnitude as the cluster benefits. By removing the collocation benefit, dollars were reapportioned to other sites. It should be mentioned that only collocated site 107 was discarded in SA1. When the collocated benefits were removed site 107 was still the only collocated site discarded. This indicates that even without the collocation benefit the collocated sampling sites are very beneficial.

Summary of Scenario 2 – SA2 analysis for determination of *n* best sampling sites

- The SA2 algorithm was configured to select a user-specified number of the *n* best sampling sites to be retained in the final monitoring network.

- The SA2 algorithm was run with the *n* best sampling sites set equal to 10, 20, 30, 40,50, 60, and 70. The SA2 algorithm was executed once for each set of *n* sites.

- It was determined that a monitoring network with a globally optimum objective function value was between *n* sites equal to 40 and 70. The maximized objective function for these configurations were fairly equal meaning the redesigned network does have some latitude and is not limited to one specified group of sampling sites.

Table 36. Sensitivity analysis of the cost-based multiplier.

Multiplier	Net Benefit	Retained sampling site changes from base
1.0	7677	-50, -52, -150, -156, -186, -190, -191, -192, -209, -210, -213, -214, -215, -221, -310, -311, -472, -473, -474, -475, -479, -480, -481, -482, -483, -484, -484
1.2	7744	BASE
1.5	34742	+1, +3, +43, +45, +46, +103, +104, +106, +107, +115, +127, +138, +336, +337
2.0	69152	All sites retained

- Graphical comparison of the results from the SA1 network optimization and the SA2 optimization shows that both forms of the SA algorithm generally agree with one another, although the SA1 algorithm is better able to maximize the objective function. Figure 14 displays the objective function values versus the number of sites for both SA optimizations.

Scenario 2 – SA2 analysis for determination of *n* best sampling sites

The SA2 algorithm, where the user specifies then *n* best sampling sites to be retained in an *n* site sampling network, was performed with *n* set equal to 10, 20, 30, 40, 50, 60, and 70. As mentioned previously, this algorithm does not require the user to pre-select the sampling sites for the initial network configuration. The user only specifies the number of sampling sites desired in the final monitoring network and then the algorithm randomly selects the initial configuration. Table 37 shows the results of the SA2 algorithm for the *n* best sampling sites for the sampling network. At each increment of 10 sites it is very insightful to note when individual sampling sites join the sampling network. The row in Table 37 labeled "Metropolis Accepted" displays the number of permutations of sampling sites out of the number of iterations that were accepted by either the Boltzmann probability distribution rule or by the objective function rule. The results of the SA1 program run in Table 30, that produced the best monitoring network based on the maximized objective function value, can be compared with the SA2 program results for *n* equal to 60 and 70 sampling sites in Table 37. It is evident that the retained sampling sites in Table 30 are listed in either column for *n* equal to 60 and 70 sampling sites of Table 37. Figure 14 displays a graph of the *n* best sampling sites and their respective maximized objective function values. It is very interesting to note how the maximized objective function for the *n* best sites in the SA2

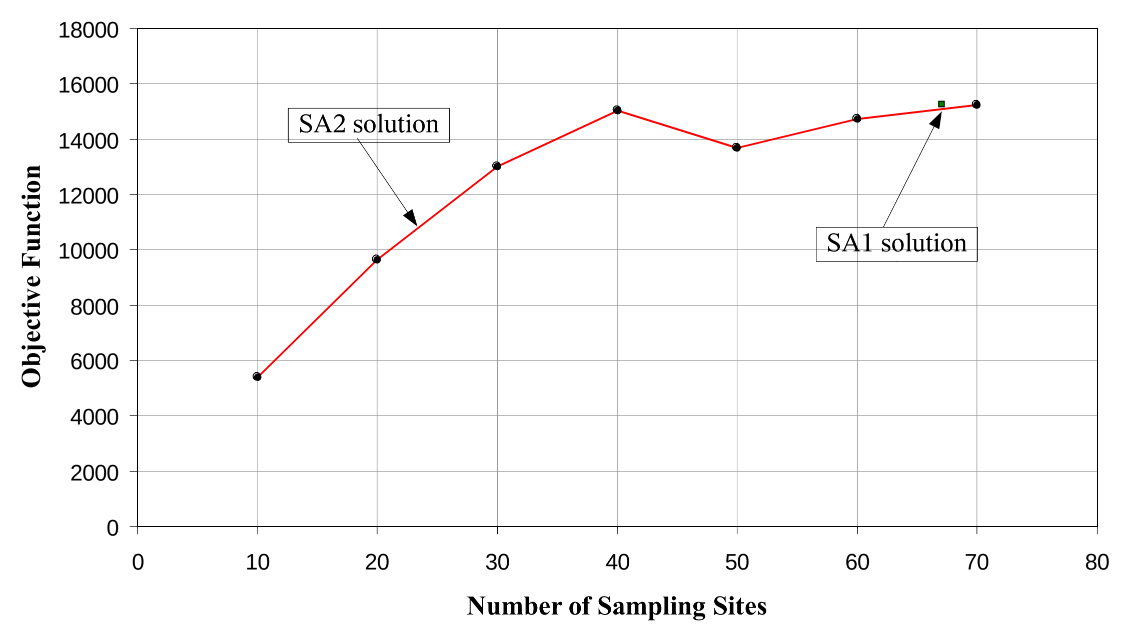

Figure 14. Comparison of SA1 and SA2 solutions.

Table 37. SA2 results.

***n* sites**	**10**	**20**	**30**	**40**	**50**	**60**	**70**
Iterations	124500	124500	124500	124500	124500	124500	124500
Runtime **(minutes)**	33	35.1	37.9	40.8	47.1	46.7	49.6
Metropolis **Accepted**	2658	3550	3578	3662	5046	5820	5468
Net Benefit	5404	9653	13009	15038	13684	14734	15250
Retained Sites	4 71 73 74 114 137 233 237 252 253	4 23 71 73 74 114 137 142 143 149 173 174 233 237 252 253 293 488 489 493	4 13 23 24 66 71 73 74 114 137 142 143 144 147 148 149 173 174 233 234 237 251 252 253 268 293 488 489 492 493	4 13 14 20 23 24 30 34 47 49 66 71 73 74 114 137 142 143 144 147 148 149 173 174 193 194 233 234 237 251 252 253 266 268 291 293 488 489 492 493	4, 13 23, 24 30, 34 47, 49 66, 71 73, 74 114, 137 142, 143 144, 147 148, 149 173, 174 191, 192 193, 194 221, 233 234, 237 251, 252 253, 266 268, 291 293, 310 311, 472 479, 480 481, 483 484, 485 488, 489 492, 493	4, 13 14, 20 23, 24 30, 34 47, 49 50, 66 71, 73 74, 114 137, 142 143, 144 147, 148 149, 173 174, 191 192, 193 194, 210 213, 214 215, 221 233, 234 237, 251 252, 253 266, 268 291, 293 310, 311 472, 474 475, 479 480, 481 482, 483 484, 485 488, 489 492, 493	4, 13 14, 20 23, 24 30, 34 45, 46 47, 49 50, 52 66, 71 73, 74 114, 137 142, 143 144, 147 148, 149 150, 156 173, 174 186, 190 191, 192 193, 194 210, 213 214, 215 221, 233 234, 237 251, 252 253, 266 268, 291 293, 310 311, 472 473, 474 475, 479 480, 481 482, 483 484, 485 488, 489 492, 493

algorithm forms a parabolic shape. The single small square shown on the graph is the result of the SA1 algorithm. The location of the SA1 point and the peak of the SA2 curve are very close to one another, thus providing a cross-validation of the SA process. Probably the most notable feature of Figure 14 is that the maximized objective function is fairly stable for networks from the 40 best sites to the 70 best sites.

Figure 15 shows a graphical representation of the results for the *n* best sites equal to 70. The top two graphs have the same interpretation as those described in Figure 13. The bottom graphs were created to show the number of monitoring network solutions accepted by the objective function rule and by the Boltzmann probability distribution rule. As expected, the number decreases for both cases as the temperature decreases. The bottom left graph shows that as the temperature decreases the acceptance of new monitoring networks by the objection function rule decreases because a better solution is more difficult to find. The bottom right graph shows the same trend because as the temperature decreases the Boltzmann probability of accepting a new monitoring network becomes less. Both of the bottom graphs have a steep linear trend with fluctuations from a temperature of 150 to around 50. The lines flatten somewhat with very small fluctuations at temperatures lower than 50.

3.7 Considering Water Quality Clusters and Elevation Classes

It was mentioned earlier that the redesigned monitoring network should also consider being able to sample at various elevations in order to measure an “elevation profile” of the water quality. Harwell (2001) found that pH exhibited significant trends at some elevation classes but not at others. His findings dictate that not only do all of

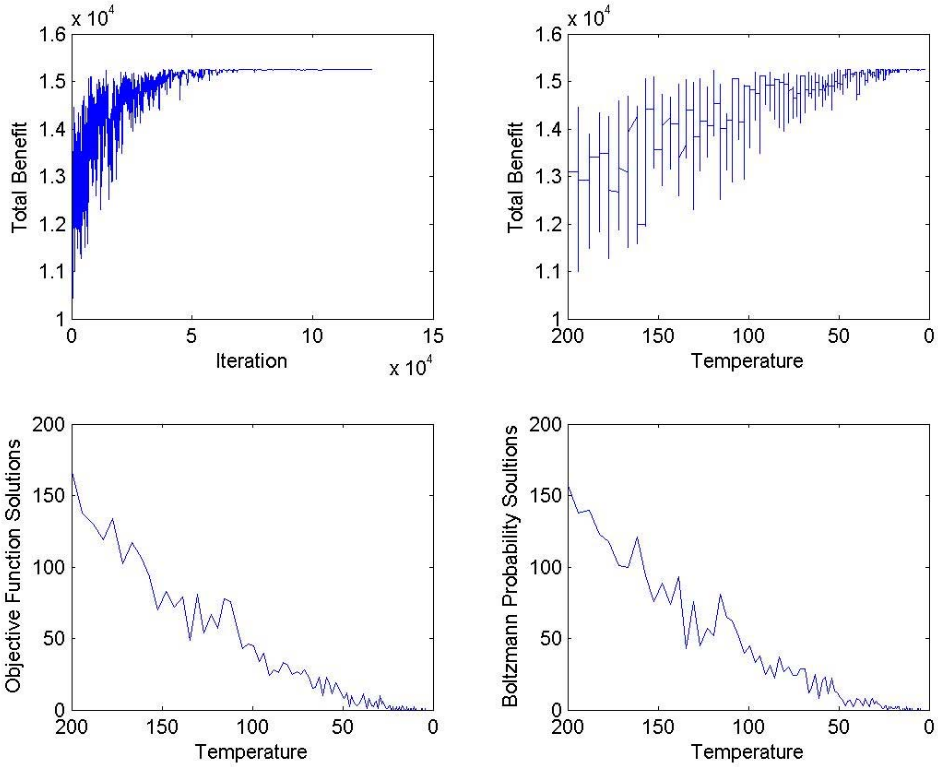

Figure 15. Objective function tracking through SA2 for *n* = 70

the original clusters of sampling sites need to be represented in the redesigned monitoring network, but that the elevation classes should also be considered in the redesign. Harwell (2001) tabulated columns one, two, and three for the elevation classes shown in Table 38. Columns four and five were added to compare the percentage of NPS area to the percentage of existing sampling sites in certain elevations ranges as shown. It can easily be seen that sampling above 3000 feet MSL is underrepresented in terms of the percentage of NPS area. If 83 sampling sites were redistributed according to the percentages of NPS area the new distribution would be:

- 36 sampling sites from the lowest elevation up to 3000 feet MSL
- 23 sampling sites between 3000 and 4000 feet MSL
- 18 sampling sites between 4000 and 5000 feet MSL
- 8 sampling sites above 5000 feet MSL

It should be mentioned that elevation classes could have been considered in the multivariate analysis of the water quality data, but it was desired to include elevation in the morphology analysis instead. Also, it was preferred to analyze the water quality data on the basis of the constituent measurements alone so that clustering would be based purely on water quality variability. Final recommendations of which sampling sites to retain and which ones to discontinue will be based upon the findings of this chapter and the elevation classes of Table 38. This will be presented in Chapter 5.

3.8 Extendibility of the Spatial Analysis

The methods used thus far in Chapter 3 are widely applicable to the assessment and redesign of other water quality monitoring networks and should therefore be

Table 38. Elevation classes.

Elevation class	Range of elevation (ft) MSL	Number of sampling sites	Percent of NPS area*	Percent of sampling sites
1	< 1000	0	43.3	65.0
2	1000 - 1500	7		
3	1500 - 2000	13		
4	2000 - 2500	16		
5	2500 - 3000	18		
6	3000 - 3500	13	27.4	20.5
7	3500 - 4000	4		
8	4000 - 4500	5	21.2	12.1
9	4500 - 5000	5		
10	5000 - 5500	1	8.1	2.4
11	> 5500	1		

*Approximate percentages based on planimetering contour map viewed as a template for similar studies. In cases where water quality data are available the methods can be directly applied if watershed information and site access data have been compiled. In the case of a new network where water quality data are not available the watershed characteristics and hypothetical sampling sites could be used. The hypothetical sites would be needed to assign access costs and to gather characteristics data about the upstream watershed. The results would provide a tool for assessing the costs and benefits of different network designs and choosing the optimum sampling sites based on watershed characteristics for the the initial network. After the network has been designed and data collected, the model could be used to evaluate the effectiveness of the monitoring network.

3.9 Summary of Findings

This section presents a summary of the results of the spatial analyses that will be the focus of the final discussion and recommendations in Chapter 5. The following list contains summary results of each sub-section in this chapter, some key points that should be taken into consideration in the redesigned network, and notable findings. The summary is as follows:

- The analysis of the water quality variables resulted in the formulation of nine clusters. Two additional clusters were formed from outliers. Cluster 10 has the membership of sites 156, 174, and 489 because of the high influence of limestone in their geology which probably causes the extremely high values of ANC. Cluster 11 has the membership of a single site (site 237) because of its extremely low ANC values probably caused by the Anakeesta geologic formation. The DA correctly classified 90 and 95 percent of the sites into their assigned clusters using the principal components and the original water quality data, respectively.

- The analysis of the geology variables resulted in the identification of 10 clusters. Cluster nine did contain sites 156, 174, and 489 listed above as being outliers in the water quality variables because of their high limestone content. The cluster (cluster three) containing site 237 had the highest percentage of Anakeesta formation. Cluster 1 contained sites with very high percentages (mean of the cluster = 97.6 percent) of Thunderhead sandstone. Cluster 2 contained sites with very high percentages (mean of the cluster = 80.8 percent) of Great Smoky group. The DA correctly classified 98.8 percent of the sites into their assigned clusters using the principal components and the original geology variables.
- The analysis of the morphology variables resulted in the formulation of five clusters. Cluster three was probably the most notable cluster with a basin area mean of 10,403 hectares. Stream elevation, stream order, and channel slope were also notable in terms of discriminating ability. The DA correctly classified 98 and 90 percent of the sites into their assigned clusters using the principal components and the original morphology data, respectively.
- The analysis of the vegetation variables identified nine clusters. It is interesting to note that sites 156, 174, and 489 were classified into a small cluster (cluster 8) in addition to sites 24 and 173. Cluster three had the highest percentage of pine. Cluster nine was the largest cluster with 24 members and was formed by high percentages of hardwoods.

- Collocated sampling sites were identified if other ongoing biological studies by the NPS were located within 100 meters of the sampling sites.
- An additional cluster analysis was performed by grouping the geology, morphology, and vegetation variables into one dataset. The analysis allowed the identification of sampling sites in overlapping regions of clusters. Sites in the overlapping regions of clusters have unusual combinations of characteristics compared to the majority of the sampling sites. These sites are: 49, 50, 52, 71, 74, 127, 147, 149, 150, 191, 192, 210, 221, and 474.
- Sampling sites in overlapping regions of the water quality clusters were also identified. These sites are: 43, 71, 115, 221, 251, 268, 293, 475, 484, and 492.
- Table 30 presents the results of the first simulated annealing algorithm (SA1) that was used to select an optimum network based on maximization of the net benefits. This resulted in 67 sites being retained and 16 sites being discontinued.
- A sensitivity analysis was performed in SA1 on the collocation benefits. The analysis showed that even without the collocation benefit the collocated sites are very beneficial. Only site 472 moved to the discarded site list when the collocation benefit was removed.
- SA2 produced the results in Table 37 that identifies the *n* best number of sites specified by the user. Table 37 is most useful for determining the order by which sampling sites were added to the network based on their benefit contribution. The table also shows that the maximized objective function is fairly equal between the 40 best sites and the 70 best sites.

- Table 38 shows the elevation classes, the number of sites in each elevation classes, and the proportion of sampling sites in certain elevation ranges. Elevations above 3000 feet MSL are underrepresented according to the percentage of park area from 3000 feet MSL and above. Each elevation class should be represented in the redesigned network and additional sampling sites may need to be added so that the percentage of sampling sites is proportional to the percentage of park area in each elevation class.
- Each water quality cluster and the clusters of the watershed characteristics should be represented in the redesigned network.
- Redesign of the network using primary, secondary, and tertiary sites based on their uniqueness in overlapping areas of clusters and the order in which the sampling sites were retained (Table 37) should be strongly considered.
- Subjective decisions by the NPS about additional sites to be retained or discontinued should consider the order of the *n* best sampling sites in Table 37.
- NPS should consider the discontinuance of high-elevation spring sampling. It has already been mentioned that this study did not address the seven high elevation springs that are included in the synoptic sampling network. An initial recommendation would be that the NPS review the information from these sites and make a determination on the need for these data.

Chapter 4 ASSESSMENT OF SAMPLING FREQUENCY

4.1 Noland Divide Data

Weekly stream sampling data from the southwestern stream at the Noland Divide watershed sampling site near Clingman's Dome were used to determine the sampling frequency needed to detect trends in the water quality data. Laboratory analysis of samples taken at this location includes tests for the following water quality variables: pH, conductivity, ANC, chloride, nitrate, sulfate, sodium, and potassium. The study period for these variables extends over a period from July 19, 1991 to January 1, 2002, or 550 weeks. These data are essential to this part of the study because of the high frequency at which the samples were taken. High-frequency sampling at this site will allow a gradual "thinning" of the data to test lower sampling frequencies.

4.2 Data Preprocessing

The water quality data from the Noland Divide site used in this portion of the study do require pre-processing before the frequency analysis begins. There were a total of 17 water quality measurements missing from weeks when samples had been collected but only partial results were reported. There were also a total of 67 weeks where no data were reported. (Many of these "missing observations" are actually the result of a change from a one-week sampling interval to a two-week sampling interval in the later years.) The missing data for these observations are estimated using a cubic spline interpolation and are needed because use of the autocorrelation function and time series decomposition require data that are evenly spaced in time. The autocorrelation functions indicate serial correlation in the data. The autoregressive pattern of the data provides a basis for using cubic spline interpolation because most adjacent data measurements are fairly close in magnitude and are not highly erratic. Descriptive statistics are used to determine if the spline interpolation had an adverse effect on the data.

It is assumed for this part of the study that the data are normally distributed or near normally distributed. By the strictest definition of normality this has certainly not been the case thus far for the water quality data. However, most of the methods used here do not make strong normality assumptions. The results of normality tests will be presented in the results section of this chapter.

4.3 Methods of Analyses

Four methods are utilized to analyze different sampling frequencies using the original time series from the Nolan Divide, southwestern stream data. The first method will determine a maximum number of samples to be collected each year so that independency is maintained in the data. The maximum number will be used as an approximate upper limit of a proposed sampling frequency. The second and third methods will compare a number of different sampling frequencies using a resampling window scheme and proven methods of nonparametric trend determination. The fourth method is a graphical technique that will be used to compare the boxplot of different sampling frequencies of the data with the boxplot of the original data. The four methods and their abbreviations are enumerated below.

- “Effective” sample method (ES)

- Sen’s slope estimation for trend (SE)
- Mann-Kendall test for trend (MK)
- Boxplot analysis (BP)

Figure 16 shows a general flowchart of these analyses. The SE and MK methods will require the construction of new time series from the original time series to represent the sampling frequencies being considered. The “moving window” approach will be used to generate the new time series by specifying a sampling interval by which data are selected from the original dataset at a constant interval or frequency. The “moving window” also includes an offset feature so that the starting observation of the new time series can be altered. Subsequent observations in the new times series will be offset by the same amount. The approach is very similar to the one recommended by Sherwani and Moreau (1975). The boxplot comparison method will compare the median, interquartile range, and outer fences of other sampling frequencies with those of the original time series.

4.3.1 “Effective” Sample Method

The “effective” sample method (ES) is used to aid in determining the maximum number of statistically independent samples that can be collected annually and then to estimate the length of record needed to reliably detect trends at specified confidence levels and test powers. The results are then compared to the results of the other methods used in this chapter. Closely spaced sampling data are usually autocorrelated (Lettenmaier 1976) and this is certainly the case with the Noland Divide data. Autocorrelated (or serially dependent) means that the data are temporally dependent.

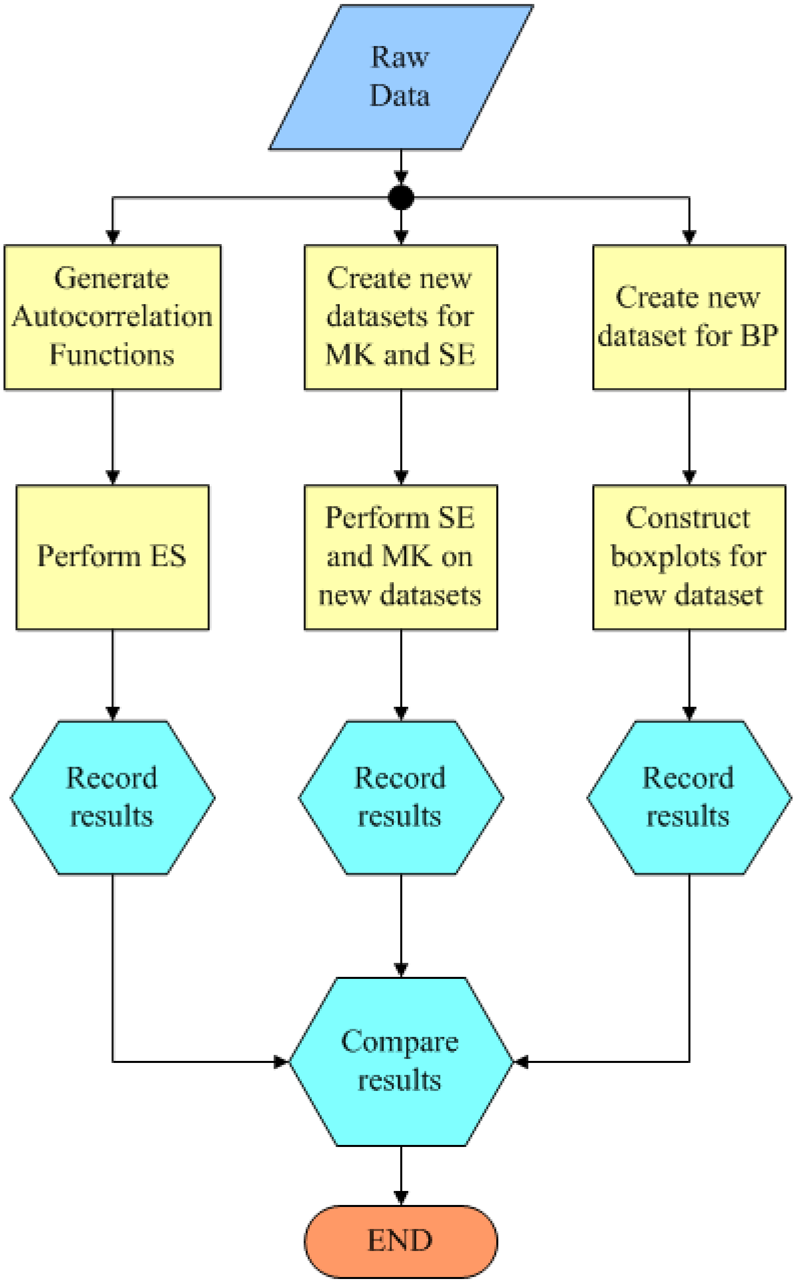
Raw Data
Generate Autocorrelation Functions
Create new datasets for MK and SE
Create new dataset for BP
Perform ES
Perform SE and MK on new datasets
Construct boxplots for new dataset
Record results
Record results
Record results
Compare results
END

Figure 16. Flowchart for analysis of sampling frequency.

Independent data contain more information about the population than dependent data and are needed to insure that the occurrence or nonoccurrence of a single event does not affect the probability of another event. Therefore, it is desired to sample in such a way that the maximum number of independent samples are collected so that accurate long-term trend detection is achievable in as short amount of time as possible. Using dependent observations for calculation of confidence intervals will underestimate confidence interval width. Since Noland Divide data are autocorrelated the number of samples must be corrected before using in statistical calculations such as confidence intervals to prevent underestimation of confidence interval width. The “effective” sample method determines a sampling frequency beyond which there is little gained by taking additional samples because of dependence among the data.

In the context of this discussion, the word “effective” is defined as the maximum number of samples that can be collected on an annual basis and maintain independence among the data. Lachance, Bobée, et al. (1989) used ES to determine the “effective” sampling frequency of water quality measurements in Lake Laflamme in Québec. They determined that the number of independent samples that could be collected annually was 10.6 and 8.4 for pH and sulfate, respectively.

The basis for the “effective” sample method (ES) was developed by Bayley and Hammersley (1946) and was later expanded upon by Lettenmaier (1976) and Sanders and Adrian (1978). It is first necessary to discuss the autocorrelation function (ACF) to gain some understanding of the basis of this method. The equation for the ACF is given as

$$ACF_k = \frac{\sum_{t=1+k}^{n} \left(Y_t - \overline{Y}\right)\left(Y_{t-k} - \overline{Y}\right)}{\sum_{t=1}^{n} \left(Y - \overline{Y}\right)^2} \qquad (4.1)$$

where $ACF_k =$ the autocorrelation coefficient for the *k*th time lag
$n =$ the number of observations
$Y_t =$ the observation at time t

The ACF_k provides a measure of correlation between the variable Y at time t, and the variable Y at time t-k. The ACF_k correlation measurements are considered statistically significant if $|ACF_k| > 2/\sqrt{n}$ meaning that the variable Y at time t, and the variable Y at time t-k are dependent to some degree, thus, samples collected for that specific time interval would indicate a degree of dependency between those samples.

Autocorrelation between measurements at different time lags is the basis for calculating the required sampling frequency to obtain roughly independent samples. The equation developed by Bayley and Hammersley (1946) for determining the number of "effective" samples from a dependent time series is given by

$$\frac{1}{n^*} = \frac{1}{n} + \frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)ACF_k \qquad (4.2)$$

where $k =$ lag number (or time steps) between samples
$n =$ number of samples taken per year based on a proposed sampling frequency
$n^* =$ "effective" number of samples taken per year based on a proposed sampling frequency
ACF$_k$ = autocorrelation coefficient for lag k

Lettenmaier (1976) used Equation 4.2 to correct for dependent samples. The number of "effective" samples for a given sampling frequency is calculated using the above equation by inserting the ACF coefficients (ACF$_k$) into the equation for the sampling frequency being used. Note that in Equation 4.2 that if ACF$_k$ = 0, then $n^* = n$ and if ACF$_k$ = 1, then $n^* = 1$ as expected. Since the Noland Divide data are on a weekly sampling interval the annual sampling frequency would be 52. An "effective" sample size per year can then be determined for each of the water quality variables. The autocorrelation coefficients for use in Equation 4.2 must be generated from a stationary time series, which is accomplished by removing trend and seasonality (LaChance, Bobée, et al. 1989). The ES method, which was used by Lettenmaier (1976) and LaChance, Bobée, et al. (1989), does make the general assumption that the data are normally distributed. However, both of these authors did accept some non-normality in their data. Using Equation 4.2 and the Noland Divide data, different values of n^* are generated from the autocorrelation function for weekly, monthly, bimonthly, and quarterly sampling intervals.

Lettenmaier (1976) and Lachance, Bobée, et al. (1989) furthered the application of the "effective" sample method using the standard error and trend magnitude of a time series to determine the length of record needed to reliably determine trend magnitudes at a specified statistical significance. Naturally, the linear trend of a time series with a low magnitude of change and a high standard deviation would be much more difficult to accurately determine than one with a high magnitude of change and a low standard deviation. Lettenmaier (1976) presents Equation 4.3, developed by Brieman (1973) for the power function of a classical *t*-test as

$$1-\beta=F_g(N_t-Z_{(1-\alpha/2)}) \qquad (4.3)$$

where $1\text{-}\beta$ = probability of not accepting the presence of a trend when a trend is, in fact, not present
F_g = cumulative distribution function of a standard normal probability distribution
N_t = measure of trend magnitude
$Z_{(1\text{-}\alpha/2)}$ = quantile of the standard normal distribution at probability *(1-α/2)*

and Equation 4.4 for the measure of the trend magnitude,

$$N_t = \frac{|Tr|\sqrt{n}}{\sigma_\epsilon \sqrt{12}} \tag{4.4}$$

where Tr = absolute value of the change in beginning and ending predicted values along a regression line for a period of study
σ_ε = standard deviation of the residuals (standard error)
n = total number of independent samples needed for a period of study

Combining Equations 4.3 and 4.4 and setting $(1\text{-}\beta) = 0.90$ yields Equation 4.5 where $N^* = n$ from Equation 4.4.

$$N^* = \frac{12(1.282 + Z_{1-\alpha/2})^2}{\left(\dfrac{Tr}{\sigma_\varepsilon}\right)^2} \tag{4.5}$$

Equation 4.5, used by Berryman, Bobée, et al. (1988) and Lachance, Bobée, et al. (1989), is a relationship between trend detection level Tr/σ_ε, and the total number of independent samples needed N^*, with the power of the test fixed at 90 percent. Equation 4.5 can be adjusted for other powers. The term Tr/σ_ε in Equation 4.5 is a ratio defining the level of trend detection desired by the user. The standard error σ_ε, can be viewed as an “average residual error” from a least-squares regression line for a time series. Therefore, it is desired to reliably detect a trend that is at least as large as the standard error, which would equate to a Tr/σ_ε value of 1.0. Values less than 1.0 would be more desirable but this comes at the cost of requiring a longer record for the same power of the test. In many cases the designer of a new monitoring network must presuppose a desired level of trend detection. In the redesign of a monitoring network, data are available so that the the trend magnitude and the standard error can be be determined. *N** (total samples) from Equation 4.5 can be divided by *n** (samples per year) from Equation 4.2 to determine the number of sampling years needed to reliably detect the existing trend with 95 percent confidence and 90 percent power. The values used in this study for Tr/σ_ε are presented in Section 4.4.2 and correspond to the trend levels that are present in the Noland Divide data. Environment Canada required that a trend level of detection be set at 1.0 for acid lakes in Québec (Berryman, Bobée, et al. 1988) meaning that it was desired to detect a trend equal to or greater than the standard error of the time series.

Power of the test is not to be confused with the confidence level as these two are inversely related. Ideally, the power of any statistical test should be known before accepting a null hypothesis (Mendenhall and Beaver 1991). Power of the test ($1-\beta$) is defined as the probability of not accepting the null hypothesis when the null hypothesis is incorrect. The power of the statistical test for trend is affected by the number of years of sampling and the magnitude of the trend (Lettenmaier 1978; Somerville and Evans 1995; Urquhart, Paulsen, et al. 1998). Longer records and greater trend magnitudes yield higher power. Increasing the sampling frequency will not necessarily yield a higher power (Lettenmaier 1978). Although, Somerville and Evans (1995) did find that the sampling frequency affected power when the length of record was between five and 15 years in length.

4.3.2 Sen's Slope Estimation

Sen's slope estimation (SE) is a nonparametric, unbiased estimator of the slope of a trend line through a set of data (Sen 1968). It allows the calculation of a two-sided confidence interval for the trend slope, which will provide a means of testing the trend slopes of different sampling intervals against the original time series. Hirsch, Alexander, et al. (1991) showed that Sen's method was more accurate at detecting monotonic trends than regression when the data were slightly non-normal. In the SE method, the trend of the time series is determined to be the median slope of all forward pairwise combinations of observations. Therefore, for a dataset of *m* observations, there are $n = m(m\text{-}1)/2$ slopes calculated. The general equation used to calculate each pairwise slope is

$$S_{ij} = \frac{x_i - x_j}{t_i - t_j} \tag{4.6}$$

where S_{ij} = Sen slope for one pair of observations at t_i and t_j for all $t_i > t_j$
x_i = water quality measurement at t_i
x_j = water quality measurement at t_j
t_i = time when water quality measurement x_i is taken
t_j = time when water quality measurement x_j is taken

An algorithm has been prepared in Matlab format to calculate Sen's slope. For each alternate sampling frequency a new time series will be generated from the original time series using the resampling window approach mentioned earlier. A confidence interval for Sen's slope will then be calculated for each new time series, which, of course, represents an alternate sampling frequency. The confidence interval of each of the alternate sampling frequencies, mentioned prior, will then be compared with the confidence interval of the original time series. It can be deduced that if the confidence interval of an alternate sampling frequency does not identify a trend slope similar to that of the confidence interval of the original time series, then that alternate sampling frequency is not an accurate estimator of the original trend. The converse is true if the slope of the alternate sampling frequency is identified similarly to the slope of the original time series.

4.3.3 Mann-Kendall Test for Trend

The Mann-Kendall (MK) test for trend is a nonparametric test used to determine if there is a negative trend, positive trend, or no trend in the data. The MK test does not specify the magnitude of the trend. The MK test was developed through a combination of efforts by Mann (1945) and Kendall (1975). The trend test that Mann developed is a special case of Kendall's test. Mann did not consider ties in the data while Kendall did take ties into consideration. Mann and Kendall also showed that by using a correction for continuity (MK test statistic, S_{MK} ± 1 depending on the sign of S_{MK} as shown in Equations 4.8 and 4.10 below) the test will approximate the normal distribution for sample sizes greater than or equal to 10. Later literature agrees that when the sample size is greater than 10 the MK test performs quite well using an approximation to the normal distribution (Hirsch, Slack, et al. 1982; EPA 1998). The MK test statistic, S_{MK}, is the sum of the number of negative and positive slopes given by

$$S_{MK} = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} sign(x_k - x_j) \tag{4.7}$$

$\mathrm{S_{MK}}$ is then used to calculate a *Z*-score using the continuity correction criteria

$$Z = \frac{S_{MK} - 1}{\sigma_S}, \textit{ if } S_{MK} > 0\,(\textit{upward trend}) \tag{4.8}$$

$$Z = 0, \textit{ if } S_{MK} = 0\,(\textit{no trend}) \tag{4.9}$$

$$Z = \frac{S_{MK} + 1}{\sigma_S}, \textit{ if } S_{MK} > 0\,(\textit{downward trend}) \tag{4.10}$$

The null hypothesis that no trend is present is tested against the alternate hypothesis that a trend does exist and is upward or downward. The standard normal *Z*-variate used for this test is ±1.96 corresponding to the two-tailed 95 percent confidence level. If the *Z*-score from Equation 4.8 or Equation 4.10 is greater than 1.96 or less than -1.96 the alternative hypothesis that a trend is present is accepted. Matlab code has been written to perform this operation. The standard deviation σ_s is calculated by

$$\sigma_s = \frac{n(n-1)(2n+5)}{18} \tag{4.11}$$

or by

$$\sigma_s = \frac{1}{18} n(n-1)(2n+5) - \Sigma_{p=1}^{g} w_p (w_p - 1)(2w_p + 5) \tag{4.12}$$

if ties occur, where *n* is the number of observations; *g* is the number of tied groups; and w_p is the number of observations in the *p*th group.

The MK test for trend is very similar to the SE method in that the positive and negative slopes of all forward pairwise combinations are enumerated. Some problems may be encountered if serial dependence or seasonality exists in the data. Serial correlation can cause inaccurate *p* values (Helsel and Hirsch 1992). Plotting the autocorrelation function for the time series in the “effective” sample method will permit easy detection of serial correlation in the data. If serial dependence is a problem, then a modification to the MK test can be made according to Hirsch and Slack (1984). If seasonality is a major concern then the seasonal Kendall test may be substituted (Hirsch, Slack, et al. 1982; Hirsch and Slack 1984; Helsel and Hirsch 1992).

The results of the MK test will be compared with the results of the SE and the BP analyses. The MK test will compare the trend detection of alternative sampling frequencies to that of the original time series. The comparison will be accomplished by extracting a subset of the original data in accordance with the proposed sampling frequency. Extraction of the data subset will be performed using the moving window approach that was mentioned earlier in the chapter. The MK test will then be performed on the subset of data for the proposed sampling frequency and compared to the MK test results of the original data. Departure of the trend detection of the proposed sampling frequency data subset from that of the original data would indicate that the proposed sampling frequency may not be an accurate estimator of the trend compared to the trend that is identified in the original data.

4.3.4 Boxplot Analysis

Boxplots (BP) will be used to visually compare the distribution of alternative time series with the original time series. Tukey (1977) introduced the boxplot as a univariate graphical analysis tool to visually display the interquartile range (IQR), median, and outliers of a set of data. This particular application of analyzing different sampling frequencies was introduced by Mueller (1989). By comparing the boxplots of data from different sampling intervals to the original time series, inference can be made about the ability of a particular sampling interval to detect the same trend pattern that exists in the parent time series.

4.4 Results of the Analyses

The results of the "effective" sample method, Sen's slope method, Mann-Kendall trend test, and the boxplot analysis are presented below. Missing data estimation, deseasonalization, and trend removal techniques were applied to these data, as described above. NCSS, SAS and Excel were used to perform the "effective" sample method. Matlab was used to perform Sen's slope method and Mann-Kendall test for trend. It should be remembered that the water quality data studied here are for one sampling site in the GRSM (Noland Divide, southwestern stream) which may not reflect the hydrologic characteristics of all 83 sampling sites. Again, the data from this site are studied because of their high sampling frequency.

4.4.1 Pre-processing and Descriptive Statistics

As mentioned earlier in this section, missing data were replaced using a cubic spline interpolation. The procedure was performed using the spline toolbox in Matlab. A value estimated by the cubic spline interpolation method is determined by the pattern of existing data before and after the missing value being estimated. Therefore, the original values are not recalculated, but remain unchanged. The results of the cubic spline imputation were validated using three methods. First, the normality of the entire data was checked before and after the spline interpolation was performed to ensure that the normality status of data had not changed. Second, descriptive statistics of mean, variance, and IQR of the water quality data before and after the spline interpolation were compared. Third, the autocorrelation functions for approximately the first five years of the original data were compared to the complete dataset after the spline imputation was performed. The first five years of the original data were used because the amount of missing data was minimal compared to the latter years.

The results of the descriptive statistics before and after spline interpolation are shown in Tables 39 through 42. Table 39 and Table 40 compare the mean, variance, and IQR for all the years of data. Tables 41 and 42 compare the same for the last three years of the data because most of the data imputation was performed within these years. In general, there is very little change in mean, variance, or IQR of the parameters. Scatterplots of the water quality variables after the spline interpolations were applied are shown in Figures 17 through 24. A least-squares line has been plotted through each of the scatterplots to show the general trend of the data. However, the least-squares lines may not be statistically significant. Outliers are most noticeable for conductivity, chloride, and potassium, and may be the cause of non-normality in these data. The normality results are not shown because the water quality variables both before and after the spline interpolation reported non-normality using the Shapiro-Wilk test and the D́Agostino Omnibus test. This may present a problem for the “effective” sample method because it does generally require that the data be normally distributed. However, examination of the histogram density trace overlaid with the normal distribution curve shows that a normality assumption might not be too unreasonable for pH, ANC, conductivity, and sodium. The normality assumption is more unreasonable

Table 39. Descriptive statistics before spline interpolation for all years.

Water Quality Variable	Mean	Variance	IQR
pH	5.83	0.04	0.27
Conductivity	13.20	6.96	1.92
ANC	11.21	39.46	7.53
Chloride	16.28	122.14	7.59
Nitrate	42.40	39.92	6.79
Sulfate	29.41	29.30	6.82
Sodium	25.65	17.91	4.24
Potassium	8.73	29.85	2.77

Table 40. Descriptive statistics after spline interpolation for all years.

Water Quality Variable	Mean	Variance	IQR
pH	5.82	0.04	0.27
Conductivity	13.22	6.95	1.99
ANC	10.84	39.65	7.45
Chloride	16.60	122.35	7.75
Nitrate	42.23	40.01	6.76
Sulfate	29.38	28.94	6.71
Sodium	25.57	18.49	4.27
Potassium	8.73	29.21	2.73

Table 41. Descriptive statistics before spline interpolation for last three years.

Water Quality Variable	Mean	Variance	IQR
pH	5.79	0.03	0.21
Conductivity	12.38	2.86	2.29
ANC	8.14	19.92	6.33
Chloride	19.2	50.92	10.56
Nitrate	40.63	21.59	5.41
Sulfate	28.75	21.05	6.46
Sodium	25.69	13.79	4.25
Potassium	9.95	31.88	4.36

Table 42. Descriptive statistics after spline interpolation for last three years.

Water Quality Variable	Mean	Variance	IQR
pH	5.79	0.03	0.21
Conductivity	12.4	2.57	2.19
ANC	8.18	19.02	6.5
Chloride	19.04	51.93	10.59
Nitrate	40.48	22.13	5.56
Sulfate	28.56	19.69	6.37
Sodium	25.71	15.66	4.23
Potassium	9.75	33.39	3.82

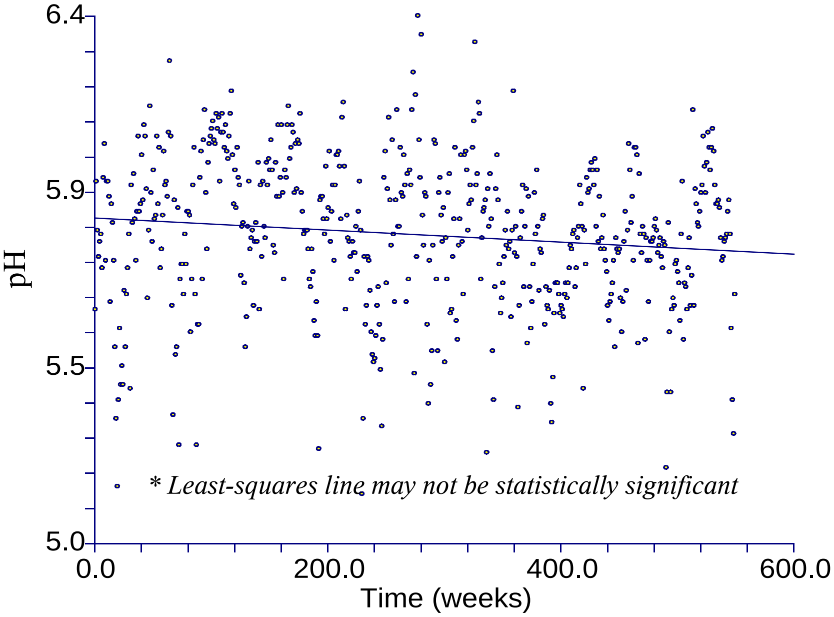

Figure 17. Time series scatterplot of pH.

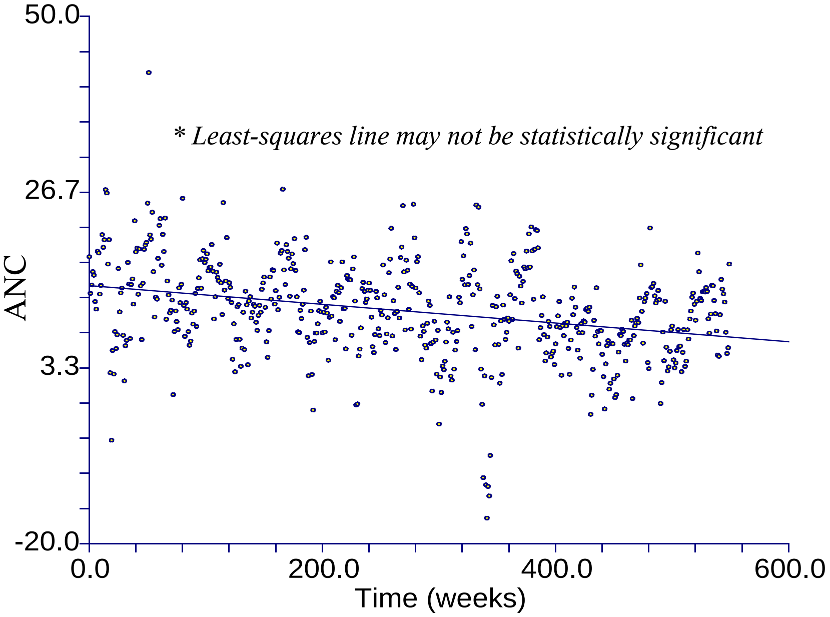

Figure 18. Time series scatterplot of ANC.

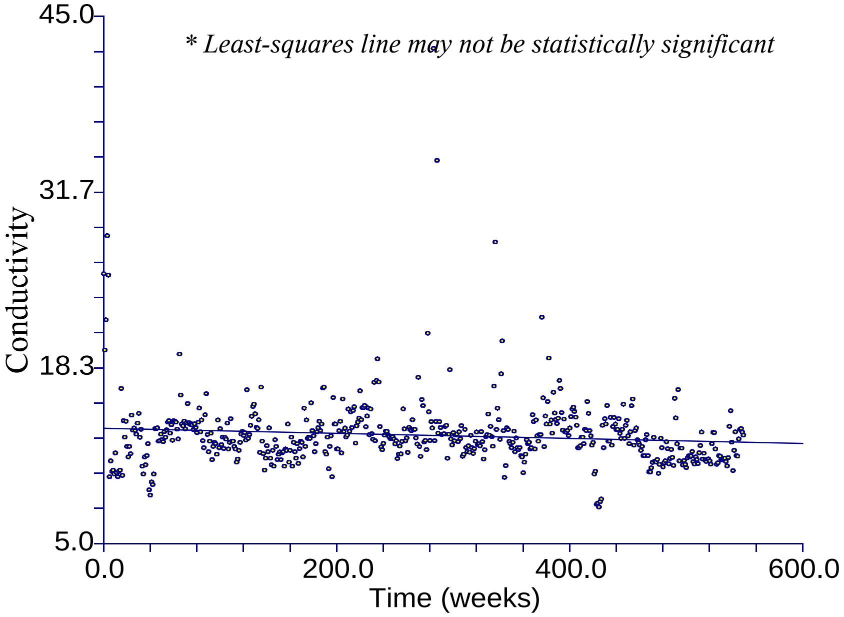

Figure 19. Time series scatterplot of conductivity.

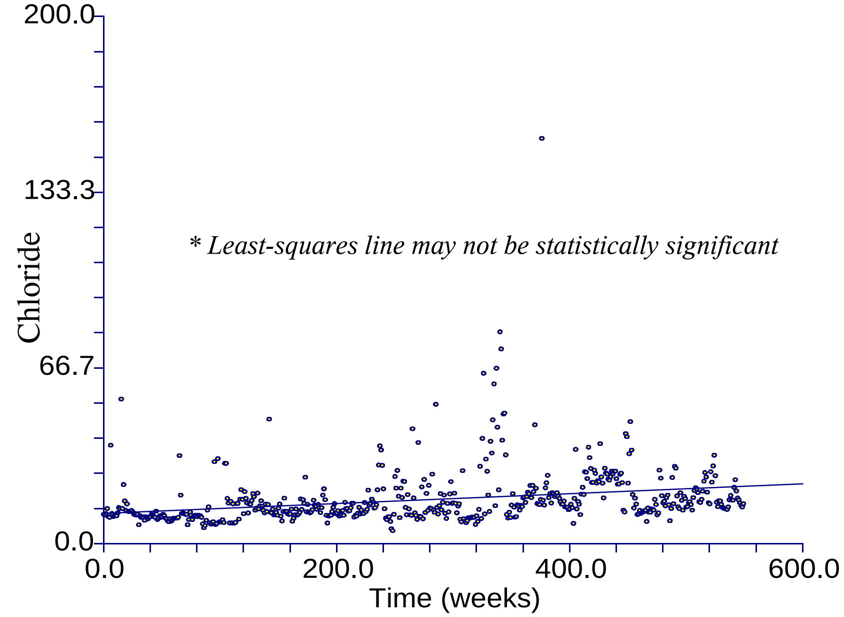

Figure 20. Time series scatterplot of chloride.

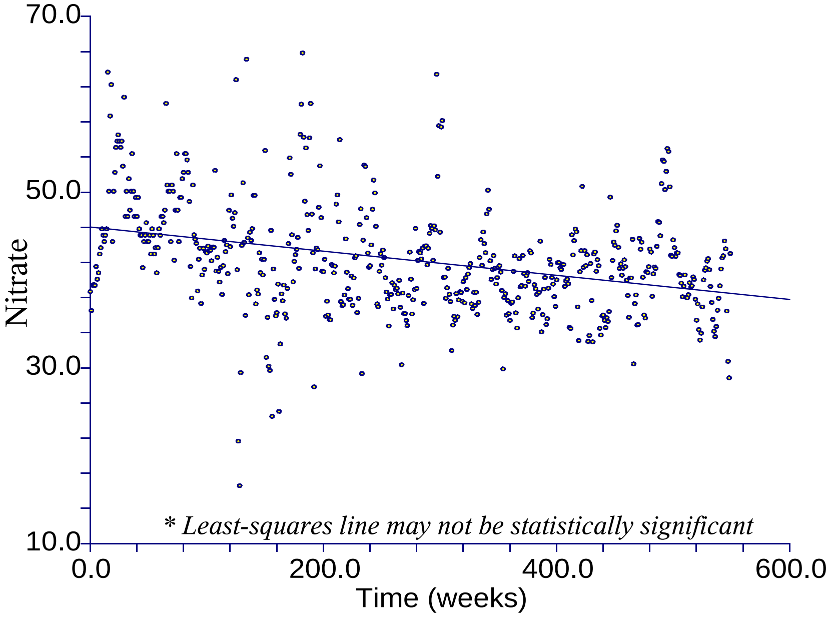

Figure 21. Time series scatterplot of nitrate.

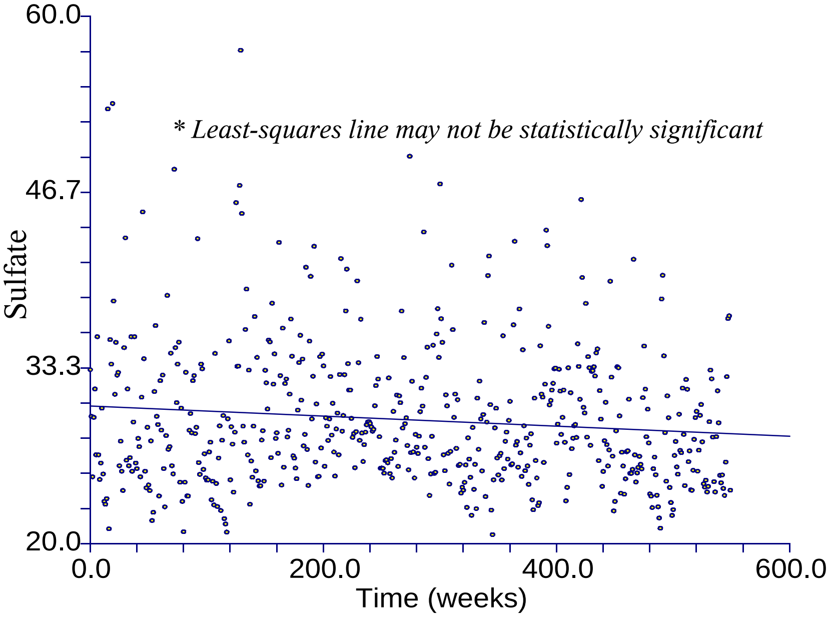

Figure 22. Time series scatterplot of sulfate.

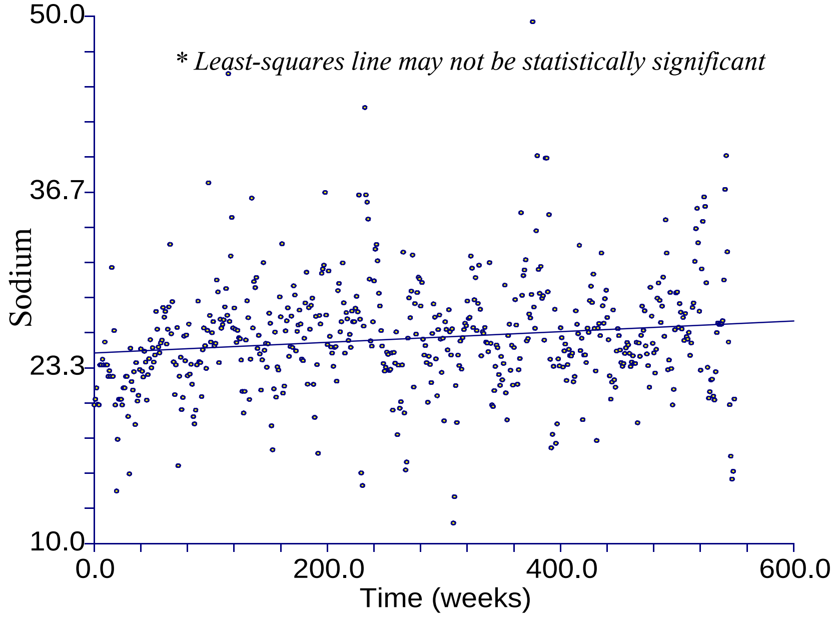

Figure 23. Time series scatterplot of sodium.

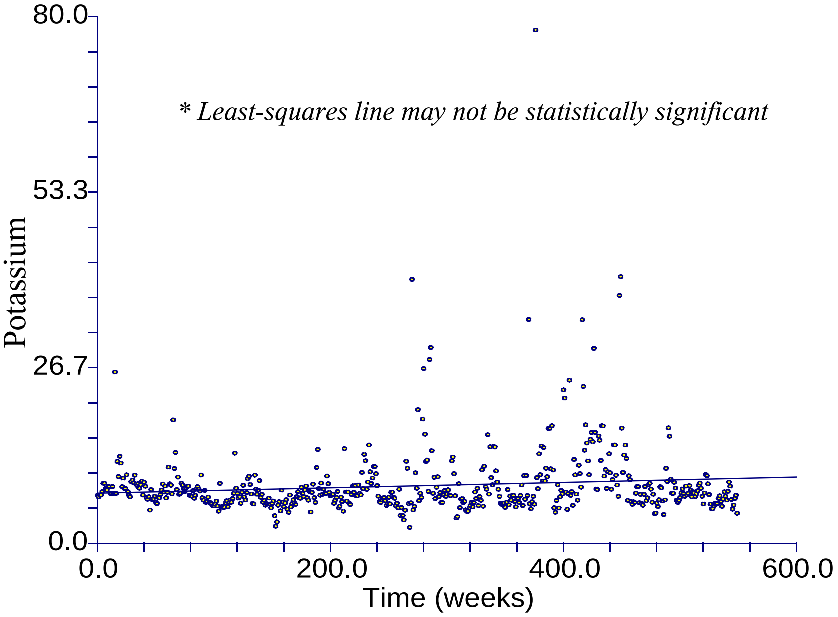

Figure 24. Time series scatterplot of potassium.

for chloride, sulfate, and potassium. Lachance, Bobée, et al. (1989) applied the normal density trace to histograms of alkalinity and calcium + magnesium data from acid lakes in Québec to justify using data that did not follow strict definitions of normality for the "effective" sample method. Therefore, this method is performed on the original data here and then compared to the results of Sen's slope estimation, Mann-Kendall test for trend, and the boxplot analysis. The latter methods do not require that the data be normally distributed.

Autocorrelation functions were plotted before and after missing data estimation. The changes were very minor and almost unnoticeable for many of the water quality variables. In fact, many of the water quality variables required a very close side-by-side inspection to identify changes.

4.4.2 Results of the "Effective" Sample Method

The "effective" sample method first required the calculation of the autocorrelation functions of the detrended, deseasonalized time series data for each of the water quality variables. The calculation was accomplished using time series decomposition, linear regression and the autocorrelation function in NCSS. Decomposition allows the identification and separation of a time series' seasonal component. After the seasonal component is removed linear regression can be used to remove the trend component of the time series. Figure 25 shows how pH at time t_0 is correlated with pH at time t_k before detrending and deseasonalization. Notice the seasonal nature of the ACF as the correlations begin with a positive correlation, then gradually fall to a negative correlation at mid-year, and then rise back to a positive

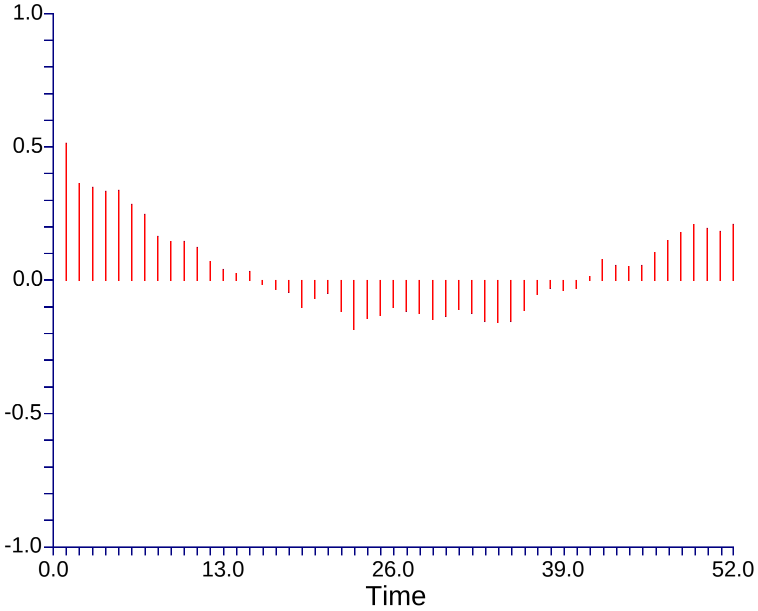

Figure 25. ACF of pH before detrending and deseasonalization

correlation at the end of the year. This same pattern is common to all of the water quality variables. Figure 26 shows the ACF of the detrended, deseasonalized time series of pH which is now absent of the seasonal variation.

Applying the "effective" sample method to the autocorrelation functions from the detrended, deseasonalized data produces the results in Table 43. The values in this table have been rounded up to the next whole number since it is impossible to obtain a fraction of a sample. This table shows the estimated maximum number of independent samples collected each year when based on the current weekly sampling interval. As an example, Table 43 shows that if pH were sampled 52 times per year the maximum number of independent samples would at most be 10. By taking dependent samples, the "effective" *n* used in statistical calculations, such as confidence intervals, would need to be reduced to 10. Obviously, a weekly sampling frequency is too high because at least 42 of the 52 samples are dependent. A more reasonable sampling frequency would be monthly where 10 of 12 samples would be independent. More frequent sampling does have the advantage of allowing separation of the data into subsets so that trends could be analyzed on more than one time series. For example, if pH were sampled 40 times per year, the data could be divided into four subsets of evenly spaced data and trends could then be determined on each subset. The four subsets could then be compared. The data could also be divided into seasonal subsets to determine seasonal trends, which may remove dependence.

The trend levels (Tr/σ_ε) calculated from the Noland Divide, southwestern stream data are shown in column two of Table 44. Remember that the trend level is the trend magnitude over the period of record divided by the standard error of the residuals.

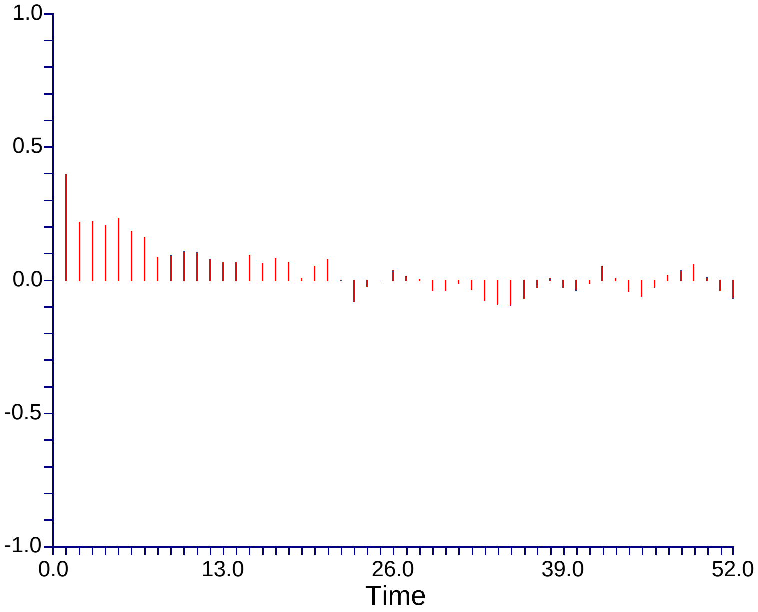

Figure 26. ACF of pH after detrending and deseasonalization.

Table 43. Maximum number of independent samples (*n**) per year.

Water Quality Variable	**Samples (*n**)**
pH	10
ANC	11
Conductivity	13
Chloride	11
Nitrate	10
Sulfate	23
Sodium	12
Potassium	8

Table 44. Current trend detection levels and sampling requirements for Noland Divide (southwestern stream).

Water Quality	**Current Trend**	**Independent**	**Sampling Years**
pH	0.438	10	69.9
ANC	1.145	11	9.0
Conductivity	0.337	13	86.1
Chloride	0.949	11	13.3
Nitrate	1.273	10	8.1
Sulfate	0.395	23	35.9
Sodium	0.523	12	39.1
Potassium	0.422	8	99.8

*Based on 95% confidence and 90% power

Using these trend detection levels to calculate *N** from Equation 4.3 and then dividing by the *n** values in Table 43, the number of years of sampling required to reliably detect trends of a magnitude equal to the standard error at a 95 percent confidence level and with 90 percent power can be calculated. The values are shown in column four of Table 44 and are comparable to those calculated by Lettenmaier (1979) and Lachance, Bobée, et al. (1989) for water quality variables at locations specified in their studies. If the standard error and the slope of the trend remain constant in future years, the trend detection level (Tr/σ_ε) increases, thus causing a decrease in the total number of samples required. Therefore, Table 44 is a "snapshot" in time. Figure 27 shows an example of this using pH and assuming a linear trend of 0.01 pH units per year with a standard error of 0.2. The confidence level and the power level are fixed at 95 percent and 90 percent, respectively. Notice that if the trend remains constant the number of samples required for trend detection, at the confidence and power specified, decreases exponentially as the length of record increases.

Figures 28, 29, and 30 represent sampling records of 15, 10, and five years in length, respectively, and assuming independence among the data. In each graph there are five curves representing five different power levels of detection from 50 percent to 99 percent. Remember that the power level is the probability of not making a Type II error, which is more serious than a Type I error (α-level). The Type II error in this case would be rejecting the presence of a trend when in reality one does exist. The y-axis represents different trend levels and the x-axis gives the number of samples that are collected each year for the record length of 15, 10, or five years. The main purpose of these plots is to show that from sampling frequencies of two samples per year to

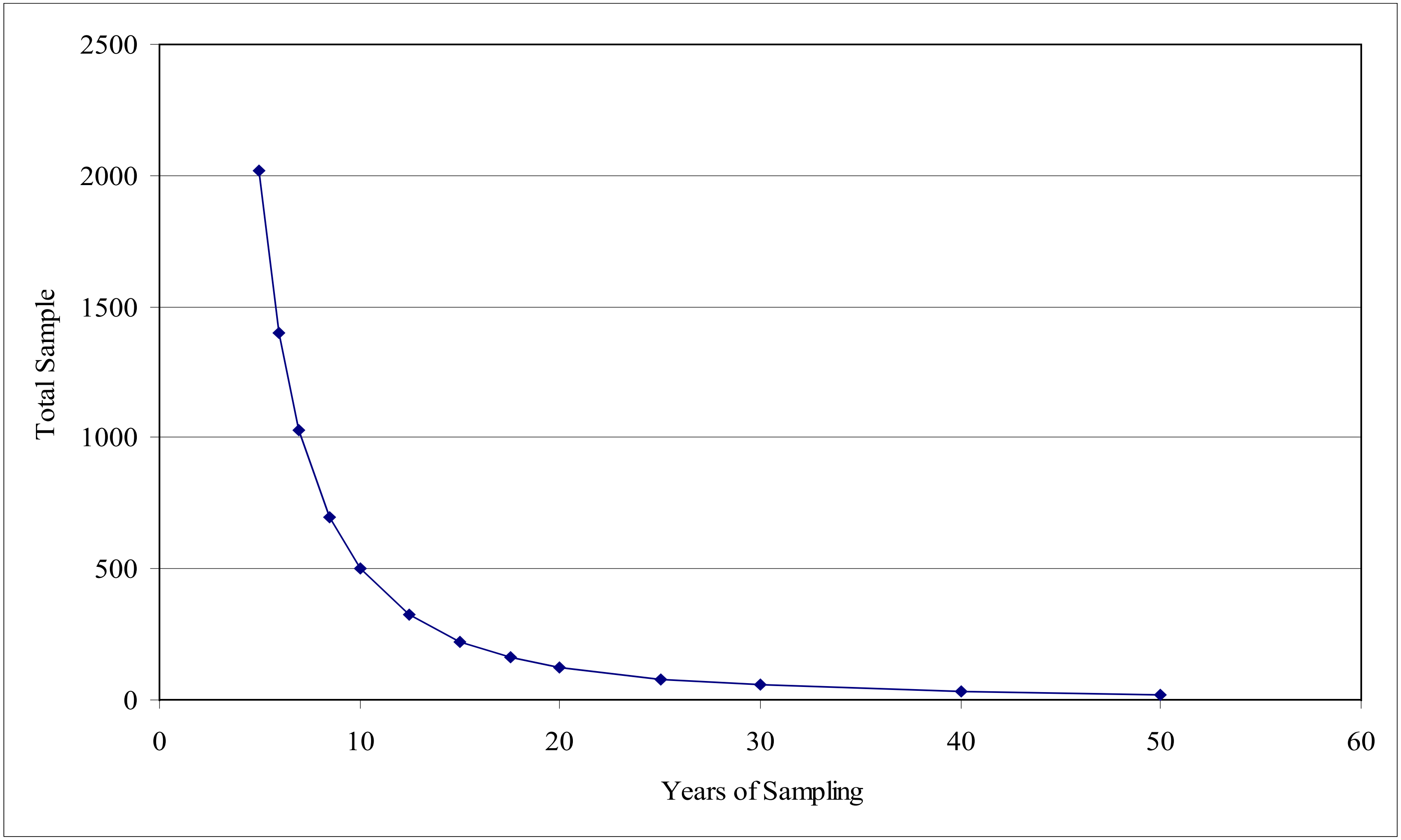

Figure 27. Total samples required given number of years of sampling for pH (α = 0.05, β = 0.10).

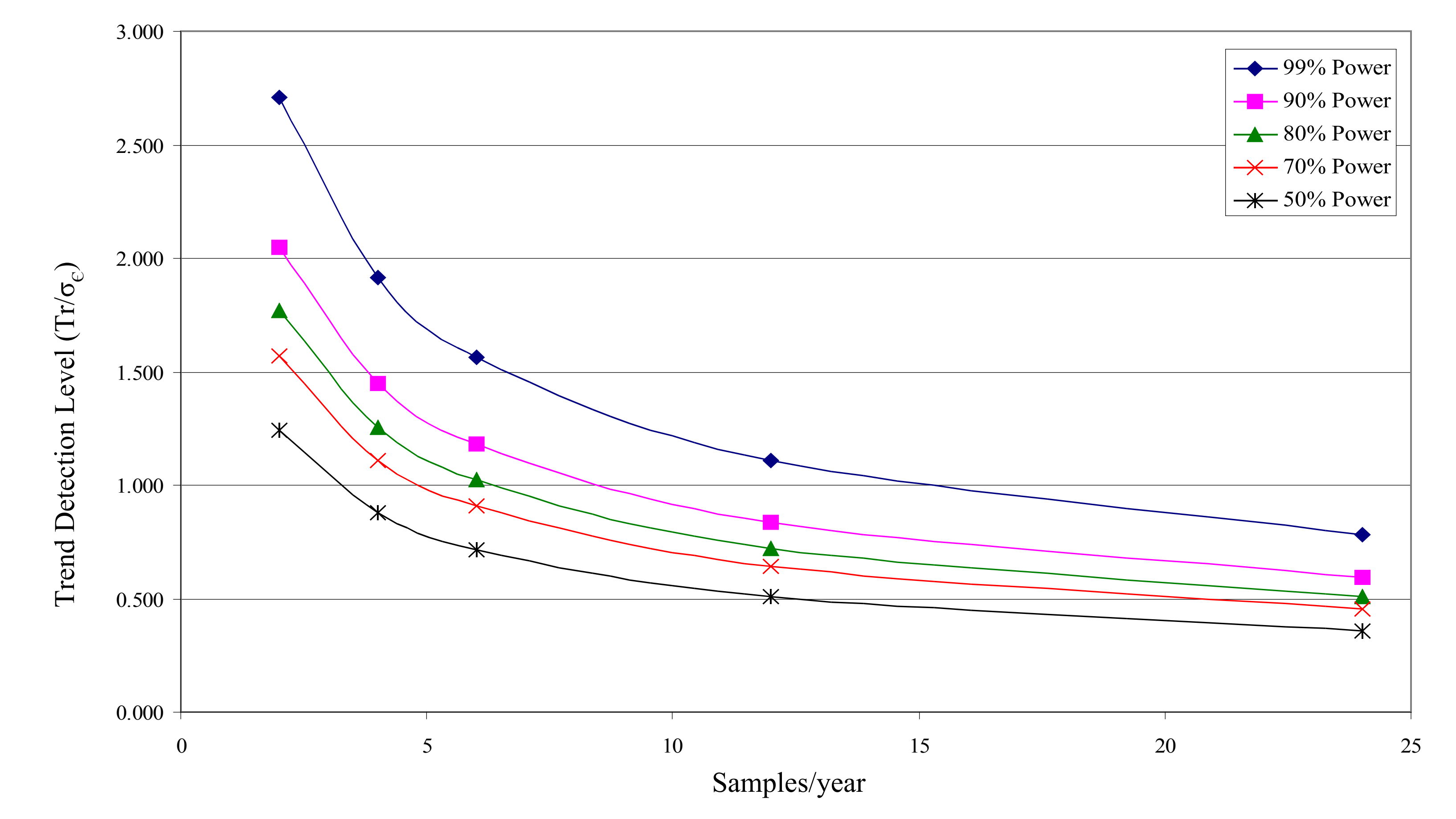

Figure 28. Power curves for 15-year sampling period (α = 0.05).

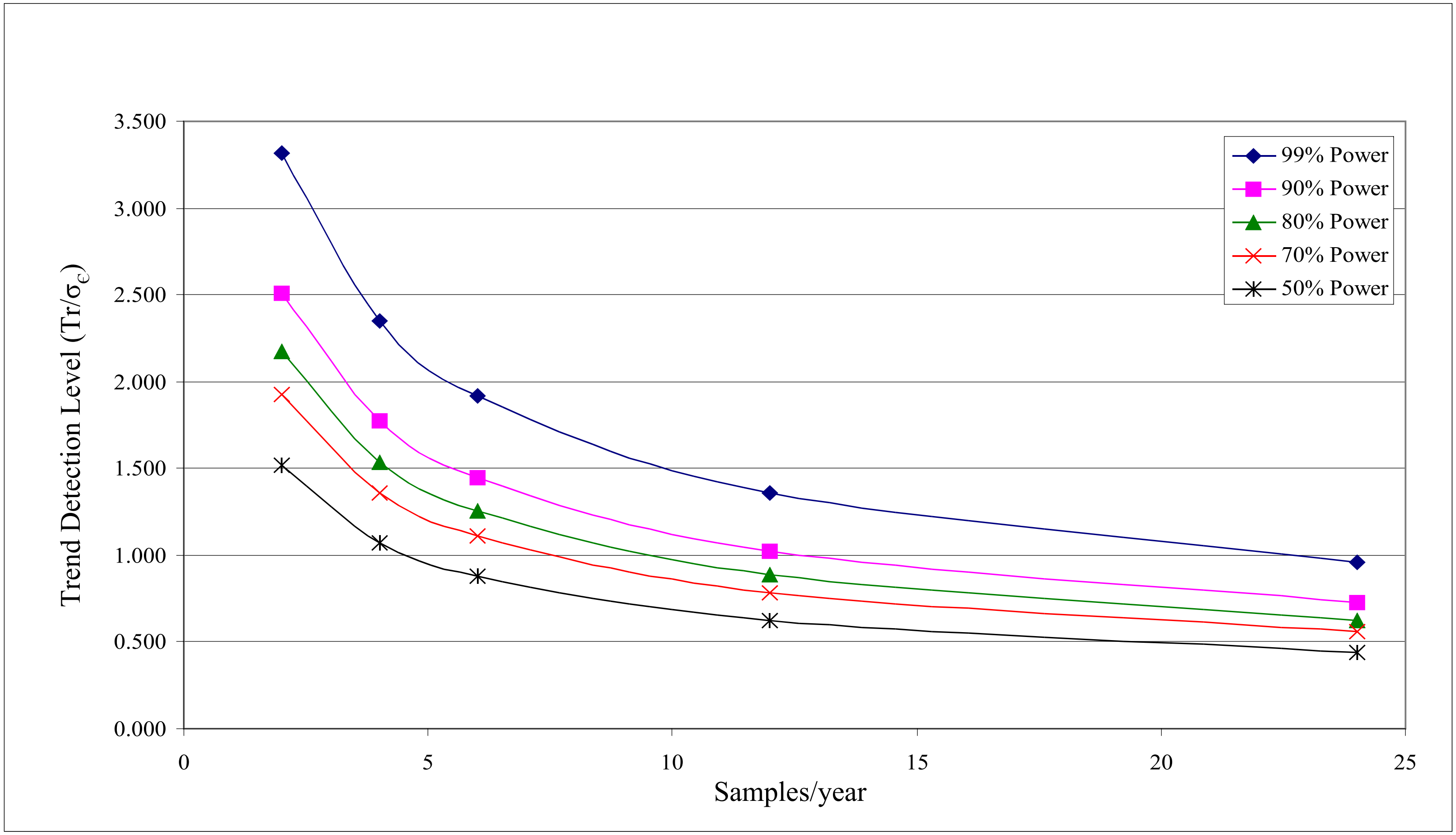

Figure 29. Power curves for 10-year sampling period (α = 0.05).

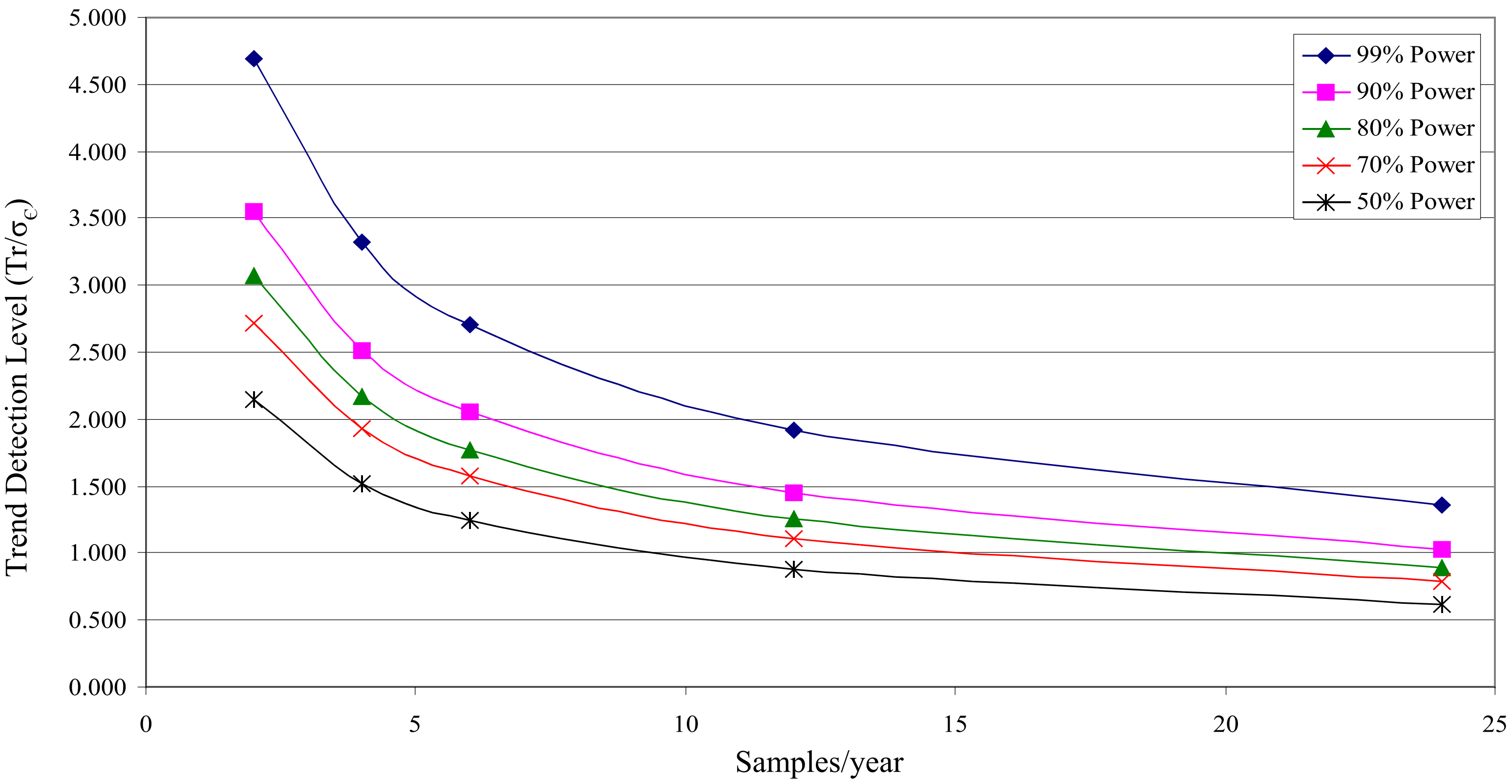

Figure 30. Power curves for 5-year sampling period (α = 0.05).

approximately 10 samples per year the power levels decrease rather rapidly. Beyond 10 samples per year the power levels are almost constant and decrease at a much smaller rate leading to the conclusion that trend detection is greatly improved when the sampling frequency is increased from two to 10 samples per year. However, greater sampling frequencies provide very little additional increase in the ability to reliably detect trends. Figures 28 through 30 also show that trend detection is most reliably obtained by increasing the length of record and not necessarily by increasing the sampling frequency. Suppose in a hypothetical situation that a total of 50 samples were to be collected from a site and two sampling schemes were proposed. The required power of the test was proposed to be 90 percent. In the first scheme, 50 samples would be collected at a frequency of five samples per year for 10 years. The second sampling scheme proposes to collect 10 samples per year for five years. Clearly the annual cost of the second sampling scheme would be twice that of the first scheme. However, according to Figures 29 and 30, the higher sampling frequency has no better trend detection level than the lower sampling frequency. Since this is a long-term trend monitoring site it would be unlikely that high-powered trend detection levels would be required in five years. Therefore, the annual cost savings from using a lower sampling frequency with a longer record would probably be more desirable than a high frequency sampling site.

Based on the trend level of detection being set at a modest value of 1.0, the power of the test is relatively high for a sampling period of 10 years and a sampling interval of four to eight weeks. Lettenmaier (1979) did a study similar to this using data from the Noorsack River near Ferndale, Washington. His study focused on step trends rather than linear trends and he found that the sampling interval should be at least monthly but not more often than biweekly. His results are similar to those produced in this study; however, inherent differences in the water quality and the resulting autocorrelation functions will affect the recommended sampling frequencies.

4.4.3 Results of Sen's Slope Estimation and the Mann-Kendall Test for Trend

The main purpose in this section was to compare larger sampling intervals with the one-week sampling interval of the original time series for each variable using Sen's slope estimation and the Mann-Kendall test for trend. The results of these techniques are reported together because they are very similar. Both methods are performed using the Matlab programs listed in Appendix F. The sampling intervals tested are two, four, eight, and 12 weeks. Thirteen datasets are constructed for each sampling interval tested using different starting points in the original data. This ensures that a number of different subsets are generated so that the results are not based on a single trial. The process is also redundant to some degree because the offset data will overlap a previous dataset at some point. The percentage of trends identified at each new sampling interval is then compared to the percentage identified in the original weekly sampling interval. In the case of Sen's slope estimation a dataset is considered a match with the original dataset if the confidence intervals are similar and bracket the original trend slope. In the case of the Mann-Kendall test for trend a dataset is considered a match with the original dataset if the Z-scores have the same significance as the original dataset. It can be inferred that a sampling interval whose percentage of recognized trends is less than the percentage in the original weekly data may not be an accurate detector of trends.

Table 45 presents the results of Sen’s slope estimation method and the Mann-Kendall test for trend. The numbers in the table represent the percentage of the 13 datasets that identify a trend. The percentage identified in the original weekly data is shown in column two. For example, all 13 datasets (100 percent) in the original one-week and two-week sampling intervals identified a trend for pH using Sen's slope estimation method and the Mann-Kendall test for trend. Eighty-five percent of the13 datasets identified trends when the sampling interval was increased to four weeks. The percentage of trends identified for pH continues to drop when the sampling interval is increased to eight weeks and to 12 weeks. Recall that pH, conductivity, and sulfate had very small trend levels identified in the “effective” sampling method. Evidence of this is shown in Table 45, as the percentage of identified trends falls rapidly, compared to the other variables, when the sampling interval is increased. The information presented in Table 45 is in agreement with the results of the “effective” sampling method from the standpoint that variables with lower trend levels require more samples compared to higher trend levels. The “effective” sample method determined that an increase in sampling frequency in the range of two to 10 samples per year provides the greatest benefit from the standpoint of sampling frequency; however, a sampling frequency greater than 10 per year provides very little increase in benefit due to data dependency. Serial dependence could be affecting some of the results in Table 45 for one-week and two-week sampling intervals. Not considering sulfate or conductivity, it would seem that a sampling interval of approximately monthly or bimonthly would be acceptable. Based on the results of MK and SE, sulfate and conductivity would need to be sampled weekly, which would be rather costly for a large sampling network.

Table 45. Percentage of 13 datasets identifying the same trend as the original sampling interval of one week.

Water Quality	**Sampling Interval (weeks)**				
Variable	**1**	**2**	**4**	**8**	**12**
PH (SE)*	100	100	85	54	38
PH (MK)**	100	100	85	46	38
Conductivity (SE)	100	31	8	8	8
Conductivity (MK)	100	38	8	8	8
ANC (SE)	100	100	100	100	92
ANC (MK)	100	100	100	100	92
Chloride (SE)	100	100	100	100	100
Chloride (MK)	100	100	100	100	100
Nitrate (SE)	100	100	100	100	85
Nitrate (MK)	100	100	100	100	85
Sulfate (SE)	100	0	0	15	8
Sulfate (MK)	100	0	0	8	8
Sodium (SE)	100	100	69	46	31
Sodium (MK)	100	100	77	46	31
Potassium (SE)	100	69	46	31	8
Potassium (MK)	100	54	46	31	8

*SE - Sen's slope estimation method
**MK – Mann-Kendall method

4.4.4 Boxplot Analysis

Boxplot analyses were performed on the water quality variables by selecting observations based on the alternative sampling frequencies. A family of boxplots for each water quality variable is shown in Figures 31 through 38. Each figure contains seven boxplots that represent weekly (X1), monthly (X4), bimonthly (X8), quarterly (X12), 16 week (X16), 20 week (X20), and 24 week (X24) sampling intervals. The boxplot method was performed to determine if visually detectable changes were apparent when the data were thinned to lower sampling frequencies. Each family of boxplots should be used by comparing the weekly (X1) boxplot to the remainder of the boxplots. The X1 boxplot represents a plot of all the original data. Subsequent boxplots are constructed from thinned data based on the sampling interval shown. Boxplots that exhibit noticeable differences from the X1 boxplot signify a change in the structure of the data as it relates to median, IQR, and observations in the outer fences. The IQR is the difference between the 25th and the 75th percentiles; these points are often called hinges. The 25th percentile is demarcated at the lower edge of the gray box in the boxplots and the 75th percentile by the upper edge. The upper and lower fence levels are calculated by the hinge values ±1.5 times IQR. The line near the midpoint of the gray box is the median.

In general, the changes were not as noticeable as one might imagine. However, there are subtle changes that signify some importance in sampling intervals. Overall, the boxplots of a monthly and bimonthly sampling interval compare best to the boxplots of the original data (X1). When the sampling interval is greater than bimonthly the degradation of similarities to X1 is more noticeable. It should be mentioned that each

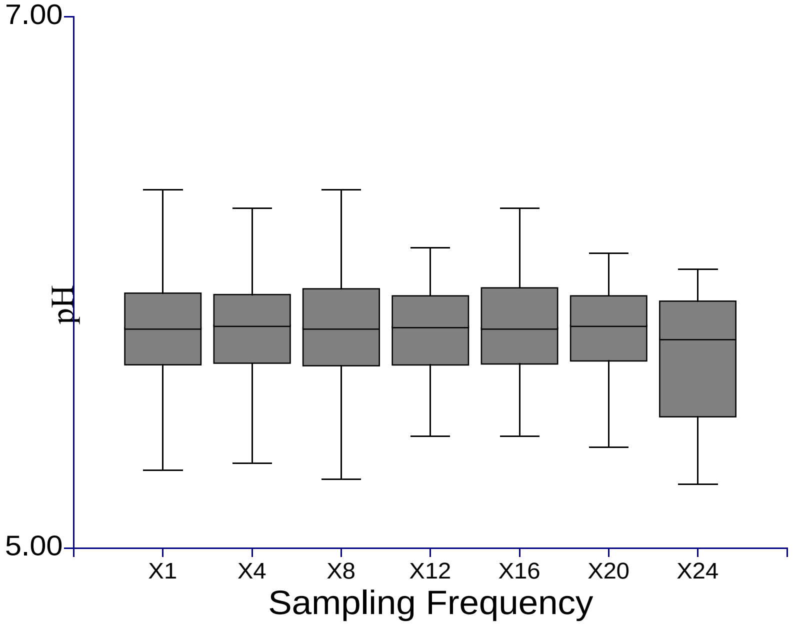

Figure 31. Sampling frequency boxplots for pH.

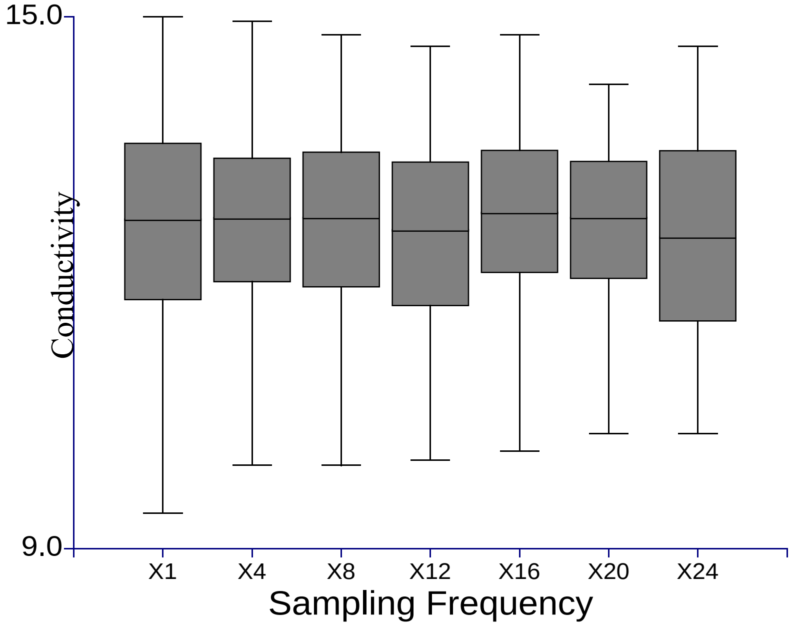

Figure 32. Sampling frequency boxplots for conductivity.

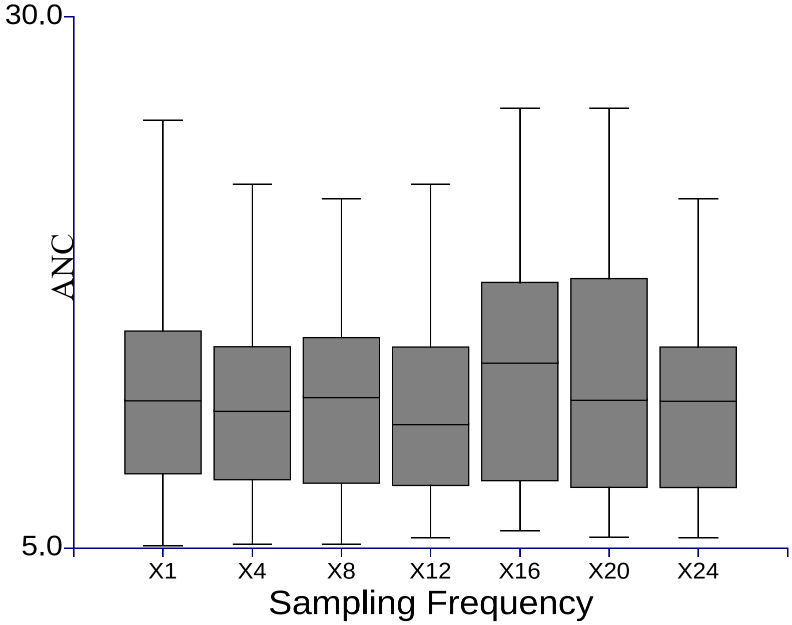

Figure 33. Sampling frequency boxplots for ANC.

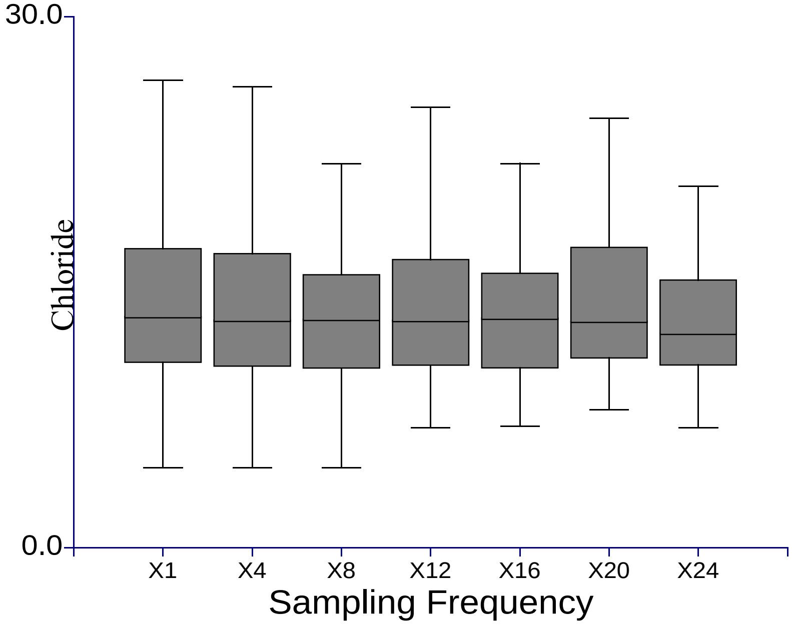

Figure 34. Sampling frequency boxplots for chloride.

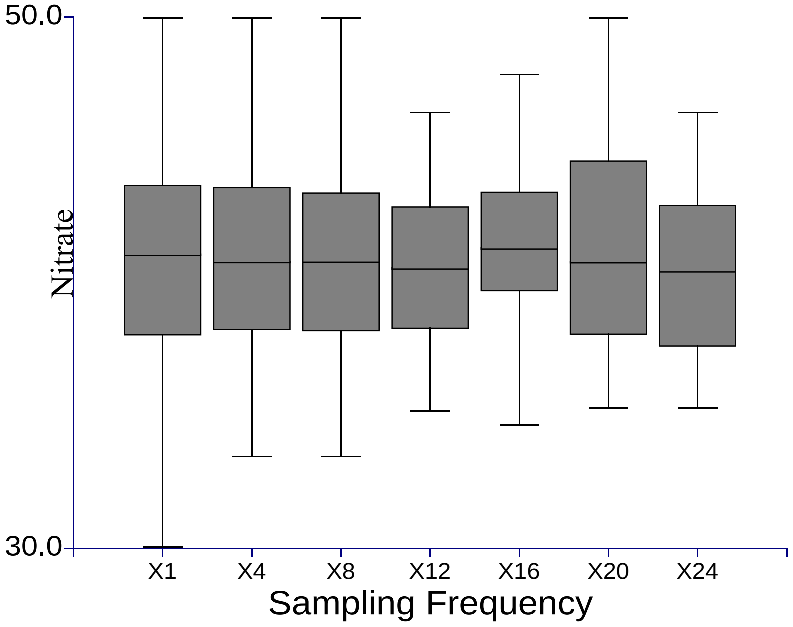

Figure 35. Sampling frequency boxplots for nitrate.

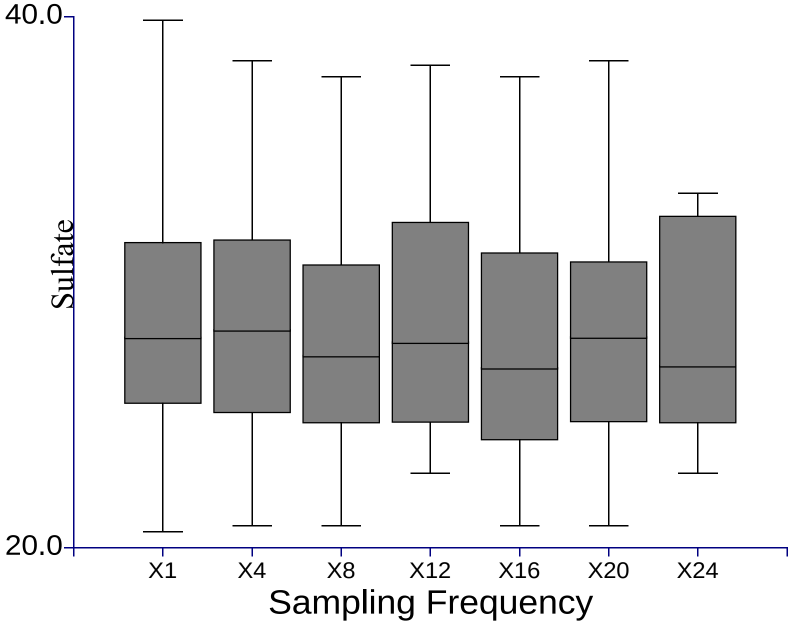

Figure 36. Sampling frequency boxplots for sulfate.

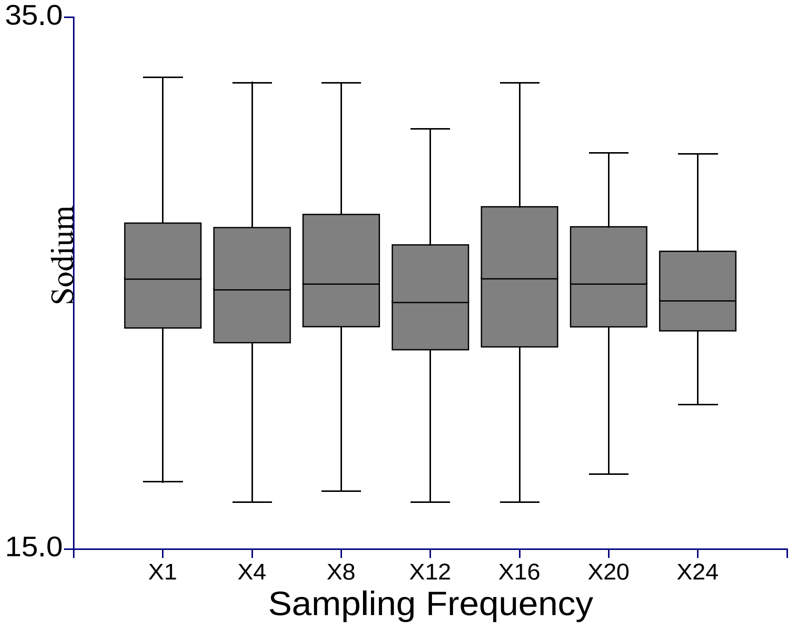

Figure 37. Sampling frequency boxplots for sodium.

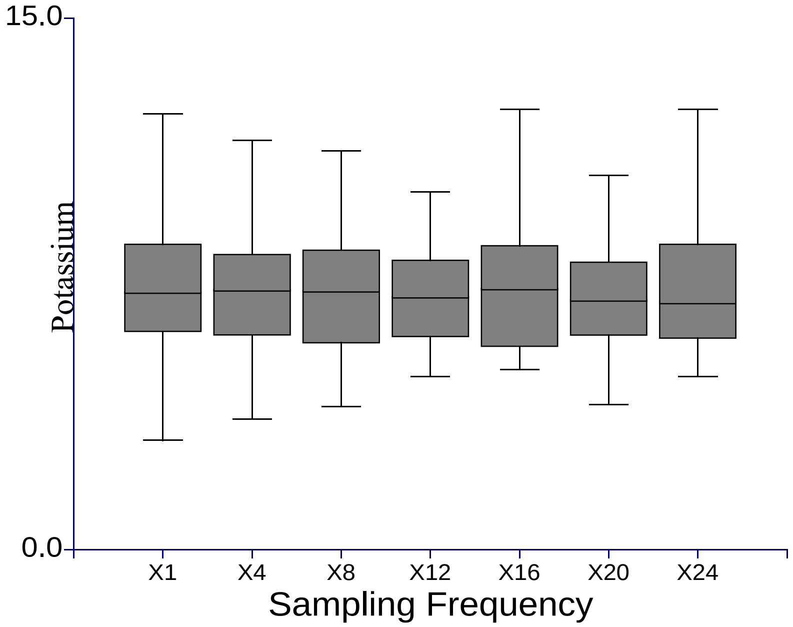

Figure 38. Sampling frequency boxplots for potassium.

boxplot is for one sampling of data, whereas earlier tests (Sen's slope estimator and Mann-Kendall test for trend) using a moving window scheme tested up to 13 different samplings of the data. Each additional sampling would produce the same number of graphs shown in Figures 31 through 38. Because one sampling is used, resulting boxplots may not represent a true picture of the population and may take on the form shown by sheer "luck of the draw." The following paragraphs provide some interpretation and comparison of the boxplots.

Figure 31 for pH shows that the boxplots for monthly (X4) and bimonthly (X8) sampling intervals more closely resemble the X1 boxplot than sampling intervals greater than eight weeks. There is a small shift in the median and the IQR but they remain fairly even. Boxplots at X12 and X20 appear to be very close comparing the median and IQR but the outer fences are quite different.

Figure 32 for conductivity shows a gradual decreasing in the range between the outer fences. This would lead one to believe that the sampling interval closest to X1 is probably the most desirable if there is a proposed change from weekly sampling. However, the median and IQR of X8 more closely resemble X1.

ANC shown in Figure 33 displays a lower fence limit that is nearly equal and an upper fence that is highly variable. The result may indicate periods of high variability possibly caused by seasonal variation. Small sampling intervals in the boxplots may mask variability because the data are lumped together, while larger sampling frequencies may isolate some of the periods when the data are highly variable. The outer fences of X16 and X20 are closer to those of X1. However, the IQRs have increased approximately 25 to 30 percent. X4 is probably the closest comparison in terms of the outer fences, but X8 is the closest in terms of the interquartile range.

Chloride, shown in Figure 34, is perhaps one of the easiest to compare to X1. X4 is much closer to X1 than any other boxplot. When the sampling frequency is increased beyond monthly, the outer fences and the IQRs become highly variable, although the median appears to remain fairly consistent.

Figure 35 for nitrate shows that X4 and X8 are very similar and compare very well with X1 except for the lower fence. X1 shows a far greater range in the lower fence. The upper fences appear to be equal. When the sampling interval is increased to quarterly and greater, the outer fences and the IQRs become highly variable.

Sulfate, in Figure 36, exhibits little difference in the fence ranges among X4, X8, X16, and X20. Overall, X4 does seem to compare better with X1 than the other boxplots. However, X4 does have a noticeable increase in the median and the IQR.

Figure 37 for sodium exhibits small differences among X1, X4, and X8. Sampling intervals greater than bimonthly exhibit more variation compared to X1. X8 appears to be more similar to X1 than X4 because of similarities in the median and the fence range.

Figure 38 for potassium shows a gradual decrease in the fence range from X1 to X12. After X12 the fence range, IQR, and the median are more variable. Again, X4 and X8 are similar to X1. The fence range and IQR of X4 are closer to X1, but the median in X8 appears to be slightly closer to X1 than X4.

4.4.5 Comparison of Sampling Frequency Methods

This section has been included to compare the results of all those methods used above to study the sampling frequency requirements. It should first be restated that the basis of the frequency analysis was to determine an optimum sampling frequency for trend detection. The current sampling frequency of the monitoring network should not be termed as the "wrong" sampling frequency because it does produce mostly independent samples according to the "effective" sampling method. Additionally, the data are viable and represent a historic record that is quite capable of detecting trend. The main question to be asked is what level of accuracy is needed in identifying long-term trends. The frequency of sampling and length of record provide the answer.

The current trend levels of Table 44 are in agreement with the results of Sen's method and the Mann-Kendall test for trend. Lower trend levels require greater sampling frequencies. However, the "effective" sampling method determined that increasing the sampling frequency is very beneficial to a certain point. The benefits are greatly outweighed by the costs when the sampling frequency is increased beyond this point (10 or 12 samples per year). The "effective" sample method also determined that, on average, a sampling frequency greater than about 10 samples per year would produce some serially dependent data. The boxplot analyses, in general, seemed to support sampling frequency of six to 12 samples per year. Sampling frequencies less than 12 samples per year usually resulted in boxplots that, from a visual standpoint, appeared significantly different or highly variable compared to the existing data.

Since sodium and potassium are not robust measures of stream health, their findings should be considered secondary. Conductivity and sulfate required the most frequent sampling (weekly), but this is unreasonable for an 83 site monitoring network, and the observations would be dependent to some degree. On the positive side, these two variables are highly correlated with other variables that do not require high sampling frequencies.

Based on the findings thus far a recommended sampling frequency would fall between six and 12 samples per year. Further discussion of this will be in Section 4.6.

4.5 Trend Detection Improvement using Multiple Sampling Sites

Most of Chapter 4 has addressed the sampling frequency needed in the monitoring network to detect long-term trends based on data from the southwestern stream at the Noland Divide monitoring site. It has already been stated that the one drawback of using these data is that it measures the water quality at only one location and this study is focusing on an 83 site synoptic network. Again, the Noland Divide data was used because of its high frequency sampling which allows a gradual thinning of the data to test lower sampling frequencies. However, some consideration should be given to the fact that whether the redesigned sampling network remains to be 83 sites or some subset of that number, the data obtained from multiple sampling locations should increase the trend detection accuracy beyond the use of only one sampling location. In statistical calculations the old saying that “there are strength in numbers” is certainly not without merit. The increase in the number of observations affects almost all statistical calculations by improving their accuracy and power. In the monitoring network studied here this should not be an exception. An example of this is presented below for sampling sites in elevation class four (between 2000' and 2500' MSL).

Elevation class four contains a total of 16 sampling sites from the monitoring network. Robinson (2003) used 15 of these sampling sites to study the change in the width of confidence intervals of the trend coefficient as the number of sites used in the calculation of the confidence intervals was decreased from 15 to one. Sampling sites were dropped from the calculation of the confidence intervals in the order of their sampling site identification number. The trend regression equation developed by Harwell (2000) for normalized H^+ in elevation class four is given by

$$normalized\, H^{+'} = 0.631 + (8.796\text{e-}02)\cos\theta + (4.686\text{e-}02)\sin\theta + (2.260\text{e-}04)\,t \qquad (4.13)$$

where θ = 2π*(fraction of a year)
t = cumulative Julian days beginning at January 1, 1991

Figure 39 displays the graphical results by Robinson (2003) for determining the confidence intervals and coefficients of the trend of the sampling sites in elevation class four. Each set of points lined up vertically from the horizontal axis represents one regression analysis. The horizontal axis shows how the sampling sites were removed in numerical order from one regression analysis to the next. The first analysis contains all 15 sampling sites. The next set of points represents the regression analysis for all sampling sites except site one, and so on. It can be seen that the confidence intervals are fairly uniform in width until five sites (sampling sites IDs less than 148) have been removed from the analysis. As additional sites are removed the confidence intervals widen at a uniform but small rate until 10 sites (sampling sites IDs less than 311) have been removed from the analysis. At this point, only five sampling sites are being used

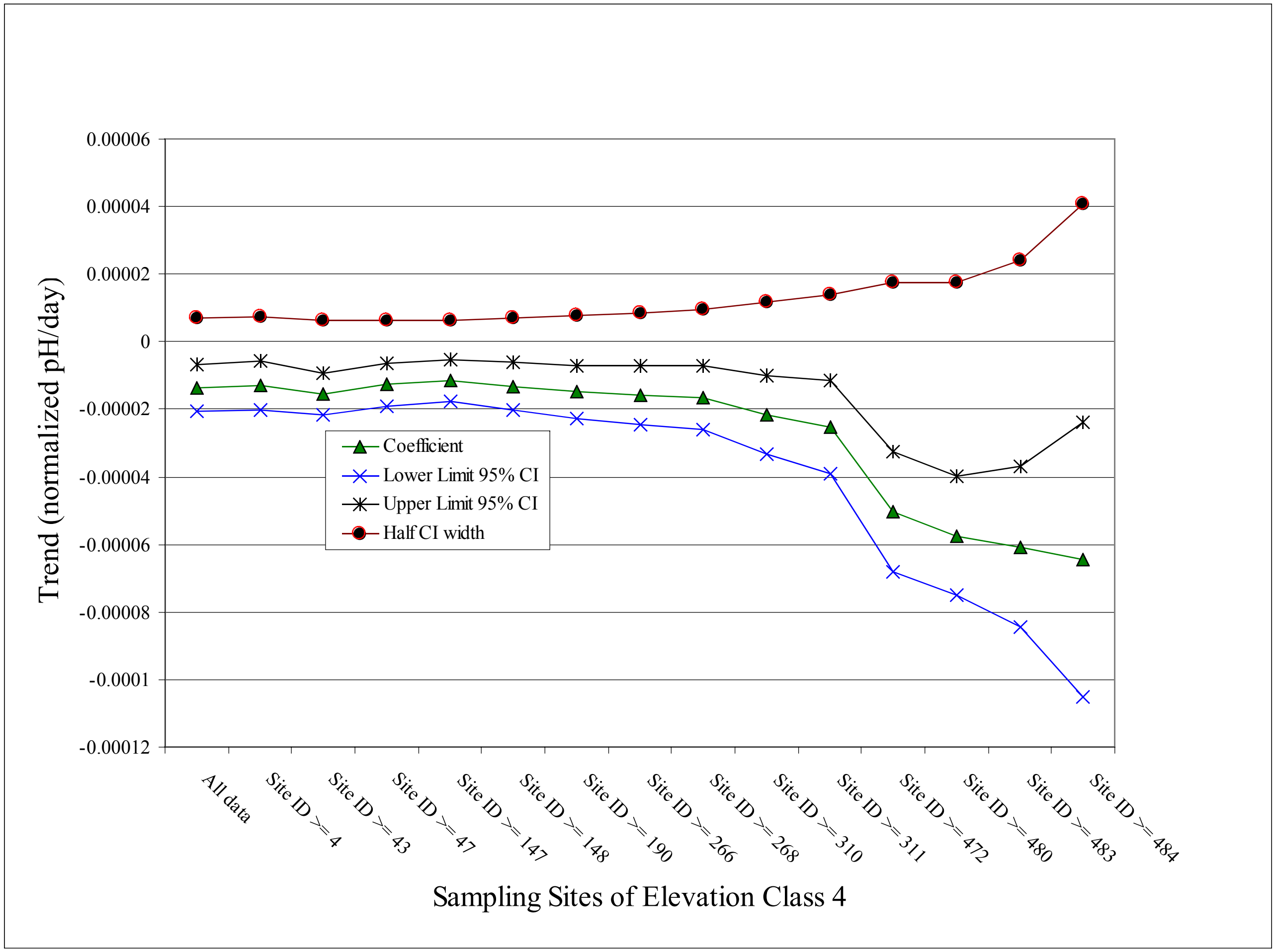

Figure 39. Confidence intervals for trend in elevation class 4 sampling sites.

in the regression analysis and the confidence intervals are still somewhat reasonable compared to the analysis with all sampling sites. However, when additional sites are removed the confidence interval widths and the coefficient of the trend component are affected drastically. For this data it seems that 10 of 15 sites provide a good comparison to that of using all 15 sampling sites; five of 15 sites still provide a good comparison but with a noticeable increase in confidence interval width; and four sites or less provides unreasonable changes in the confidence interval width and in the trend coefficient.

These results have two implications. The first issue is that of including a number of sampling sites from each elevation class in the redesigned monitoring network. The issue was briefly addressed in Section 3.7 and will be addressed further in Section 4.6. The second issue is the confidence of the detected trend. The confidence interval width using only one sampling site is more than twice the confidence interval width using five sampling sites. This very simply displays the increase in confidence of having multiple sites for trend detection. The elevation classes of Table 38 should be used in the final network redesign to ensure that some minimal degree of redundancy is kept that will increase the confidence in the trend identification.

Urquhart, Paulsen, et al. (1998) proposed a method to determine regional trends that may be applicable in the GRSM. Their method was rather simple and involved performing time series regression on the means of multiple sampling sites to determine a regional trend. They calculated the means among the sites for each time period and then performed time series regression using the means. It is easy to understand that this method requires a relatively high number of sites to be accurate. The method could also be easily influenced by outliers in the data. In this case the medians could be used. This method could possibly be applied to all the sampling sites in the GRSM. The sampling sites could be divided into elevation classes or into the clusters that were formulated in this study. Another issue that has not been touched on is the possibility of spatial dependency, which could definitely be a factor when using this method or any other method that uses multiple sites to determine trends.

4.6 Summary of Findings

This section presents a summary of the findings from the frequency analyses for determining the adequacy of the current sampling frequency to detect long-term trends. The following list contains summary results of each sub-section in this chapter, some key points that should be taken into consideration in the redesigned network, and notable findings. A final discussion and recommendations will be presented in Chapter 5. The summary is as follows:

- The “effective” sampling method determined that, on average, a sampling frequency greater than approximately 10 samples per year may result in serial dependence among the data. It was also determined that length of record is much more beneficial than increasing the sampling frequency, especially when the data becomes dependent. At the point where dependency is a factor, increasing data collection results in very little improvement of statistical power.
- Sen's slope estimator and the Mann-Kendall test for trend both showed that sampling intervals of monthly and bimonthly were still relatively accurate in detecting the trends of the weekly data. However, it must be remembered that the weekly data is

serially dependent which could also affect the actual trend in the data.

- The boxplot analysis revealed that for most water quality variables a monthly or bimonthly sampling interval compares well with the weekly boxplots. Sampling intervals greater than bimonthly exhibit much more noticeable changes in the medians, IQRs, and outer fences.
- The use of multiple sites decreases the width of confidence intervals for determining the trend. Regression analysis for trends in elevation class four showed that seven sampling sites produced confidence intervals that were only slightly larger than if 15 sites were used. When fewer than seven sites were used the confidence interval width increased rapidly. Since some water quality variables have produced different trends at different elevation classes it would be prudent to preserve elevation classes in the final design so that multiple sights might be used to decrease confidence interval widths of trend estimates or to possibly decrease the time required to reliably detect trends. Multiple sites for trend detection improvement should however be checked for spatial dependence.

Chapter 5 Final Discussion and Recommendations

5.1 Spatial Analyses

The spatial analyses were performed by deriving monetary benefit scores using multivariate statistical methods and then optimizing the monitoring network using a simulated annealing algorithm that considered the monetary benefit scores. In general, those sampling sites that explained the greatest amount of variability within their respective cluster received the highest benefit while those sites that explained the least amount of variability in their cluster the lowest benefit. The costs of obtaining the samples from the field were also considered.

The first simulated annealing optimization (SA1) resulted in the recommendation that 16 of 83 sampling sites be discontinued. The list of sites to be retained and discontinued are shown in Table 30. The list of sites retained in Table 30 is the optimized network based on maximized net benefit. However, there are some other considerations to take into account based on the Summary of Findings in Section 3.8.

If the sampling sites in overlapping regions of clusters, that were discontinued in SA1, were moved to the retained group this would result in a redesigned network of 70 sampling sites. Although these results are probably the most beneficial they do not represent a significant change from the original network. Perhaps a better way of approaching the selection would be to combine the information about sites in overlapping clusters, cluster memberships, elevation classes, and the simulated annealing method of finding the *n* best sites (SA2) in Table 37. This can lead to the identification of primary, secondary, and tertiary sampling sites as follows (the addition of sites to the following lists are determined by the order in which they are added to the network in Table 37 to ensure that the most beneficial sites are chosen):

1. Primary Sites – [4, 13, 14, 20, 23, 24, 30, 34, 47, 49, 66, 71, 73, 74, 114, 137, 142, 143, 144, 147, 148, 149, 173, 174, 191, 193, 194, 233, 234, 237, 251, 252, 253, 266, 268, 291, 293, 488, 489, 492, and 493] This list of sites is the set of the *n* = 40 best sites. These sites are considered primary because their maximized objective function in SA2 is not much different from the objective function of the optimum network in SA1. This group of sites includes 10 of 22 sites in the overlapping regions of clusters based on water quality and watershed characteristics and 17 of 24 collocated sites. All clusters and elevation classes are represented in the set of primary sites.
2. Secondary Sites – [115, 156, 184, 192, 221, 310, 311, 337, 472, 473, 479, 480, 481, 482, 483, 484, and 485] Adding the secondary sites to the primary sites would guarantee that all elevation classes, water quality clusters, and clusters of the watershed characteristics are represented by at least three membership sites. In elevation classes or clusters where there are less than three sites the remaining sampling sites from those classes or clusters are added. The *n* = 50 best sites would be included in this list. The secondary sites also add three more of the collocated sites for a total of 20 out of 24 and four more sites in overlapped areas of clusters for a total of 14 out of 22.
3. Tertiary Sites - [50, 52, 150, 186, 190, 209, 210, 213, 214, 215, 474, and 475] The addition of the tertiary sites would include all the sites from the SA1 optimization. Secondary sites include the addition of three more collocated sites for a total of 23 out of 24 and five more sites in overlapping areas of clusters for a total of 19 out of

22.

4. Assuming that primary and secondary sites are included in the redesigned network for a total of 58 sampling sites, new sites should be added to the network in the elevation ranges of 3000-4000 feet MSL, 4000-5000 feet MSL, and above 5000 feet MSL to distribute the network proportionally according to the park areas at these elevations. The redesigned network would include eight new sites between 3000 and 4000 feet MSL, six new sites between 4000 and 5000 feet MSL, and three new sites above 5000 feet MSL. This would bring the redesigned network to a total of 75 sampling sites.
5. Sites recommended for discontinuance – [1, 3, 43, 45, 46, 103, 104, 106, 107, 127, 138, 200, and 336] These sites represent the balance of the sites not included in the primary, secondary, or tertiary lists and are recommended for discontinuance. All of these sites were all discarded by the SA1 optimization. An evaluation may be furthered by the NPS to determine if any sites are needed from this list for other purposes not mentioned in this study.

The sampling sites listed in No. 5 should not be viewed as the only sites that can be discontinued. These sites are simply the ones that remained after the selection of the primary, secondary, and tertiary sites. Discontinuation of additional sites should be contemplated but it is recommended that their priority (primary, secondary, or tertiary) and order of being retained (Table 37) be considered as guidelines. The primary, secondary, tertiary, and remaining sites are shown in the tree diagram of Figure 40 and on the plot of the GRSM in Figure 41.

This study did not address the seven sampling sites that are taken from high-

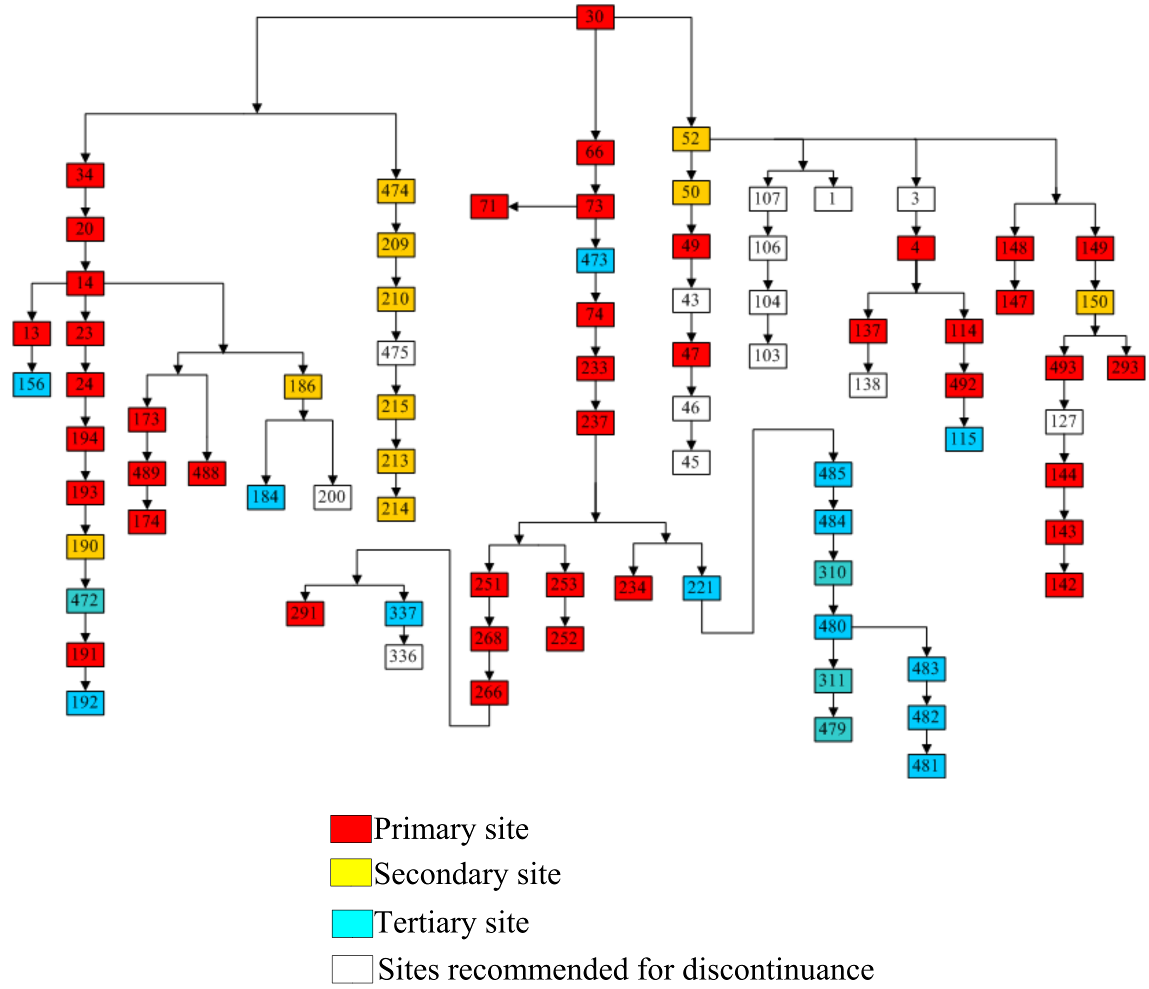

Figure 40. Identification of primary, secondary, and tertiary sampling sites.

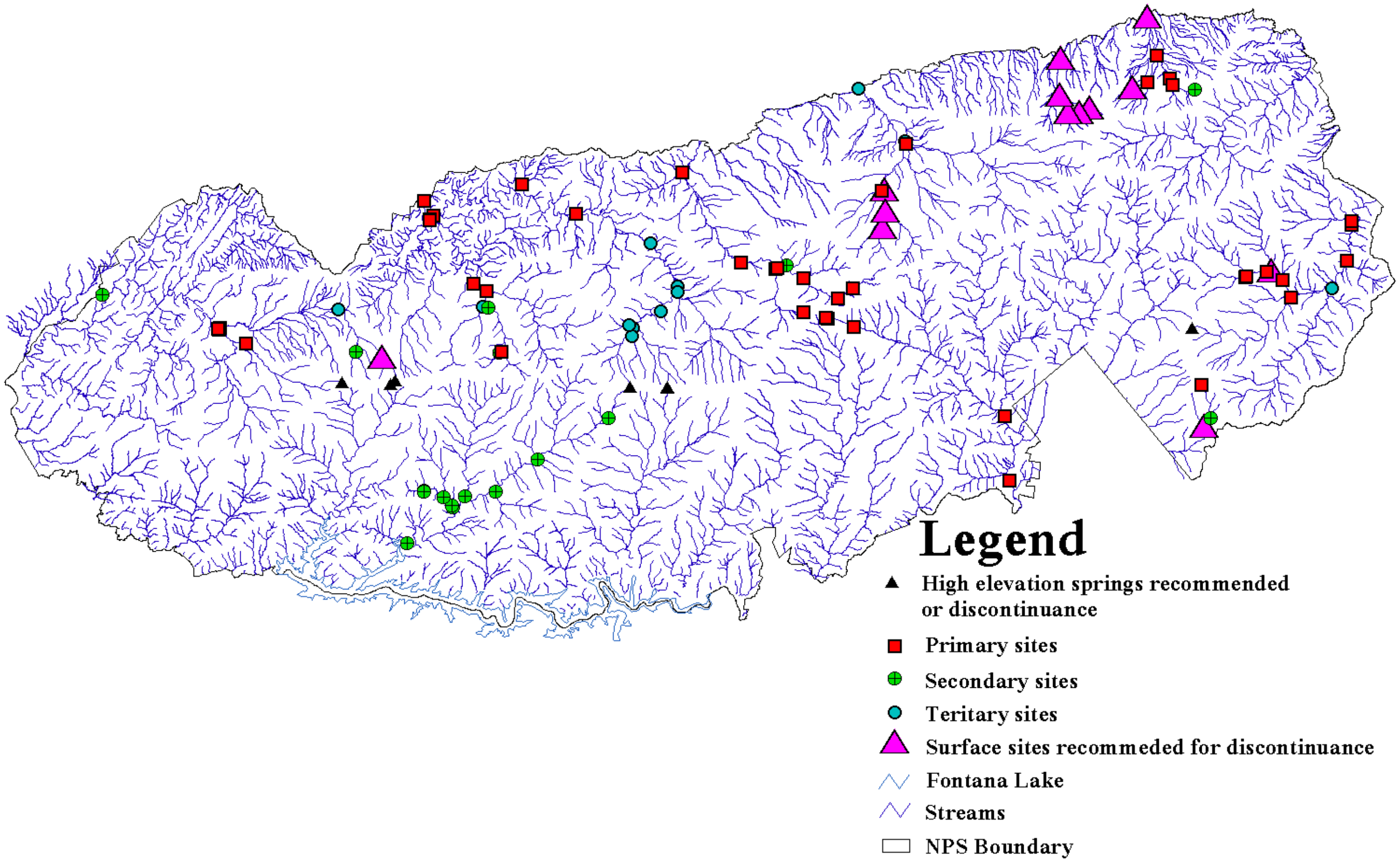
Legend
▲ High elevation springs recommended or discontinuance
🟥 Primary sites
⊕ Secondary sites
● Teritary sites
▲ Surface sites recommeded for discontinuance
Fontana Lake
Streams
NPS Boundary

Figure 41. Plot of primary, secondary, and tertiary sampling sites.

elevation springs. However, these sites are sampled at the same time samples are collected for the 83 sites focused on by this study. On average, the seven high-elevation springs are probably higher in access cost because they are located near the North Carolina-Tennessee border. It has already been mentioned that the chemistry is usually different for springs compared to surface water and the same difference would be expected here. It is suggested that the NPS review the high-elevation springs and consider discontinuance so that funds can be redirected to additional surface sites or to the existing surface sites.

A final recommendation for a redesigned network that incorporates sampling frequency will be given in Section 5.3 after stating the conclusions of the frequency analyses in Section 5.2.

5.2 Frequency Analyses

The frequency analyses employed the "effective" sample method, Sen's slope estimator, Mann-Kendall test for trend, and boxplots to study the effectiveness of smaller sampling frequencies relative to the weekly sampling data from the Noland Divide, southwestern stream. The sampling frequency does need to be increased if the time for determining a trend with a specified confidence and statistical power is to be reduced. However, the frequency should not be increased to a point where serial dependency becomes a problem. Based on the findings of Chapter 4 it is recommended that the sampling frequency be increased from four samples per year to 6-12 samples per year. If long-term trend detection is the only focus then six samples per year may be adequate. As mentioned earlier, this study has not focused on the detection of seasonality. If it is desired to also capture seasonal trends then 12 samples per year may be better. After review of past literature and observing the results of the frequency analyses the word “optimum” may be difficult to define, however, because the answer depends on the length of record and the desired level of statistical power. A general opinion from the literature review has revealed that in most cases a sampling frequency of 12 samples per year is recommended, especially when there is no data available to base a sampling frequency.

It is also realized that any increase in sampling frequency is subject to the feasibility of the current operating budget. The variable cost breakdown has shown that the laboratory costs, data interpretation, overhead, and other miscellaneous expenses are more expensive than the costs associated with retrieving the sample. The cost of one sample on average for retrieval is approximately $58 while the costs for laboratory work, data interpretation, overhead, and miscellaneous expenses for one sample is approximately $134. These costs combined with the recommendations of a redesigned network and an increase in sampling frequency will be discussed in Section 5.3.

5.3 Combining the Analyses

This section combines the discussions of the spatial and frequency analyses into the final recommendation for a redesigned monitoring network. First, from a spatial perspective it would seem prudent to include all of the primary sites in a new network. These sites are the most important from the standpoints of benefit and uniqueness. Second, from a temporal perspective primary stations could be divided into monthly and a bimonthly sampling intervals. The monthly sampling sites would consist of: 4, 13, 14, 20, 23, 24, 30, 34, 49, 66, 71, 73, 74, 144, 147, 148, 149, 173, 174, 193, 194, 233, 237, 251, 252, 253, 266, 268, 291, 293, 488, and 489. These are primary sites that are located next to roads so vehicle access could mainly be used. The balance of the primary sites (47, 114, 137, 142, 143, 191, 234, 492, and 493) would be sampled bimonthly. Based on an average access cost of $58 per sample and $134 for laboratory expenses, etc., the total annual variable cost would be $84,096.

It is also recommended that the secondary sites (115, 156, 184, 192, 221, 310, 311, 337, 472, 473, 479, 480, 481, 482, 483, 484, and 485) be sampled six times per year because these sites guarantee that there is representation from all clusters and all elevation classes in the new network. The total annual variable cost for the secondary sites would be $19,584. A combined total for primary and secondary sites would be $103,680. If new sites were added to correct the proportionality to park area within all elevation ranges this would add an additional 17 sites that could be sampled bimonthly for an additional variable cost of $19,584. The new network would consist of 75 sampling sites for a total variable cost of $123,264. If tertiary sites were added to the network the recommended sampling frequency would be six samples per year. Each site added would cost an additional $1,152 per year. The addition of the 12 tertiary sites would equal an increase of $13,824. Again it should be made clear that these costs are the variable costs and do not include the fixed costs for the sampling program. The redesigned network including fixed costs for primary, secondary, and new sampling sites would be $193,674. This represents a net increase of $54,099 from the existing network.

There are other considerations that must be taken into account. Much of the sampling is performed by volunteers. Their willingness and personal schedules may be strained by the increased frequency. Organization could be even more complicated depending on how the bimonthly schedule is arranged around holiday seasons. Additionally, accessing sampling sites in higher elevations during the winter months can be treacherous because of snow and ice. An increase in sampling frequency will also increase the exposure of sample collectors during the wintry months.

LIST OF REFERENCES

LIST OF REFERENCES

Aarts, E. and J. Korst (1989). Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing. New York, John Wiley and Sons.

Anderberg, M. R. (1973). Cluster Analysis for Applications. New York, Academic Press.

Baker, F. B. and L. J. Hubert (1975). "Measuring Power of Hierarchical Cluster Analysis." Journal of the American Statistical Association **70**(349): 31-38.

Bayley, G. V. and J. M. Hammersley (1946). "The "Effective" Number of Independent Observations in an Autocorrelated Time Series." Supplement to the Journal of the Royal Statistical Society **8**(2): 184-197.

Ben-Jemaa, F., M. A. Marino, et al. (1995). "Sampling Design for Contaminant Distribution in Lake Sediments." Journal of Water Resources Planning and Management-Asce **121**(1): 71-79.

Berryman, D., B. Bobée, et al. (1988). "Nonparametric Tests for Trend Detection in Water Quality Time Series." Water Resources Bulletin **24**(3): 545-556.

Biswas, A. K. (1997). Water Resources - Environmental Planning, Management and Development. New York, McGraw-Hill.

Brieman, L. (1973). Statistics with a View Toward Application. Boston, Houghton-Mifflin.

Brown, C. E. (1998). Applied Multivariate Statistics in Geohydrology and Related Sciences. Berlin, Springer-Verlag.

Cattell, R. B. (1966). "The Scree Test for the Number of Factors." Multivariate Behaviorial Research **1**: 245-276.

Cattell, R. B. and J. Jaspers (1967). "A General Plasmode for Factor Analytic Exercises and Research." Multivariate Behaviorial Research Monographs **67-3**: 1-212.

Champley, S. and S. Doledec (1997). "How to Separate Long-term Trends from Periodic Variation in Water Quality Monitoring." Water Research **31**(11): 2849-2875.

Chen, J. H. and J. L. Liu (1999). "Mixture principal component analysis models for process monitoring." Industrial & Engineering Chemistry Research **38**(4): 1478-1488.

Christensen, E. R., W. Phoomiphakdeephan, et al. (1997). "Water quality in Milwaukee, Wisconsin versus intake crib location." Journal of Environmental Engineering-Asce **123**(5): 492-498.

Crisp, N. H. (1989). Regional Surface Water Quality Characteristics of Nebraska. International Symposium on the Design of Water Quality Information Systems, Colorado State University, Colorado State University.

Dietz, E. J. and T. J. Killeen (1981). "A Nonparametric Multivariate Test for Monotone Trend with Pharmaceutical Applications." Journal of the American Statistical Association **76**(373): 169-174.

Dixon, W., G. K. Smyth, et al. (1999). "Optimized Selection of River Sampling Sites." Water Research **33**(4): 971-978.

Dunteman, G. H. (1989). "Principal Components Analysis." Sage University Papers **07-069**.

EPA, U.S. (1998). Guidance for Data Quality Assessment – Practical Methods for Data Analysis – EPA QA/G-9. Washington, D.C., U.S. EPA Office of Research and Development: 169.

Everitt, B. S. and G. Dunn (1991). Applied Multivariate Data Analysis. London, Edward Arnold, a division of Hodder and Stoughton.

Eynon, B. P. (1988). "Statistical Analysis of Precipitation Chemistry Measurements over the Eastern United States. Part II: Kriging Analysis of Regional Patterns and Trends." Journal of Applied Meteorology **27**(12): 1334-1343.

Flury, B. (1988). Common Principal Components and Related Multivariate Models. New York, John Wiley & Sons.

Flury, B. and H. Riedwyl (1988). Multivariate Statistics: A Practical Approach. New York.

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Boston, Addision-Wesley.

Halliday, D. and R. Resnick (1981). Fundamentals of Physics. New York, John Wiley & Sons.

Hannah, D. M., B. P. G. Smith, et al. (2000). "An approach to hydrograph classification." Hydrological Processes **14**(2): 317-338.

Harmancioglu, N. B. and N. Alpaslan (1992). "Water Quality Monitoring Network Design: A Problem of Multi-Objective Decision Making." Water Resources Bulletin **28**(1): 179-192.

Harmancioglu, N. B., O. Fistikoglu, et al. (1999). Water Quality Monitoring Network Design. Dordrecht, Kluwer Academic Publishers.

Hartigan, J. (1975). Clustering Algorithms. New York, John Wiley.

Harwell, G. (2001). Water quality characteristics, temporal trends, and influencing factors for selected streams in the Great Smoky Mountains National Park. Civil and Environmental Engineering Department. Knoxville, TN, University of Tennessee**:** 279.

Helsel, D. R. and R. M. Hirsch (1992). Statistical Methods in Water Resources. Amsterdam, Elsevier.

Hillis, W. D. (1990). "Co-evolving Parasites Improve Simulated Evolution as an Optimization Procedure." Physica D **42**: 228-234.

Hillis, W. D. (1992). Co-evolving Parasites Improve Simulated Evolution as an Optimization Procedure. Artificial Life II. C. G. Langton, C. Taylor, J. D. Farmer and S. Rasmussen. New York, Addision-Wesley.

Hintze, J. (2001). NCSS Help System. Kaysville, Utah, NCSS (Number Cruncher Statistical Systems).

Hirsch, R. M. (1988). "Statistical Methods and Sampling Design for Estimating Step Trends in Surface Water Quality." Water Resources Bulletin **24**(3): 493-503.

Hirsch, R. M., R. B. Alexander, et al. (1991). "Selection of Methods for the Detection and Estimation of Trends in Water Quality." Water Resources Research **27**(5): 803-813.

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., J. R. Slack, et al. (1982). "Techniques of Trend Analysis for Monthly Water Quality Data." Water Resources Research **18**(1): 107-121.

Holland, J. H. (1995). Hidden Order: How Adaptation Builds Complexity. Cambridge, Massachusetts, Perseus Books.

Holland, J. H. (2001). Adaptation in Natural and Artificial Systems. Cambridge, Massachusetts, MIT Press.

Hotelling, H. (1933). "Analysis of a Complex of Statistical Variables into Principal Components." Journal of Educational Psychology **24**: 417-441, 498-520.

Huang, W. C. and F. T. Yang (1998). "Streamflow estimation using Kriging." Water Resources Research **34**(6): 1599-1608.

Hughes, J. P. and D. P. Lettenmaier (1981). "Data Requirements for Kriging - Estimation and Network Design." Water Resources Research **17**(6): 1641-1650.

Husain, T. (1989). "Hydrologic Uncertainty Measure and Network Design." Water Resources Bulletin **25**(3): 527-534.

Isaaks, E. H. and R. H. Srivastava (1989). An Introduction to Applied Geostatistics. New York, Oxford University Press.

Jackson, J. E. (1991). A User's Guide to Principal Components Analysis. New York, John Wiley & Sons.

Jager, H. I., M. J. Sale, et al. (1990). "Cokriging to Assess Regional Stream Quality in the Southern Blue Ridge Province." Water Resources Research **26**(7): 1401-1412.

Jobson, J. D. (1992). Applied Multivariate Data Analysis Volume II: Categorical and Multivariate Methods. New York, Springer-Verlag.

Johnson, D. E. (1998). Applied Multivariate Methods for Data Analysts. Pacific Grove, CA, Duxbury Press.

Jolliffe, I. T. (1972). "Discarding Variables in a Principal Component Analysis, I: Artificial Data." Applied Statistics **21**: 160-173.

Jolliffe, I. T. (2000). Principal Component Analysis. New York, Springer-Verlag.

Kaiser, H. F. (1960). "The Application of Electronic Computers to Factor Analysis." Educational and Psychological Measurement **20**: 141-151.

Kaufman, L. and P. J. Rousseeuw (1990). Finding Groups in Data. New York, John Wiley.

Kendall, M. G. (1975). Rank Correlation Methods Fourth Edition. Charles Griffin, London.

King, P. B., R. B. Neuman, et al. (1968). Geology of the Great Smoky Mountains National Park, Tennessee and North Carolina. Geological Survey Professional Paper 587, U. S. Government Printing Office**:** 23.

Kirkpatrick, S., C. D. Gelatt, et al. (1983). "Optimization by Simulated Annealing." Science **220**(4598): 671-680.

Knuth, D. E. (1973). The Art of Computer Programming. New York, Addison-Wesley.

Koontz, W. L. G. and K. Fukunaga (1972a). "A Nonparametric Valley-Seeking Technique for Cluster Analysis." IEEE Transactions on Computers **C-21**: 171-178.

Koontz, W. L. G. and K. Fukunaga (1972b). "Asymptotic Analysis of a Nonparametric Clustering Technique." IEEE Transactions on Computers **C-21**: 967-674.

Laarhoven, P. J. M. V. and E. Aarts (1987). Simulated Annealing: Theory and Applications. Boston, Kluwer Academic Publishers.

LaChance, M., B. Bobée, et al. (1989). Methodology for the Planning and Operation of a Water Quality Network with Temporal and Spatial Objectives: Application to Acid Lakes in Québec. International Symposium on the Design of Water Quality Information Systems, Colorado State University, Colorado State University.

Langbein, W. B. (1979). "Overview of Conference on Hydrologic Data Networks." Water Resources Research **15**(6): 1867-1871.

Lettenmaier, D. P. (1976). "Detection of Trends in Water Quality Data from Records with Dependent Observations." Water Resources Research **12**(5): 1037-1046.

Lettenmaier, D. P. (1978). "Design Considerations for Ambient Stream Quality Monitoring." Water Resources Bulletin **14**(4): 884-902.

Lettenmaier, D. P. (1979). "Dimensionality Problems in Water Quality Network Design." Water Resources Research **15**(6): 1692-1700.

Lettenmaier, D. P. (1988). "Multivariate Nonparametric Tests for Trend in Water Quality." Water Resources Bulletin **24**(3): 505-512.

Lettenmaier, D. P., D. E. Anderson, et al. (1984). "Consolidation of a Stream Quality Monitoring Network." Water Resources Bulletin **20**(4): 473-481.

Lettenmaier, D. P., L. L. Conquest, et al. (1982). Routine Streams and Rivers Quality Trend Monitoring Review. Seattle, WA, University of Washington, Department of Civil Engineering, Charles W. Harris Hydraulics Laboratory**:** 233.

Liebetrau, A. M. (1979). "Water-Quality Sampling - Some Statistical Considerations." Water Resources Research **15**(6): 1717-1725.

Mackenzie, M. C., R. N. Palmer, et al. (1987). "Analysis of Statistical Monitoring Network Design." Journal of Water Resources Planning and Management-Asce **113**(5): 599-615.

Mackenzie, M. D. (1993). The Vegetation of the Great Smoky Mountains National Park: Past, Present, and Future. Knoxville, University of Tennessee**:** 154.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

Mann, H. B. (1945). “Non-parametric test against trend.” Econometrica **13**: 245-259.

Mather, P. (1976). Computational Methods of Multivariate Analysis in Physical Geography. New York, John Wiley & Sons.

McHenry, C. (1978). "Multivariate Subset Selection." Journal of the Royal Statistical Society, Series C **27**(23): 291-296.

McLeod, A. I., K. W. Hipel, et al. (1983). "Trend Assessment of Water Quality Time-Series." Water Resources Bulletin **19**(4): 537-547.

McNeil, J. H., A. G. McNeil, et al. (1989). Development of a Water Quality Monitoring System in Queensland. International Symposium on the Design of Water Quality Information Systems, Colorado State University, Colorado State University.

Mendenhall, William and Robert J. Beaver (1991). Introduction to Probability and Statistics, Eighth Edition. Boston, PWS-Kent.

Messer, J. J. (1989). Why Monitor? International Symposium on the Design of Water Quality Information Systems, Colorado State University, Colorado State University.

Metropolis, N., A. Rosenbluth, B. Rosenbluth, A. Teller, and E. Teller (1953). “Equation of state calculation by fast computing machines.” Journal of Chemical Physics **21**: 1087-1092

Michalewicz, Z. and D. B. Fogel (2002). How to Solve It: Modern Heuristics. New York, Springer-Verlag.

Mitchell, M. (2001). An Introduction to Genetic Algorithms. Cambridge, Massachusetts, MIT Press.

Morin, G., J. P. Fortin, et al. (1979). "Use of Principal Component Analysis to Identify Homogeneous Precipitation Stations for Optimal Interpolation." Water Resources Research **15**(6): 1841-1850.

Moss, M. E. (1979). "Some basic considerations in the design of hydrologic data networks." Water Resources Research **15**(6): 1673-1676.

Moss, M. E., D. P. Lettenmaier, et al. (1978). "On the Design of Hydrologic Data Networks." Transactions-American Geophysical Union **59**(8): 772-775.

Mueller, D. K. (1989). Use of Boxplots and Trend Analyses to Evaluate Sampling Frequency at Water-Quality Monitoring Sites. International Symposium on the Design of Water Quality Information Systems, Colorado State University, Colorado State University.

Mueller, D. K., W. W. Lapham, et al. (2002). Changes in Water Quality over Time. Water Resources Impact. **4:** 22-25.

Palmer, R. N. and M. C. Mackenzie (1985). "Optimization of Water Quality Monitoring Networks." Journal of Water Resources Planning and Management **3**(4): 478-493.

Pardo-Iguzquiza, E. (1998). "Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing." Journal of Hydrology **210**(1-4): 206-220.

Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode." Annals of Mathematical Statistics **33**: 1065-1076.

Peck, R., L. Fisher, et al. (1989). "Approximate Confidence Intervals for the Number of Clusters." Journal of the American Statistical Association **84**(405): 184-191.

Piechota, T. C., J. A. Dracup, et al. (1997). "Western US streamflow and atmospheric circulation patterns during El Nino Southern Oscillation." Journal of Hydrology **201**(1-4): 249-271.

Press, W. H., S. A. Teukolsky, et al. (1992). Numerical Recipes in C: The Art of Scientific Computing. Cambridge, Cambridge University Press.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications, Second Edition. New York, John Wiley & Sons.

Rauch, H. W. and W. B. White (1970). "Lithologic Controls on the Development of Solution Porosity in Carbonate Aquifers." Water Resour Res **6**(4): 1175-1192.

Reeve, A. S., D. I. Siegel, et al. (1996). "Geochemical controls on peatland pore water from the Hudson Bay Lowland: A multivariate statistical approach." Journal of Hydrology **181**(1-4): 285-304.

Reinelt, L. E., R. R. Horner, et al. (1988). "Nonpoint Source Pollution Monitoring Program Design." Journal of Water Resources Planning and Management-Asce **114**(3): 335-352.

Riola, R. L. (1992). "Survival of the Fittest Bits." Scientific American **267**(1): 114-116.

Robinson, R. B. (2003). Personal communication.

Rodrigues, M. R. D. and A. J. B. Anjo (1993). On Simulating Thermodynamics. Applied Simulated Annealing. V. V. R. Vidal. Berlin, Springer-Verlag**:** 45-60.

Rosenblatt, M. (1956). "Remarks on Some Nonparametric Estimates of a Density Function." Annals of Mathematical Statistics **27**: 832-837.

Salman, S. R. and Y. H. Abu Ruka'h (1999). "Multivariate and principal component statistical analysis of contamination in urban and agricultural soils from north Jordan." Environmental Geology **38**(3): 265-270.

Sanders, T. G. and D. D. Adrian (1978). "Sampling Frequency for River Quality Monitoring." Water Resources Research **14**(4): 569-576.

Sanders, T. G., R. C. Ward, et al. (1983). Design of Networks for Monitoring Water Quality. Littleton, CO, Water Resources Publications.

SAS Institute, I. (1999). Version 8 Online Documentation, SAS Institute, Inc.

Sharp, W. E. (1970). "Stream Order as a Measure of Sample Source Uncertainty." Water Resources Research **6**(3): 919-&.

Sharp, W. E. (1971). "A Topologically Optimum Water Sampling Plan for Rivers and Streams." Water Resources Research **7**(6): 1641-&.

Sherwani, J. K. and D. H. Moreau (1975). Strategies for Water Quality Monitoring. Chapel, NC, University of North Carolina**:** 137.

Skaggs, R. L., L. W. Mays, et al. (2001). "Simulated Annealing with Memory and Directional Search for Ground Water Remediation Design." Journal of the American Water Resources Association **37**(4): 853-866.

Smeyers-Verbeke, J., J. C. Denhartog, et al. (1984). "The Use of Principal Components Analysis for the Investigation of an Organic Air Pollutants Data Set." Atmospheric Environment **18**(11): 2471-2478.

Smeyers-Verbeke, J., J. C. Denhartog, et al. (1984). "Clustering Applied to an Organic Air Pollutants Data Set." Analusis **12**(10): 486-489.

Smith, D. G. and G. B. McBride (1990). "New Zealand's National Water Quality Monitoring Network - Design and First Years Operation." Water Resources Bulletin **26**(5): 767-775.

Somerville, M. C. and E. G. Evans (1995). "Effect of Sampling Frequency on Trend Detection for Atmospheric Fine Mass." Atmospheric Environment **29**(18): 2429-2438.

Spath, H. (1985). Cluster Dissection and Analysis. New York, Halsted Press.

Symons, F. and A. Ringele (1976). "Study of Time-Related Variability within Genus Astarte (Bivalvia): Multivariate Approach." Journal of the International Association for Mathematical Geology **8**(2): 113-136.

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

TOPO! (1999). Interactive Maps on CD-ROM North Georgia, Great Smoky Mountains, and Atlanta. San Francisco, National Geographic/Wildflower Productions.

Tukey, J. W. (1977). Exploratory Data Analysis. Reading, MA, Addison-Wesley.

Urquhart, N. S., S. G. Paulsen, et al. (1998). "Monitoring for Policy-Relevant Regional Trends over Time." Ecological Applications **8**(2): 246-257.

Varekamp, J. C., G. B. Pasternack, et al. (2000). "Volcanic lake systematics II. Chemical constraints." Journal of Volcanology and Geothermal Research **97**(1-4): 161-179.

Venkatram, A. (1988). "On the Use of Kriging in the Spatial Analysis of Acid Precipitation Data." Atmospheric Environment **22**(9): 1963-1975.

Vidal, Réne V. V. (1993). Applied Simulated Annealing. Berlin, Springer-Verlag.

Ward, R. C. (1989). Water Quality Monitoring - A Systems Approach to Design. International Symposium on the Design of Water Quality Information Systems, Colorado State University, Colorado State University.

Ward, R. C. (1996). "Water quality monitoring: Where's the beef?" Water Resources Bulletin **32**(4): 673-680.

Ward, R. C. and J. C. Loftis (1986). "Establishing Statistical Design Criteria for Water Quality Monitoring Systems: Review and Synthesis." Water Resources Bulletin **22**(5): 759-767.

Weissberg, B. G. and W. A. Singers (1982). "Trace Elements and Provenance of Phosphate Rocks." New Zealand Journal of Science **25**(2): 149-154.

Whitfield, P. H. (1983). "Evaluation of Water Quality Sampling Locations on the Yukon River." Water Resources Bulletin **19**(1): 115-121.

Whitfield, P. H. (1988). "Goals and Data Collection Designs for Water Quality Monitoring." Water Resources Bulletin **24**(4): 775-780.

Wong, M. A. (1982). "A Hybrid Clustering Method for Identifying High-Density Clusters." Journal of the American Statistical Association **77**(380): 841-847.

Wotling, G., C. Bouvier, et al. (2000). "Regionalization of extreme precipitation distribution using the principal components of the topographical environment." Journal of Hydrology **233**(1-4): 86-101.

Yang, Y. J. and D. H. Burn (1994). "An Entropy Approach to Data Collection Network Design." Journal of Hydrology **157**(1-4): 307-324.

APPENDICES

APPENDIX A. Breakdown of Monitoring Network Costs

A. Summary of Total Program Costs

FIXED costs of the synpotic network	$70,410.00	
VARIABLE costs of the synoptic network	$69,165.00	
TOTAL SYNOPTIC NETWORK COSTS		$139,575.00
TOTAL COST of NOLAND DIVIDE		$60,025.00
TOTAL MONITORING NETWORK COSTS (This includes the synoptic network and the Noland Divide project)		$199,600.00

B. Breakdown of synoptic network costs

			SUBTOTAL	TOTAL
	NPS Costs			
1	Direct costs from NPS (70% of total NPS program cost)	$55,300.00		
2	Indirect costs @ 15%	$8,300.00		
	TOTAL COSTS TO NPS (1,2)			$63,600.00
	Cost share by UT			
3	Uncompensated time of RBR including FB	$19,300.00		
4	Uncompensated time of grad students	$4,500.00		
5	Misc (CEE instrument use, secretarial, CEE lab technician)	$1,500.00		
	SUBTOTAL DIRECT COST SHARING BY UT (3,4,5)		$25,300.00	
6	Indirect costs @ 45% of direct cost share	$11,385.00		
7	Cost sharing of indirect costs for NPS contract ((0.3*$55,300)	$16,590.00		
	SUBTOTAL (6,7)		$27,975.00	
	TOTAL COST SHARING BY UT (3,4,5,6,7)			$53,275.00

B. Continued

			SUBTOTAL	TOTAL
	Steve Moore, vehicle, and Trout Unlimited			
8	Steve Moore (NPS in-kind contribution for CY 2003)	$11,300.00		
9	Vehicle	$4,200.00		
10	Trout Unlimited (4*60 hrs*$30/hr)	$7,200.00		
	TOTAL (8,9,10)			$22,700.00
	TOTAL COSTS			**$139,575.00**

C. Separation of variable synoptic network costs for use in optimization

			SUBTOTAL	TOTAL
11	Technician	$22,050.00		
12	Grad student: 70% of [0.5($5,000 tuition + 1.5*$15000)]	$9,625.00		
13	Hourly	$3,500.00		
14	Supplies	$2,800.00		
15	Instrumental use cost share	$560.00		
16	Vehicle	$4,200.00		
	SUBTOTAL		$42,735.00	
17	overhead @ 45% of $42,735	$19,230.00		
18	Trout Unlimited	$7,200.00		
	SUBTOTAL		$26,430.00	
	TOTAL VARIABLE COSTS			**$69,165.00**

APPENDIX B. Water Quality Data Means

Table 46. Means of water quality variables for period from 1996-2001.

SiteID	pH	ANC	Conductivity	Chloride	Nitrate	Sulfate	Sodium	Potassium
1	6.18	32.75	15.11	15.13	25.00	45.91	35.60	13.10
3	6.46	75.97	16.03	15.65	15.00	32.50	50.05	9.39
4	6.23	58.82	13.42	13.73	14.65	28.74	44.44	10.78
13	6.61	110.58	17.46	16.78	5.65	34.26	42.57	12.87
14	6.58	101.47	17.49	16.74	6.82	33.76	42.26	13.62
20	6.55	81.65	15.17	15.66	8.24	32.69	39.16	13.30
23	6.56	97.68	16.60	18.90	7.16	31.21	38.05	11.35
24	6.7	147.14	22.57	18.42	3.94	42.93	55.90	13.37
30	6.45	64.76	20.89	21.14	24.41	63.63	35.38	9.24
34	6.41	69.31	15.00	19.04	10.86	33.48	37.35	12.80
43	5.82	13.30	15.26	18.02	24.91	62.97	26.31	6.56
45	5.63	5.81	19.36	18.80	29.30	91.10	24.89	5.80
46	5.74	5.75	16.79	17.06	28.35	75.25	28.40	4.79
47	5.89	19.24	11.53	16.70	21.26	39.40	25.25	9.65
49	6.29	58.11	16.86	17.02	17.03	50.86	36.00	8.05
50	6.25	43.07	14.90	17.06	18.65	48.46	35.59	9.52
52	6.42	56.00	15.82	18.94	15.31	46.42	38.68	8.83
66	6.2	41.50	19.53	17.53	32.19	70.51	25.84	7.59
71	6.14	33.12	15.21	17.74	31.20	45.91	29.59	9.39
73	6.21	33.64	20.18	18.41	33.11	77.54	25.79	8.87
74	6.38	66.76	24.97	21.08	34.07	90.43	27.64	6.24
103	5.49	0.96	19.68	19.15	55.59	65.09	30.77	12.43
104	5.65	7.32	17.76	17.98	50.49	56.65	30.60	13.28
106	5.93	17.72	18.47	16.42	42.43	57.65	27.63	16.41
107	5.99	19.03	16.33	16.93	35.64	51.72	29.91	12.83
114	6.24	36.39	17.00	19.78	38.65	45.32	35.76	9.92
115	6.23	38.47	16.63	18.24	33.15	49.84	34.99	8.60
127	6.34	47.77	11.07	18.85	15.30	20.43	37.83	11.19
137	5.84	11.09	14.63	16.76	31.81	51.33	31.77	9.48
138	5.55	2.91	13.82	17.81	35.99	36.84	29.59	10.20
142	6.45	62.31	10.91	18.34	6.87	15.90	40.92	12.69
143	6.36	48.49	10.51	16.35	7.04	21.58	37.75	11.26
144	6.39	53.42	11.20	18.40	7.32	19.61	38.49	12.00
147	6.59	86.20	14.50	16.88	6.69	21.47	51.80	14.72

Table 46. Continued.

SiteID	pH	ANC	Conductivity	Chloride	Nitrate	Sulfate	Sodium	Potassium
148	6.66	125.11	16.83	16.66	4.21	19.90	71.54	15.89
149	6.53	81.10	13.94	16.90	7.76	21.89	51.38	14.72
150	6.49	80.49	13.84	16.35	8.98	21.03	49.32	13.71
156	7.14	434.04	51.15	23.03	3.72	41.69	52.42	16.24
173	6.59	158.06	20.05	19.91	5.48	26.61	44.54	12.11
174	7.37	1103.47	106.23	25.80	11.55	48.31	57.16	16.36
184	6.22	33.55	12.94	21.04	24.53	26.46	39.79	11.48
186	6.37	49.06	13.24	18.41	16.70	27.64	40.81	11.45
190	6.23	34.37	11.95	16.42	13.77	30.42	35.01	9.36
191	6.05	18.78	10.28	17.44	16.45	25.65	28.34	8.91
192	6.07	22.48	12.84	14.89	21.47	40.61	27.62	7.10
193	6.34	52.48	11.75	18.38	5.80	24.52	36.04	10.41
194	6.34	49.72	11.96	16.96	8.63	28.66	36.86	10.15
200	6.1	27.88	12.77	21.60	26.71	25.59	37.51	12.70
209	6.15	32.86	10.09	16.97	5.44	24.56	34.27	9.91
210	6.35	55.06	14.13	16.43	13.55	39.05	32.72	10.86
213	6.03	19.59	11.22	16.58	19.74	29.82	28.34	9.09
214	6.23	37.57	11.35	17.84	13.27	22.63	33.39	10.29
215	6.36	55.60	13.46	18.96	11.60	34.83	38.67	9.72
221	6.14	25.90	10.57	17.96	23.99	18.27	33.63	10.61
233	6.02	24.41	23.35	18.91	36.41	107.27	24.71	4.98
234	5.86	21.91	14.98	18.77	53.21	28.25	32.91	6.57
237	4.91	-11.39	19.49	15.94	32.51	72.57	18.81	4.23
251	6.08	22.81	29.42	23.26	39.37	156.98	43.15	11.35
252	5.69	19.35	38.35	23.34	54.05	242.15	51.77	12.00
253	6.41	81.29	23.35	24.33	55.18	41.62	54.72	9.07
266	6.45	74.78	13.61	17.57	7.06	26.16	46.62	13.23
268	6.43	66.03	13.56	17.56	13.48	32.52	42.25	11.45
291	6.07	27.81	12.79	20.95	35.92	24.03	37.00	11.34
293	6.51	85.66	16.10	19.53	17.19	24.67	56.61	13.55
310	6.49	75.68	12.63	15.94	3.96	21.79	41.79	10.23
311	6.44	68.22	12.12	19.00	4.48	18.42	42.67	11.35
336	6.39	51.19	11.23	18.76	24.49	10.79	40.84	13.46
337	6.24	41.50	12.37	19.79	26.11	16.89	37.98	11.17
472	6.18	32.58	11.85	17.15	14.09	32.96	32.60	8.86

Table 46. Continued.

SiteID	pH	ANC	Conductivity	Chloride	Nitrate	Sulfate	Sodium	Potassium
473	6.06	36.03	20.64	19.00	33.34	80.44	23.31	5.66
474	6.32	52.37	13.06	17.40	11.39	33.23	34.77	11.12
475	6.31	36.47	11.65	19.27	13.99	28.57	32.41	11.16
479	6.51	68.89	11.91	16.68	3.25	17.72	42.04	11.63
480	6.52	82.99	13.54	15.69	1.77	21.35	46.49	12.45
481	6.51	90.02	17.75	16.14	0.83	50.73	44.89	13.60
482	6.49	90.98	13.65	16.42	2.51	20.83	46.79	11.54
483	6.54	87.79	14.50	15.84	1.26	27.83	49.61	14.30
484	6.41	56.39	10.81	16.78	6.24	17.05	38.13	10.73
485	6.48	72.45	11.10	15.95	2.42	15.20	39.78	11.02
488	6.37	46.23	11.64	18.98	9.68	22.84	40.19	10.19
489	7.39	985.63	97.25	27.20	10.85	46.35	54.86	15.95
492	6.16	30.27	17.10	17.76	45.26	41.93	35.77	9.23
493	6.45	71.15	12.52	17.12	6.74	17.91	46.85	12.47

APPENDIX C. Data Screening Results

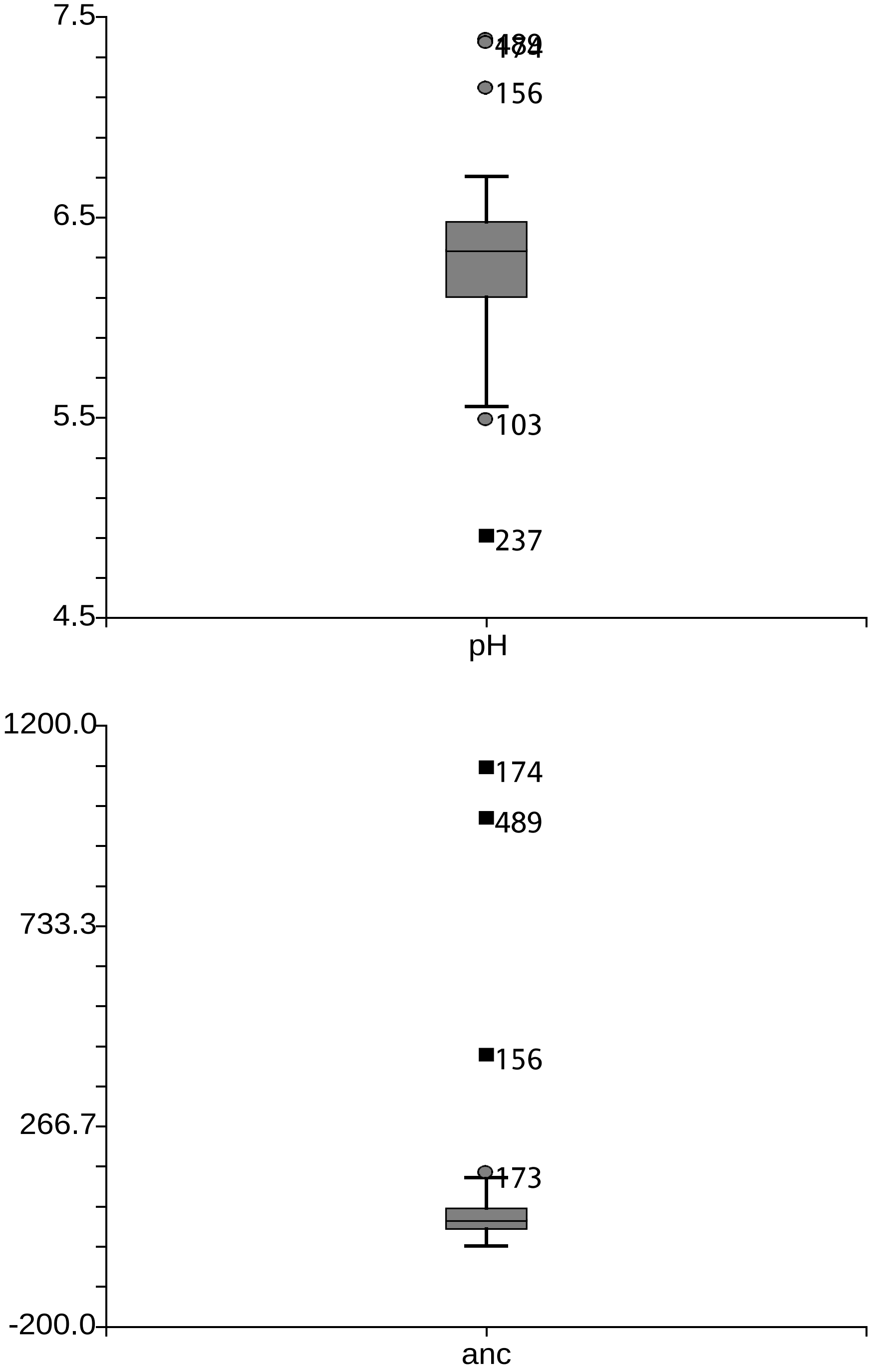

Figure 42. Boxplots of water quality variables pH and ANC.

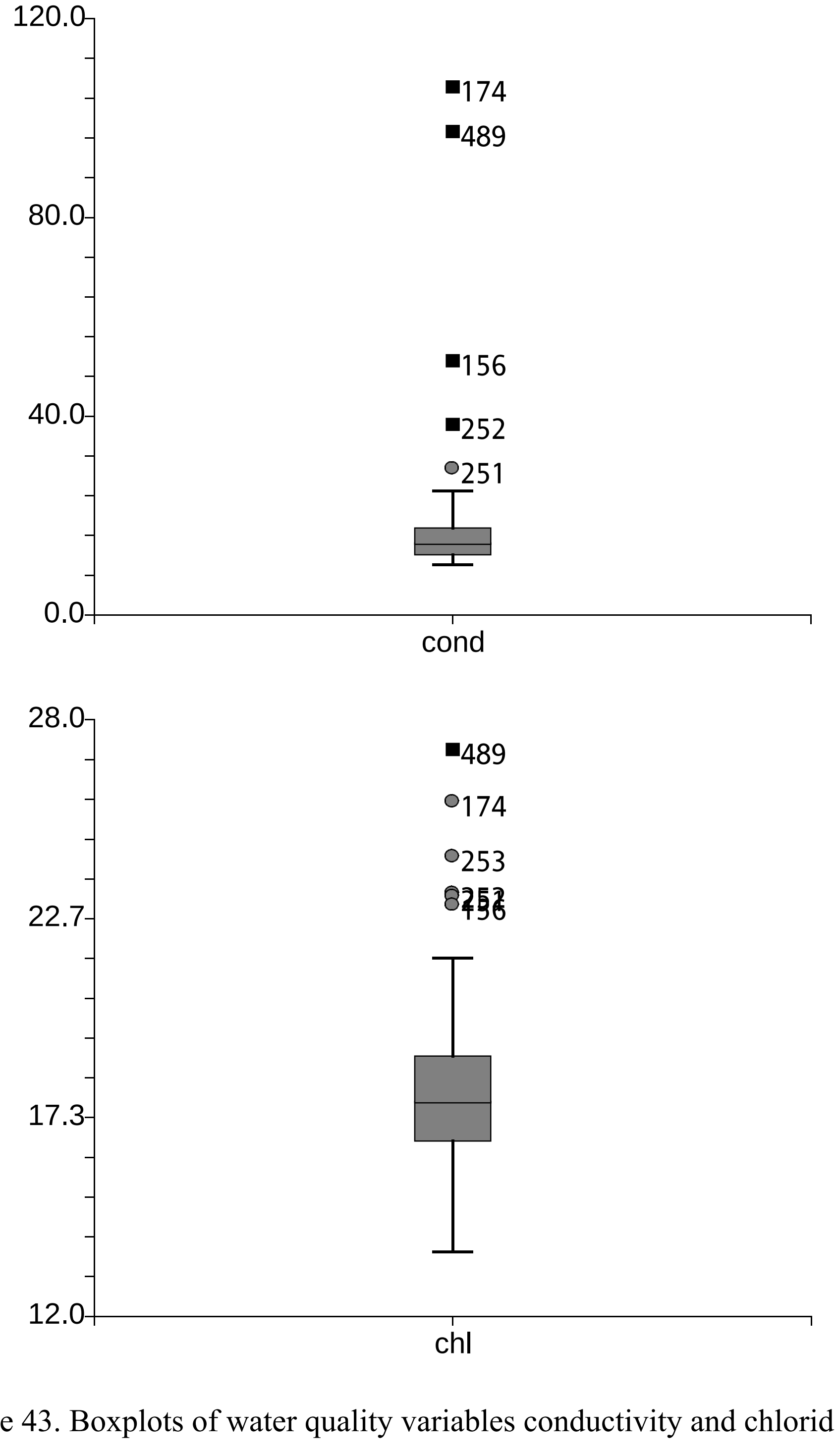

Figure 43. Boxplots of water quality variables conductivity and chloride.

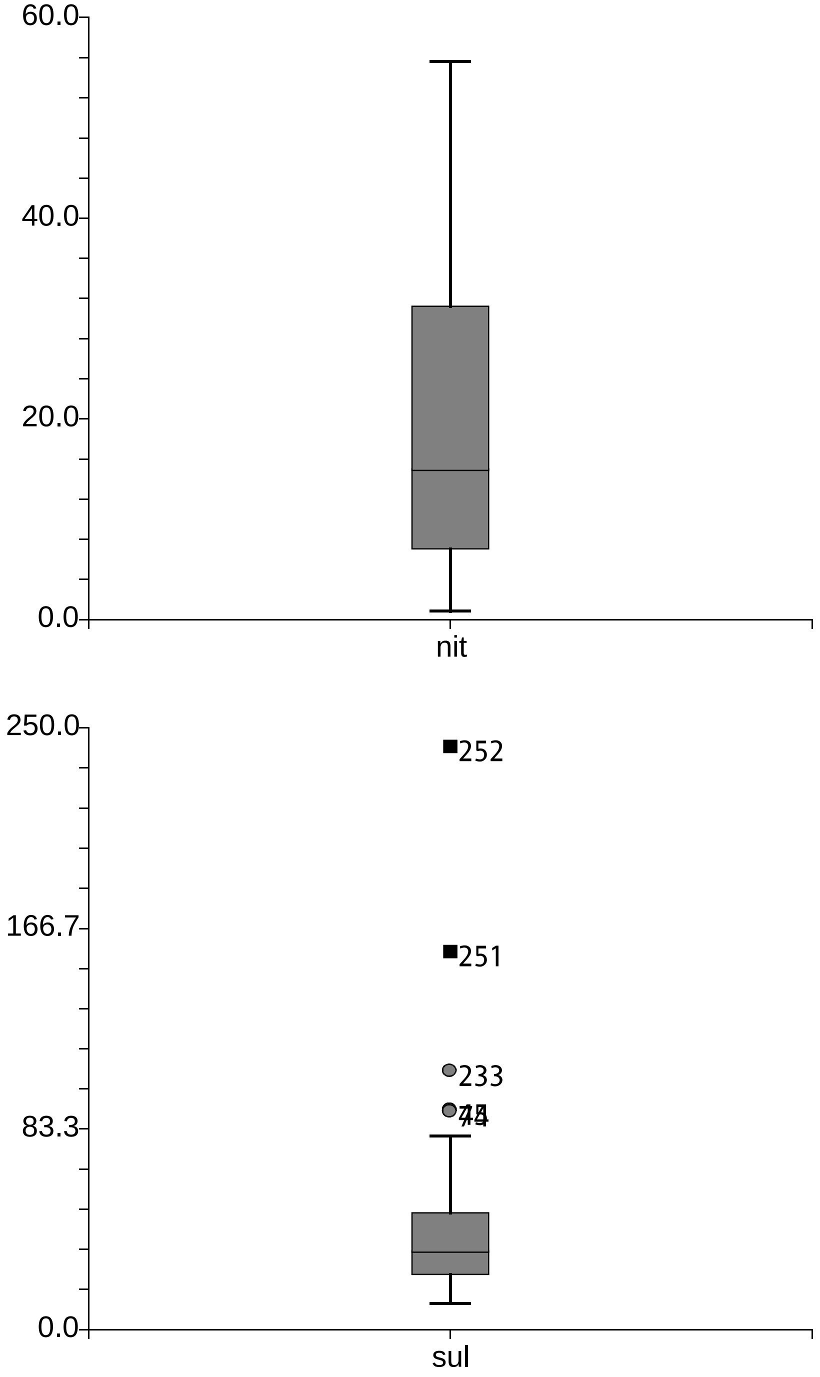

Figure 44. Boxplots of water quality variables nitrate and sulfate.

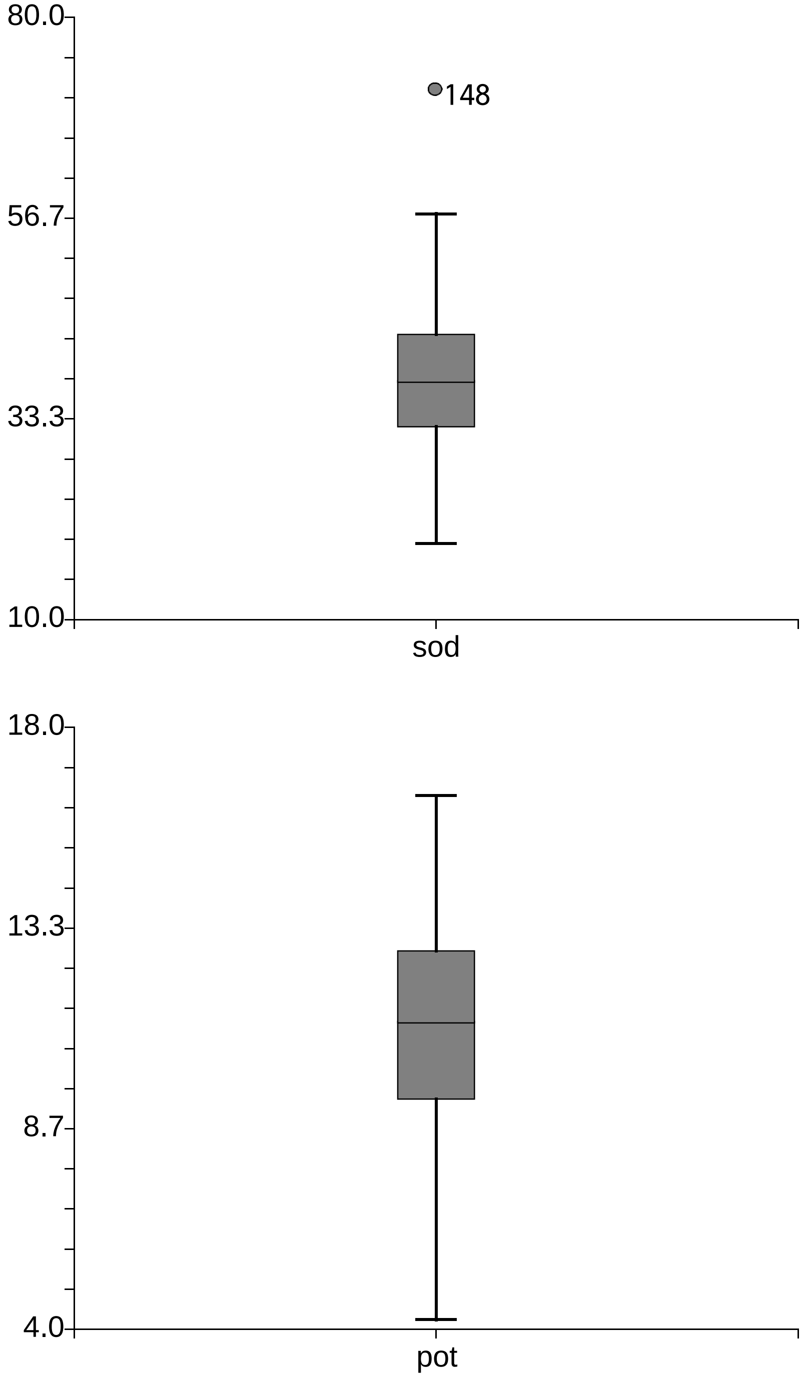

Figure 45. Boxplots of water quality variables sodium and potassium.

Table 47. Correlation matrix for geology variables.

	Thunderhead sandstone	Limestone	Cades Cove sandstone	Anakeesta Formation	Elkmont sandstone	Basement complex	Great Smoky Group
Thunderhead sandstone	**1.0**						
Limestone	**-0.228** *0.038*	**1.0**					
Cades Cove sandstone	**-0.271** *0.013*	**0.639** *0.000*	**1.0**				
Anakeesta Formation	**-0.346** *0.001*	**-0.129** *0.245*	**-0.155** *0.162*	**1.0**			
Elkmont Sandstone	**-0.354** *0.001*	**0.310** *0.004*	**0.414** *0.000*	**-0.221** *0.045*	**1.0**		
Basement complex	**0.144** *0.194*	**-0.053** *0.633*	**-0.066** *0.556*	**-0.130** *0.242*	**-0.106** *0.339*	**1.0**	
Great Smoky group	**-0.535** *0.000*	**0.011** *0.919*	**-0.003** *0.979*	**-0.351** *0.001*	**-0.115** *0.300*	**0.007** *0.947*	**1.0**

***correlations are in bold print**
*** p-values are shown in italics, p-values shown as 0.000 are actually <0.0001*

Table 48. Correlation matrix for morphology variables.

	Stream elevation	Mean basin elevation	Stream order	Max. channel length	Basin length	Basin area	Stream density	Mean basin slope	Channel slope	Basin width
Stream elevation	**1.0**									
Mean basin elevation	**0.858**	**1.0**								
	0.000									
Stream order	**-0.757**	**-0.526**	**1.0**							
	0.000	*0.000*								
Max. channel length	**-0.678**	**-0.499**	**0.752**	**1.0**						
	0.000	*0.000*	*0.000*							
Basin length	**-0.737**	**-0.540**	**0.834**	**0.956**	**1.0**					
	0.000	*0.000*	*0.000*	*0.000*						
Basin area	**-0.603**	**-0.454**	**0.737**	**0.940**	**0.915**	**1.0**				
	0.000	*0.000*	*0.000*	*0.000*	*0.000*					
Stream density	**-0.506**	**-0.458**	**0.360**	**0.286**	**0.317**	**0.245**	**1.0**			
	0.000	*0.000*	*0.001*	*0.009*	*0.004*	*0.026*				
Mean basin slope	**0.005**	**0.261**	**-0.063**	**-0.209**	**-0.200**	**-0.210**	**-0.037**	**1.0**		
	0.967	*0.017*	*0.573*	*0.058*	*0.069*	*0.057*	*0.739*			
Channel slope	**0.479**	**0.411**	**-0.660**	**-0.643**	**-0.686**	**-0.569**	**-0.204**	**0.339**	**1.0**	
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.064*	*0.002*		
Basin width	**-0.630**	**-0.452**	**0.844**	**0.865**	**0.878**	**0.939**	**0.249**	**-0.164**	**-0.646**	**1.0**
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.023*	*0.138*	*0.000*	

***correlations are in bold print**

*** p-values are shown in italics, p-values shown as 0.000 are actually <0.0001*

Table 49. Correlation matrix for vegetation variables.

	Spruce -fir	Northern hardwood	Cove hardwood	Mesic Oak	Mixed-mesic hardwood	Tulip poplar	Pine	Heath bald	Xeric Oak	Pine-oak
Spruce-fir	**1.0**									
Northern hardwood	**0.295**	**1.0**								
	0.007									
Cove hardwood	**-0.230**	**-0.098**	**1.0**							
	0.037	*0.380*								
Mesic Oak	**-0.510**	**-0.311**	**0.163**	**1.0**						
	0.000	*0.004*	*0.140*							
Mixed-mesic hardwood	**-0.417**	**-0.581**	**-0.405**	**0.294**	**1.0**					
	0.000	*0.000*	*0.000*	*0.007*						
Tulip poplar	**-0.213**	**-0.396**	**0.086**	**-0.262**	**0.047**	**1.0**				
	0.053	*0.000*	*0.438*	*0.017*	*0.675*					
Pine	**-0.276**	**-0.406**	**-0.481**	**-0.184**	**0.349**	**0.429**	**1.0**			
	0.012	*0.000*	*0.000*	*0.097*	*0.001*	*0.000*				
Heath bald	**-0.005**	**-0.027**	**0.371**	**-0.036**	**-0.320**	**0.205**	**-0.230**	**1.0**		
	0.961	*0.806*	*0.001*	*0.744*	*0.003*	*0.063*	*0.037*			
Xeric oak	**-0.343**	**-0.519**	**-0.568**	**0.015**	**0.553**	**0.405**	**0.774**	**-0.171**	**1.0**	
	0.002	*0.000*	*0.000*	*0.890*	*0.000*	*0.000*	*0.000*	*0.121*		
Pine-oak	**-0.272**	**-0.472**	**-0.582**	**0.098**	**0.705**	**0.179**	**0.544**	**-0.289**	**0.798**	**1.0**
	0.013	*0.000*	*0.000*	*0.377*	*0.000*	*0.105*	*0.000*	*0.008*	*0.000*	

***correlations are in bold print**

*** p-values are shown in italics, p-values shown as 0.000 are actually <0.0001*

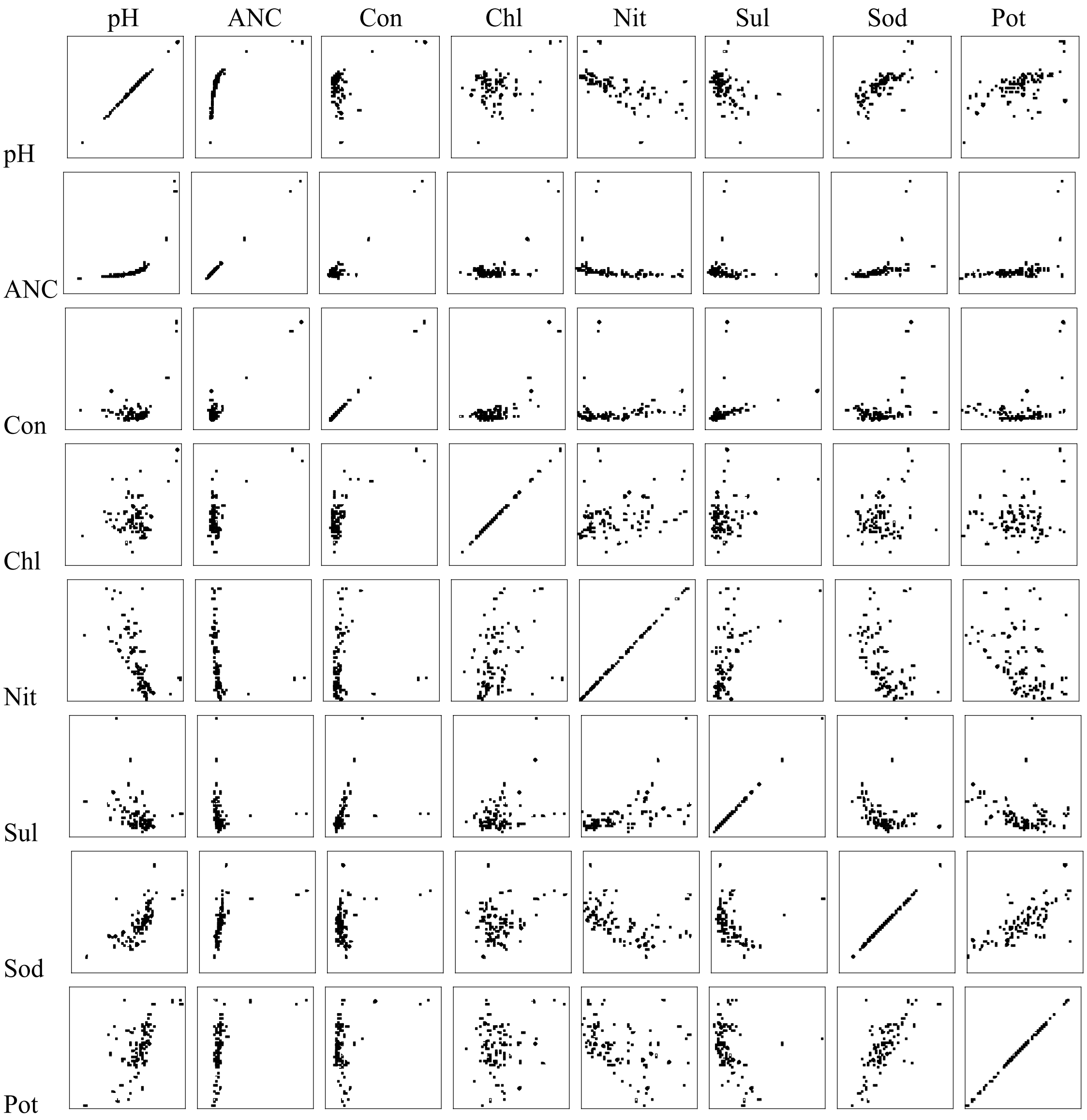

Figure 46. Scatterplot matrix of water quality variables.

APPENDIX D. Geology, Morphology, and Vegetation Data

Table 50. Geology characteristics of the sampling site watersheds.

SiteID	Thunderhead sandstone	Limestone	Cades Cove sandstone	Anakeesta formation	Elkmont sandstone	Basement complex	Great Smoky group
1	81.50	0.00	0.00	0.00	0.00	0.00	18.50
3	53.70	0.00	0.00	0.00	0.00	0.00	46.30
4	88.20	0.00	0.00	0.00	0.00	0.00	11.80
13	51.50	0.40	1.80	14.50	15.80	0.00	16.00
14	64.30	0.00	1.00	13.00	11.80	0.00	9.90
20	66.90	0.00	0.00	15.40	12.60	0.00	5.10
23	45.20	0.00	0.70	27.70	9.50	0.00	16.90
24	19.80	2.30	5.90	0.00	39.40	0.00	32.60
30	49.30	0.00	0.00	34.00	4.80	0.00	11.90
34	68.20	0.00	0.00	18.30	10.60	0.00	2.90
43	46.60	0.00	0.00	53.40	0.00	0.00	0.00
45	3.80	0.00	0.00	96.20	0.00	0.00	0.00
46	24.80	0.00	0.00	75.20	0.00	0.00	0.00
47	97.10	0.00	0.00	2.90	0.00	0.00	0.00
49	48.70	0.00	0.00	36.60	13.80	0.00	0.90
50	67.30	0.00	0.00	22.70	9.40	0.00	0.60
52	56.10	0.00	0.00	18.70	8.90	0.00	16.30
66	43.40	0.00	0.00	56.50	0.00	0.00	0.10
71	77.40	0.00	0.00	22.50	0.00	0.00	0.10
73	22.20	0.00	0.00	77.80	0.00	0.00	0.00
74	7.90	0.00	0.00	92.10	0.00	0.00	0.00
103	100.00	0.00	0.00	0.00	0.00	0.00	0.00
104	100.00	0.00	0.00	0.00	0.00	0.00	0.00
106	100.00	0.00	0.00	0.00	0.00	0.00	0.00
107	100.00	0.00	0.00	0.00	0.00	0.00	0.00
114	97.00	0.00	0.00	0.00	0.00	0.00	3.00

Table 50. Continued.

SiteID	Thunderhead sandstone	Limestone	Cades Cove sandstone	Anakeesta formation	Elkmont sandstone	Basement complex	Great Smoky group
115	100.00	0.00	0.00	0.00	0.00	0.00	0.00
127	98.30	0.00	0.00	0.00	0.00	0.00	1.70
137	99.70	0.00	0.00	0.00	0.00	0.00	0.30
138	100.00	0.00	0.00	0.00	0.00	0.00	0.00
142	53.80	0.00	0.00	0.00	0.00	6.30	39.90
143	98.30	0.00	0.00	0.00	0.00	0.00	1.70
144	74.60	0.00	0.00	0.00	0.00	3.30	22.10
147	60.30	0.00	0.00	0.00	0.00	1.30	38.40
148	14.30	0.00	0.00	0.00	0.00	0.40	85.30
149	63.20	0.00	0.00	0.00	0.00	1.30	35.50
150	67.90	0.00	0.00	0.00	0.00	0.70	31.40
156	0.20	6.90	33.20	0.00	34.40	0.00	25.30
173	0.00	0.90	17.50	0.00	67.80	0.00	13.80
174	0.70	21.00	13.50	0.00	38.50	0.00	26.30
184	0.00	0.00	0.00	0.00	93.40	0.00	6.60
186	3.30	0.00	0.00	0.00	92.10	0.00	4.60
190	58.40	0.00	0.00	39.10	0.00	0.00	2.50
191	0.00	0.00	0.00	6.40	0.00	0.00	93.60
192	0.00	0.00	0.00	67.50	0.00	0.00	32.50
193	43.30	0.00	0.00	40.00	0.00	0.00	16.70
194	46.90	0.00	0.00	35.70	0.70	0.00	16.70
200	18.90	0.00	0.00	0.00	71.80	0.00	9.30
209	76.00	0.00	0.00	24.00	0.00	0.00	0.00
210	81.70	0.00	0.00	15.50	0.00	0.00	2.80
213	83.10	0.00	0.00	15.30	0.00	0.00	1.60

Table 50. Continued.

SiteID	Thunderhead sandstone	Limestone	Cades Cove sandstone	Anakeesta formation	Elkmont sandstone	Basement complex	Great Smoky group
214	96.60	0.00	0.00	0.00	0.00	0.00	3.40
215	0.00	0.00	0.00	91.40	0.00	0.00	8.60
221	0.00	0.00	0.00	0.00	0.00	0.00	100.00
233	0.00	0.00	0.00	100.00	0.00	0.00	0.00
234	0.00	0.00	0.00	100.00	0.00	0.00	0.00
237	0.00	0.00	0.00	100.00	0.00	0.00	0.00
251	48.80	0.00	0.00	51.20	0.00	0.00	0.00
252	31.30	0.00	0.00	68.80	0.00	0.00	0.00
253	22.20	0.00	0.00	77.80	0.00	0.00	0.00
266	75.20	0.00	0.00	16.40	0.00	4.70	3.70
268	75.70	0.00	0.00	20.90	0.00	3.40	0.00
291	100.00	0.00	0.00	0.00	0.00	0.00	0.00
293	60.50	0.00	0.00	0.00	0.00	0.00	39.50
310	0.00	0.00	0.00	0.00	0.00	0.00	100.00
311	0.00	0.00	0.00	0.00	0.00	0.00	100.00
336	100.00	0.00	0.00	0.00	0.00	0.00	0.00
337	100.00	0.00	0.00	0.00	0.00	0.00	0.00
472	19.20	0.00	0.00	46.80	0.00	0.00	34.00
473	16.80	0.00	0.00	83.20	0.00	0.00	0.00
474	76.40	0.00	0.00	18.90	0.50	0.00	4.20
475	66.80	0.00	0.00	24.80	0.00	0.00	8.40
479	0.00	0.00	0.00	0.00	0.00	0.00	100.00
480	0.00	0.00	0.00	0.00	0.00	0.00	100.00
481	0.00	0.00	0.00	0.00	0.00	0.00	100.00
482	0.00	0.00	0.00	0.00	0.00	0.00	100.00

Table 50. Continued.

SiteID	Thunderhead sandstone	Limestone	Cades Cove sandstone	Anakeesta formation	Elkmont sandstone	Basement complex	Great Smoky group
483	0.00	0.00	0.00	0.00	0.00	0.00	100.00
484	0.00	0.00	0.00	0.00	0.00	0.00	100.00
485	0.00	0.00	0.00	0.00	0.00	0.00	100.00
488	0.00	0.00	2.30	0.00	90.40	0.00	7.30
489	0.40	11.70	13.60	0.00	54.60	0.00	19.70
492	100.00	0.00	0.00	0.00	0.00	0.00	0.00
493	72.80	0.00	0.00	0.00	0.00	1.90	25.30

Table 51. Morphology characteristics of the sampling site watersheds.

SiteID	Stream elevation	Mean basin elevation	Stream order	Maximum channel length	Basin length	Basin area	Stream density	Mean basin slope	Channel slope	Baisin width
1	671	1314	3	6.54	5.20	866	1.5	49.1	18.3	1.7
3	511	1032	4	8.62	6.40	2664	2.4	43.3	10.5	4.1
4	634	1241	2	5.23	4.80	378	2.2	52.5	19.7	0.8
13	335	979	5	42.76	19.30	27528	1.8	42.0	1.9	14.3
14	347	1041	4	41.62	18.20	15592	1.8	41.7	2.0	8.6
20	518	1116	4	28.85	15.80	13154	1.7	42.2	2.6	8.3
23	351	1003	5	19.11	11.50	7152	1.6	43.7	4.8	6.2
24	351	748	4	14.02	9.80	4670	1.8	39.9	6.2	4.8
30	436	1221	4	21.17	14.10	6267	1.4	49.2	5.8	4.4
34	597	1176	4	22.54	12.40	11067	1.6	42.9	3.4	8.9
43	713	1306	4	7.19	5.60	2564	1.4	61.7	11.2	4.6
45	975	1366	2	3.38	2.80	377	1.3	72.1	16.5	1.4
46	838	1291	3	5.52	4.30	867	1.4	65.3	13.1	2.0
47	732	1329	2	4.89	4.50	506	1.4	54.1	22.4	1.1
49	509	1186	4	11.03	8.60	4664	1.4	55.1	8.7	5.4
50	507	1245	5	13.93	11.90	9717	1.4	50.3	11.1	8.2
52	411	1145	5	19.24	15.30	11778	2.0	49.0	8.4	7.7
66	817	1460	4	13.60	9.20	3770	1.3	53.5	5.7	4.1
71	1036	1515	3	6.06	5.40	893	1.3	45.4	13.9	1.7
73	1024	1505	3	10.36	6.90	2361	1.4	54.3	5.3	3.4
74	1164	1523	2	7.91	5.30	1069	1.3	51.4	6.3	2.0
103	1320	1614	1	1.70	1.50	131	0.7	52.2	31.2	0.9
104	1280	1578	1	1.62	1.40	54	2.1	50.5	35.7	0.4
106	1170	1555	3	2.35	2.00	233	1.9	54.4	33.6	1.1

Table 51. Continued.

SiteID	Stream elevation	Mean basin elevation	Stream order	Maximum Channel length	Basin length	Basin area	Stream density	Mean basin slope	Channel slope	Baisin Width
107	927	1458	3	3.90	3.20	642	1.6	53.0	25.5	2.0
114	765	1201	3	3.79	3.00	597	1.6	56.6	18.4	2.0
115	1170	1332	1	0.63	0.60	25	0.6	55.0	49.7	0.4
127	908	1386	3	6.46	5.30	1115	1.8	45.2	10.7	2.1
137	838	1320	2	3.35	3.20	320	2.1	58.0	26.3	1.0
138	1058	1435	2	2.34	2.30	155	2.0	46.8	30.4	0.7
142	1006	1351	3	6.92	5.40	1149	1.3	41.6	6.9	2.1
143	1000	1422	2	6.43	4.90	847	1.7	43.5	11.7	1.7
144	911	1359	3	8.58	6.60	2183	1.4	43.0	6.8	3.3
147	750	1214	4	19.26	12.00	12742	1.6	41.4	3.2	10.7
148	754	1152	4	7.44	5.80	2188	1.6	40.8	10.6	3.8
149	777	1227	4	16.44	11.50	12156	1.6	41.4	3.9	10.5
150	799	1245	4	14.03	10.80	11157	1.5	41.9	4.5	10.3
156	338	745	5	31.27	19.20	15775	1.8	33.6	1.4	8.2
173	523	918	4	11.04	7.90	4290	1.4	36.6	7.1	5.4
174	523	773	4	15.35	11.50	5056	1.9	31.2	3.7	4.4
184	863	1236	1	2.60	2.40	168	1.2	44.7	24.3	0.7
186	599	1076	3	6.66	5.80	1038	1.5	44.6	12.5	1.8
190	646	1160	3	6.66	5.40	1133	1.0	49.8	14.3	2.1
191	975	1299	2	2.99	2.40	288	1.6	44.9	16.8	1.2
192	975	1273	2	2.49	2.00	245	1.4	53.8	17.0	1.2
193	585	1140	4	7.69	6.00	2358	1.3	50.4	11.2	3.9
194	518	1109	5	11.79	7.50	5539	1.4	45.0	6.7	7.4
200	1012	1299	2	2.09	1.80	183	1.8	44.8	23.1	1.0

Table 51. Continued.

SiteID	Stream elevation	Mean basin elevation	Stream order	Maximum Channel length	Basin length	Basin area	Stream density	Mean basin slope	Channel slope	Baisin Width
209	823	1047	1	2.89	2.10	145	1.9	41.4	12.7	0.7
210	835	1350	3	8.78	7.00	3536	1.4	47.9	11.8	5.1
213	1018	1406	3	5.98	4.30	735	1.9	43.5	10.8	1.7
214	1049	1332	2	3.78	3.10	342	1.4	44.7	14.0	1.1
215	1030	1249	2	2.14	1.80	169	1.6	45.1	12.6	0.9
221	1219	1449	2	2.99	2.70	299	1.1	42.9	15.8	1.1
233	1297	1596	2	5.14	3.70	594	1.3	51.2	8.3	1.6
234	1524	1621	1	0.52	0.40	16	0.0	38.4	20.7	0.4
237	1378	1636	2	4.04	2.60	429	1.0	50.0	9.2	1.7
251	1222	1442	2	2.71	2.30	211	1.8	45.9	13.5	0.9
252	1426	1550	1	0.87	0.50	29	0.6	44.6	20.3	0.5
253	1451	1562	1	0.73	0.40	22	0.0	42.5	17.9	0.5
266	614	1150	5	22.02	15.30	13480	1.7	49.9	2.9	8.8
268	666	1205	5	16.74	11.30	10574	1.6	51.3	3.8	9.4
291	1603	1716	1	0.93	0.90	34	0.5	34.6	24.0	0.4
293	840	1231	3	10.62	7.50	2735	1.4	43.9	6.8	3.6
310	683	1093	3	10.52	8.00	3088	1.6	46.6	5.1	3.9
311	657	1129	4	16.57	13.70	10062	1.5	45.2	4.5	7.3
336	1451	1559	1	3.29	2.70	160	1.5	25.3	4.3	0.6
337	1445	1598	2	3.45	3.10	306	1.4	31.3	9.3	1.0
472	671	1165	3	6.52	5.00	1071	1.5	51.5	11.1	2.1
473	1106	1506	3	9.65	6.20	2152	1.4	54.7	5.5	3.4
474	732	1267	4	15.04	9.20	7930	1.6	46.0	4.9	8.6
475	899	1305	4	9.07	4.80	2737	1.7	45.6	7.4	5.7
479	530	1089	4	23.86	17.40	11556	1.6	44.9	3.5	6.6
480	666	970	3	7.59	6.40	1310	1.6	44.1	9.5	2.1

Table 51. Continued.

SiteID	Stream elevation	Mean basin elevation	Stream order	Maximum Channel length	Basin length	Basin area	Stream density	Mean basin slope	Channel slope	Baisin Width
481	774	942	1	1.70	1.80	59	2.5	42.9	22.1	0.3
482	774	937	2	2.12	1.20	212	1.5	43.6	9.7	1.7
483	707	903	2	3.78	2.80	430	1.5	43.2	12.6	1.5
484	754	1211	4	13.01	11.00	5281	1.5	44.8	5.5	4.8
485	872	1184	2	7.06	5.30	752	1.4	43.0	7.7	1.4
488	546	1034	3	7.95	6.80	1095	1.4	38.6	11.7	1.6
489	521	839	5	15.79	11.60	9356	1.7	33.6	3.6	8.1
492	832	1260	1	3.28	2.70	255	0.9	52.3	19.5	0.9
493	866	1332	4	10.14	7.70	3756	1.6	43.4	6.2	4.9

Table 52. Vegetation characteristics of the sampling site watersheds.

SiteID	Spruce-fir	Northern hardwood	Cove hardwood	Mesic oak	Mixed-mesic hardwood	Tulip poplar	Pine	Heath bald	Xeric oak	Pine-oak
1	10.2	22.5	46.8	1.5	16.1	0.6	1.1	1.2	0.0	0.0
3	0.6	13.6	36.6	4.3	22.9	3.5	11.0	1.9	4.5	0.9
4	2.1	20.6	53.3	1.3	16.3	2.8	1.9	0.9	0.4	0.4
13	1.4	7.9	38.2	4.5	12.4	7.6	12.1	1.4	12.5	1.5
14	2.5	7.9	41.4	4.9	13.0	6.2	10.4	1.5	10.5	1.3
20	3.0	9.4	47.7	5.8	12.8	6.0	5.2	1.7	7.6	0.6
23	0.0	10.6	43.1	6.0	7.5	10.3	9.1	1.7	10.5	0.6
24	0.0	4.1	20.8	1.0	17.9	8.5	21.7	0.5	22.1	2.8
30	14.0	15.7	39.7	2.4	10.0	6.5	5.4	1.8	3.5	0.6
34	3.6	11.1	52.4	6.5	10.7	5.1	2.5	2.1	5.8	0.2
43	15.5	9.7	52.1	4.9	9.4	3.5	0.3	3.2	1.0	0.1
45	22.9	9.4	54.5	3.2	5.6	1.1	0.0	2.4	0.0	0.0
46	15.3	10.0	59.0	4.4	4.2	6.2	0.0	0.8	0.0	0.1
47	8.6	11.7	42.8	10.7	9.7	0.8	1.0	9.2	5.2	0.2
49	10.7	9.8	50.2	4.2	11.4	6.0	1.2	2.3	3.9	0.1
50	10.6	10.9	53.3	3.6	9.4	6.6	0.8	1.4	2.9	0.2
52	8.7	9.1	46.7	3.5	13.0	7.5	4.2	1.2	5.0	0.8
66	22.8	24.3	43.1	1.6	3.0	1.9	0.2	2.5	0.2	0.1
71	21.8	38.1	35.8	0.5	2.2	0.5	0.0	1.0	0.1	0.0
73	27.1	23.4	40.4	1.0	2.3	1.6	0.2	3.2	0.0	0.1
74	31.1	32.2	33.2	1.1	0.2	0.2	0.2	0.9	0.0	0.2
103	13.0	37.3	48.4	0.0	0.0	0.0	0.0	1.2	0.0	0.0
104	23.1	36.9	38.5	0.0	0.0	0.0	0.0	1.5	0.0	0.0
106	24.8	35.7	39.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
107	14.0	28.9	51.5	2.0	2.2	0.1	0.0	1.3	0.0	0.0
114	0.5	22.2	56.1	5.7	12.1	0.4	0.0	3.0	0.0	0.0

Table 52. Continued.

SiteID	Spruce-fir	Northern hardwood	Cove hardwood	Mesic oak	Mixed-mesic hardwood	Tulip poplar	Pine	Heath bald	Xeric oak	Pine-oak
115	0.0	23.3	60.0	13.3	3.3	0.0	0.0	0.0	0.0	0.0
127	2.0	7.0	68.4	14.3	6.8	0.1	0.1	0.3	0.9	0.0
137	2.5	22.6	60.4	1.5	9.9	1.5	0.0	1.0	0.5	0.0
138	3.1	28.0	57.5	8.8	1.0	0.0	0.0	1.6	0.0	0.0
142	0.4	22.3	57.2	18.4	1.2	0.0	0.0	0.4	0.2	0.0
143	2.9	11.8	60.3	20.2	3.5	0.4	0.0	0.9	0.0	0.0
144	1.4	16.3	57.2	19.1	4.3	0.3	0.3	0.5	0.5	0.0
147	1.0	9.4	45.1	9.9	27.3	0.5	4.4	0.2	1.2	0.2
148	0.1	0.7	31.4	29.7	33.3	0.2	3.9	0.0	0.6	0.1
149	1.1	9.8	46.0	10.0	26.2	0.5	4.2	0.2	1.1	0.1
150	1.2	10.6	46.7	10.6	24.7	0.4	3.7	0.2	1.1	0.2
156	0.0	3.7	12.8	2.2	14.9	3.2	28.7	0.0	26.5	2.4
173	0.0	5.6	25.1	5.9	20.9	4.2	20.1	0.0	15.4	1.3
174	0.0	5.1	12.8	1.5	14.4	3.0	23.7	0.0	21.3	2.0
184	0.0	44.6	37.6	0.0	15.5	1.9	0.5	0.0	0.0	0.0
186	0.0	18.8	35.7	4.7	19.6	5.0	10.4	0.2	4.4	0.9
190	0.0	23.0	47.7	8.6	2.0	5.0	0.0	4.4	9.0	0.0
191	0.0	35.3	53.3	7.8	0.0	0.0	0.0	3.3	0.0	0.0
192	0.0	28.1	42.4	24.5	0.0	0.0	0.0	5.0	0.0	0.0
193	0.0	19.4	49.0	8.5	4.3	5.8	1.3	3.4	7.8	0.1
194	0.0	13.7	52.7	7.7	5.7	8.4	1.7	2.2	7.5	0.1
200	0.0	36.6	52.9	4.8	2.2	0.9	0.0	0.4	0.0	0.0
209	0.0	0.0	73.2	4.5	6.1	5.0	4.5	2.8	3.9	0.0
210	9.5	15.1	61.4	3.8	1.7	5.4	0.0	2.7	0.2	0.0
213	7.4	26.8	54.9	2.4	1.0	2.5	0.0	4.7	0.1	0.0

Table 52. Continued.

SiteID	Spruce-fir	Northern hardwood	Cove hardwood	Mesic oak	Mixed-mesic hardwood	Tulip poplar	Pine	Heath bald	Xeric oak	Pine-oak
214	1.4	36.9	48.0	8.7	1.7	0.2	0.0	3.1	0.0	0.0
215	0.0	1.4	64.3	23.3	0.5	3.8	0.0	5.7	1.0	0.0
221	0.0	81.2	14.0	4.8	0.0	0.0	0.0	0.0	0.0	0.0
233	41.2	29.3	24.6	1.4	0.3	0.3	0.3	1.4	0.0	0.0
234	45.0	50.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
237	49.7	26.7	20.0	0.8	0.2	0.4	0.2	1.1	0.0	0.0
251	6.2	27.3	54.6	6.9	0.0	0.4	0.0	0.0	0.0	0.0
252	18.8	28.1	43.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
253	22.2	29.6	40.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
266	2.9	13.6	45.6	16.5	13.2	0.4	0.8	0.2	1.2	1.3
268	3.7	17.1	51.0	15.6	9.8	0.5	0.4	0.3	0.8	0.5
291	51.2	14.0	34.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
293	2.3	10.7	37.4	8.3	34.9	0.6	3.0	0.3	1.2	0.3
310	0.0	10.4	40.8	21.6	17.0	0.8	0.4	0.4	8.2	0.4
311	0.0	16.0	42.7	18.1	15.4	0.6	0.7	0.2	5.9	0.4
336	2.5	58.5	27.0	11.0	0.5	0.0	0.0	0.0	0.0	0.0
337	10.2	28.1	47.1	13.6	0.8	0.0	0.0	0.0	0.0	0.3
472	0.0	18.5	53.4	9.5	6.2	4.6	1.7	2.8	3.1	0.1
473	27.2	23.9	40.6	1.0	1.5	1.5	0.2	3.3	0.0	0.1
474	5.0	14.8	59.3	6.2	5.0	4.6	0.3	2.7	1.7	0.1
475	2.2	22.6	55.4	8.6	3.2	2.8	0.1	4.0	1.2	0.0
479	0.0	14.0	39.7	16.6	18.3	0.5	2.3	0.2	6.6	1.8
480	0.0	0.8	24.8	18.6	36.8	1.1	2.5	0.0	13.7	1.7
481	0.0	0.0	8.5	15.5	43.7	0.0	4.2	0.0	25.4	2.8
482	0.0	0.0	21.3	16.0	42.2	0.8	2.3	0.0	12.5	4.9

Table 52. Continued.

SiteID	Spruce-fir	Northern hardwood	Cove hardwood	Mesic oak	Mixed-mesic hardwood	Tulip poplar	Pine	Heath bald	Xeric oak	Pine-oak
483	0.0	0.0	15.7	11.5	49.2	1.5	2.1	0.0	16.2	3.8
484	0.0	24.2	49.3	16.9	7.1	0.2	0.0	0.1	2.0	0.0
485	0.0	13.0	51.8	24.3	9.9	0.4	0.0	0.0	0.4	0.1
488	0.0	12.4	36.8	7.3	19.4	3.5	14.0	0.0	5.4	0.7
489	0.0	6.1	19.8	3.7	17.2	3.5	21.3	0.0	17.5	1.6
492	0.6	20.0	57.1	7.0	9.2	0.6	0.0	5.4	0.0	0.0
493	1.4	11.6	57.5	17.3	10.5	0.3	0.3	0.4	0.7	0.0

APPENDIX E. Collocation Benefit Scores, Total Benefit Scores, and Sampling Site Cost Allocations

Table 53. Rank and Benefit Scores of the Clustering Results.

SiteID	Water Quality	Geology	Morphology	Vegetation	Collocation	Benefit Score
1	13	1	13	4	0	44
3	5	18	2	7	0	37
4	17	2	16	15	30	97
13	14	8.5	1	8	20	65.5
14	12	6	13	12	0	55
20	3	10	21	15	0	52
23	17	11.5	20	16	0	81.5
24	15	14	12	17	0	73
30	15	12	9	5	0	56
34	3	16	22	15	0	59
43	4	16	3	12	0	39
45	14	18	13	9	0	68
46	7	16.5	14	10	0	54.5
47	15	3	21	8	20	82
49	12	10.5	4	18	0	56.5
50	7	17	5	14	0	50
52	10	7	6	11	0	44
66	11	14	8	15	0	59
71	16	13	17	17	20	99
73	10	17.5	15	9	20	81.5
74	16	13	12	12	20	89
103	17	8.5	21	10	0	73.5
104	5	8.5	15	18	0	51.5
106	3	8.5	22	13	0	49.5
107	8	8.5	23	11	20	78.5
114	17	18	19	12	20	103
115	17	8.5	18	12	0	72.5

Table 53. Continued.

SiteID	Water Quality	Geology	Morphology	Vegetation	Collocation	Benefit Score
127	8	16.5	7	1	0	40.5
137	9	14	20	13	20	85
138	15	8.5	17	5	0	60.5
142	14	15	10	17	30	100
143	15	16.5	19	7	30	102.5
144	10	16	11	16	0	63
147	11	17	19	18	0	76
148	16	9.5	9	15	0	65.5
149	6	18	17	17	20	84
150	1	16	14	14	0	46
156	16	18	4	14	0	68
173	14	13	7	15	20	83
174	16	17	3	16	20	88
184	13	15	11	17	0	69
186	4	16	21	9	0	54
190	5	13	6	18	0	47
191	16	12.5	20	11	0	75.5
192	17	14.5	18	14	20	100.5
193	9	18	17	13	0	66
194	10	15	16	13	0	64
200	14	17	8	2	0	55
209	11	12	7	8	0	49
210	8	9	23	16	0	64
213	14	8	9	16	0	61
214	9	15	23	10	30	96
215	13	14	22	9	30	101
221	14	13.5	14	16	0	71.5
233	16	17	13	17	0	79

Table 53. Continued.

SiteID	Water Quality	Geology	Morphology	Vegetation	Collocation	Benefit Score
234	12	17	20	15	0	76
237	17	17	6	18	0	75
251	16.5	17	15	10	0	75
252	16.5	13.5	23	14	0	83.5
253	17	17.5	22	16	0	89.5
266	10	18	11	8	0	57
268	14	17	10	18	0	73
291	15	8.5	19	16	0	73.5
293	17	14	16	13	0	77
310	9	13.5	22	6	0	59.5
311	14	13.5	18	6	0	65.5
336	15	8.5	3	18	0	59.5
337	16	8.5	4	3	0	47.5
472	4	11	19	11	20	69
473	6	15.5	12	8	0	47.5
474	6	15	15	17	20	79
475	12	11	20	17	0	72
479	13	13.5	23	10	0	72.5
480	4	13.5	18	17	0	56.5
481	15	13.5	12	14	0	69.5
482	2	13.5	18	18	0	53.5
483	7	13.5	16	16	0	59.5
484	15	13.5	10	14	0	67.5
485	16	13.5	21	11	20	97.5
488	15	18	5	4	20	77
489	16	16	8	18	20	94
492	13	8.5	5	15	20	74.5
493	8	15	14	9	20	74

Table 54. Site access information.

SiteID	**Vehiclev(miles)* (normal)**	**Vehicle (CC) (miles)****	**All-terrain (miles)*****	**Hiking (miles)**	**Sampling time at site (10 min)**	**Dollars per year******
1				0.45	0.167	92.04
3	0.5				0.167	22.71
4	1.6				0.167	28.57
13	0.7				0.167	23.77
14	7.67				0.167	60.95
20	3.71				0.167	39.83
23	0.24				0.167	21.32
24	0.01				0.167	20.15
30	0				0.167	20.04
34	0.6				0.167	23.24
43	1.82			0.57	0.167	120.95
45				0.74	0.167	138.44
46				0.89	0.167	162.44
47				0.16	0.167	45.64
49	0.09				0.167	20.52
50	2.96				0.167	35.83
52	8.25				0.167	64.04
66	4.78				0.167	45.53
71				0.11	0.167	37.64
73	2.55			0.08	0.167	46.44
74	1.01				0.167	25.43
103				0.85	0.167	156.04
104				0.62	0.167	119.24

Table 54. Continued.

SiteID	Vehiclev(miles)* (normal)	Vehicle (CC) (miles)**	All-terrain (miles)***	Hiking (miles)	Sampling time at site (10 min)	Dollars per year****
106				1.32	0.167	231.24
107				1.58	0.167	272.84
114				0.24	0.167	58.44
115				1.62	0.167	279.24
127	0.85				0.167	24.57
137	0.74			0.18	0.167	52.79
138				0.93	0.167	168.84
142				0.09	0.167	34.44
143				1.04	0.167	186.44
144				0.18	0.167	48.84
147	0.17				0.167	20.95
148	1.35				0.167	27.24
149	0.74				0.167	23.99
150	1.35				0.167	27.24
156	24.2				0.167	149.11
173	0.49			0.09	0.167	37.05
174				0.09		34.44
184				0.64		122.44
186		0.5			0.167	32.04
190				0.72	0.167	135.24
191				2.15	0.167	364.04
192				0.14	0.167	42.44
193	0.82				0.167	24.41

Table 54. Continued.

SiteID	Vehiclev(miles)* (normal)	Vehicle (CC) (miles)**	All-terrain (miles)***	Hiking (miles)	Sampling time at site (10 min)	Dollars per year****
194	4.23				0.167	42.6
200				1.27	0.167	223.24
209				2.19	0.167	370.44
210				0.26	0.167	61.64
213				0.2	0.167	52.04
214				0.42	0.167	87.24
215				1.9	0.167	324.04
221	5.62			6.38	0.167	1070.81
233	1.67				0.167	28.95
234				0.29	0.167	66.44
237	0.86				0.167	24.63
251	4.41				0.167	43.56
252				0.11	0.167	37.64
253				0.14	0.167	42.44
266	3				0.167	36.04
268	7.24				0.167	58.65
291	2.61				0.167	33.96
293	0.25				0.167	21.37
310			1.54		0.167	125.64
311			0.06		0.167	24.15
336				0.53	0.167	104.84
337				0.31	0.167	69.64
472				0.33	0.167	72.84

Table 54. Continued.

SiteID	Vehiclev(miles)* (normal)	Vehicle (CC) (miles)**	All-terrain (miles)***	Hiking (miles)	Sampling time at site (10 min)	Dollars per year****
473	0.63				0.167	23.4
474	2.28			1.49	0.167	273.48
475				1.21	0.167	213.64
479			4.17		0.167	305.98
480			0.7		0.167	68.04
481			0.02		0.167	21.41
482			0.86		0.167	79.01
483			0.48		0.167	52.95
484			2.34		0.167	180.5
485				3.55	0.167	588.04
488	0.75			0.08	0.167	36.84
489				0.09	0.167	34.44
492				0.42	0.167	87.24
493	2.38				0.167	32.73

* Normal is specified to mean all roads except those around Cades Cove.

** CC is specified to indicate those distances around Cades Cove. The average vehicle speed was reduced to 10 mph around Cades Cove.

***All-terrain miles denotes those trails on Hazel Creek where the NPS transports personnel using an all-terrain vehicle.

****Dollars per year are based on one person collecting 4 samples per year at a cost of $30 per man-hour. Two people are assumed to be on each trip so this number would be double in the annealing algorithm.

APPENDIX F. Matlab Program Listings

SA1 Program listing

```
%SA Algorithm for Optimization of GSMNP Monitoring Network Version 1
%MAIN PROGRAM
%Kenneth R. Odom      October 15, 2002
%
datafiles;         %site analysis data matrices
sampfreq = 4;      %number of times samples per year
labc = 602.40;    %lab, admin, interp costs for one sample
T = 200;           %initial temperature (200)
k = 0.97;          %temperature decay factor (0.95)
of = [ ];          %objective function array
shuffle = [ ];    %random site generation array
t = 0;             %iteration time
nsites = 83;       %number of sites
numBP = 0;         %number removed by Boltzmann probability rule in local search
numOF = 0;         %number removed by objective function rule in local search
tnumBP = 0;        %number removed by Boltzmann probability rule in global search
tnumOF = 0;        %number removed by objective function rule in global search
nint = 14;         %number of intersections
time = [ ];        %time storage
oftemp = [ ];      %objective function storage
temp = [ ];        %temperature storage
ntemps = 0;        %number of temperature tries
numit = [ ];       %number of iterations
count = 0;         %counter
nlimit = 10 * nsites;    %limiting factor 1
nconfigs = 10 * nsites;  %limiting factor 2
nsucc = 0;         %number of successes
iter = 0;          %number of iterations
c1=fix(clock);     %Clock time at beginning of SA
discarded = [ ];   %matrix for sites discarded
retained = [ ];    %matrix for sites retained
ebcount = 0;       %counter
%
%    Begin Simulated Annealing
%
To = T;
OFcur = objfcn_v1(nsites, nint, labc, AA, BB, CC, DD, EE);
while ntemps < 150
   while count < nconfigs
      shuffle = randperm(nsites);
      siteselect = shuffle(1,1);
      x1 = find(AA(:,1) == siteselect);
      AAhold = AA;
      if AA(x1,3) == 1
         AA(x1,3) = 0;
         OFnew = objfcn_v1(nsites, nint, labc, AA, BB, CC, DD, EE);
      else
         AA(x1,3) = 1;
         OFnew = objfcn_v1(nsites, nint, labc, AA, BB, CC, DD, EE);
      end
      if OFnew > OFcur
```

SA1 Program listing continued

```
            OFcur = OFnew;
            numOF = numOF + 1;
         elseif  exp((OFnew-OFcur)/T) > rand(1)   % Boltzmann probability
            OFcur = OFnew;
            numBP = numBP + 1;
         else
            AA = AAhold;
         end
         t = t + 1;
         of(t) = OFcur;
         time(t) = t;
         temp(t) = T;
         oftemp(t) = OFcur;
         count = count + 1;
         iter = iter +1;
         nsucc = numOF + numBP;
         if nsucc > nlimit
            break
         end
      end
      tnumOF = tnumOF + numOF;
      tnumBP = tnumBP + numBP;
      numOF = 0;
      numBP = 0;
      count = 0;
      ntemps = ntemps + 1;
      T = T * k;
   end
   c2 = fix(clock);
   %
   %%%%%%%%%%%%%%%%%% OUTPUT %%%%%%%%%%%%%%%%%%%%%%%
   %
   q = 0;
   r = 0;
   for i = 2 : nsites+1
      if AA(i,3) == 1
         q = q +1;
         retained(q) = AA(i,2);
      else
         r = r + 1;
         discarded(r) = AA(i,2)
      end
   end
   rundate = date;
   ret = retained';
   dis = discarded';
   maxof = max(of);
   minof = min(of);
   elapsed_time = etime(c2,c1)/60;
   fprintf('Simulated Annealing Report for Sampling Network Optimization \n\n');
```

SA1 Program listing continued

```
fprintf('SA Run Date:  %s\n' ,rundate);
fprintf('SA Run Begin Time:  %i:%i:%i\n' ,c1(1,4),c1(1,5),c1(1,6));
fprintf('SA Run End Time:  %i:%i:%i\n' ,c2(1,4),c2(1,5),c2(1,6));
fprintf('SA Run Elapsed Time (min):  %f\n\n' ,elapsed_time);
fprintf('Total Number of Iterations:  %i\n' ,iter);
fprintf('Initial Temperature:  %f\n' ,To);
fprintf('Temperature Decay:  %f\n' ,k);
fprintf('Final Temperature:  %f\n' ,T);
fprintf('Number of Permutations accepted by Objective Function Rule:  %i\n' ,tnumOF);
fprintf('Number of Permutations accepted by Boltzmann Probability Rule:  %i\n' ,tnumBP);
fprintf('Minimum value of the Objective Function:  %f\n' ,minof);
fprintf('Maximum value of the Objective Function:  %f\n\n' ,maxof);
fprintf('Retained Sampling Site:  %i\n' ,ret);
fprintf('Discarded Sampling Site:  %i\n' ,dis);
%
%Plotting
%
subplot(2,1,1); plot(time, of);
xlabel('Iteration');
ylabel('Objective Function');
subplot(2,1,2); plot(temp, oftemp);
xlabel('Temperature Steps');
ylabel('Objective Function');
set(gca,'XDir','reverse')

function OF = objfcn_v1(nsites, nint, labc, AA, BB, CC, DD, EE)
%Function for calculation of the objective function
%Kenneth R. Odom
%Dissertation Project  Version 11.30.02
%
A = [ ];
j = 0;
sitelist = [ ];
q = 0;
for i = 2 : nsites+1
    if AA(i,3) == 1
      j=j+1;
      A(j) = AA(i,2);
   end
end
B = [ ];
[m n] = size(A);
B = (AA(:,2))';
for i = 1 : n
    z = find(BB(:,1) == A(i));
    B = cat(1,B,BB(z,:));
end
Bsum = sum(B(2:end,2:end));
sumsdist = 0;
for i = 1 : nsites
```

SA1 Program listing continued

```
    if Bsum(1,i) >= 1
        sumsdist = sumsdist + CC(i,2);
        q = q +1;
        sitelist(q) = CC(i,1);
    end
end
statbenefit = 0;
for i = 1 : n
    c = find(CC(:,1) == A(i));
    statbenefit = statbenefit  + CC(c,3);
end
D = [ ];
D = (DD(1,:));
for i = 1 : n
    z = find(DD(:,1) == A(i));
    D = cat(1,D,DD(z,:));
end
Dsum = sum(D(2:end, 2:end));
sumidist = 0;
for i = 1 : nint
    if Dsum(1,i) >= 1
        sumidist = sumidist + EE(i,2);
    end
end
labcost = n * labc;
OF = statbenefit + 1.2*(sumsdist + sumidist) - labcost - sumsdist - sumidist;
return;
```

SA2 Program listing

```
%SA Algorithm for Optimization of GSMNP Monitoring Network Version 2
%Matlab Program for calculating best sites from a predefined number
%Kenneth R. Odom      October 15, 2002
%
fid=fopen('bestnets.m','w');
for numnet = 10:10:70
%
datafiles;              %site analysis data matrices
sampfreq = 4;           %number of times samples per year
labc = 602.40;          %lab, admin, interp costs for one sample
T = 200;                %initial temperature (200)
k = 0.97;               %temperature decay factor (0.95)
of = [ ];               %objective function array
shuffle = [ ];          %random site generation array
t = 0;                  %iteration time
nsites = 83;            %number of sites
numBP = 0;              %number removed by Boltzmann probability rule in local search
numOF = 0;              %number removed by objective function rule in local search
tnumBP = 0;             %number removed by Boltzmann probability rule in global search
tnumOF = 0;             %number removed by objective function rule in global search
nint = 14;              %number of intersections
time = [ ];             %time storage
oftemp = [ ];           %objective function storage
temp = [ ];             %temperature storage
ntemps = 0;             %number of temperature tries
numit = [ ];            %number of iterations
count = 0;              %counter
nlimit = 10 * nsites;   %limiting factor 1
nconfigs = 10 * nsites; %limiting factor 2
nsucc = 0;              %number of sucesses counter
iter = 0;               %iteration counter
c1=fix(clock);          %Clock time at beginning of SA
discarded = [ ];        %matrix for sites discarded
retained = [ ];         %matrix for sites retained
ebcount = 0;            %counter
ofattemp = [ ];         %objective function array
bpattemp = [ ];         %Boltzmann probability array
temps = [ ];            %temperature array
bestsites = numnet;     %holder for best subset of sites
%
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Begin Simulated Annealing  %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
shuffle = randperm(nsites);  % randomly choose n bestsites
inset = shuffle;
inset(bestsites+1:nsites) = [ ];
outset = shuffle;
outset(1:bestsites) = [ ];
[m,n] = size(inset);
```

SA2 Program listing continued

```
[r,s] = size(outset);
for i = 1:n
   XX(1,i) = i;
   XX(2,i) = inset(i);
end
for i = 1:s
   YY(1,i) = i;
   YY(2,i) = outset(i);
end
XX = XX';
YY = YY';
To = T;
OFcur = objfcn_v2(bestsites,nsites, nint, labc, AA, BB, CC, DD, EE, XX);
while ntemps < 100
   while count < nconfigs
      shuffle1 = randperm(bestsites);
      shuffle2 = randperm(nsites-bestsites);
      siteselect1 = shuffle1(1,1);
      siteselect2 = shuffle2(1,1);
      x1 = find(XX(:,1) == siteselect1);
      y1 = find(YY(:,1) == siteselect2);
      hold1 = XX(x1,2);
      XX(x1,2) = YY(y1,2);
      YY(y1,2) = hold1;
      OFnew = objfcn_v2(bestsites,nsites, nint, labc, AA, BB, CC, DD, EE, XX);
      if OFnew > OFcur
         OFcur = OFnew;
         numOF = numOF + 1;
      elseif  exp((OFnew-OFcur)/T) > rand(1)   % Boltzmann probability
         OFcur = OFnew;
         numBP = numBP + 1;
      else
         hold1 = XX(x1,2);
         XX(x1,2) = YY(y1,2);
         YY(y1,2) = hold1;
      end
      t = t + 1;
      of(t) = OFcur;
      time(t) = t;
      temp(t) = T;
      oftemp(t) = OFcur;
      count = count + 1;
      iter = iter +1;
      nsucc = numOF + numBP;
      if nsucc > nlimit
         break
      end
   end
   tnumOF = tnumOF + numOF;
   tnumBP = tnumBP + numBP;
   ntemps = ntemps + 1;
```

SA2 Program listing continued

```
    ofattemp(ntemps) = numOF;
    bpattemp(ntemps) = numBP;
    temps(ntemps) = T;
    numOF = 0;
    numBP = 0;
    count = 0;
    T = T * k;
end
c2 = fix(clock);    %Clock time at end of SA
%
%%%%%%%%%%%%%%%%%%%% OUTPUT %%%%%%%%%%%%%%%%%%%%%%%%%%
%

for i = 1 : bestsites
    x1 = find(AA(:,1) == XX(i,2));
    retained(i) = AA(x1,2);
end
rundate = date;
ret = sort(retained);
%dis = discarded';
maxof = max(of);
minof = min(of);
elapsed_time = etime(c2,c1)/60;
fprintf(fid,'\n\nSimulated Annealing Report for Sampling Network Optimization \n\n');
fprintf(fid,'SA Run Date:  %s\n' ,rundate);
fprintf(fid,'SA Run Begin Time:  %i:%i:%i\n' ,c1(1,4),c1(1,5),c1(1,6));
fprintf(fid,'SA Run End Time:  %i:%i:%i\n' ,c2(1,4),c2(1,5),c2(1,6));
fprintf(fid,'SA Run Elapsed Time (min):  %f\n\n' ,elapsed_time);
fprintf(fid,'Total Number of Iterations:  %i\n' ,iter);
fprintf(fid,'Initial Temperature:  %f\n' ,To);
fprintf(fid,'Temperature Decay:  %f\n' ,k);
fprintf(fid,'Final Temperature:  %f\n' ,T);
fprintf(fid,'Number of Permutations accepted by Objective Function Rule:  %i\n' ,tnumOF);
fprintf(fid,'Number of Permutations accepted by Boltzmann Probability Rule:  %i\n' ,tnumBP);
fprintf(fid,'Minimum value of the Objective Function:  %f\n' ,minof);
fprintf(fid,'Maximum value of the Objective Function:  %f\n\n' ,maxof);
fprintf(fid,'Retained Sampling Sites:');
fprintf(fid,'%i, ' ,ret);
%
%Plotting: set plott=1 for plots and plot=2 to suppress
plott=1;
if plott == 1
    subplot(2,2,1); plot(time, of);
    xlabel('Iteration');
    ylabel('Total Benefit');
    subplot(2,2,2); plot(temp, oftemp);
    xlabel('Temperature');
    ylabel('Total Benefit');
    set(gca,'XDir','reverse');
    subplot(2,2,3); plot(temps, ofattemp);
    xlabel('Temperature');
```

SA2 Program listing continued

```
   ylabel('Objective Function Solutions');
   set(gca,'XDir','reverse');
   subplot(2,2,4); plot(temps, bpattemp);
   xlabel('Temperature');
   ylabel('Boltzmann Probability Soultions');
   set(gca,'XDir','reverse');
else
   fprintf(fid,'Plotting Suppressed');
end
end
fclose(fid);


function OF = objfcn_v2(bestsites,nsites, nint, labc, AA, BB, CC, DD, EE, XX)
%Function for calculation of the objective function
%Kenneth R. Odom
%Dissertation Project  Version 12.18.02
%
A = [ ];
j = 0;
sitelist = [ ];
q = 0;
for i = 1 : bestsites
   z = find(AA(:,1) == XX(i,2));
   A(i) = AA(z,2);
end
B = [ ];
[m n] = size(A);
B = (AA(:,2))';
for i = 1 : n
   z = find(BB(:,1) == A(i));
   B = cat(1,B,BB(z,:));
end
Bsum = sum(B(2:end,2:end));
sumsdist = 0;
for i = 1 : nsites
   if Bsum(1,i) >= 1
      sumsdist = sumsdist + CC(i,2);
      q = q +1;
      sitelist(q) = CC(i,1);
   end
end
[r s] = size(sitelist);
statbenefit = 0;
for i = 1 : n
   c = find(CC(:,1) == A(i));
   statbenefit = statbenefit  + 1*CC(c,3) ;
end
D = [ ];
D = (DD(1,:));
for i = 1 : n
```

SA2 Program listing continued

```
   z = find(DD(:,1) == A(i));
   D = cat(1,D,DD(z,:));
end
Dsum = sum(D(2:end, 2:end));
sumidist = 0;
for i = 1 : nint
   if Dsum(1,i) >= 1
      sumidist = sumidist + EE(i,2);
   end
end
labcost = n * labc;
OF = statbenefit + 1.2*(sumsdist + sumidist) - labcost - sumsdist - sumidist;
return;
```

SA3 Program listing

```
%SA Algorithm for Optimization of GSMNP Monitoring Network Version 3
%Batch mode of version 1
%Kenneth R. Odom      October 15, 2002
%
% *************** USER RUN PREFERENCE *******************
%
% NOTE: datafiles.m must be checked to ensure proper format
% enter batmode = 1 for batch mode
% enter batmode = 2 for single run mode
batmode = 2;
fid=fopen('sensanaly.m','a');
% ******************* PLOT OFF/ON **********************
plotoo = 1;  %1=off 2=on
% *************** BATCH FILE OPERATION *****************
nvars = 5;         %number of variables(water quality,geology,morphology,vegetation,collocation)
nsampyr = 4;       %number of samples per year
labc = 150.60;     %lab, admin, interp. costs for one site/yr at one sample/yr
nsites = 83;       %number of sites
countiter = 0;     %count number of iterations in batch file operation
datafiles;         %site analysis data matrices
%if batmode == 1
benfact = [1.2];
weights = [2.0] ;
[bx by] = size(benfact);
[wx wy] = size(weights);
for xx = 1:by
   totbfit = labc * nsampyr * nsites * benfact(xx);
   for yy = 1:wy
      for qq = 1:nvars
         scoreben = 0;
         datafiles; %site analysis data matrices
         JJ = CC;
         countiter = countiter + 1;
         CC(:,1) = JJ(:,1);
         CC(:,2) = JJ(:,2);
         if qq == 1
           CC(:,3)=weights(1,yy)*JJ(:,3)+JJ(:,4)+JJ(:,5)+JJ(:,6)+JJ(:,7);
         elseif qq == 2
           CC(:,3)=JJ(:,3)+weights(1,yy)*JJ(:,4)+JJ(:,5)+JJ(:,6)+JJ(:,7);
         elseif qq == 3
           CC(:,3)=JJ(:,3)+JJ(:,4)+weights(1,yy)*JJ(:,5)+JJ(:,6)+JJ(:,7);
         elseif qq == 4
           CC(:,3)=JJ(:,3)+JJ(:,4)+JJ(:,5)+weights(1,yy)*JJ(:,6)+JJ(:,7);
         elseif qq == 5
           CC(:,3)=JJ(:,3)+JJ(:,4)+JJ(:,5)+JJ(:,6)+weights(1,yy)*JJ(:,7);
         else
         end
         scoreben = sum(CC(:,3));
         CC(:,3) = (CC(:,3)/scoreben)*totbfit;

         %elseif batmode == 2
```

SA3 Program listing continued

```
    %proceed
    %else
    %proceed
    %end
%
% **************** Begin Simulated Annealing *****************
%
T = 200;           %initial temperature (200)
k = 0.97;          %temperature decay factor (0.95)
of = [ ];          %objective function array
shuffle = [ ];     %random site generation array
t = 0;             %iteration time
numBP = 0;         %number removed by Boltzmann probability rule in local search
numOF = 0;         %number removed by objective function rule in local search
tnumBP = 0;        %number removed by Boltzmann probability rule in global search
tnumOF = 0;        %number removed by objective function rule in global search
nint = 14;         %number of intersections
time = [ ];        %time storage
oftemp = [ ];      %objective function storage
temp = [ ];        %temperature storage
ntemps = 0;        %number of temperature tries
numit = [ ];       %number of iterations
count = 0;         %counter
nlimit = 10 * nsites;    %limiting factor 1
nconfigs = 10 * nsites;  %limiting factor 2
nsucc = 0;         %number of successes
iter = 0;          %number of iterations
c1=fix(clock);     %Clock time at beginning of SA
discarded = [ ];   %matrix for sites discarded
retained = [ ];    %matrix for sites retained

To = T;
OFcur = objfcn_v3(nsites, nint, labc, nsampyr, AA, BB, CC, DD, EE);
while ntemps < 150
   while count < nconfigs
      shuffle = randperm(nsites);
      siteselect = shuffle(1,1);
      x1 = find(AA(:,1) == siteselect);
      AAhold = AA;
      if AA(x1,3) == 1
         AA(x1,3) = 0;
         OFnew = objfcn_v3(nsites, nint, labc, nsampyr, AA, BB, CC, DD, EE);
      else
         AA(x1,3) = 1;
         OFnew = objfcn_v3(nsites, nint, labc, nsampyr, AA, BB, CC, DD, EE);
      end
      if OFnew > OFcur
         OFcur = OFnew;
         numOF = numOF + 1;
      elseif  exp((OFnew-OFcur)/T) > rand(1)   % Boltzmann probability
         OFcur = OFnew;
```

SA3 Program listing continued

```
         numBP = numBP + 1;
      else
         AA = AAhold;
      end
      t = t + 1;
      of(t) = OFcur;
      time(t) = t;
      temp(t) = T;
      oftemp(t) = OFcur;
      count = count + 1;
      iter = iter +1;
      nsucc = numOF + numBP;
      if nsucc > nlimit
         break
      end
   end
   tnumOF = tnumOF + numOF;
   tnumBP = tnumBP + numBP;
   numOF = 0;
   numBP = 0;
   count = 0;
   ntemps = ntemps + 1;
   T = T * k;
end
c2 = fix(clock);
%
%%%%%%%%%%%%%%%%%% OUTPUT %%%%%%%%%%%%%%%%%%%%%%%
%
q = 0;
r = 0;
for i = 2 : nsites+1
   if AA(i,3) == 1
      q = q +1;
      retained(q) = AA(i,2);
   else
      r = r + 1;
      discarded(r) = AA(i,2);
   end
end
rundate = date;
maxof = max(of);
minof = min(of);
elapsed_time = etime(c2,c1)/60;
fprintf(fid,'Simulated Annealing Report for Sampling Network Optimization \n\n');
fprintf(fid,'SA Run Date:  %s\n' ,rundate);
fprintf(fid,'SA Run Begin Time:  %i:%i:%i\n' ,c1(1,4),c1(1,5),c1(1,6));
fprintf(fid,'SA Run End Time:  %i:%i:%i\n' ,c2(1,4),c2(1,5),c2(1,6));
fprintf(fid,'SA Run Elapsed Time (min):  %f\n\n' ,elapsed_time);
fprintf(fid,'Total Number of Iterations:  %i\n' ,iter);
fprintf(fid,'Initial Temperature:  %f\n' ,To);
fprintf(fid,'Temperature Decay:  %f\n' ,k);
```

SA3 Program listing continued

```
fprintf(fid,'Final Temperature:  %f\n' ,T);
fprintf(fid,'Number of Permutations accepted by Objective Function Rule:  %i\n' ,tnumOF);
fprintf(fid,'Number of Permutations accepted by Boltzmann Probability Rule:  %i\n' ,tnumBP);
fprintf(fid,'Minimum value of the Objective Function:  %f\n' ,minof);
fprintf(fid,'Maximum value of the Objective Function:  %f\n\n' ,maxof);
fprintf(fid,'Benefit Factor: %f\n', benfact(xx));
fprintf(fid,'Weight Factor: %f\t on Variable: %i\n', weights(yy),qq);
fprintf(fid,'Retained Sampling Sites:');
fprintf(fid,'%i\n' ,retained);
fprintf(fid,'Discarded Sampling Sites:');
fprintf(fid,'%i\n' ,discarded);
%
%Plotting
%
if plotoo == 2
   subplot(2,1,1); plot(time, of);
   xlabel('Iteration');
   ylabel('Objective Function');
   subplot(2,1,2); plot(temp, oftemp);
   xlabel('Temperature Steps');
   ylabel('Objective Function');
   set(gca,'XDir','reverse')
end
countiter;

end
end
end
fclose(fid);

function OF = objfcn_v3(nsites, nint, labc, nsampyr, AA, BB, CC, DD, EE)
%Function for calculation of the objective function
%Kenneth R. Odom
%Dissertation Project  Version 11.30.02
%
A = [ ];
j = 0;
sitelist = [ ];
q = 0;
for i = 2 : nsites+1
   if AA(i,3) == 1
      j=j+1;
      A(j) = AA(i,2);
  end
end
B = [ ];
[m n] = size(A);
B = (AA(:,2))';
for i = 1 : n
   z = find(BB(:,1) == A(i));
   B = cat(1,B,BB(z,:));
```

SA3 Program listing continued

```
end
Bsum = sum(B(2:end,2:end));
sumsdist = 0;
for i = 1 : nsites
   if Bsum(1,i) >= 1
      sumsdist = sumsdist + CC(i,2);
      q = q +1;
      sitelist(q) = CC(i,1);
   end
end
statbenefit = 0;
for i = 1 : n
   c = find(CC(:,1) == A(i));
   statbenefit = statbenefit  + CC(c,3);
end
D = [ ];
D = (DD(1,:));
for i = 1 : n
   z = find(DD(:,1) == A(i));
   D = cat(1,D,DD(z,:));
end
Dsum = sum(D(2:end, 2:end));
sumidist = 0;
for i = 1 : nint
   if Dsum(1,i) >= 1
      sumidist = sumidist + EE(i,2);
   end
end
labcost = n * nsampyr * labc;
OF = statbenefit + 1.2*(sumsdist + sumidist) - labcost - sumsdist - sumidist;
return;
```

Sen's Slope Estimation program listing

```
% Sen's Slope Estimation for Time Trends
%   with resampling windows and data offsets
% Kenneth R. Odom  12/24/02

%Resampling window sequence
varspec = 9;   %pH=2, con=3, ANC=4, chl=5, nit=6, sul=7, sod=8, pot=9
ind = 0;       %set counter
sf = [ ];      %initialize array
zval = [ ];    %initialize array
offs = [ ];    %initialize array
offset = 0;    %offset sampling window by this number
icnt = 0;      %counter
jcnt = 0;      %counter
kcnt = 0;      %counter
[m n] = size(finalsw); %size of finalsw
numoffset = 12;  %user specified number of offsets
             %i.e. 4 equals 0 through 4 week offsets
sampfcount = 12; %user specified number of sampling frequencies in 1-week periods
             %i.e. 4 equals 1 through 4 week frequencies

for p = 1:sampfcount
   sampfreq = p;
for r = 0:numoffset
   offset = r;
   [m n] = size(finalsw);
   new2 = [ ];    %initialize array
   icnt = 0;      %counter
   jcnt = 0;      %counter
   kcnt = 0;
for i = 1:sampfreq:m
   icnt = icnt + 1;
   kcnt = i + offset;
   if kcnt >= (m-numoffset), break, end;
   new2(icnt,1) = finalsw(kcnt,1);
   new2(icnt,2) = finalsw(kcnt,2);
   new2(icnt,3) = finalsw(kcnt,3);
   new2(icnt,4) = finalsw(kcnt,4);
   new2(icnt,5) = finalsw(kcnt,5);
   new2(icnt,6) = finalsw(kcnt,6);
   new2(icnt,7) = finalsw(kcnt,7);
   new2(icnt,8) = finalsw(kcnt,8);
   new2(icnt,9) = finalsw(kcnt,9);
end

numtie=0;
A=new2(:,varspec);       %variable array
T=new2(:,1);      %time array
% Get dimension for loop control
[m n] = size(A);
% Transpose A
B=A';
```

Sen's Slope Estimation program listing continued

```
slp = [ ];
% Calculate Sen's Slope
numslope = m*(m-1)/2;
ctr=0;
for j=1:m-1
   for k=j+1:m
      ctr=ctr+1;
      slp(ctr)=(A(k,1)-A(j,1))/(T(k,1)-T(j,1));
   end
end

medslope=median(slp);
% Calculate number of ties
numtie=0;
tie = [ ];
for i=1:m
   for j=i+1:m
      if A(j)==A(i)
         numtie=0;
         for k=i:m
            if A(k)==A(i)
               numtie=numtie+1;
            end
            tie(i)=numtie;
         end
      end
   end
end
for i=1:m
   for j=i+1:m
      if A(j)==A(i)
         tie(j)=0;
      end
   end
end

[m n] = size(tie);
tiemat = [ ];
tcnt = 0;
for i = 1:n
   if tie(i) > 0
      tcnt = tcnt + 1;
      tiemat(tcnt) = tie(i);
   end
end
[m n] = size(tiemat);
% Calculate Z-value for Mann-Kendall
sumties=0;
for i=1:n
   w(i) = tiemat(i) * (tiemat(i)-1) * (2*tiemat(i) + 5);
   sumties=sumties+w(i);
```

Sen's Slope Estimation program listing continued

```
end

[g h] = size(new2);
vars = (1/18) * (g * (g-1) * (2*g + 5)- sumties);

% Calculate confidence limits
CC=1.645*sqrt(vars);
m1=(numslope-CC)/2;
m2=((numslope+CC)/2)+1;
C=sort(slp);

for i=1:numslope
   if i<=m1
      lcl=C(i);
   end
end
ucltrigger=0;
for i=1:numslope
   if i>=m2
      ucl=C(i);
      ucltrigger=1;
      break
   end
end
if ucltrigger==0
   ucl=C(numslope);
end
ind = ind + 1;
sf(ind) = p;
offs(ind) = r;
fprintf('\nFrequency=%i\t Offset=%i\t lcl=%f\t slp=%f\t ucl=%f',sampfreq,offset,lcl,medslope,ucl);
end
end
```

Mann-Kendall program listing

```
% Mann-Kendall procedure using Normal Approximation
%   with resampling windows and data offsets
% Kenneth R. Odom  12/24/01

%Resampling window sequence
varspec = 9;    %pH=2, con=3, ANC=4, chl=5, nit=6, sul=7, sod=8, pot=9
ind = 0;
sf = [ ];
zval = [ ];
offs = [ ];
icnt = 0;       %counter
jcnt = 0;       %counter
kcnt = 0;       %counter
[m n] = size(finalsw);
new2 = [ ];
numoffset = 12;     %user specified number of offsets
              %i.e. 4 equals 0 through 4 week offsets
sampfcount = 12;    %user specified number of sampling frequencies in 1-week periods
              %i.e. 4 equals 1 through 4 week frequencies

for p = 1:sampfcount
   sampfreq = p;
for r = 0:numoffset
   offset = r;
   [m n] = size(finalsw);
   new2 = [ ];
   icnt = 0;       %counter
   jcnt = 0;       %counter
   kcnt = 0;
for i = 1:sampfreq:m
   icnt = icnt + 1;
   kcnt = i + offset;
   if kcnt >= (m-numoffset), break, end;
   new2(icnt,1) = finalsw(kcnt,1);
   new2(icnt,2) = finalsw(kcnt,2);
   new2(icnt,3) = finalsw(kcnt,3);
   new2(icnt,4) = finalsw(kcnt,4);
   new2(icnt,5) = finalsw(kcnt,5);
   new2(icnt,6) = finalsw(kcnt,6);
   new2(icnt,7) = finalsw(kcnt,7);
   new2(icnt,8) = finalsw(kcnt,8);
   new2(icnt,9) = finalsw(kcnt,9);
end


%Mann-Kendall sequence

numtie=0;
A=new2(:,varspec);
% Get dimension for loop control
[m n] = size(A);
```

Mann-Kendall program listing continued

```
% Transpose A
B=A';

% Generate sign matrix
for j=1:m
    for i=1:m
        s(i,j)=(A(j,1)-B(1,i));
    end
end

% Total positives and negatives by row
countneg=0;
countpos=0;
for j=1:m
    for i=1:j
        if s(i,j)>0
            countpos=countpos+1;
        elseif s(i,j)<0
            countneg=countneg+1;
        end
    end
end

% Calculate Mann-Kendall S-statistic
S=countpos-countneg;

% Calculate number of ties
numtie=0;
tie = [ ];
for i=1:m
    for j=i+1:m
        if A(j)==A(i)
            numtie=0;
            for k=i:m
                if A(k)==A(i)
                    numtie=numtie+1;
                end
                tie(i)=numtie;
            end
        end
    end
end
for i=1:m
    for j=i+1:m
        if A(j)==A(i)
            tie(j)=0;
        end
    end
end

[m n] = size(tie);
```

Mann-Kendall program listing continued

```
tiemat = [ ];
tcnt = 0;
for i = 1:n
   if tie(i) > 0
       tcnt = tcnt + 1;
       tiemat(tcnt) = tie(i);
   end
end
[m n] = size(tiemat);
% Calculate Z-value for Mann-Kendall
sumties=0;
for i=1:n
   w(i) = tiemat(i) * (tiemat(i)-1) * (2*tiemat(i) + 5);
   sumties=sumties+w(i);
end
[g h] = size(new2);
vars = (1/18) * (g * (g-1) * (2*g + 5)- sumties);

if S < 0
   zvalue = (S+1)/sqrt(vars);
elseif S == 0
   zvalue = 0;
elseif S > 0
   zvalue = (S-1)/sqrt(vars);
else
end
ind = ind + 1;
sf(ind) = p;
offs(ind) = r;
zval(ind) = zvalue;
fprintf('\nFrequency = %i  Offset = %i   Z-score = %f',sampfreq,offset,zvalue);
end
end
outmat = ([sf; offs; zval])';
```

VITA

Kenneth Ray Odom was born in Fayette, Alabama, on January 21, 1965 and is the son of Bobby R. Odom and Brenda C. Odom of Bankston, Alabama. He attended Berry Elementary School and Berry High School in Berry, Alabama. Following high school Kenneth attended Brewer State Junior College in Fayette, Alabama where he graduated with an Associate degree in pre-engineering and was awarded the Outstanding Mathematics Student award in 1986. Kenneth briefly attended the University of Alabama in Huntsville where he worked with TRW, Inc., and the US Army Missile Command on the Patriot Missile project.

In August 1986 Kenneth was enrolled in the civil engineering program at the University of Alabama in Tuscaloosa. Kenneth was invited to join Chi Epsilon, Tau Beta Pi, and the Golden Key Honor Society. Kenneth received a Bachelor of Science in Civil Engineering degree and graduated with honors. He was also named the 1989 Outstanding Senior Civil Engineering Student. Upon completion of college he began his engineering career working with Mobil Oil Corporation in Beaumont, Texas. In September of 1990 Kenneth began working with Almon Associates, Inc., in Tuscaloosa, Alabama as a consulting engineer. While working at Almon Associates Kenneth enrolled in the Graduate School at the University of Alabama in 1994 and later received a Master of Science degree in Civil Engineering in 1999. Kenneth also received his professional engineering license in Alabama in February 1994.

In August 1994 Kenneth accepted the position of assistant city engineer for Maryville, Tennessee. In January of 2000 Kenneth enrolled in the doctoral program at the University of Tennessee in Knoxville, Tennessee. Kenneth was promoted to Director of Engineering and Development for the City of Maryville in November of 2001.

Kenneth currently resides in Maryville, Tennessee with his wife Shanda, of 19 years, and their two children Nolan and Garrett Odom.