2013

# Rating Curve Development And Multivariate Statistical Analyses Of Stream Water Quality In Greensboro, North Carolina

Bhasker Jha
*North Carolina Agricultural and Technical State University*

# Rating Curve Development and Multivariate Statistical Analyses of Stream Water Quality in Greensboro, North Carolina

Bhasker Jha

North Carolina A&T State University

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department: Civil, Architectural and Environmental Engineering

Major: Civil Engineering

Major Professor: Dr. Manoj Kumar Jha

Greensboro, North Carolina

2013

School of Graduate Studies

North Carolina Agricultural and Technical State University

This is to certify that the Master's Thesis of

Bhasker Jha

has met the thesis requirements of

North Carolina Agricultural and Technical State University

Greensboro, North Carolina

2013

Approved by:

| | |
|---|---|
| Dr. Manoj K. Jha | Dr. Shoou-Yuh Chang |
| Major Professor | Committee Member |
| Dr. Ahmed Megri | Dr. Godfrey Gayle |
| Committee Member | Committee Member |
| Dr. Sameer Hamoush | Dr. Sanjiv Sarin |
| Department Chair | Dean, The Graduate School |

# Biographical Sketch

Bhasker Jha was born on 11th of January, 1986 in Basatpur, Rautahat, Nepal. He was brought up in the Kathmandu valley, which is also the capital of Nepal. His late father was an electrical engineer and mother is a housewife. He has four elder sisters and one younger brother.

Bhasker completed his schooling from Siddhartha Vanasthali Institute, Kathmandu. He did his undergraduate in Civil Engineering from Institute of Engineering, Pulchowk campus. During his undergraduate studies he developed interest towards the water related courses. He was interested in the underlying mechanism of watershed hydrology and complex environmental processes. In order to pursue his interest and desire to experience the world class education he started his graduate studies at North Carolina A&T State University from the fall of 2011 under the guidance of Dr. Manoj K. Jha. It was completely different experience for him, as he got to know one of the best education system and research environments in the world.

Acknowledgements

First of all, I would like to acknowledge the contribution of my major professor Dr. Manoj K. Jha throughout my graduate program. This thesis would not have been possible without his patient help and advice. It was a great experience to have him as a mentor to guide me through not only academic but personal hurdles. His presence was instrumental in making my stay in USA comfortable. I believe the teaching that I received from him during my graduate program will be useful not only in the career but throughout my life. I am also thankful to other committee members Dr. Shoou-Yuh Chang, Dr. Ahmed Megri, and Dr. Godfrey Gayle for their constructive comments on my research and thesis.

I would also like to acknowledge the contribution of my loving parents; Late Mr. Kaushalendra Jha and Mrs. Asha Jha. I would like remember all my siblings; Ms. Archana Jha, Ms. Bandana Jha, Ms. Sadhana Jha, Ms. Anjana Jha and my brother Mr. Prabhat Jha. My eldest brother-in-law Dr. Pramod K. Thakur has always been a great inspiration and source of comfort while studying alone here in USA. My brother-in-laws Mr. Bimalesh Mishra, Mr. Sanjeev K. Jha and Mr. Rajeev Jha have always been in my thoughts. Finally, I would like to thank all my extended family members for their wonderful presence in my life.

Last but not the least; I would like to express my heartfelt gratitude to Mr. and Mrs. Eric Yvette Howard for being such a nice and helping couple. In them, I find God's true messenger. I would like to acknowledge the help and company I received from my friends and fellow students Somsubhra Chattopadhyay, Anup Saha and Torupallab Ghosal. Special note of appreciation goes to Mr. Ashraf Al Fadiel for being such a nice friend and fun to be around.

Table of Contents

List of Figures

List of Tables

## Abstract

A suite of regression models were tested for the construction of rating curves and constituent load estimation for 17 water quality parameters monitored at 16 stations regularly since 1999 by the City of Greensboro in North Carolina. Best models were selected based on the statistical evaluation within the framework of the LOAD ESTimator (LOADEST) model. The constituent prediction varied from the "true load" by –6% to 16% for Nitrate; -14% to +12% for Nitrite; -6% to 0% for Total Dissolved Solids (TDS); -2% to 9% for Total Kjeldahl Nitrogen (TKN); -22% to 9% for Total Phosphorus (TP); and -51% to 23% for Total Suspended Solids (TSS). There was a systematic bias towards under-prediction for TDS, TP and TSS whereas nitrate and TKN were over predicted and none for Nitrite. The predicted loads were compared with five interpolation methods (M1, M2, M3, M4 and M5) in the following pattern: for nitrate, TDS and TSS, load estimated by M3, M4 and M5 > LOADEST > M1 and M2; for nitrite, TKN and TP: LOADEST > M3, M4 and M5 > M1 and M2.

Multivariate analyses used cluster analysis (CA), factor analysis (FA) and principal component analysis (PCA) on all parameters at all stations. CA grouped the water quality station into four spatially similar clusters. PCA/FA was applied on the entire dataset of entire watershed and spatially similar stations. Combination of FA/PCA and CA reduced the size of the dataset by 71% and represented the 64% of the total variance.

# CHAPTER 1

# Introduction

## 1.1 Background

Rivers are one of the major inland sources of the water. Rivers are primarily used for drinking, irrigation, power generation, industrial processes, recreational activities, etc. Rivers are also used to assimilate the industrial and municipal waste as well as to receive surface runoffs from agricultural lands. Waters from industries, cities and agricultural land brings with it different pollutants which degrades the water quality of the rivers. Surface runoffs are seasonal and depend on the climate of the river basin. Seasonal variations in precipitation, surface run-off, interflow, groundwater flow and pumping directly affect river discharge and in turn affect the concentration of pollutants in river water (Vega, Pardo, Barrado and Deban, 1998). Rivers are also vulnerable to open dumping, atmospheric deposition, leachate from the landfill sites etc. So, water quality of the river system is a composite effect of many factors and is very difficult to manage. To regulate and maintain the sustainability of rivers system proper management is required. Therefore monitoring systems are setup in river networks to collect the samples to determine the level of pollutant. It is important for sustaining the aquatic ecosystem as well as use of river water for anthropogenic purposes.

Health of the water bodies depend on the amount of nutrients, dissolved oxygen, pH, TDS value, etc. present in them at any point of time. Accurate estimation of nutrient loads in the river and streams is very necessary for many applications, including determining sources of nutrient loads in the watersheds (Alexander et al., 2008; Preston, Bierman and Sillman, 1989), calibrating and validating watershed models (Ullrich and Volk, 2010; Jha, Arnold and Gassman, 2007; Jha, Schilling, Gassman and Wolter, 2010a; Jha, Wolter, Gassman and Schilling, 2010b)

and evaluation of long-term trends in the loads (Littlewood, Watts and Custance, 1998; Schilling and Zhang, 2004). The instantaneous load which can be found by multiplying the nutrient concentration C (t) and discharge Q (t) for given time t and the load over an extended period of time T is given by

$$\boldsymbol{L_T = \int_0^T C(t)Q(t)dt} \quad (1)$$

Due to constraint in the resources, the continuous measurement of concentration and discharge is not carried out. Even though daily discharge measurement is available from the United States Geological survey (USGS), concentrations of the nutrients are measured less frequently and gap between measurements can be from weeks to months. For the estimation of load for extended period of time continuous data is necessary. So, there is a need to convert the weekly or monthly data into daily data. It is well known that load estimation of nutrients is subjected to many potential sources of error and uncertainty (Guo and Demissie, 2002) and rating curve generation is one of them.

There are several methods of load estimation. Many studies have compared these methods (Guo and Demissie, 2002; Aulenbach and Hooper, 2006; Moatar and Meybeck, 2005; Li, Zhang, Schilling and Skopec, 2006; Zamyadi, Gallichand and Duchemin, 2007; Ullrich and Volk, 2010) and various techniques were applied to measure the performances of the model. In some studies, under sampling against a true load to evaluate load uncertainty and model performance (Guo and Demissie, 2002; Li et al., 2006) is used, while in others, different algorithm method were applied to same dataset (Moatar and Meybeck, 2005; Zamyadi et al., 2007). These studies show that there is lots of variability in the nutrients load estimates. Errors on the estimated annual phosphorus load are 30% (Robertson and Roerish, 1999) and 34% (Moatar and Meybeck, 2005). Annual nitrate loads have differed by as much as 64% depending

on the sampling strategy, load estimation method and monitoring period used (Ullrich and Volk, 2010).

Among these various methods, this study is exclusively focused on the accuracy of one method, the widely used multiple regression model developed by Cohn et al.(1992), to evaluate how the model estimated nutrient loads at Greensboro watershed. This model is incorporated in LOADEST, a computer program for load estimation, which is widely used to estimate nitrogen and phosphorus loads in the rivers (Aulenbach and Hooper, 2006; Goolsby, Battaglin, Aulenbach and Hooper, 2000; Goolsby and Battaglin, 2001; Hooper et al., 2001; Maret et al., 2008). The model has been utilized by the USGS to estimate nutrient flux in the major rivers flowing to the Gulf of Mexico (USGS, 2009a) and to calculate "observed" loads in the USGS SPARROW model (USGS, 2009b). Both of these applications of the Cohn et al. (1992) regression model have a great deal of significance for agricultural states such as North Carolina states that are major contributor of the nutrient loads.

In view of the spatial and temporal variations in hydrochemistry of rivers, regular monitoring programs are required for reliable estimates of the water quality. Monitoring program for water quality in river system running for considerable time period produces large amount of data. These large datasets with many physico-chemical parameters are complex in nature and their interpretation for meaningful conclusions are very difficult. Different watersheds have different physico-chemical parameters which play major role in affecting the water quality. So, for each watershed it is very important to find out these physic-chemical parameters affecting water quality. By recognizing important parameters, amount of data could be reduced, making the analysis and interpretation easier. Among many measured, it is very difficult to determine which is the most important parameters without any mathematical analysis and interpretation.

The application of multivariate techniques, such as cluster analysis (CA) helps to reduce the number of data by reducing the monitoring point. Similarly, principal component analysis (PCA), and factor analysis (FA) helps to achieve the same by identifying the important parameters to be measured.

These techniques helps in the interpretation of complex data matrices to better understand the water quality and ecological status of the studied systems, allows the identification of possible factors/ sources that influence water systems and offers a valuable tool for reliable management of water resources as well as rapid solution to pollution problems (Adams, Titus, Pietesen, Tredoux, and Harris, 2001; Lee, Cheon, Lee, Lee and Lee, 2001; Reghunath et al., 2002; Simeonova et al., 2003; Vega et al., 1998; Wunderlin et al., 2001). Multivariate statistical techniques have been applied to characterize and evaluate surface and freshwater quality, and it is useful in verifying temporal and spatial variations caused by natural and anthropogenic factors linked to seasonality (Helena et al., 2000; Singh, Malik, Mohan, and Sinha, 2004).

**1.2 Problem Statement and Objective**

A comprehensive water quality monitoring program is generally needed to regulate the water quality of a river system for its sustainable use and healthy ecosystem. However, limited resources and impracticality of installing monitoring stations on every tributary pose a challenge. Even when the location is decided, the constituent concentration in the stream is not often measured on a daily basis. It is controlled by many factors like resources availability and objective and scope of the project. Infrequent sampling necessitates developing interpolation methods or rating curves that can convert intermittent data into daily data with confidence (low error). Additionally, for stations which monitors multiple parameters and for longer duration,

comprehensive statistical analyses such as principal component analysis (PCA), factor analysis (FA), and cluster analysis (CA) are needed.

It therefore requires a very efficient and effective monitoring system which includes determining representative water quality parameters and suitable location of the monitoring stations which can effectively represent the water quality of the streams in the region. This study aims to understand the variability of stream water quality through a set of statistical techniques applied on a large datasets developed for City of Greensboro from 16 monitoring stations each monitoring 17 water quality parameters. The specific objectives are as follows:

- Develop rating curves of water quality constituents using multiple regression models in LOADEST
- Compare and contrast 6 methods of constituent load estimation from available intermittent water quality data
- Identify/rank most to least important water quality parameters for their variabilities and significance in the health of the stream
- Cluster monitoring stations based on spatial variability's of water quality parameters

**1.3 Scope of This Study**

Any study involved with the rating curve construction for loadings, analysis and interpretations are directly affected by the frequency of data measurement. The outcome of this study was limited to the monthly (2009-2012) and bi-monthly (1999-2008) water quality constituents data monitored at 16 locations in the City of Greensboro, North Carolina. The data frequency was very low for the load estimation and rating curve development. According to Cohn et al. (1992), at least 75 data points per year may be adequate for load estimation from intermittent data points. In the study presented here, we were constrained by only 6 to 12 data

per year which may have significantly affected the output. In contrast, monthly data seem to be adequate for multivariate analyses such as PCA.

# CHAPTER 2

# Literature Review

## 2.1 Rating Curve

Alexander et al. (2008) modeled Mississippi and Atchafalaya River Basin using SPARROW water quality model to analyze seasonal hypoxia in northern Gulf of Mexico. This study revealed different sources and transport process controlling the loading of nitrogen and phosphorus. Agricultural sources are found to be major contributor of nitrogen and phosphorus with more than 70% of contribution. Among different agricultural sources, for nitrogen corn and soybean (52%) was highest contributor, whereas for phosphorus animal manure contributed the most (37%). Atmospheric nitrogen also had about 16% contribution to loading into Gulf. In stream removal of nutrient depends upon the size of streams involved. For smaller streams removal rate of nutrients was higher in comparison to bigger streams. Therefore, bigger streams delivered larger flux in Gulf. So, targeting sources near big streams or close to big stream was found to be very effective way to reduce the loading. In stream removal rate of TP was 69% of that of TN. In stream removal rate of TP was one- third of that of removal rate in reservoir. Removal rate of TP was much higher than that of TN in the case of reservoir.

Ulrich and Volk (2010) studied the impact of different sampling strategy and load estimation methods on calibration and validation of Soil and Water Assessment Tool (SWAT). SWAT was modeled for Parthe watershed (315 $km^2$) located in Central Germany. For the study period deviation of annual loading of nitrate deviated from the mean loading for all the methods from 9.8% to 15.7% for daily and 24.9% to 67% for sub-monthly and monthly sampling strategy. Mean loading was calculated by averaging the annual loading of all the methods. Nash-Sutcliffe efficiency (NSE) calculated using daily, sub-monthly and monthly dataset was 0.52,

0.42 and 0.31 respectively. From the above results it was proposed the evaluation of existing sampling strategy in terms of spatial distribution and temporal resolution depending on the substance being monitored and hydrological conditions. Monitoring during storms was also found to be very important due to presence of high variability. Monthly monitoring for above site was found to be ineffective and more intense sampling was proposed. Presence of economic constraint was acknowledged and it was advised to use multiple methods to estimate loading. Importance of calculation of uncertainty was also pointed out.

Schilling and Zhang (2004) analyzed the baseflow contribution of nitrate-nitrogen in Racoon River watershed in west-central Iowa. Long term data of stream flow and nitrate concentration data from 1972-2000 were used. Hydrograph separation technique was used to find the contribution of base flow from stream flow. It was found that base flow contributed close to two-third of the annual nitrate (17.3 kg/ha) loading (26.1kg/ha). From the study of season and annual patterns of nitrate loss, it was found that base flow contributed up to 80% of total transport in spring and late fall. New term baseflow enrichment ratio (BER) was introduced to describe connection between base flow water and base flow nitrate loadings. Long term BER for Raccoon River was found to be 1.23 which suggested preferential leaching of nitrate to baseflow. But the value of BER was less than 1 for crop growing season. It indicated that plant absorbed the nitrate in the soil decreasing the amount of nitrate in the baseflow water. This study showed the importance of studying long term flow helps to identify non-point sources.

Goolsby and Battaglin (2001) studied long term nitrogen loading in Mississippi river basin and its effect on the size of hypoxic zone in Gulf of New Mexico. By analyzing historical data, it was found that nitrogen concentration in lower Mississippi River Basin have heavily increased in last century with major increase since 1970s. Nitrate was found to be sole

contributor to this increase. The average annual nitrogen flux from lower Mississippi to Gulf of Mexico for the period of (1980-99) was found to be three times that of period 1955-70 and was one of the reasons of increase in size of hypoxic region. Of the current nitrogen loading to Gulf, nitrate contributes 62% and remaining came mostly from organic nitrogen. The increase in nitrogen flux was due to increase in use of fertilizer, variability in precipitation and increase in stream flow.

Guo and Demissie (2002) calculated the nitrate-N loading for an agricultural watershed in central Illinois. Methods used to calculate loading in this study were rating curve method, ratio estimator and weighted average method. Bias correction technique called Minimum Variance Unbiased Estimator (MVUE) and smearing estimator was applied for rating curve. Six year of data was used in this study and Monte Carlo simulation was used to generate different sampling scenario from weekly to bi-monthly. For example from daily data for six years Monte Carlo simulation will randomly choose any data within a week as weekly data sampling data using weekly monitoring process. Sampling duration was chosen 1, 2, 3, and 6 years. The results demonstrated that a desired accuracy of the estimates could be achieved either by sampling more frequently or by monitoring the site longer. Although the ratio and the flow-weighted average estimators had a small negative bias, in most cases rating curve estimators were positively biased when applied to the study site. Also, neither of the two bias correction techniques, MVUE and smearing estimator, decreased this positive bias. On the contrary, those techniques produced a higher bias, which resulted in increased root-mean-square error (RMSE). The rating curve uncorrected for bias, the simple ratio, and the flow-weighted estimator had a significantly smaller RMSE for all sampling frequencies and all periods of record than the bias-corrected rating curve.

**2.2 Multivariate Statistical Analyses**

Singh et al. (2004) applied different multivariate techniques for water quality data mining and interpretation of data obtained from monitoring Gomti River, tributary of river Ganges, located in the Northern India. Eight monitoring stations which were involved in this study were divided into three groups namely least polluted, moderately polluted and most polluted. These categories were divided on the basis of location of these stations. Least polluted was located upstream of major city on the bank of river, most polluted downstream of city and moderately polluted further away from the city. Twenty four water quality parameters were measured in these stations on a regular basis for duration of five years (1994-1998). Multivariate techniques used in this study were cluster analysis (CA), factor analysis/ principle component analysis (FA/PCA) and discriminant analysis (DA). Before raw data were used by these processes they pretreated by Z-transformation except for discriminant analysis which uses raw data. Non-normal distribution of water quality parameters were accounted for by studying correlation structure between variables using Spearman R method. Temporal variation was studied by dividing seasons into three seasons. They were summer, winter and monsoon. PCA helped reduce the data by 40% i.e. 14 from 24 with 71% representation in variability. CA grouped the stations into three groups which were exactly similar to spatial distribution and helped in reducing the number of station. But DA was most successful as it used only five parameter to discriminate between seasons with 88% correct assignment (80% reduction) and nine parameter for spatial discrimination with 91% correct assignment (63% reduction).

Xu, Xu, Wu and Tang (2012) studied the spatio-temporal variation of water quality of rivers in the Zhangweinan River Basin, China. Data from nineteen water quality monitoring sites monitoring eleven water quality parameters were used in this study. The duration of data used

was from 2001 to 2009. Fuzzy comprehensive logic was used to evaluate water quality and statistical techniques like system cluster analysis and seasonal Mann Kendall tests were used to analyze the spatiotemporal variation of the water quality data. Clustering analysis was done spatially as well as temporally. Temporally duration of 2001 to 2009 was divided into flood season and non-flood season. Results showed that water quality in the Zhangweinan river basin was divided into two areas on the basis of pollution. One is Zhang river basin in northwest of the Zhangweinan, where water quality is good and another is eastern plains and Wei river where water quality was very bad. Flood season was divided into three period of 2002-2003, 2004-2006, 2001 and 2007-2009 according to pollution level. Pollution level in first period is worst, with comparative improvement in second and third respectively. Similar results were observed by seasonal Mann Kendall test.

Helena et al. (1999) studied the temporal evolution in the groundwater composition in an alluvial aquifer located in Pisuerga River, Spain by the method of principal component analysis. In this study temporal evolution was studied by using the data of survey done on the two periods. First period was October, 1994 at the end of irrigation season and low water period just before autumn rain whereas second period was April-May 1995 at the beginning of irrigation and high water period after recharge of aquifers. Sixteen parameters of water quality were used in this study consisting of data from thirty-two monitoring sites (20 wells and 12 natural springs). Study area was divided into five zone left bank, right bank and three independent areas on the left bank. The experimental 64×16 matrix was analyzed by Principal Component Analysis (PCA), and the resulting Principal Components (PCs) and Varimax rotated PCs (VFs) analyzed by means of box and bivariate plots. PCA showed the existence of up to five significant PCs which account for

71.39% of the variance. Two of them can be initially assigned to mineralization' whereas the other PCs are built from variables indicative of pollution.

Boyacioglu, Boyacioglu, and Gunduz (2005) applied factor analysis in the assessment of surface water quality in Buyuk Menderes River Basin. This study attempted to identify pollution indicators for domestic and agricultural pollution. In this study, factor analysis was applied on the standardized data and factors were chosen which had eigen value greater than one. There were two factors representing 84.5% of the total variance represented by the water quality data. First factor represented 63.39% of variance and higher loading for electrical conductivity, sulfate, sodium and TKN. These parameters represented the discharge from the agricultural land and were considered as “inorganic contamination”. Factor 2 explained 21% of variance and had high loading for COD, BOD and total coliform. Factor 2 represented discharges from domestic sources and considered as “organic contamination”. This study recommended possible use of marker parameter from each of these two groups to be used as pollution indicator.

# CHAPTER 3

# Methodology

## 3.1 Background

This study was concerned with construction of rating curve, annual load calculation and application of multivariate statistical methods for interpretation and analysis of the large water quality datasets. Among various methods, this study was exclusively focused on the accuracy of one method for the rating curve construction. The widely used multiple regression model developed by Cohn et al. (1992) was used and its accuracy was evaluated for the stream water quality in the City of Greensboro, North Carolina (Figure 1). Table 1 gives the information about water quality stations and number of concentration point used in this study. This model was incorporated in LOADEST (USGS LOADEST, 2004), a computer program for load estimation, which was widely used to estimate nitrogen and phosphorus loads in the rivers (Aulenbach and Hooper, 2006; Goolsby et al., 2000; Goolsby and Battaglin, 2001; Hooper et al., 2001; Maret et al., 2008). The model was utilized by the USGS to estimate nutrient flux in the major rivers flowing to the Gulf of Mexico (USGS, 2009a) and to calculate "observed" loads in the USGS SPARROW model (USGS 2009b). Both of these applications of the Cohn et al. (1992) regression model have a great deal of significance for agricultural states such as North Carolina states that are major contributor of the nutrient loads.

In addition to LOADEST, five interpolation methods were used to calculate annual loading in the watershed. These methods were fairly accurate when there was enough water quality data available and had been used in many studies for comparison between many methods.

Multivariate statistical techniques were very powerful techniques which has important application in water quality data interpretation and analysis. These techniques helped in the

interpretation of complex data matrices to better understand the water quality and ecological status of the studied systems, allows the identification of possible factors/ sources that influence water systems and offers a valuable tool for reliable management of water resources as well as rapid solution to pollution problems (Adams et al.2001; Lee et al., 2001; Reghunath et al., 2002; Simeonova et al., 2003; Vega et al., 1998; Wunderlin et al., 2001). Multivariate statistical techniques has been applied to characterize and evaluate surface and freshwater quality, and it is useful in verifying temporal and spatial variations caused by natural and anthropogenic factors linked to seasonality (Helena et al., 2000; Singh et al., 2005). So, techniques like factor analysis (FA), principal component analysis (PCA), and cluster analysis (CA) are used in this study for spatio-temporal evaluation of water quality datasets.

**3.2 Study Area and Data Analysis**

Figure 1 shows the location of the city of Greensboro within the Haw and Deep River watersheds in North Carolina, USA. Haw River drains into the Cape Fear River basin. Greensboro is one of the top four largest city of state of North Carolina and consists of many streams and lakes. The city area is 283 $km^2$ of which 4.16% is covered by water bodies. The rivers and lakes are monitored regularly by the city.

Various water quality parameters have been measured on bi-monthly and monthly basis at 16 water quality monitoring locations (Table1) (http://www.greensboro-nc.gov/index.aspx?page=2301). Fourteen stations fall within Haw River Watershed while 2 stations are in Deep River Watershed.

This study used 17 parameters (Table 2) from all monitoring sites. The data were available on a bi-monthly basis for a period of 1999-2008 and on a monthly basis for a period of 2009-2012.

Table 1

*Water Quality Stations*

| Water Quality Stations | | | | | |
|---|---|---|---|---|---|
| a. | 16th St. | g. | Fleming Rd. | m. | Old Oak Ridge Rd. |
| b. | Aycock St. | h. | Friendship Church | n. | Pleasant Ridge Rd. |
| c. | Battleground Avenue | i. | Kivett Dr. | o. | Randleman Rd. |
| d. | Bluff Run Rd. | j. | Mackay Rd. | p. | Rankin Mill Rd. |
| e. | Church St. | k. | McConnell Rd. | q. | W. JJ Dr. |
| f. | Fieldcrest Dr. | l. | Merritt Dr. | r. | White St. |

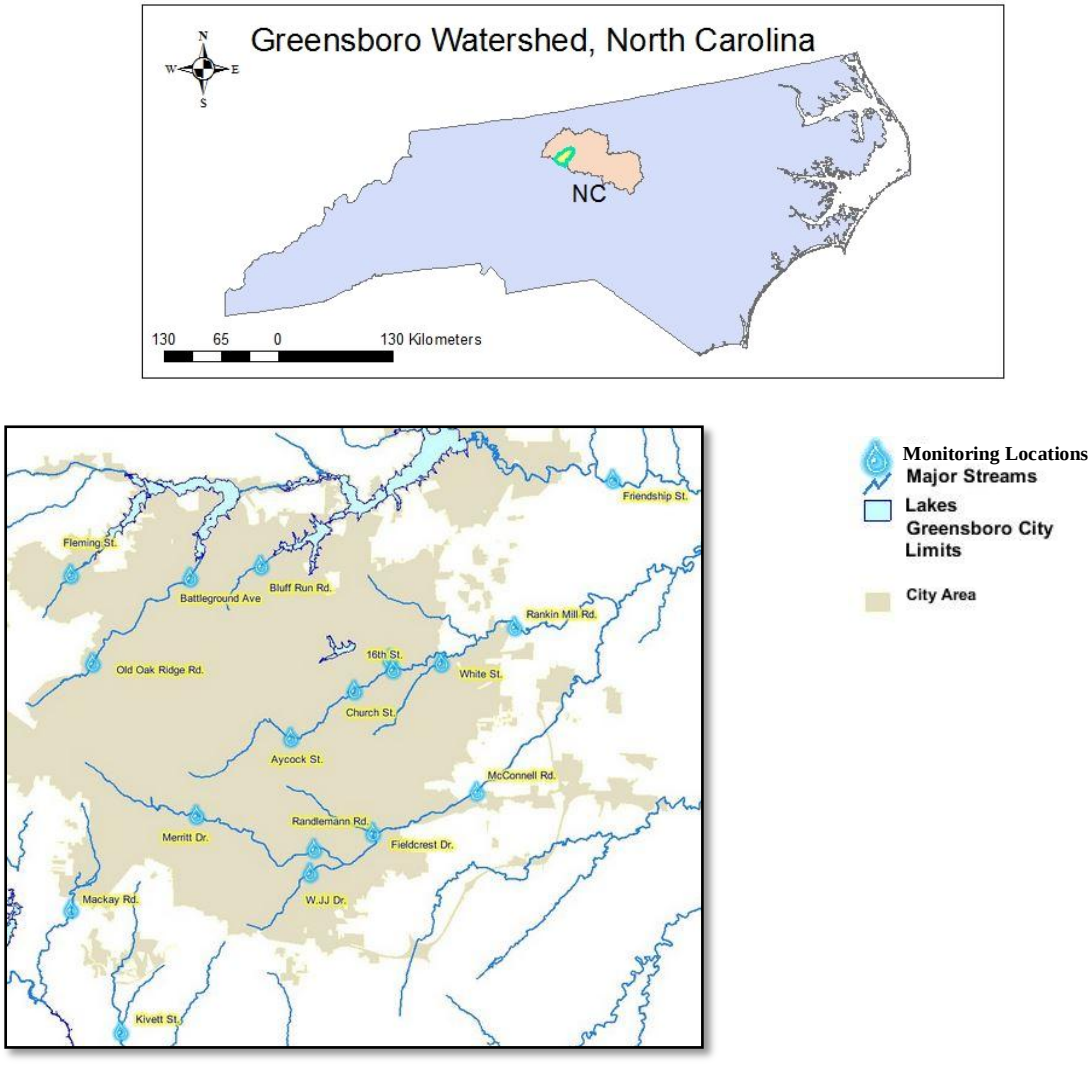


*Figure 1.* Location of the 18 water quality monitoring locations within the city of Greensboro in Haw River watershed in North Carolina

Table 2

*Water Quality Parameters and number associated with them for the analysis*

| S.N. | Parameters | S.N. | Parameters | S.N. | Parameters |
|---|---|---|---|---|---|
| 1 | Cadmium | 7 | COD | 13 | TSS |
| 2 | Copper | 8 | Fecal Coliform | 14 | TKN |
| 3 | Lead | 9 | Hardness | 15 | Total Phosphorus |
| 4 | Zinc | 10 | Nitrate | 16 | Turbidity |
| 5 | Alkalinity | 11 | Nitrite | 17 | Conductivity |
| 6 | BOD | 12 | TDS | | |

**3.2.1 Data characteristics.** Data characteristics that were explored before using them are as follows:

1. Magnitude of difference of measured concentrations of the parameters may not indicative of their relative importance. For e.g. a parameter with high magnitude of concentration should not automatically indicate its high importance. Similarly parameters with small magnitude of concentration do not necessarily indicate low importance.

2. Technical variations; this originates from sample collection and preparation and analytical errors.

3. Heteroscedasticity: It is the possible absence of constant variance and symmetry around zero among sub-population.

**3.2.2 Data preprocessing and pretreatment.** Data obtained from the City of Greensboro couldn't be directly used in this study. Data was first checked for possible anomalies like missing data, zeroes or unequal number of data for different parameters, cleaning of outliers etc. Grubb's two method test was used for finding the outliers and it was replaced by median value of that parameter. Then eight different method of pretreatment were used for transformation of data. They were as follows:

Centering: It focuses on the differences and not the similarity of the data.

$$\tilde{x}_{ij} = x_{ij} - \overline{x}_i \quad (2)$$

Scaling:

Auto-Scaling: It compares parameters based on the correlation.

$$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x}_i}{s_i} \quad (3)$$

Range-Scaling: It compares parameters relative to the hydrologic response.

$$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x}_i}{(x_{i_{max}} - x_{i_{min}})} \quad (4)$$

Pareto-Scaling: It reduces the relative importance of large data values, but keeps the data structure partially intact.

$$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x}_i}{\sqrt{s_i}} \quad (5)$$

Vast-Scaling: It focuses on the parameters that show small fluctuations.

$$\tilde{x}_{ij} = \frac{(x_{ij} - \overline{x}_{i)}}{s_i} * \frac{\overline{x}_i}{s_i} \quad (6)$$

Level-Scaling: It focuses on relative response.

$$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x}_i}{\overline{x}_i} \quad (7)$$

Transformations:

Log-Transformation: It corrects for heteroscedasticaity, pseudo-scaling and make multiplicative model additive.

$$\tilde{x}_{ij} = \log_{10}(x_{ij}) \quad (8)$$

$$\hat{x}_{ij} = \tilde{x}_{ij} - \overline{\tilde{x}}_i \quad (9)$$

Power-Transformation: It corrects for heteroscedasticity and pseudo-scaling.

$$\tilde{x}_{ij} = \sqrt{x_{ij}} \quad (10)$$

$$\hat{x}_{ij} = \tilde{x}_{ij} - \overline{\tilde{x}}_i \quad (11)$$

### 3.3 Rating Curve

Among many methods, method used in this study for the construction of rating curve is LOADEST (2004). Along with the construction of the rating curve, it is also used to calculate the annual loading. In addition to that, five different averaging methods were also used to calculate the annual loading.

**3.3.1 LOADEST.** LOAD ESTimator (LOADEST) provides a suite of multiple regression models for estimating constituent loads in streams and rivers. Given a time series of streamflow, additional data variables, and constituent concentrations, LOADEST assists the user in developing a regression model (USGS LOADEST, 2004). Mean load estimates, standard errors, and 95 percent confidence intervals are developed on a monthly and (or) seasonal basis. The details on modeling assumptions, calibration and estimation procedures and error quantification can be found in the User's Manual (USGS LOADEST, 2004). Some basics of the formulation and procedures are provided in following paragraphs.

Total mass loading over an arbitrary time period, τ, is given by:

$$L_\tau = \int_0^\tau QCdt \quad (12)$$

Where *C* is concentration [M/L3], Lτ is total load [M], *Q* is instantaneous stream flow [L3/T], and *t* is time [T]. Equation 12 cannot be used directly as continuous estimates of *Q* and *C* was rarely available. Although discrete values of *Q* were readily available for many locations, values of *C* are considerably less common due to the expense of sample collection and analysis. Load estimates are therefore more commonly given by:

$$\hat{L}_\tau = \Delta t \sum_{i=1}^{NP} \left(\widehat{QC}\right) i = \Delta t \sum_{i=1}^{NP} \hat{L}_\tau \quad (13)$$

Where $\widehat{L_\iota}$an estimate of instantaneous load [M/T] is, $\widehat{L_\tau}$ is an estimate of total load [M], *NP* is the number of discrete points in time, and Δ*t* is the time interval represented by the instantaneous load [T]. Mean load for time period τ is then given by

$$\bar{L} = \frac{\widehat{L_\tau}}{\Delta t\, NP} \tag{14}$$

Where is the mean load [M/T]. Calculation of loads using equations (13) and (14) is contingent upon two assumptions. First, each estimate of instantaneous load is assumed to represent the mass load over the discrete time interval (Δ*t*). Second, the discrete time interval is constant; all of the *NP* discrete points in time will have the same Δ*t*. For example, a common application is to calculate the mean load for a calendar year, by using daily estimates of stream flow. Under this application, there will be 365 discrete points in time (*NP*=365) with a time interval of one day (Δ*t*=1 day). Each estimate of instantaneous load represents average conditions for a given day.

As described by Cohn (1995), several techniques are available for estimating total load. Of these techniques, one based on linear regression is used within LOADEST. In its simplest form, the regression approach proceeds as follows. First, a linear model is formed in which the log of instantaneous load is related to one or more explanatory variables:

$$\ln(\hat{\mathrm{L}}) = \mathrm{a}_0 + \sum_{\mathrm{j}=1}^{\mathrm{NV}} \mathrm{a_j X_j} \tag{15}$$

Where $a_0$ and $a_j$ are model coefficients, *NV* is the number of explanatory variables, and *Xj* is an explanatory variable. Equation (15) is then exponentiated to yield an estimate of instantaneous load:

$$\hat{L}_{RC} = exp\left(a_0 + \sum_{j=1}^{M} a_j X_j\right) \tag{16}$$

Where $\hat{L}_{RC}$ is a "rating curve" estimate of instantaneous load. Development of load estimates using equations (15) and (16) is thus a 3-step process:

Model Formulation: The form of the linear model (the right-hand side of equation 15) is determined based on the user's knowledge of the hydrologic and biogeochemical system. Each explanatory variable ($X_j$) is a function of a data variable (streamflow or time, for example) that is thought to influence instantaneous load. The number and form of explanatory variables is highly dependent on the system under study and the constituent of interest. A simple model with a single explanatory variable (log streamflow) is often sufficient for prediction of suspended-sediment load (Crawford, 1991), whereas a model with six explanatory variables based on various functions of streamflow and time is often applicable to nutrients (Cohn et al. 1992). Additional guidance on model formulation is provided elsewhere (Judge, Hill, Griffiths, Lutkepohl, and Lee, 1988; Draper and Smith, 1998; Helsel and Hirsch, 2002).

Model Calibration: Given the form of the regression model, a time series of constituent load and the explanatory variables is used to develop the model coefficients ($a_0$ and $a_j$, equation 15) by using ordinary least squares (OLS) regression. The regression equation then is used to calculate estimates of log load [ln($L$)] for each observation in the time series (the calibration data set).Residual error for each observation is equal to the difference between observed and estimated values of log load [ln($L$) –ln(L)'].

Load Estimation**:** Estimates of the instantaneous load are obtained using the retransformed version of the regression model (equation 16) and a time series of explanatory variables (the estimation data set). Individual estimates of instantaneous load then are used to determine the total (equation 13) or mean (equation 14) load. As outlined above, estimation of constituent loads using the regression approach is theoretically straightforward. Several statistical complications arise, however, when dealing with real-world data. Load calculations within LOADEST are therefore more complex than the calculations described above. Three of

these complicating factors (retransformation bias, data censoring, and non-normality) are described below, where the three load estimation methods used within LOADEST are detailed. Additional issues that are germane to all three methods are described below. The load estimation process is complicated by retransformation bias, data censoring, and non-normality. As noted by Ferguson (1986), rating curve estimates (equation 16) of instantaneous load are biased; estimates may underestimate the true load by as much as 50 percent. This retransformation bias is addressed by introducing bias correction factors for the calculation of instantaneous load. Data censoring occurs when one or more observations used in the calibration step have constituent concentrations that are less than the laboratory detection limit. Although substitution (setting $C$ equal to one half the detection limit, for example) appears to be a simple remedy for the replacement of less-than values, none of the substitution methods commonly used yield adequate results (Helsel and Cohn, 1988). A more rigorous treatment of censored data is therefore required. A final complication is the assumption of OLS regression that the model residuals are normally distributed. Alternate methods for estimating model coefficients are applicable when model residuals do not follow a normal distribution. Because of these complications, LOADEST provides three methods for load estimation; each method is described below.

a) Maximum Likelihood Estimation (MLE): As an alternative to OLS regression, model coefficients ($a_0$ and $a_j$, equation 15) may be calculated using the method of maximum likelihood (MLE). When the calibration data set includes censored data, implementation of MLE also is known as tobit regression (Helsel and Hirsch, 2002). As with OLS, tobit regression assumes that model residuals are normally distributed with constant variance. Given the model coefficients provided by regression, estimates of instantaneous load may be obtained by retransforming equation 15. When the calibration data set is uncensored, the bias correction factor of Bradu and

Mundlak (1970) provides a minimum variance unbiased estimate (MVUE) of instantaneous load (Cohn, Delong, Gilroy, Hirsch and Wells, 1989):

$$\hat{L}_{MVUE} = exp\left(a_0 + \sum_{j=1}^{M} a_j X_j\right) g_m(m, s^2, V) \quad (17)$$

Where $\hat{L}_{MVUE}$ is the MLE estimate of instantaneous load, *m* is the number of degrees of freedom, $s^2$ is the residual variance, and *V* is a function of the explanatory variables (Cohn et al, 1989). The model coefficients in equation 17 ($a_0$ and $a_j$) are estimated by maximum likelihood; the bias correction factor [$g_m$ ($m$,$s^2$,$V$)] is an approximation of the infinite series. Within LOADEST, $g_m$ ($m$,$s^2$,$V$) is replaced by a similar function, phi (Likes, 1980).Under the MLE method, estimates of instantaneous load are developed for all of the observations in the estimation dataset using equation 17. Mean load estimates for various time periods then are calculated using equation 14. Standard errors reflecting the uncertainty in each estimate of mean load are calculated by using the method described by Likes (1980) and Gilroy, Hirsch, and Cohn (1990).

b) Adjusted Maximum Likelihood Estimation (AMLE): For the case of censored data, model coefficients estimated by tobit regression exhibit first-order bias. In addition, the Bradu-Mundlak bias correction factor (*gm*, equation 17) results in biased estimates of instantaneous load. By using adjusted maximum likelihood estimation (AMLE, Cohn 1988; Cohn, Gilroy, and Baier, 1992b), first order bias in the model coefficients is eliminated using the calculations given in Shenton and Bowman (1977). A "nearly unbiased" (Cohn, 1988) estimate of instantaneous load then is given by:

$$\hat{L}_{AMLE} = exp\left(a_0 + \sum_{j=1}^{M} a_j X_j\right) H(a, b, s^2, \alpha, k) \quad (18)$$

Where is the AMLE estimate of instantaneous load, *a* and *b* are functions of the explanatory variables (Cohn et al., 1992b), α and κ are parameters of the gamma distribution, and

$s^2$ is the residual variance. The model coefficients in equation 18 ($a_0$ and $aj$) are maximum likelihood estimates corrected for first-order bias; the bias correction factor [$H$ ($a$, $b$, $s^2$, $\alpha$, $\kappa$)] is an approximation of the infinite series given in Cohn et al. (1992b).

Under AMLE, estimates of instantaneous load are developed for all of the observations in the estimation data set using equation 18. Mean load estimates for various time periods then are calculated using equation 14. The uncertainty associated with each estimate of mean load is expressed in terms of the standard error (SE) and the standard error of prediction (SEP). The SE for each mean load estimate (Cohn et al., 1992b) represents the variability that may be attributed to the model calibration (parameter uncertainty). Calculation of the SEP begins with an estimate of parameter uncertainty (the SE) and adds the unexplained variability about the model (random error). Because SEP incorporates parameter uncertainty and random error, it is larger than SE and provides a better description of how closely estimated loads correspond to actual loads. The SEP is therefore the preferred method of describing uncertainty in loads and is used within LOADEST to develop 95 percent confidence intervals for each estimate of mean load.

c) Least Absolute Deviation (LAD): All of the regression methods discussed thus far (OLS, MLE, AMLE) assume the model residuals are normally distributed with constant variance. When model residuals do not conform to the assumption, alternate techniques may be appropriate. One such technique, the least absolute deviation (LAD) method, is implemented within LOADEST. Model coefficients for LAD are developed using the regression method of Powell (1984), as implemented by Buchinsky (1994).

Given the model coefficients, estimates of instantaneous load are developed using the "smearing" approach of Duan (1983):

$$\hat{L}_{AMLE} = exp\left(a_0 + \sum_{j=1}^{M} a_j X_j\right) \frac{\sum_{k=1}^{n} \exp\,(e_k)}{n} \qquad (19)$$

Where is the LAD estimate of instantaneous load, $a0$ and $aj$ are model coefficients developed by the LAD regression, $e$ is the residual error, and $n$ is the number of uncensored observations in the calibration data set. LAD estimates of instantaneous load are developed for all of the observations in the estimation data set using equation 19. Mean load estimates for various time periods then are calculated using equation 14. Standard errors reflecting the uncertainty in each estimate of mean load are calculated using the jackknife method described by Efron (1982).

Summary of MLE, AMLE, and LAD for Load Estimation**:** The primary load estimation method used within LOADEST is AMLE. AMLE has been shown to have negligible bias when the calibration data set is censored. For the special case where the calibration data set is uncensored, the AMLE method converges to MLE, resulting in a minimum variance unbiased estimate of constituent loads. MLE estimates are provided as a check on AMLE results and as a means of comparing LOADEST results with standard statistical packages that implement MLE. AMLE and MLE results are contingent upon the assumption that model residuals are normally distributed. Following model formulation and calibration, AMLE residuals should be examined to see if the normality assumption is valid. Checks for normality include calculation of the probability plot correlation coefficient (Vogel, 1986) and Turnbull-Weiss likelihood ratio (Turnbull and Weiss, 1978) statistics, construction of a normal-probability plot (Helsel and Hirsch, 2002), and examination of standardized residuals. If the residuals do not adhere to the assumption of normality, AMLE (and MLE) results for censored data may not be optimal. Load estimates from the LAD method should therefore be considered in lieu of AMLE, as the LAD load estimates are not dependent on the normality assumption.

Multicollinearity and Centering: Multicollinearity arises when one of the explanatory variables (equation 15) is related to one or more of the other explanatory variables (Helsel and Hirsch, 2002). The presence of collinear explanatory variables is undesirable because it confounds interpretation of model coefficients and tests of their significance. Causes of multicollinearity include natural phenomena, such as a positive relation between explanatory variables based on streamflow and precipitation, as well as mathematical artifacts, when one explanatory variable is a function of another explanatory variable. This latter cause is common in load estimation problems, when quadratic terms based on decimal time or log streamflow are included in the regression model. In such a case, explanatory variables may be centered to eliminate the collinearity. The center of the calibration data, is given by (Cohn et al, 1992a):

$$\tilde{T} = \bar{T} + \frac{\sum_{k=1}^{N}(T-\bar{T})^3}{\sum_{k=1}^{N}(T-\bar{T})^3} \quad (20)$$

Where *N* is the number of observations in the calibration data set, is the mean of the data, and *T* is the quantity to be centered (decimal time or log streamflow). Within LOADEST, is subtracted from *T*, and the resulting "centered" values are used to develop the linear (decimal time, log streamflow) and quadratic (decimal time squared, log streamflow squared) explanatory variables. As a result, the linear and quadratic terms are orthogonal and no longer collinear.

Model Selection**:** LOADEST includes several predefined models that specify the form of the regression equation (the right-hand side of equation 15). These models may be selected by the user based on the user's knowledge of the hydrologic and biogeochemical system. Alternatively, the software provides an automated model selection option that selects the "best" model from the set of predefined models. Under this option, AMLE is used to determine model coefficients and estimates of log load (equation 15) for each predefined model.

Two statistics, the Akaike Information Criterion (AIC) and the Schwarz Posterior Probability Criterion (SPPC), then are computed for the calibrated model (Judge et al., 1988). The predefined model with the lowest value of the AIC statistic then is selected for use in load estimation (values of SPPC are provided for comparative purposes only and are not used directly in the model selection process)

**3.3.2 Interpolation method.** Five different methods were used in this study to calculate annual loading using intermittent concentration and flow data. They are as follows:

(1). Method 1: Product of means of sampled $C_i$ and $Q_i$

$$Load = K(\sum_{i=1}^{n} \frac{C_i}{n})(\sum_{i=1}^{n} \frac{Q_i}{n}) \quad (21)$$

(2). Method 2: Mean of instantaneous fluxes

$$Fi = Ci * Qi \quad (22)$$

$$Load = K \sum_{i=1}^{n} \frac{C_i Q_i}{n} \quad (23)$$

(3). Method 3: Constant concentration hypothesis around sample

$$Load = K \sum_{i=1}^{n} C_i \overline{Q_{pi}} \quad (24)$$

(4). Method 4: Product of means of sampled $C_i$ and annual discharge $\overline{Q_r}$

$$Load = K(\sum_{i=1}^{n} \frac{C_i}{n}) \overline{Q_r} \quad (25)$$

(5). Method 5: Flow-weighted mean concentration

$$Load = \frac{K \sum_{i=1}^{n} (C_i Q_i)}{\sum_{i=1}^{n} Q_i} \overline{Q_r} \quad (26)$$

Where, K=conversion factor to take account of period of record

$C_i$ = instantaneous concentration associated with individual samples ($m^3$/l)

$Q_i$ = instantaneous discharge at time of sampling ($m^3$/s)

$\overline{Q_r}$ = mean discharge for period of record ($m^3$/s)

$\overline{Q_{pi}}$ =mean discharge for intervals between samples ($m^3/s$)

**3.3.3 Performance evaluation.** Three performance evaluators were used in this study and they are follows:

a) Partial Load Factor (PLF): The PLF is obtained by dividing long term average estimated data by long term average measured data. The performance is also measured by comparing the estimated data with the observed data.

b) Nash-Sutcliffe Efficiency (NSE): The NSE coefficient is used to evaluate the forecasting accuracy of hydrological models. It is defined as:

$$E = 1 - \frac{\sum_{t=1}^{T}(Q_0^t - Q_m^t)^2}{\sum_{t=1}^{T}(Q_0^t - \overline{Q_0})^2} \quad (27)$$

Where $Q_o$ is observed discharge, and $Q_m$ is modeled discharge. $Q_o^t$ is observed discharge at time *t*.

The value of NSE can range from $-\infty$ to 1. An efficiency of 1 ($E = 1$) corresponds to a perfect match of modeled discharge to the observed data. An efficiency of 0 ($E = 0$) indicates that the model predictions are as accurate as the mean of the observed data, whereas an efficiency less than zero ($E < 0$) occurs when the observed mean is a better predictor than the model or, in other words, when the residual variance (described by the numerator in the expression above), is larger than the data variance (described by the denominator).

c) Coefficient of Determination ($R^2$): This evaluator describes the degree of collinearity between simulated and measured data. $R^2$ describes the proportion of the variance in measured data explained by the model. $R^2$ ranges from 0 to 1, with higher values indicating less error variance, and typically values greater than 0.5 are considered acceptable (Santhi, Arnold, Williams, Dugas, and Hauck, 2001; Van Liew, Arnold, and Garbrecht, 2003). Although $R^2$ have been widely used for model evaluation, its statistics is oversensitive to high extreme values

(outliers) and insensitive to additive and proportional differences between model predictions and measured data.

### 3.4 Multivariate Techniques

**3.4.1 Factor analysis and principal component analysis (FA/PCA).** Factor Analysis is one of the methods to reduce the dimension of the datasets by choosing parameters contributing the high percentage of variance. PCA was one of the techniques used for extraction of factors and therefore used in this study.

PCA is a data reduction technique used to reduce the dimension of data without losing much of its variation. PCA is affected by different scaling technique so; in this study effect of different scaling method on PCA was also examined. But as studies (Singh et al. 2004) used z-transformation as method of scaling, data scaled by z-scaling was used for analysis by PCA.

PCA is designed to transform the original variables into new, uncorrelated variables (axes), called the principal components, which are linear combinations of the original variables. The new axes lie along the directions of maximum variance. PCA provides an objective way of finding indices of this type so that the variation in the data can be accounted for as concisely as possible (Sarbu and Pop, 2005). PC provides information on the most meaningful parameters, which describes a whole data set affording data reduction with minimum loss of original information (Helena et al., 2000).

All Principal components were used to find the most to least important parameters. Ranking of the parameters varied with use of different data pretreatment method to treat the data. Overall ranking of parameters of parameter for each location were calculated by doing cumulative average of results from each data pretreatment method. Overall ranking of parameters in the Greensboro watershed was calculated by doing the cumulative average for all

monitoring locations. Percentage contribution represented by individual parameter for each monitoring locations were calculated for each data pretreatment method as well as for overall data pretreatment methods.

**3.4.2 Cluster analysis.** Cluster analysis consists of a number of different algorithm and methods for grouping objects of similar kind into respective categories. In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Among many clustering method, the method used in this study for clustering water quality data was Agglomerative Hierarchical Clustering. Euclidean distance was used as the method of finding the distance and Ward's Method was used for linkage.

In Ward's minimum-variance method, the distance between two clusters is the *ANOVA* sum of squares between the two clusters added up over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give proportions of variance (squared semi-partial correlations). Ward's method tends to join clusters with a small number of observations, and it is strongly biased toward producing clusters with the same shape and with roughly the same number of observations. It is also very sensitive to outliers. Ward's method joins clusters to maximize the likelihood at each level of the hierarchy under the following assumptions:

- Multivariate normal mixture
- Equal spherical covariance matrices
- Equal sampling probabilities

# CHAPTER 4

## Results and Discussion

### 4.1 Rating Curve

**4.1.1 LOADEST.** LOADEST model was used in this study for construction of rating curve and load estimation. It is a model based on logarithmic regression of multiple parameters like time, concentration of parameter, discharge etc. Three main statistical performance evaluators were used in this study. They were coefficient of determination, partial load factor and Nash- Sutcliffe efficiency factor. Performance of LOADEST was evaluated for each parameter separately and values for each water quality stations are presented below. In addition to LOADEST, five more interpolation method were used for calculation of loading annually and entire time period.

There were sixteen water quality monitoring stations involved in this study but only nine water quality stations were used for load calculation and rating curve construction. Water quality data were obtained by the method of "grab sample" and these locations were chosen close to some water quality monitoring site measuring discharge on the daily basis. So, in this section of the study only data for water quality monitoring site with discharge data were used and there were nine such sites.

Six water quality parameters were used for this study. They were nitrate, nitrite, total phosphorus (TP), total dissolved solute (TDS), total suspended solute (TSS) and Total Kjeldahl Nitrogen (TKN). There were 102 water quality data points for each parameter. These data were collected on the periodic sampling strategy with one data per couple of months from 1999-2008 and monthly sampling from 2009-2012.

Table 3 summarizes the performance of LOADEST model for all the parameters for all water quality monitoring stations. The estimation of all the parameters by the model was described as below.

Table 3

*Summary of performance indicator for all water quality station*

| Station | Parameter | Nitrate | Nitrite | TDS | TKN | TP | TSS |
|---|---|---|---|---|---|---|---|
| | $R^2$ | 0.82 | 0.85 | 0.92 | 0.83 | 0.79 | 0.70 |
| | PLF | 1.16 | 1.12 | 0.94 | 1.09 | 0.90 | 0.63 |
| Aycock | NSE | 0.69 | -0.59 | 0.83 | 0.74 | 0.62 | 0.53 |
| | $R^2$ | 0.80 | 0.90 | 0.96 | 0.88 | 0.74 | 0.77 |
| | PLF | 1.06 | 0.96 | 0.99 | 1.00 | 0.94 | 0.84 |
| Battleground | NSE | 0.82 | 0.22 | 0.95 | 0.85 | 0.47 | 0.42 |
| | $R^2$ | 0.80 | 0.86 | 0.92 | 0.83 | 0.72 | 0.80 |
| | PLF | 1.13 | 0.96 | 0.97 | 1.04 | 1.02 | 0.73 |
| Church | NSE | 0.87 | 0.23 | 0.88 | 0.78 | 0.60 | 0.12 |
| | $R^2$ | 0.82 | 0.88 | 0.96 | 0.86 | 0.69 | 0.80 |
| | PLF | 0.99 | 0.95 | 0.99 | 1.02 | 0.95 | 1.24 |
| Fleming | NSE | 0.78 | 0.89 | 0.97 | 0.91 | 0.82 | 0.88 |
| | $R^2$ | 0.79 | 0.85 | 0.94 | 0.89 | 0.82 | 0.80 |
| | PLF | 1.02 | 1.04 | 0.98 | 0.99 | 0.95 | 0.53 |
| McConnell | NSE | 0.75 | -0.18 | 0.91 | 0.73 | 0.72 | 0.28 |
| | $R^2$ | 0.82 | 0.89 | 0.93 | 0.88 | 0.78 | 0.81 |
| | PLF | 1.10 | 1.06 | 0.99 | 1.07 | 0.78 | 0.65 |
| Merritt | NSE | 0.88 | -0.04 | 0.84 | 0.87 | 0.60 | 0.53 |
| | $R^2$ | 0.75 | 0.85 | 0.90 | 0.84 | 0.71 | 0.87 |
| | PLF | 1.08 | 0.96 | 1.00 | 1.05 | 1.00 | 0.94 |
| Pleasant | NSE | 0.77 | 0.21 | 0.94 | 0.81 | 0.65 | 0.78 |
| | $R^2$ | 0.79 | 0.87 | 0.94 | 0.85 | 0.77 | 0.81 |
| | PLF | 1.09 | 1.02 | 0.98 | 1.08 | 0.86 | 0.72 |
| Randleman | NSE | 0.77 | 0.11 | 0.89 | 0.76 | 0.33 | 0.10 |
| | $R^2$ | 0.33 | 0.51 | 0.87 | 0.75 | 0.61 | 0.78 |
| | PLF | 0.95 | 0.86 | 0.98 | 0.99 | 1.09 | 0.49 |
| Rankin | NSE | 0.27 | 0.31 | 0.77 | 0.77 | 0.55 | 0.53 |
| | $R^2$ | 0.82 | 0.89 | 0.94 | 0.91 | 0.77 | 0.79 |
| | PLF | 1.05 | 1.10 | 0.96 | 0.99 | 0.93 | 1.02 |
| W.JJ | NSE | 0.72 | -0.44 | 0.93 | 0.79 | 0.50 | 0.36 |

Nitrate: $R^2$ value for nitrate ranged from32.82% to 82.49% (Figure 2). Only one water quality station, Rankin, had bad $R^2$ value of 32.82% otherwise all other water quality station had value higher than 75%. Partial load factor, measure of accuracy of load estimation, varied from 0.94 to 1.15. So, the range of estimation varying from the true loading was -6% to 16%. LOADEST overestimated the loading of nitrate in 80% of water quality station and rest of the time it under predicted the loading. So, mostly it showed the bias towards over prediction. Nash-Sutcliffe efficiency (NSE) varied from 26.53% to 88.42%. Only water quality station at Rankin had very poor value of 26.53% otherwise NSE value for other water quality station varied from 68.52% to 88.42%. Water quality station at Rankin seemed to have poor results for $R^2$ and NSE as well as only one of station which under predicted the loading.

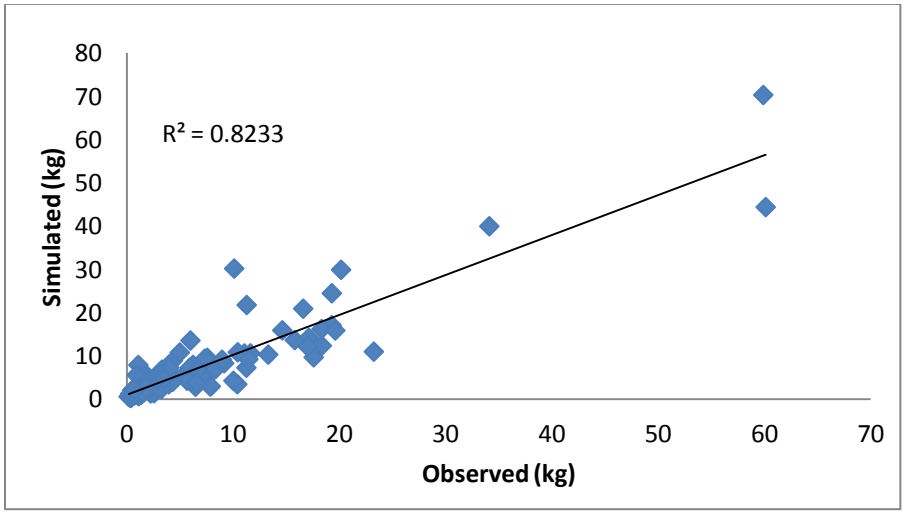


*Figure 2.* Nitrate: scatter plot for observed value vs. simulated value for water quality station at Battleground Avenue

Nitrite: $R^2$ values for nitrite were in mid to high eighties (Figure 3) with the exception of water quality station at Rankin, where $R^2$ value was 50.78%. For PLF, range varied from 0.86 to 1.12, so the range of estimation differing from the true loading was between -14% to +12%. LOADEST did not give definite trend of bias for nitrite as 50% of time load was underestimated

and remaining 50% of the time it overestimated the load. NSE value for nitrite was very poor for all of the water quality station except water quality station at Fleming St. NSE value for all the monitoring station with PLF value less than one was negative and opposite was true for PLF with value more than one. Range of NSE was -58.91% to 89.2%. Minimum value of NSE was for the water quality station with highest value of PLF.

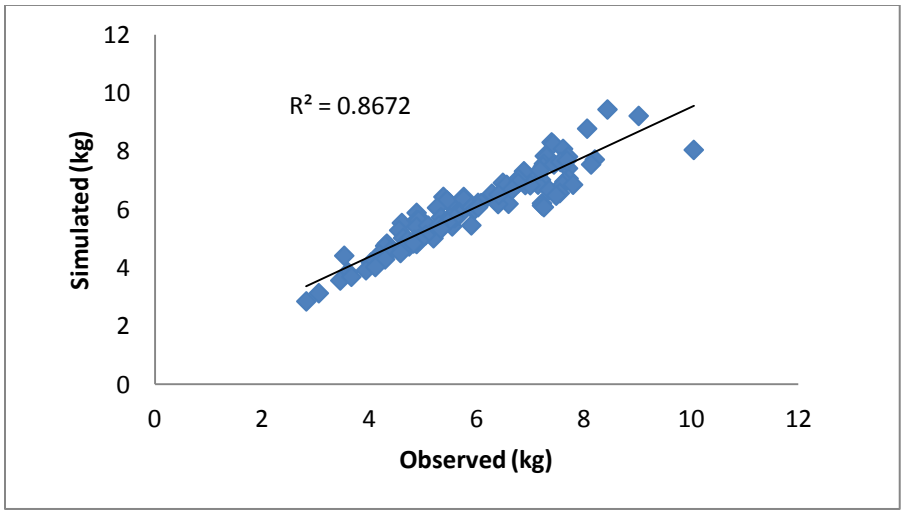


*Figure 3.* Nitrite: scatter plot for observed value vs. simulated value for water quality station at Battleground Avenue

Total Dissolved Solids (TDS): Performance of LOADEST for TDS was very good with almost all performance evaluator performing very well. Coefficient of determination indicated good performance for TDS by LOADEST. Its value ranged from 87.38% to 96.33% (Figure 4). NSE values were also fairly good with the range of 77.17% to 96.96%. PLF indicated estimated load was close to the true load with value ranging from 0.93 to 1. So, estimated value differed from true value by -6% to 0%. Except from one water quality station at Pleasant ridge all the other station had underestimated the load.

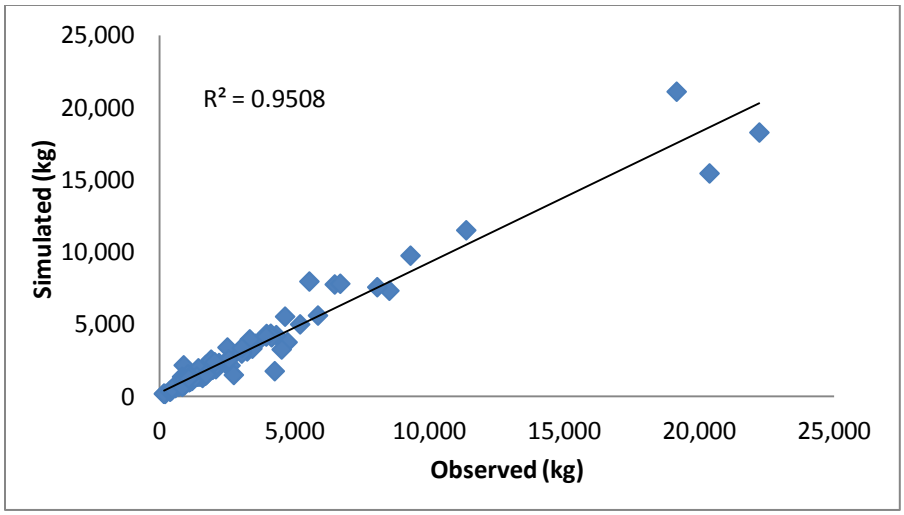


*Figure 4.* TDS: scatter plot for observed value vs. simulated value for water quality station at Battleground Avenue

Total Kjeldahl Nitrogen (TKN): $R^2$ value for TKN was very good with value ranging from 75.34% to 90.86%. Figure 5 showed the scatter plot between observed and simulated loading for water quality station at Battleground Avenue. NSE value fell in the interval of 72.51% to 90.53%. TKN had very good values for both $R^2$ and NSE simultaneously and represented very good performance of LOADEST. For PLF values varied from 0.98 to 1.09. So, estimated value varied from true value by -2% and 9%. Model underestimated the loading 30% of the time and for the rest of 70% of time it overestimated the annual loading.

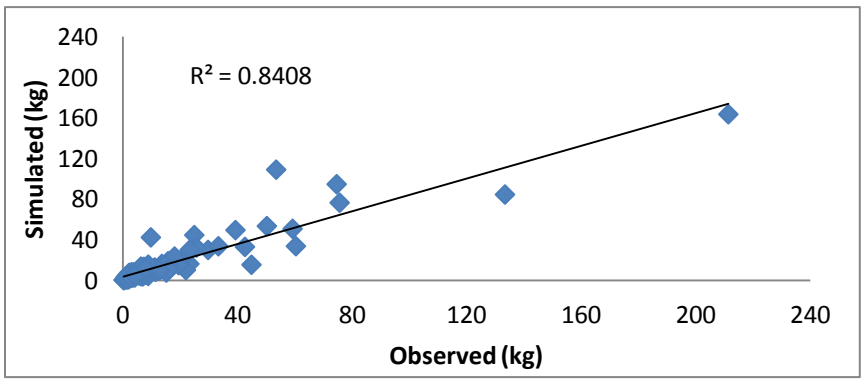


*Figure 5.* TKN: scatter plot for observed value vs. simulated value for water quality station at Battleground Avenue

Total Phosphorus (TP): $R^2$ value varied from 60.65% to 81.69%. Figure 6 shows scatter plot between natural log of observed and simulated value for water quality station at Battleground Avenue. NSE value ranged from poor 32.86% to fairly good value of 81.88%. PLF showed that estimated loading varied from true loading in the range of 0.78 to 1.09. Estimated loading varied from true loading by -22% to 9%. Eight out of ten water quality stations LOADEST under predicted annual loading and two stations overestimated the values.

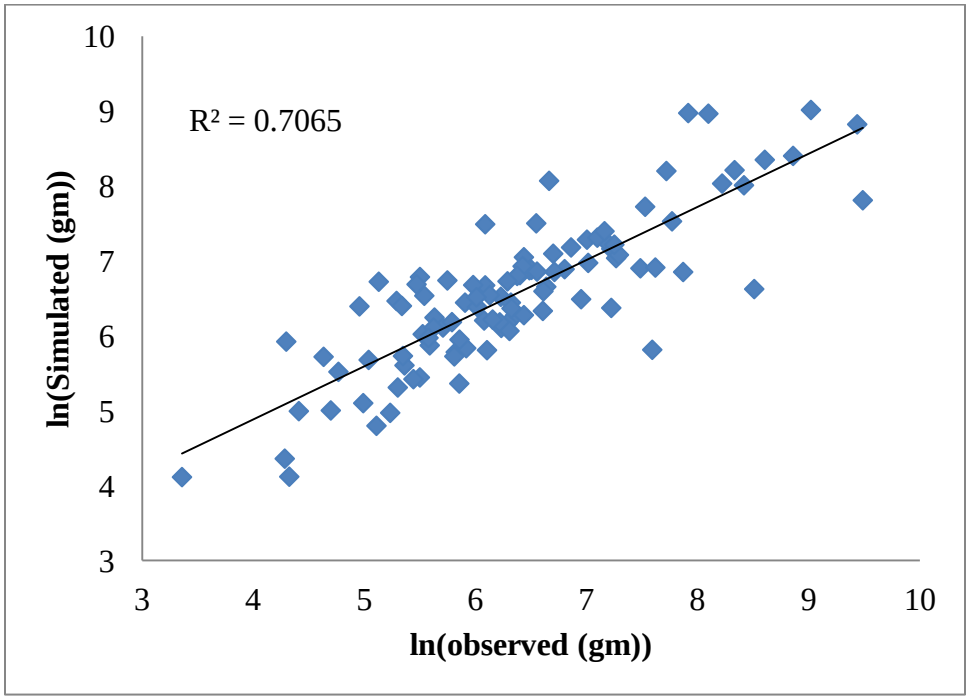


*Figure 6.* TP: scatter plot for observed value vs. simulated value for water quality station at Battleground Avenue

Total Suspended Solids (TSS): Performance of LOADEST for TSS was pretty poor in view of the performance indicator. Even though value of $R^2$ was fairly good with the range of 70.31% to 87.1%, value of NSE and PLF were poor. NSE varied from 10.37% to 87.64% while PLF had a range of 0.49 to 1.23. TSS varied from the true loading by -51% to 23%. LOADEST in general underestimated the loading with 80% of time estimated value being lower than observed loadings.

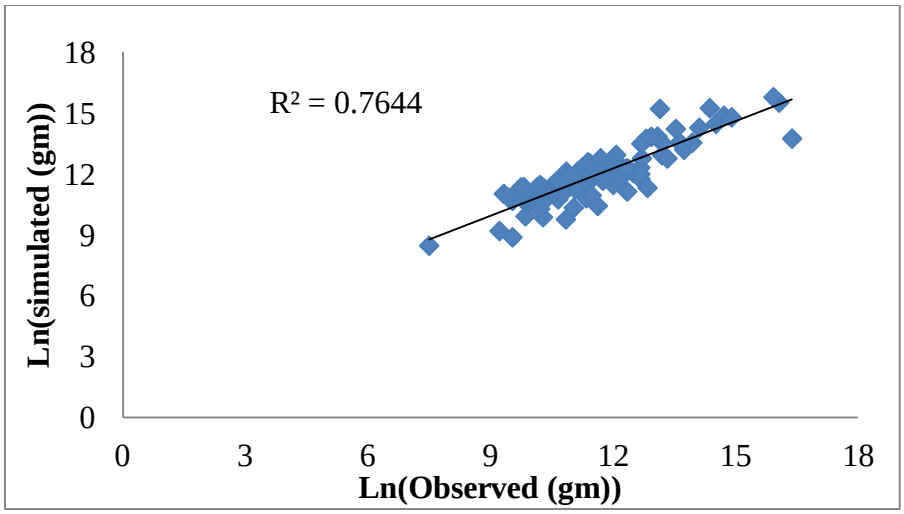


*Figure 7.* TKN: scatter plot for observed value vs. simulated value for water quality station at Battleground Avenue

**4.1.2 Correlation.** Three performance indicators were used to evaluate the performance of the LOADEST model in this study. So, possible correlations between these indicators were studied by plotting the graph. Figure 8(a), (b), (c) and Table 4 summarizes and presents the correlation between these indicators.

Table 4

*Correlation between performance indicators*

| Parameter | NSE-$R^2$ | PLF-$R^2$ | NSE-PLF |
|---|---|---|---|
| Nitrate | 0.86 | 0.05 | 0.35 |
| Nitrite | 0.03 | 0.38 | 0.65 |
| TDS | 0.46 | 0 | 0.13 |
| TKN | 0.04 | 0 | 0 |
| TP | 0 | 0.51 | 0.04 |
| TSS | 0.02 | 0.07 | 0.24 |

Nitrate: Correlations between all three statistical performance evaluator were studied by plotting the value of one against another. Graphs showed there was no correlation between PLF and coefficient of determination and $R^2$ value between them was close to zero. Correlation

between PLF and NSE was also very poor with $R^2$ value of 0.35 where as correlation between coefficient of determination and NSE was strongest with value for $R^2$ being 0.91.

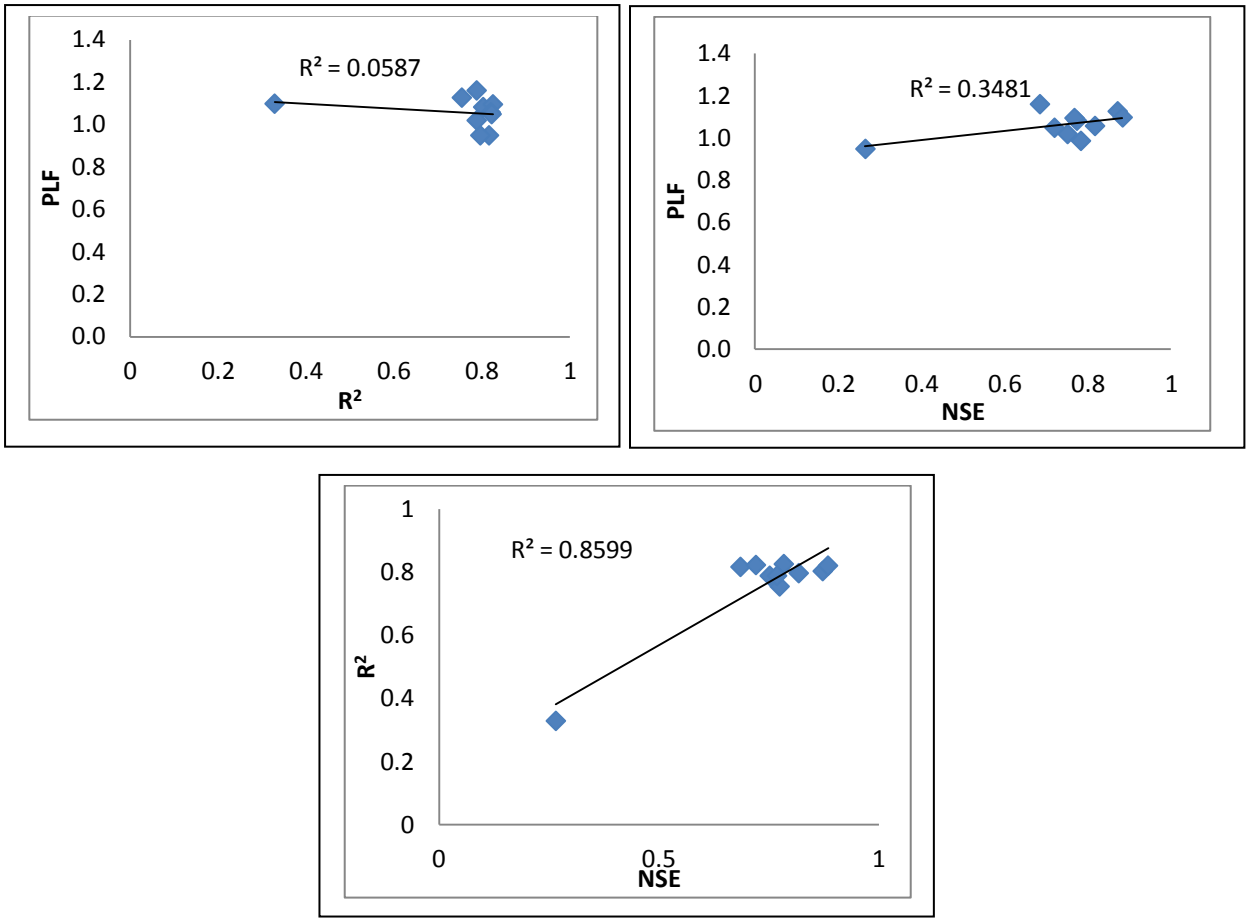


*Figure 8.* Scatter plot for nitrate between (a) PLF Vs $R^2$ (b) PLF Vs NSE (c) NSE Vs $R^2$

Nitrite: Scatter plot of performance evaluators were studied for correlation between them. NSE and PLF showed strong correlation with $R^2$ value of 0.65. Correlation between PLF and $R^2$ was fairly weak with value of 0.34 for coefficient of determination but there was almost no correlation between NSE and $R^2$. For PLF and NSE, the trend line suggested that, with increase in value of PLF resulted in decrease in value of NSE.

Total Dissolved Solids (TDS): PLF value was plotted against NSE and $R^2$ values to study its correlation with them. PLF did not showed any correlation with them with value of coefficient of determination close to zero. NSE and $R^2$ showed positive correlation with value of coefficient of determination of 0.46.

Total Kjeldahl Nitrogen (TKN): For TKN, there was no correlation between the different performance indicators. All indicators were independent of each other with values of coefficient of determination being zero for them.

Total Phosphorus (TP): PLF and $R^2$ showed average degree of correlation with the value of 0.51 for coefficient of determination. There was no correlation between NSE-$R^2$ and NSE-PLF with value of coefficient of correlation being zero.

Total Suspended Solids (TSS): Performance indicators for TSS were independent of each other with value of coefficient of determination close zero for them.

**4.1.3 Time series analysis.** For the visual inspection of performance of the rating curve, time series graph was plotted for all the parameters for the water quality station located at Battleground Avenue. The results and conclusion from visual analysis were presented as follows:

Nitrate: For nitrate at Battleground Ave. all performance indicator were fairly good. Figure 9 shows time series plot between AMLE and observed concentration and it showed approximately half of the observation points falling on the rating curve constructed by the LOADEST model. Points not falling on the rating curve were evenly distributed on the top and bottom of rating curve. But in latter part of the rating curve between 2007 and 2012, more observed value were below the rating curve than above it with some extreme value. From the visual analysis rating curve and observed value it is clear that with time concentration of nitrate in stream was decreasing.

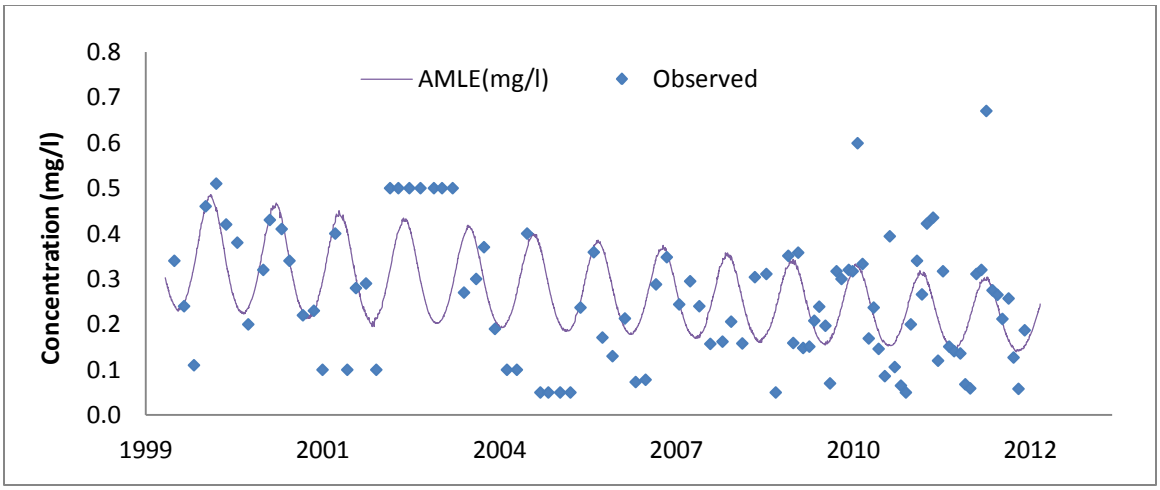


*Figure 9.* Nitrate: time series plot between AMLE and observed Value

Nitrite: Even though PLF and $R^2$ were very good, NSE value was poor for nitrite and got reflected in the time series graph, Figure 10, which showed many points falling outside the rating curve. One of the big reasons for the poor match between the observed value and rating curve was quality of data which just dropped from 0.1 to 0.01 mg/l and stayed there for remaining of the time period.

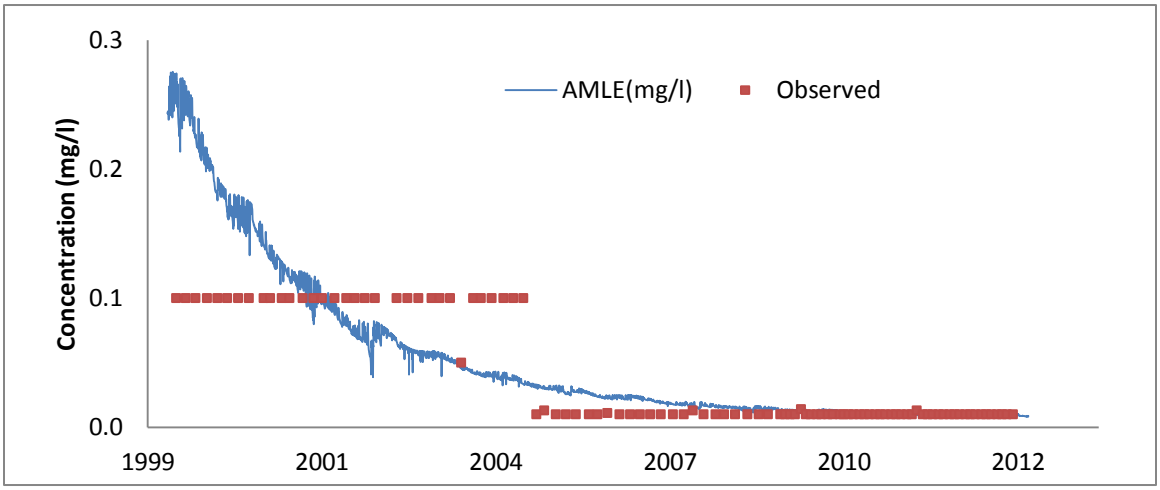


*Figure 10.* Nitrite (a) scatter plot for observed value vs. simulated value (b) time series plot between AMLE and observed value

Rating curve which is a function of discharge, varied with time but observed value of concentration remained constant irrespective of value of discharge which pointed to issue with quality of data. If we look at the rating curve we can see the concentration of nitrite in the stream was decreasing with the passage of time.

Total Dissolved Solids (TDS): **P**erformance indicator for TDS was excellent with all value close to one. It was also reflected on time series graph (Figure 11). Very high percentage of observed value fell on the rating curve, which was very wavy with many undulations. Points which did not fell on the curve were evenly distributed on top and bottom of rating curve. Visual inspection of the graph shows clear gradual increase in concentration of TDS in the river with time.

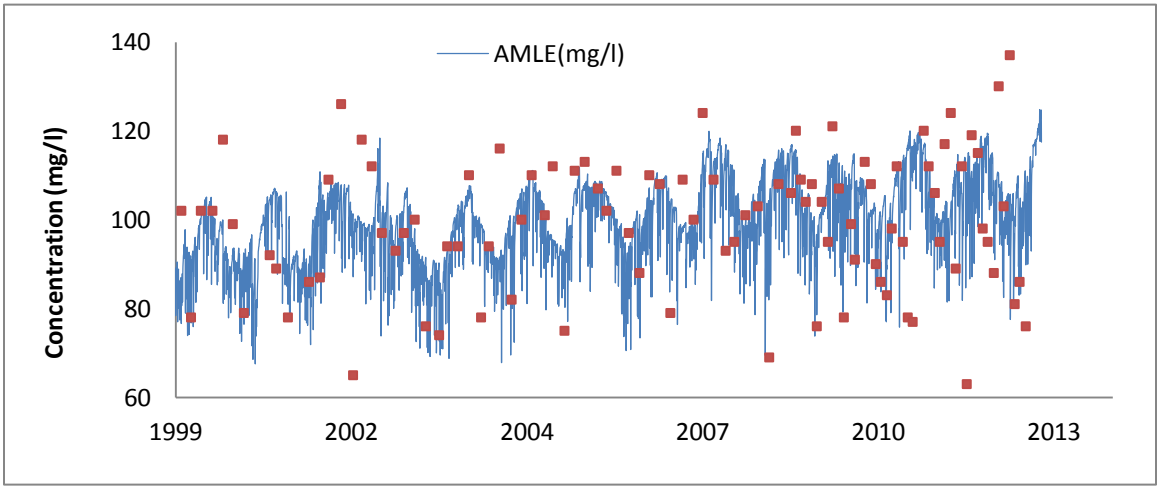


*Figure 11.* TDS: time series plot between AMLE and observed value

Total Kjeldahl Nitrogen (TKN): All the performance indicators for TKN were excellent. High percentage of observed value fell around rating curve. From Figure 12, we can see that from 1999 up to 2006, concentration of TKN was gradually decreasing. After 2006, concentration slowly started to increase. By the time of end of 2012, the concentration had risen

above the initial value of concentration at 1999. The performance of LOADEST seemed to improve after 2006 with more observed point falling on the rating curve.

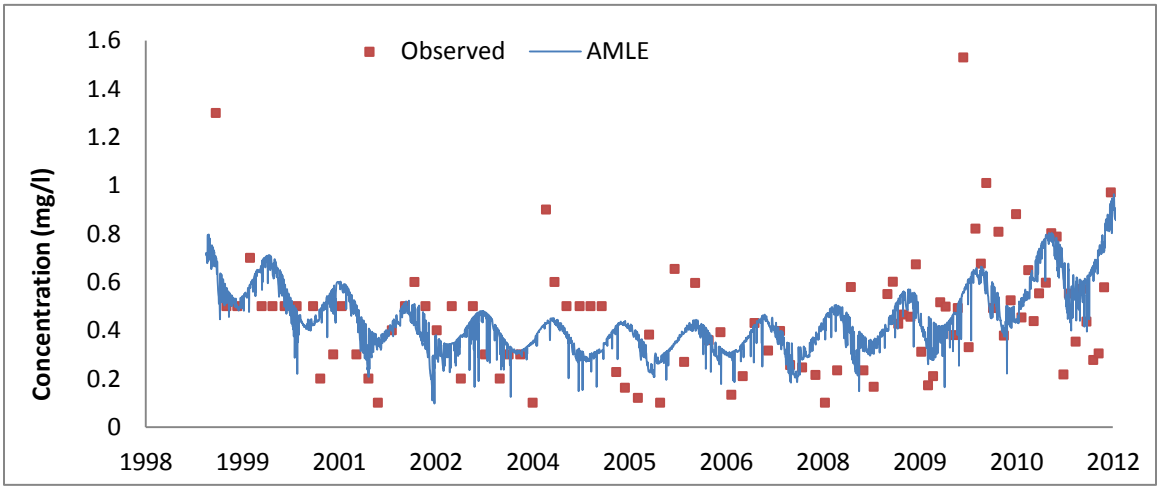


*Figure 12.* TKN: time series plot between AMLE and observed value

Total Phosphorus (TP): All three performance indicators were poor for TP. PLF showed under prediction and but from the Figure 13, it could be seen that, before 2006, rating curve predominantly over predicted the value whereas after that many observed value was below the rating curve with few extremely high value above rating curve. From trend analysis of the rating curve, it could be seen that concentration of TP was decreasing with time.

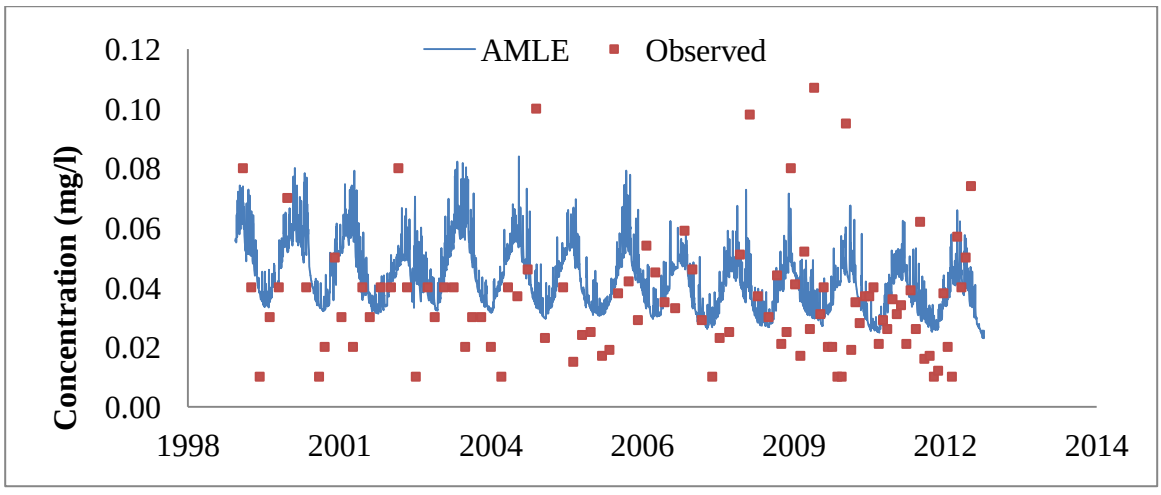


*Figure 13.* Total Phosphorus: time series plot between AMLE and observed value

Total Suspended Solids (TSS): Time series graph showed that most of the time observed value was below the rating curve, so the rating curve was overestimating the concentration of TSS (Figure 14). But from the PLF value we know that overall loading calculated by LOADEST was below the observed loading. This shows the hazards of using the monthly data to calculate the annual data. Trend analysis showed that with time concentration of TSS was increasing.

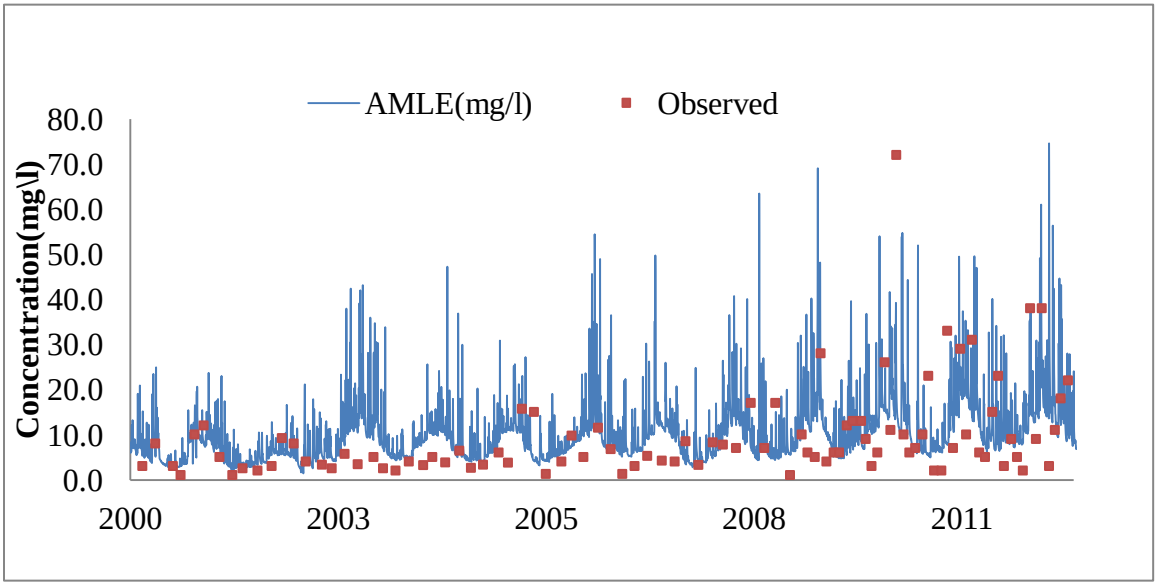


*Figure 14.* TSS: time series plot between AMLE and observed value

**4.2 Comparison of Different Methods.**

LOADEST and five different averaging methods were used for the calculation of the loading for particular time period. In this study annual loadings were calculated for entire time period for all the six parameters. Summary of the results for parameters of water quality station at Aycock St. were presented in the Table 5. Tables for the remaining sites are in Appendix 2. For the visual study, boxplots were plotted for water quality station at Aycock St. for all the parameters and presented below.

Table 5

*Summary of min, max and standard deviation of all parameters estimated by all methods for water quality station at Aycock St. (Other stations are in appendix)*

| Parameters | | M1 (kg) | M2 (kg) | M3 (kg) | M4 (kg) | M5 (kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std | 1249 | 1372 | 2852 | 2331 | 2354 | 703 |
| | Min | 162 | 160 | 975 | 1070 | 1057 | 2392 |
| Nitrate | Max | 4503 | 4419 | 12623 | 10384 | 10396 | 5119 |
| | Std | 236 | 272 | 1248 | 932 | 1297 | 774 |
| | Min | 13 | 13 | 79 | 76 | 76 | 127 |
| Nitrite | Max | 833 | 872 | 4554 | 3296 | 4711 | 2470 |
| | Std | 733785 | 550910 | 566574 | 452289 | 564089 | 183154 |
| | Min | 156076 | 185968 | 920348 | 907976 | 646469 | 732222 |
| TDS | Max | 3055246 | 2408827 | 3229533 | 2620629 | 2628076 | 1475844 |
| | Std | 151663 | 340565 | 213994 | 119543 | 210352 | 433852 |
| | Min | 2456 | 4740 | 15211 | 16184 | 16110 | 312000 |
| TSS | Max | 580815 | 994731 | 840609 | 393533 | 802303 | 2062736 |
| | Std | 3419 | 4447 | 4940 | 6318 | 7266 | 3760 |
| | Min | 686 | 525 | 3876 | 3140 | 3261 | 7235 |
| TKN | Max | 11285 | 13841 | 22381 | 28564 | 30780 | 22603 |
| | Std | 367 | 696 | 496 | 389 | 314 | 400 |
| | Min | 96 | 78 | 368 | 351 | 356 | 491 |
| TP | Max | 1201 | 2135 | 1881 | 1531 | 1361 | 1808 |

Nitrate: Estimation of M1 and M2 were similar. The range for M1 was 162-4503 kg per year with the standard deviation of 1249 kg per year and for M2 it was 160-4419 kg per year with the standard deviation of 1372 kg per year. Similarly, M4 and M5 had estimation value close to each other. Estimation from M4 varied between 2331-10384 kg per year with standard deviation of 1070 kg per year. For M5 estimation was between 2354-10396 kg per year with standard deviation of 1057 kg per year. M3 had the biggest range of 2852-12623 kg per year with standard deviation of 12623 kg per year. LOADEST had a very small range and smallest standard deviation. The range of estimation was 2392-5119 kg per year with the standard deviation of 703 kg per year.

Box plot, Figure 15, showed that methods M3, M4, M5 and LOADEST had similar distribution loading. From the bar chart, Figure 16, it could be seen that annual estimation made by M3, M4, M5, and LOADEST were similar and M1 and M3 were close to each other. Ranking of standard deviation for all the methods from highest to lowest were M3, M5, M4, M2, M1 and LOADEST.

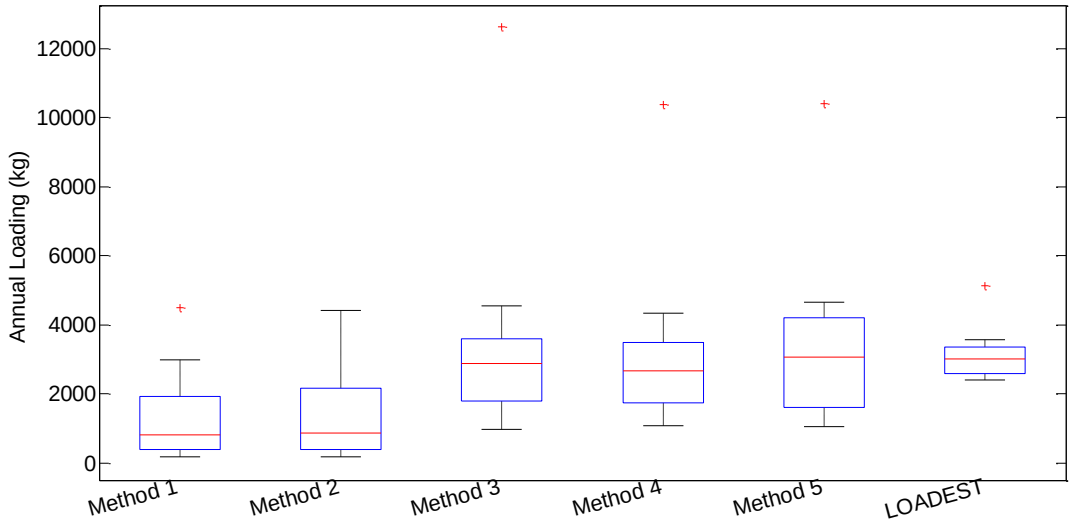


*Figure 15.* Box plot of loading calculated by all six methods for Nitrate

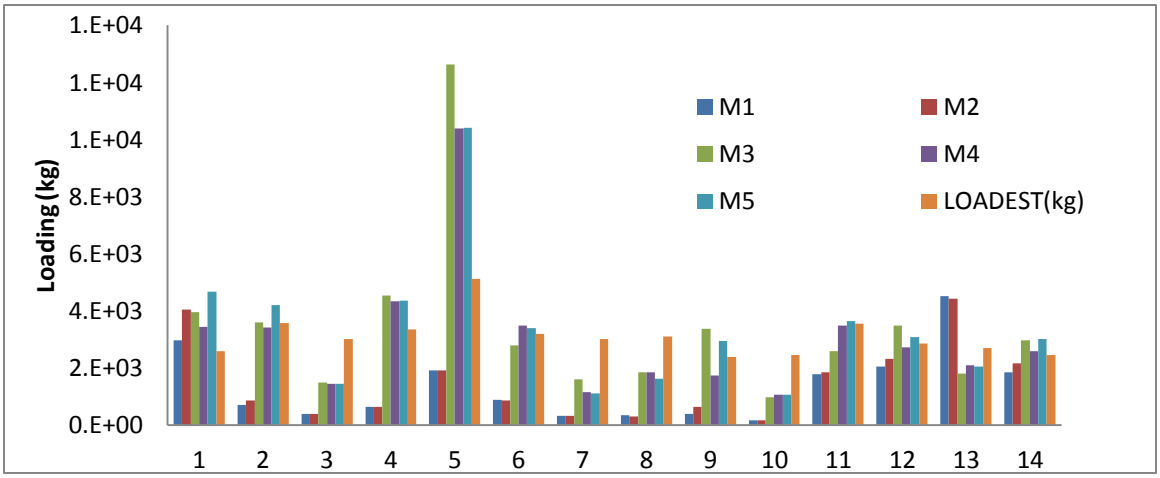


*Figure 16.* Bar chart of loading calculated by all six methods for Nitrate

Nitrite: Annual loadings of nitrite was calculated by LOADEST and five averaging methods. Range of loading calculated by M1 and M2 were similar and were between 13-833 kg

per year with the standard deviation of 236 kg per year and between 13-272 kg per year with the standard deviation of 272 kg per year respectively. Similarly M3 and M5 had similar range of estimation with the range of 79-4554 kg per year with standard deviation of 1248 kg per year and 76-4711 kg per year with standard deviation of 1297 kg per year. For M4, range was 76-3296 kg per year with standard deviation of 932 kg per year. LOADEST had a range of 774-2470 kg per year with the standard deviation of 127 kg per year.

Loadings calculated by M1 and M2 were of similar range and are of most conservative estimate. For both of them, difference between maximum and minimum value was smallest in comparison to other methods, so the result was more robust. Box plot, Figure 17, showed none of the estimation calculated by all the methods was normally distributed. From the bar chart, Figure 18, and box plot, it could be seen that annual estimation made by M3, M4, M5, and LOADEST were similar and M1 and M3 were close to each other.

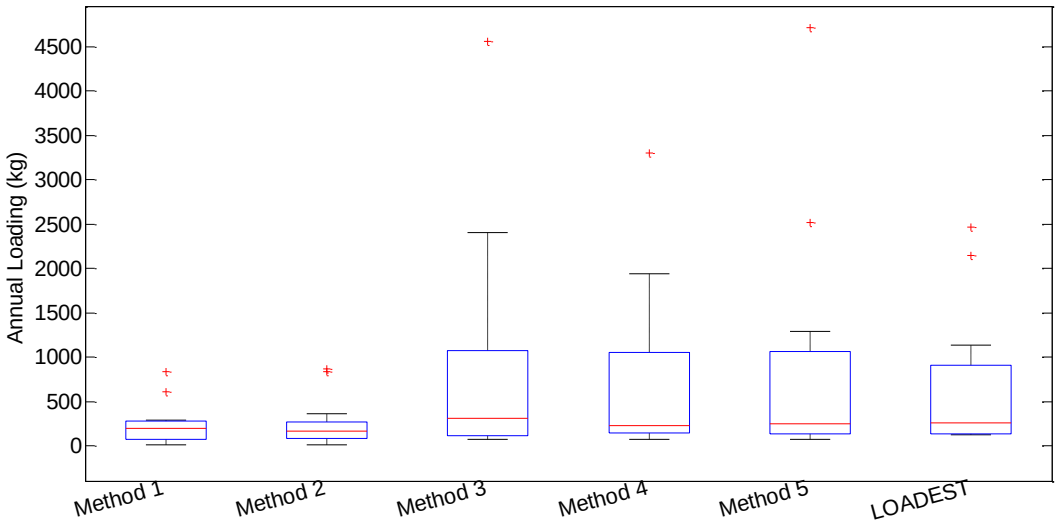


*Figure 17.* Box plot of loading calculated by all six methods for Nitrite

Load estimation of M1 and M2, and M3 and M5 were similar, with similar range and standard deviation. Ranking of standard deviation from maximum to minimum for all the

methods were M5>M3>M4>LOADEST>M2>M1. From the bar chart and box plot it could be seen estimation from all methods was similar when it fell between zeros to 50$^{th}$ percentile.

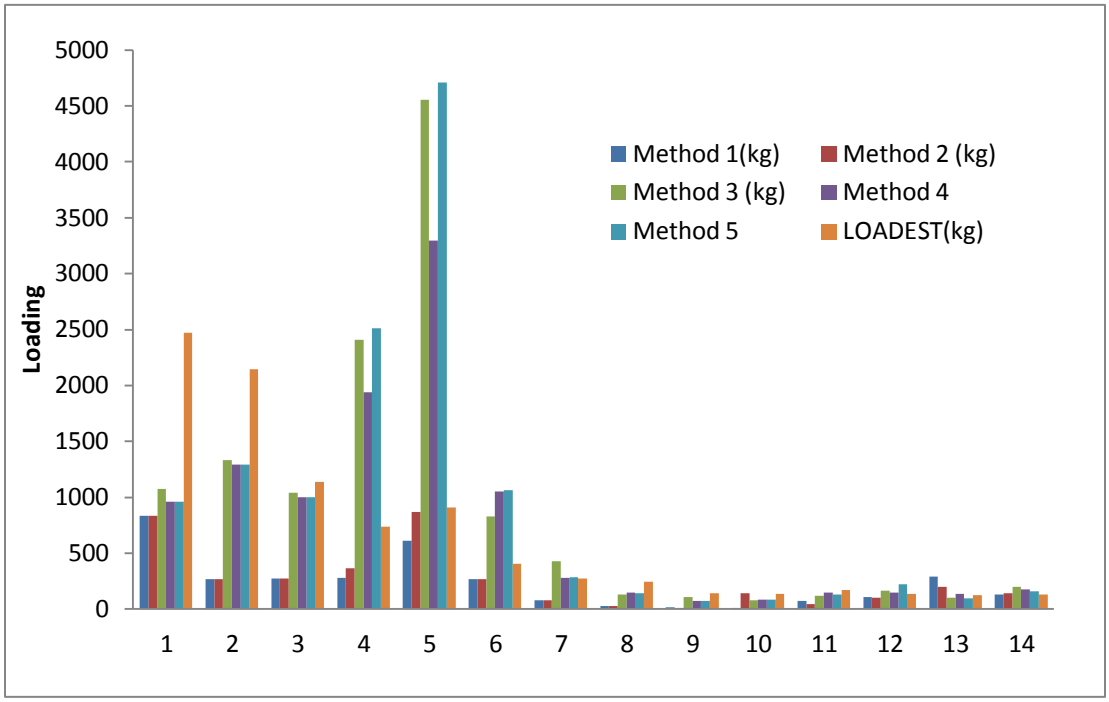


*Figure 18.* Bar Charts of loading calculated by all six methods for Nitrite

Total Dissolved Solids (TDS): Range of estimate for M1 was between 156-3055 tons per year with standard deviation of 734 tons per year. For M2, range was between 186-2409 tons per year with standard deviation of 551 tons per year. M3, M4 and M5 had similar results and their range was 186-2409, 920-3230, and 908-2621 tons per year respectively. Their standard deviations were 567, 452, and 564 tons per year respectively. LOADEST had the smallest range and standard deviation. Its range was 732-1476 tons per year with standard deviation of 184 tons per year.

From box plot, Figure 19, and bar chart, Figure 20, it could be seen that estimation of M3, M4 and M5 were similar and greater than other method most of the time followed by

LOADEST, M2 and M1. The ranking of standard deviation from maximum to minimum for all the method were in the following range: M1> M5> M4> M3> M2>LOADEST.

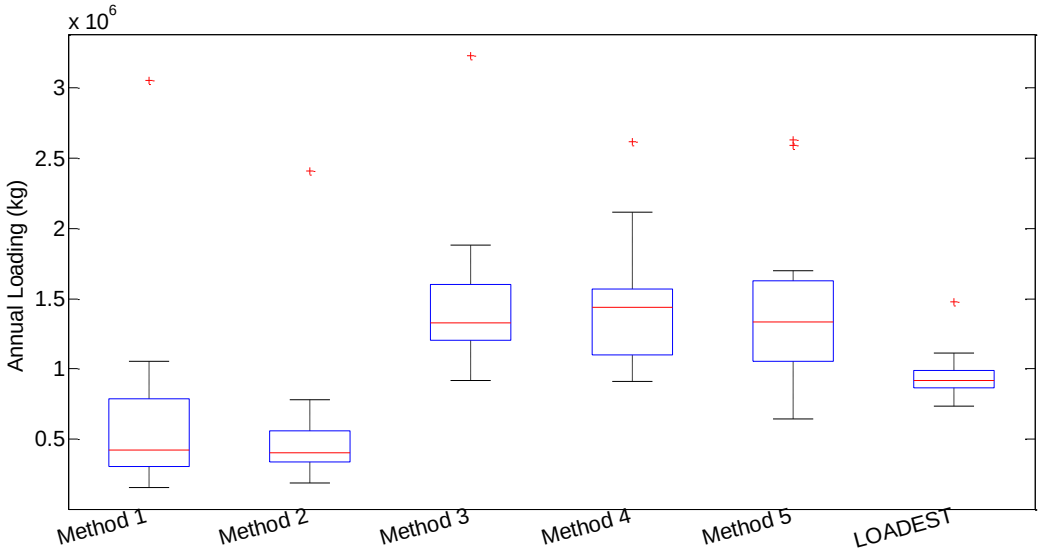


*Figure 19.* Box plot of loading calculated by all six methods for TDS

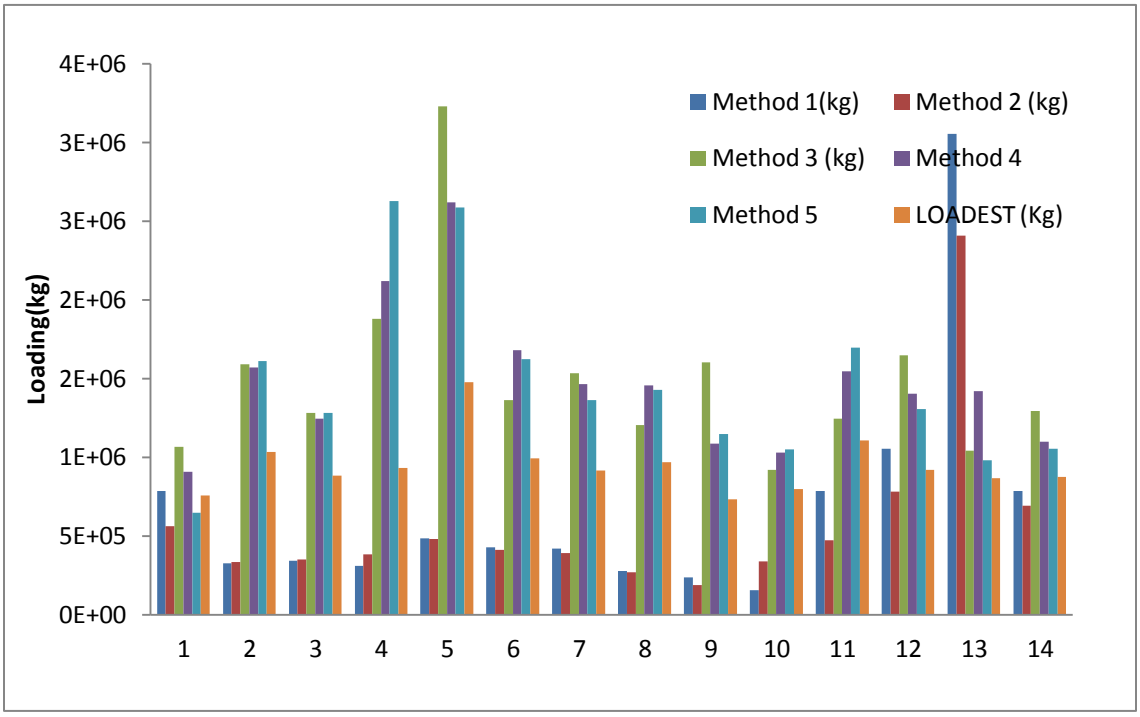


*Figure 20.* Box plot of loading calculated by all six methods for TDS

Total Kjeldahl Nitrogen (TKN): From the visual inspection of box plot (Figure 21) and bar chart (Figure 22), it could be seen that M1 and M2 had similar result and lowest estimates. M3, M4 and M5 had results similar to each other which was in the middle with LOADEST being the method with highest estimation. Range of estimation for M1 was between 686-11285 kg per year with the standard deviation of 3419 kg per year. M2 had the range 525-13841kg per year for its prediction with the standard deviation of 4447 kg per year.

M3, M4 and M5 range was between 3876-22381kg per year with standard deviation of 4940 kg per year, 3140-28564 kg per year with the standard deviation of 6318 kg per year and 3261-30780 kg per year with standard deviation of 7266 kg per year respectively. For LOADEST range of prediction was between 7235-22603 kg per year with standard deviation of 3760 kg per year.

Box plot showed that results from M3, M4 and LOADEST were normally distributed whereas for M1, M2 and M5 were skewed towards the bottom. Extreme value were also present in the estimation of M3, M4, M5 and LOADEST.

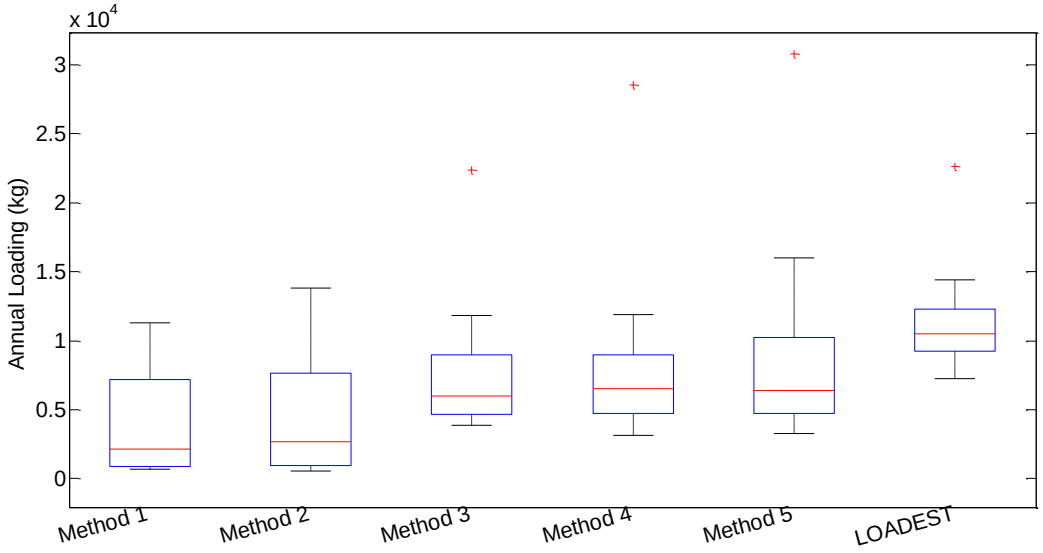


*Figure 21.* Box plot of loading calculated by all six methods for TKN

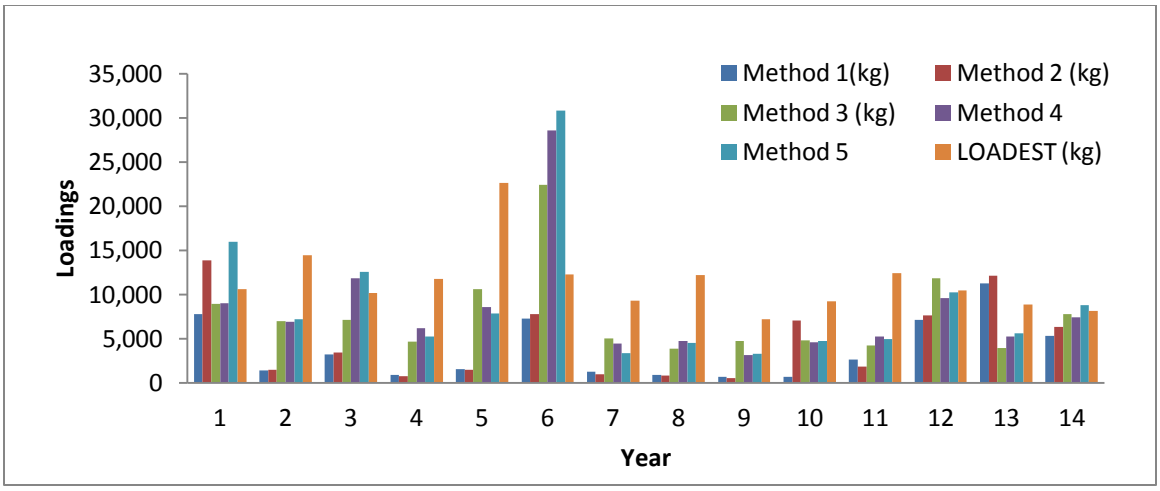


*Figure 22.* Bar Chart of loading calculated by all six methods for TKN

Total Phosphorus (TP): From the visual inspection of box plot (Figure 23), and bar chart (Figure 24), it can be seen M1 and M2 were in the same range and same was true for M3, M4 and M5. LOADEST had the maximum estimate excluding the outliers shown in the graph. The value of estimate for M1, varied from 96-1201 kg per year with standard deviation of 367 kg per year. For M2 range of estimate was between 78-2135 kg per year and its standard deviation was 696 kg per year.

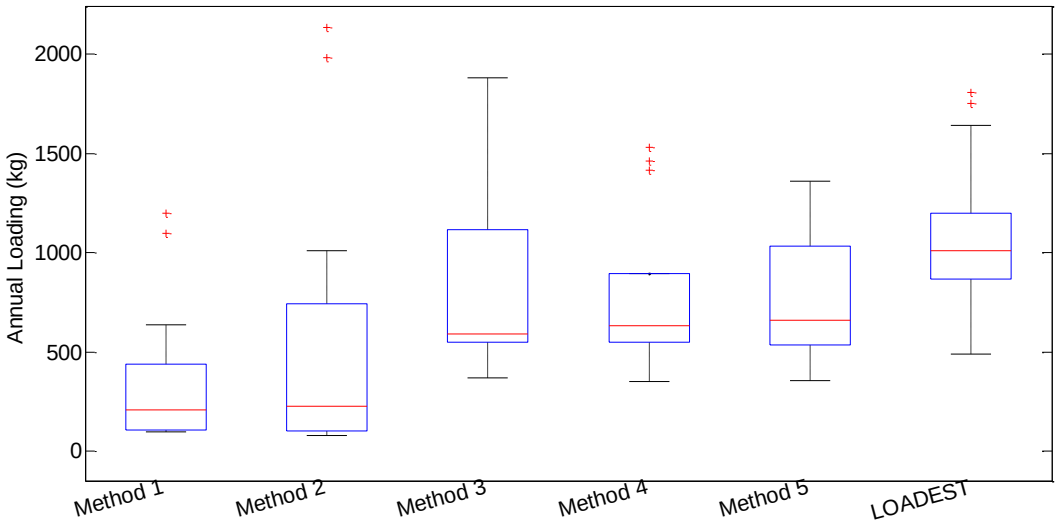


*Figure 23.* Box plot of loading calculated by all six methods for TP

M3 prediction varied between 368-1881 kg per year with standard deviation of 496 kg per year. M4 and M5 had similar range of estimation with the value of 351-1531 kg per year with standard deviation of 389 kg per year and 491-1808 kg per year with the standard deviation respectively. For LOADEST the range was 491-1808 kg per year with standard deviation of 400 kg per year. From the bar chart it can be that LOADEST had highest estimate of load for majority of the time and M1 had lowest predicted load. Except LOADEST, none of the loadings estimated by other methods were normally distributed.

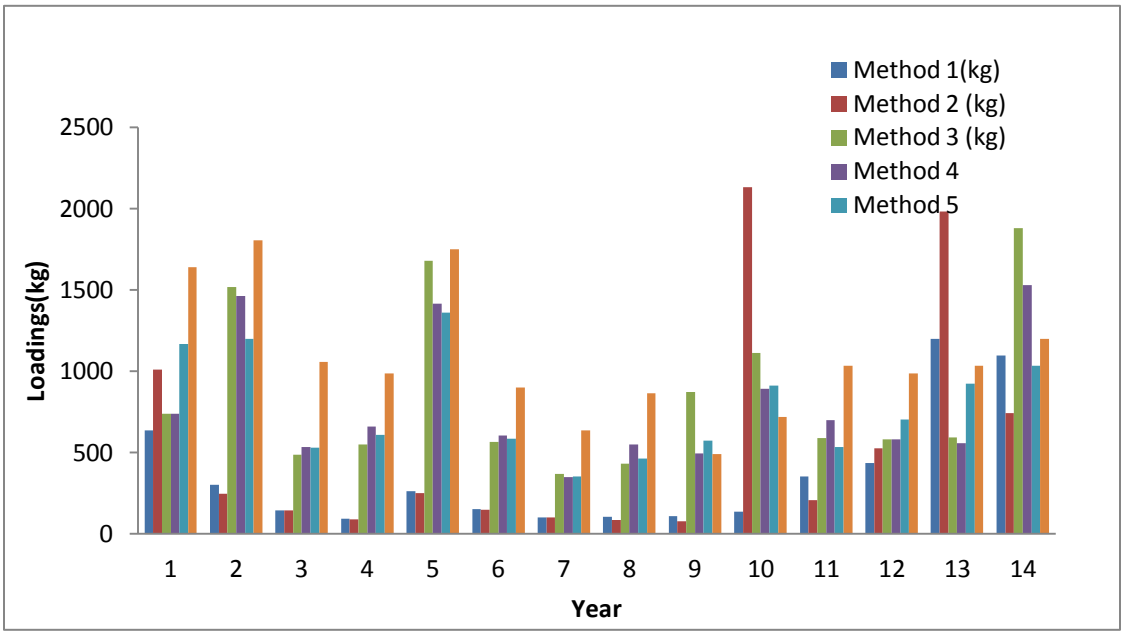


*Figure 24.* Bar Chart of loading calculated by all six methods for TP

Total Suspended Sediments (TSS): From the visual inspection of box plot, Figure 25, and bar chart, Figure 26, it can be seen that all the averaging method had very small range and value of estimate in comparison to LOADEST. So, the range of estimate for M1-5 was 2,456-994,731 kg per year but for LOADEST the range was 312,000-2,062,736 kg per year with the standard deviation of 433,852 kg per year. Estimates of none of the methods were normally distributed. All the five method M1-M5 had number of outliers.

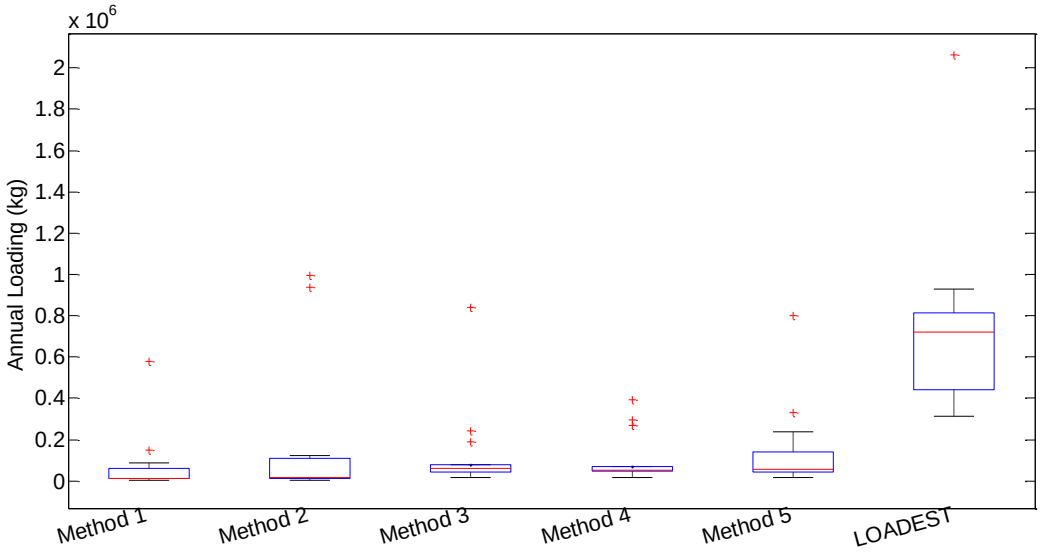


*Figure 25.* Box plot of loading calculated by all six methods for TSS

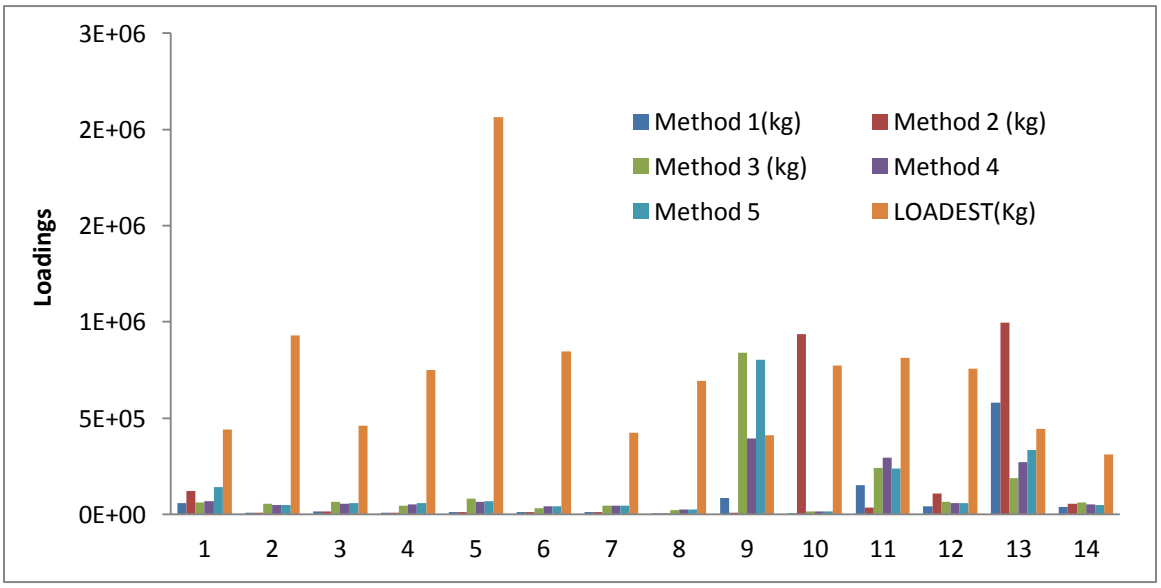


*Figure 26.* Bar Chart of loading calculated by all six methods for TSS

## 4.3 Multivariate

**4.3.1 Data pretreatment.** Data pretreatment was carried out to bring out the importance of each parameter regardless of their numerical magnitude. Different data pretreatment method brought out different aspect of data. Centering removed the offset; auto-scaling, range-scaling

made all parameters equally important, pareto-scaling remains closer to original structure of the datasets whereas transformation methods reduced heteroscedasticity.

First of all, when original dataset went through different data pretreatment processes only one fourth of the time it retained its original unit like centering and pareto-scaling. Other times either it becomes unit less likes in the case of different scaling methods or got transformed. These scaling methods are auto-scaling, range-scaling, vast-scaling and level-scaling. And the data pretreatment methods which transformed original units were log-transformation and power-transformation. In original datasets, fecal coliform had numerically very high value in comparison to heavy metals like cadmium and lead.

Statistical analyses were influenced by the relative magnitude of the participating parameters. Therefore, if the original dataset had been used directly, parameters with higher numerical magnitude would have suppressed other parameters with numerically smaller magnitude. Figure 27 presented the original data and value of highest point after going through all the data pretreatment procedures. Table 6 gave the number representing all the parameters involved in the study. Scaling methods like auto-scaling and range-scaling produced fairly well distributed datasets with no peaks. Data pretreatment methods vast-scaling and log-transformation produces datasets with few peaks without suppressing other parameters. But in vast scaling value of fecal coliform becomes lowest whereas in log-transformation its value is fairly large in comparison to other parameters.

Data pretreatment method like centering, pareto-scaling, level-scaling and power transformations had one big peak suppressing all other parameters. Peaks in the pretreated data from centering, pareto-scaling, and power transformation were consistent with original data.

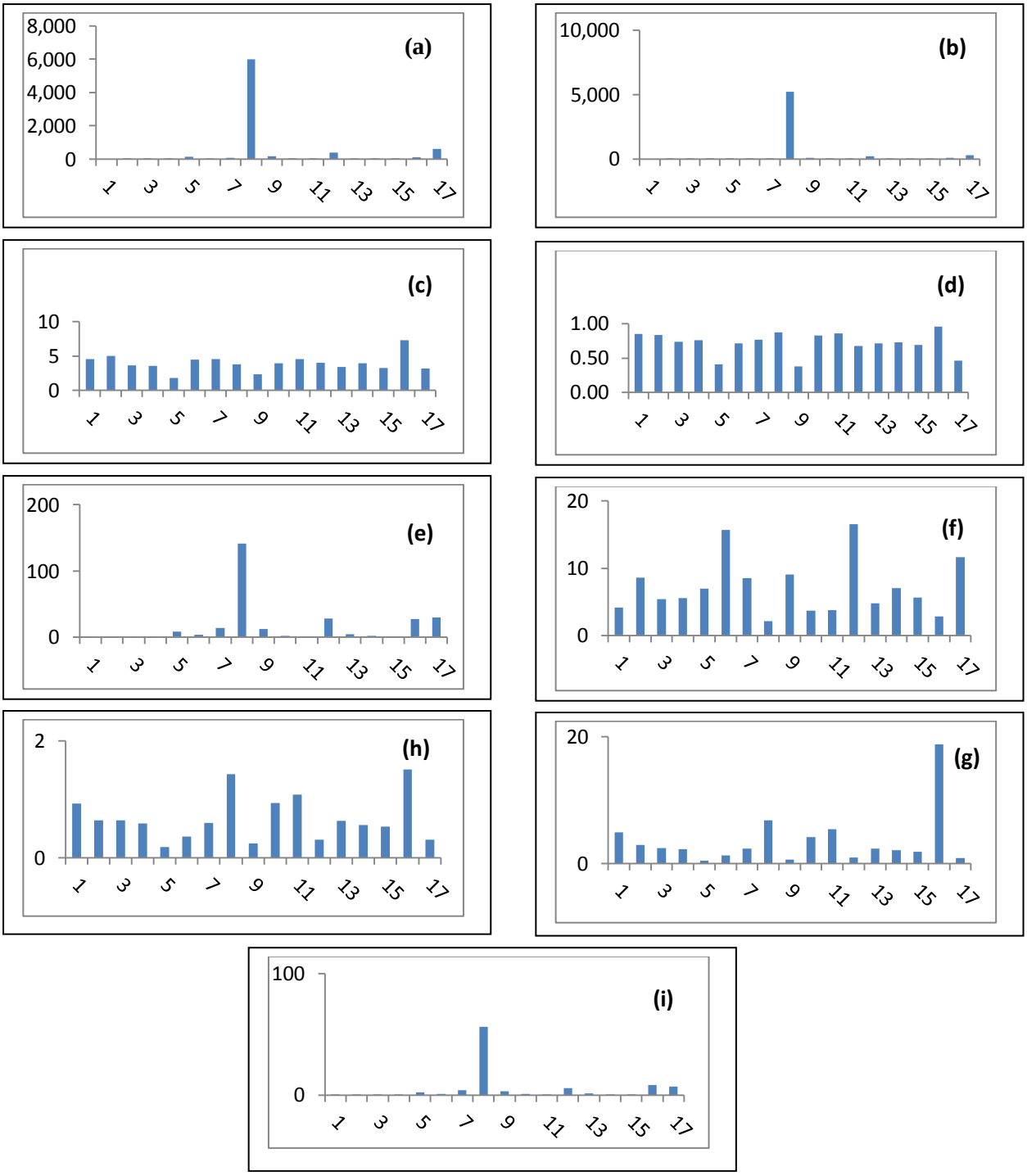


*Figure 27.* Maximum Values of parameters before and after data pretreatment (a) data; (b) centering; (c) auto-scaling; (d) range-scaling; (e) pareto-scaling; (f) vast-scaling; (g) level-scaling; (h) log-transformation and (i) power-transformation

Table 6

*Number associated with parameter*

| S.N | Parameter | S.N | Parameter | S.N | Parameter |
|---|---|---|---|---|---|
| 1 | Cadmium | 7 | COD | 13 | TSS |
| 2 | Copper | 8 | Fecal Coliform | 14 | TKN |
| 3 | Lead | 9 | Hardness | 15 | Total Phosphorus |
| 4 | Zinc | 10 | Nitrate | 16 | Turbidity |
| 5 | Alkalinity | 11 | Nitrite | 17 | Conductivity |
| 6 | BOD | 12 | TDS | | |

**4.3.2 PCA: one location at a time.** In this method, importance of the entire variable was obtained by using all the principal components of the PCA. It was obtained by multiplying percentage represented by each principal component with principal components coefficients matrix. And the cumulative ranking was obtained by adding individual scores obtained by each parameter in each of the different data pretreatment method. Cumulative ranking is presented in Table 7. Ranking of water quality parameters varied for each monitoring station according to different data pretreatment method, for e.g. for monitoring location at White Street, turbidity is the most important parameters according to centering, auto-scaling, pareto-scaling and power transformation whereas for range-scaling, hardness is the most important parameters. Likewise ranking varied for other methods.

Ranking of parameters also varied among different monitoring location. For example for monitoring location at W.JJ Street, fecal coliform was most important parameter whereas for monitoring location at White Street, conductivity was most important parameter. Overall ranking of parameters for entire Greensboro watershed shows that fecal coliform, conductivity and TSS are the most important parameter in the respective order and zinc, copper and cadmium in the

same order as least important parameter. It was obtained by adding the individual scores of each parameters of each water quality station for entire study area.

Table 7

*Ranking of parameter obtained by using data pretreated by different method*

| Rank | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 8 | 8 | 7 | 8 | 8 | 12 | 13 | 8 | 17 | 17 | 9 | 8 | 8 | 8 | 17 |
| 2 | 8 | 13 | 17 | 8 | 7 | 9 | 16 | 8 | 16 | 8 | 8 | 8 | 10 | 10 | 7 | 8 |
| 3 | 7 | 17 | 7 | 13 | 17 | 17 | 13 | 7 | 17 | 5 | 16 | 13 | 17 | 13 | 9 | 9 |
| 4 | 14 | 12 | 13 | 16 | 9 | 12 | 5 | 17 | 12 | 9 | 7 | 1 | 13 | 17 | 13 | 5 |
| 5 | 6 | 16 | 14 | 12 | 13 | 13 | 7 | 10 | 13 | 7 | 5 | 17 | 9 | 7 | 16 | 12 |
| 6 | 17 | 5 | 5 | 14 | 12 | 14 | 8 | 12 | 9 | 13 | 13 | 14 | 16 | 5 | 12 | 7 |
| 7 | 13 | 7 | 10 | 5 | 5 | 16 | 9 | 9 | 5 | 12 | 9 | 10 | 7 | 15 | 14 | 16 |
| 8 | 16 | 10 | 12 | 9 | 16 | 5 | 14 | 5 | 10 | 3 | 14 | 6 | 5 | 16 | 15 | 13 |
| 9 | 9 | 15 | 16 | 3 | 6 | 10 | 10 | 16 | 14 | 4 | 12 | 11 | 12 | 12 | 5 | 14 |
| 10 | 12 | 4 | 9 | 11 | 10 | 15 | 17 | 14 | 7 | 16 | 15 | 15 | 11 | 6 | 10 | 10 |
| 11 | 4 | 6 | 6 | 10 | 3 | 11 | 6 | 15 | 6 | 14 | 3 | 16 | 3 | 9 | 17 | 11 |
| 12 | 11 | 9 | 11 | 17 | 11 | 7 | 11 | 4 | 11 | 6 | 10 | 7 | 14 | 14 | 3 | 6 |
| 13 | 10 | 3 | 2 | 15 | 14 | 6 | 3 | 6 | 3 | 10 | 4 | 5 | 6 | 3 | 11 | 4 |
| 14 | 2 | 11 | 3 | 4 | 2 | 4 | 15 | 2 | 4 | 11 | 6 | 12 | 15 | 11 | 6 | 15 |
| 15 | 15 | 14 | 15 | 6 | 4 | 1 | 4 | 11 | 15 | 15 | 11 | 4 | 2 | 4 | 4 | 2 |
| 16 | 3 | 2 | 1 | 1 | 15 | 3 | 1 | 3 | 2 | 2 | 1 | 3 | 4 | 1 | 1 | 1 |
| 17 | 1 | 1 | 4 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 3 |

Table 8 presents percentage variance represented by each parameter obtained by using all principal components for different data pretreatment method. Almost 90% of the total water quality variance is represented by fecal coliform when the data pretreated by centering method was used. For this method, there was no representation of the heavy metals, nutrients and next highest representative of the variance was conductivity. For auto-scaling, range-scaling, vast-scaling, level-scaling and log-transformation variance represented by almost all the parameters were similar and between the range of 4.94-6.57%, 4.14-7.90%, 1.21-9.97%, 2.90-13.81%, and

3.17-8.51% respectively. For pareto-scaling and power-transformation fecal coliform represented 43.96% and 50.84% respectively followed by conductivity at 11.98% and 9.75%. Heavy metal and nutrients had represented almost no variance.

Table 8

*Percentage Variance represented by each parameter obtained by using all principal components for different data pretreatment method*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Parameter |
|---|---|---|---|---|---|---|---|---|
| 0.00 | 5.59 | 6.30 | 0.03 | 4.65 | 5.48 | 5.40 | 0.03 | 1 |
| 0.00 | 5.77 | 5.48 | 0.11 | 6.35 | 4.71 | 4.88 | 0.09 | 2 |
| 0.00 | 5.54 | 4.38 | 0.09 | 4.79 | 6.24 | 5.63 | 0.08 | 3 |
| 0.00 | 6.32 | 5.38 | 0.23 | 5.22 | 6.37 | 6.31 | 0.21 | 4 |
| 0.98 | 6.57 | 7.90 | 6.81 | 9.97 | 2.90 | 3.73 | 5.24 | 5 |
| 0.06 | 5.73 | 6.12 | 1.75 | 7.42 | 4.64 | 3.17 | 1.22 | 6 |
| 0.31 | 6.33 | 6.90 | 3.67 | 7.64 | 4.22 | 5.59 | 3.34 | 7 |
| 89.97 | 5.98 | 5.44 | 43.96 | 1.91 | 9.52 | 8.51 | 50.84 | 8 |
| 1.36 | 6.18 | 6.84 | 7.74 | 9.51 | 3.07 | 5.70 | 6.74 | 9 |
| 0.01 | 5.35 | 6.25 | 0.64 | 4.75 | 5.26 | 7.62 | 0.88 | 10 |
| 0.00 | 5.62 | 6.23 | 0.39 | 3.61 | 5.93 | 7.34 | 0.47 | 11 |
| 2.46 | 6.38 | 5.59 | 10.07 | 8.95 | 3.86 | 3.46 | 8.07 | 12 |
| 0.86 | 4.94 | 4.14 | 5.68 | 1.21 | 13.81 | 8.18 | 6.05 | 13 |
| 0.02 | 5.70 | 6.35 | 0.96 | 7.11 | 4.89 | 4.93 | 0.77 | 14 |
| 0.00 | 6.03 | 5.41 | 0.34 | 6.33 | 5.70 | 6.01 | 0.32 | 15 |
| 0.77 | 5.54 | 5.47 | 5.55 | 2.06 | 9.83 | 7.12 | 5.90 | 16 |
| 3.20 | 6.44 | 5.81 | 11.98 | 8.53 | 3.56 | 6.44 | 9.75 | 17 |
| 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | |

It showed that the PCA was susceptible to different data pretreatment method. But data pretreatment methods (auto-scaling, range-scaling, vast-scaling, level-scaling and log-transformation) which produced data with no peaks had variance distributed among parameters fairly equally where as data pretreatment method (centering, pareto-scaling, and power transformation) which produced data with peak had uneven representation of variance by parameter with very high peak value suppressing peak value of other parameters. This showed

that scaling method like centering, pareto-scaling and power transformations were unsuitable method of data pretreatment method.

Distribution of percentage of parameters obtained by PCA using data from different pretreatment method by different method was similar to data distribution obtained from different pretreatment method. Parameters consistently featuring in top ten in all the water quality parameters were fecal coliform, conductivity, TSS, COD, hardness, alkalinity, turbidity, TDS, nitrate and TKN respectively. The three most important and least parameters for whole study area were fecal coliform, conductivity, TSS and cadmium, copper and zinc respectively.

**4.3.3 PCA: all locations.** FA/PCA was conducted on all data of all locations obtained through auto-scaling or Z-transformation. Missing data, outliers and extreme values were replaced by the median value. Factors were obtained by FA using PCA as the method of factor extraction. Then obtained factor was orthogonally rotated using normalized varimax transformation. Among the uncorrelated factors obtained, five factors with eigen values greater than one (Table 9) were chosen. These factors represented 64.32% of the total variance represented by water quality parameters. Factors loadings of parameter greater than (> 0.7) were chosen as principal components (Liu, Lin and Kuo, 2003). They were presented in Table 10. Figure 28 was a screeplot of eigen values.

Table 9

*Eigen Values for the entire watershed*

| | **Eigenvalue** | **Percentage Total variance** | **Cumulative (Eigen value)** | **Cumulative (%)** |
|---|---|---|---|---|
| 1 | 3.28 | 19.29 | 3.28 | 19.29 |
| 2 | 2.97 | 17.47 | 6.25 | 36.77 |
| 3 | 1.67 | 9.84 | 7.92 | 46.61 |
| 4 | 1.55 | 9.15 | 9.48 | 55.76 |
| 5 | 1.46 | 8.56 | 10.93 | 64.32 |

Table 10

*Factor scores for most important factors for the entire watersheds*

| Parameter | **Factor (1)** | **Factor (2)** | **Factor (3)** | **Factor (4)** | **Factor (5)** |
|---|---|---|---|---|---|
| Cadmium (mg/L) | 0.04 | 0.07 | -0.03 | 0.88 | -0.02 |
| Copper (mg/L) | 0.03 | 0.45 | 0.12 | -0.03 | 0.20 |
| Lead (mg/L) | -0.05 | 0.69 | -0.02 | 0.49 | -0.04 |
| Zinc (mg/L) | 0.14 | 0.49 | 0.49 | -0.01 | -0.04 |
| Alkalinity (mg/L) | 0.84 | -0.10 | -0.25 | 0.07 | -0.01 |
| BOD (mg/L) | 0.01 | 0.33 | 0.04 | 0.00 | 0.75 |
| COD (mg/L) | 0.02 | -0.03 | 0.08 | 0.21 | 0.82 |
| Fecal Coliform CFU/100 mL | -0.05 | 0.10 | 0.05 | -0.10 | 0.60 |
| Hardness (mg/L) | 0.89 | -0.12 | -0.08 | 0.02 | 0.01 |
| Nitrate Nitrogen (mg/L) | 0.06 | -0.08 | 0.85 | 0.04 | -0.04 |
| Nitrite Nitrogen (mg/L) | 0.03 | -0.12 | 0.16 | 0.73 | 0.08 |
| TDS (mg/L) | 0.78 | 0.01 | 0.33 | -0.04 | -0.04 |
| TSS (mg/L) | -0.12 | 0.84 | 0.00 | -0.06 | 0.12 |
| TKN (mg/L) | -0.09 | 0.07 | 0.57 | 0.06 | 0.33 |
| Total Phosphorus (mg/L) | 0.13 | 0.12 | 0.76 | 0.06 | 0.07 |
| Turbidity (ntu) | -0.22 | 0.79 | -0.01 | -0.09 | 0.12 |
| Conductivity (µmhos/cm) | 0.80 | -0.05 | 0.32 | 0.03 | -0.02 |

Factor (1) represented 19.29% of total variance and its principal factors were alkalinity, hardness, TDS and conductivity. Its eigen value was 3.28. Factor (1) represented the physical component of the water quality and it showed that it was the most dominant factor affecting the water quality. Factor (2) represented 17.47% of total variance and its principal factors are TSS and turbidity. Its eigen value is 2.97. It represented sediments in water quality. Factor (3) represented 9.84% of total variance with nitrate and total phosphorus as its principal factors and represented nutrients. Its eigen value was 1.67. Factor (4) constituted 9.15% of total variance and its main contributor was cadmium and nitrite and did not represent any particular component of water quality but was mixture of nutrient and heavy metal. Its eigen value was 1.55. Finally factor (5) represented 8.56% of total variance. BOD and COD were its major contributor and represented chemical component of water quality. Its eigen value was 1.46. So, the number of important parameters reduced from 17 to 12 which lead to 29.41% reduction in data.

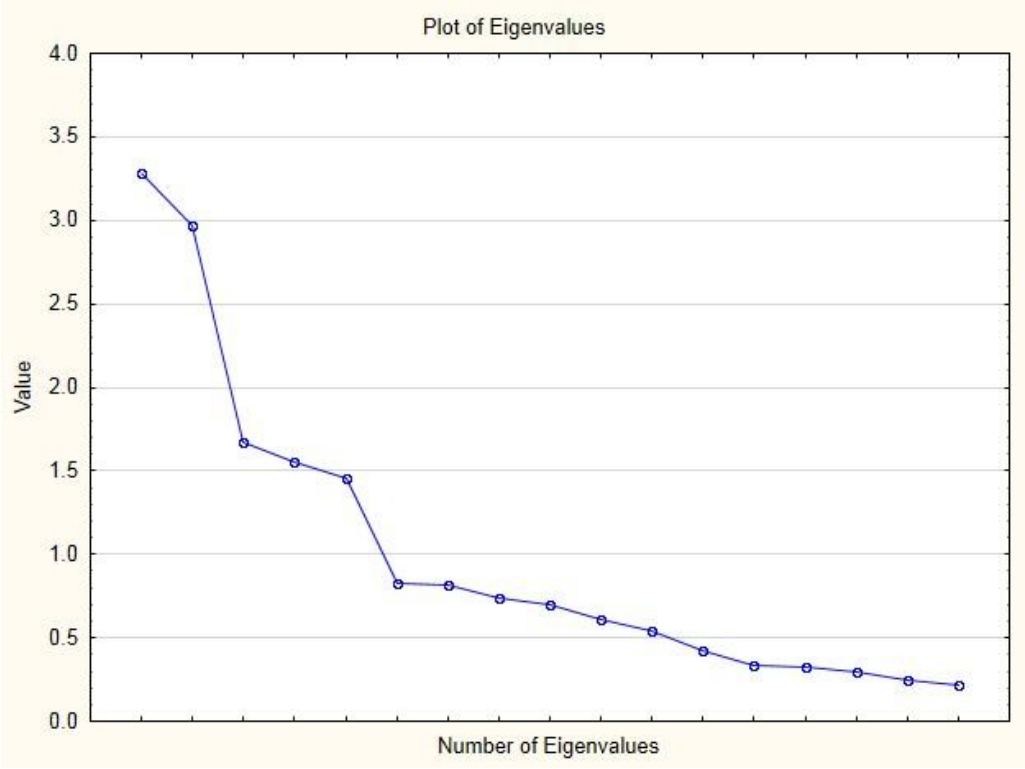


*Figure 28.* Screeplot of eigen values of principal factors of entire watershed

It is important to note that parameters with high eigen value mean high variability in their magnitude which in turn means they are considered relatively important water quality components. This can potentially lead to misconception that other parameter which have lower or none eigen value are not significant or important parameters. For example, fecal coliform in the city of Greensboro is of major concern as can be seen from their magnitude at each of the monitoring locations; consistently exceeding the standards. However, it lacks variability in measured data over last 14 years of monthly or bi-monthly data (1999-2012), and was not considered as an important parameter in the PCA analysis. Using the technique presented here should not be the only criterion to determine significant or important parameters but should be combined with other analyses. Fecal coliform was found to be “not important” parameter in these analyses, but need to measure consistently but less frequently since it is the “important” water quality parameter for the health of the stream.

**4.3.4 Cluster analysis on water quality stations.** The City of Greensboro is high on the Haw River Watershed with small area toeing in the Deep Watershed. Fourteen out of sixteen water quality monitoring location included in this study fell in Haw River Watershed whereas remaining was located in Deep Watershed. River in city of Greensboro was called Buffalo creek. It had basically three tributaries which combine to form Buffalo Creek. Buffalo Creek drains into Haw River. Two water quality monitoring stations monitored the tributaries of Deep River, which was located in Deep Watershed.

Buffalo Creek had three tributaries on which fourteen water quality monitoring stations were located. On the top of the study area there was Buffalo Lakes. It was in the middle of upstream river and water drained through downstream river. Four water quality monitoring stations were located there. Monitoring stations located at Battleground Avenue, Bluff Run Rd., Friendship St., and Old Oak Ridge Rd. Similarly on the middle and lower tributaries five water quality monitoring stations each were located. Remaining two was located near the edge of Haw and Deep Watershed on the side of Deep Watershed. The middle tributary consisted water quality location at 16$^{th}$ St., Aycock St., Church St., White St. and Rankin Mill Rd. Similarly for third tributaries they were at McConnell Rd., Merritt Dr., Randleman Rd., W.JJ Dr. and Fieldcrest Dr. And water quality monitoring location at Kivett St. and Mackay Rd. were located in Deep Watershed.

Cluster Analysis grouped sixteen water quality monitoring stations into two loose groups for (Dlink/Dmax)* 100 > 55 (Figure 29). These two groups represented two distinct river tributaries in the city of Greensboro. First cluster was called cluster (1) and it was further sub-divided into two distinct sub-groups, namely cluster 1(a) and cluster 1(b). Cluster (1) had six water quality monitoring stations, four from top tributary and the remaining two from Deep

watershed. Cluster 1(a) consisted of water quality stations at Old Oak Ridge Rd., Mackay Rd. and Kivett St. All three were located close to the edge of the watershed. Cluster 1(b) consisted of water quality monitoring station at Battleground Avenue, Bluff Run Rd. and Friendship St. Water quality monitoring stations at Battleground Avenue and Bluff Run Rd. were located upstream of Buffalo Lake on different river branches whereas water quality monitoring station at Friendship St. was downstream of the lake.

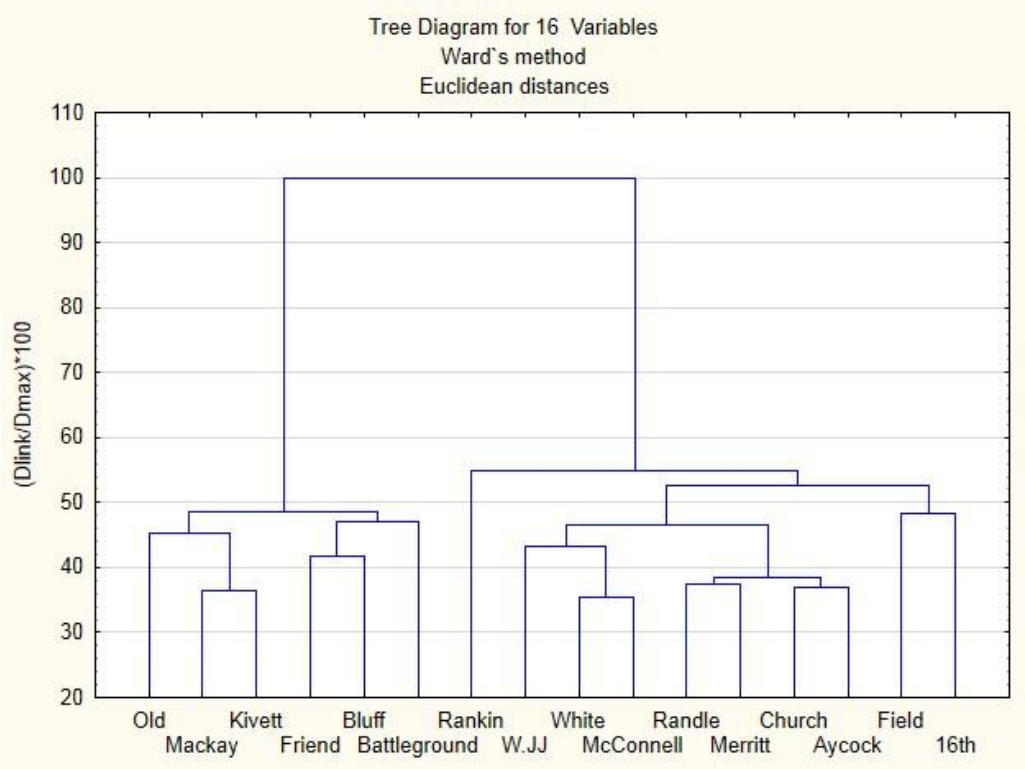


*Figure 29.* Cluster of water quality stations according to spatial similarity between the stations

The second group consisted of ten water quality station and represented middle and lower tributary of Buffalo Creek which were spatially closer to each other. For (Dlink/Dmax)* 100 > 50, second group was divided into three clusters. Cluster (2) consisted of only one water quality monitoring station at Rankin Mill Rd. It was located at the lower reach of middle tributary. Cluster (3) got divided into two groups namely 3(a) and 3(b). Cluster 3(a) consisted of water quality monitoring stations at McConnell Rd., W.JJ Dr. and White St. Cluster 3 (b) consisted of

water quality at Merritt Dr., Randleman Rd., Aycock St. and Church St. Water quality station at Merritt and Randleman were from same river branch and same was true for remaining. Cluster (4) consisted of water quality station at 16th St. and Fieldcrest Dr.

**4.3.5. PCA: spatial clusters of stations.** Cluster Analysis (CA) grouped water quality stations according to their spatial similarity. So, the sixteen water quality stations got grouped into four groups. Since the first group and third group consisted of six and seven water quality stations respectively as its member, it was sub divided into two groups each. These sub-groups were treated as one separate clustering units and FA/PCA was carried out on them. The results of application of FA/PCA are presented in tables and graph as follows.

Cluster 1(a): Cluster 1(a) consisted of three water quality stations. They were Kivett St., Mackay Rd., and Old Oak Ridge. Cluster 1(a) had six factors with eigen value greater than one (>1) representing 63.55% (Table 11) of total variance. Principal factors for factor (1) were alkalinity, hardness, TDS and conductivity and they represented the physical component of water quality. Factor (1) represented 19.59% of water quality with cumulative eigen value of 3.33. Factor (2) represented 11.15% of total variance with eigen value of 1.89 and represented sediments with TSS and turbidity as the principal factors.

Factor (3) had cadmium and lead as its principal components representing 10.42% of total variance with eigen value of 1.77. It represented heavy metal. Factor (4) and (5) consisted of total phosphorus, nitrate and nitrite. They represented nutrients in water with more than 15% of total variance. Factor (6) represented chemical component with BOD (Table 12). In total, there were twelve principal factors reducing the data size by 29.34%.When one representative water quality station will be chosen, it would reduce the size of datasets by 76.5%. Figure 30 is a screeplot of eigen values.

Table 9

*Eigen Values for cluster 1(a)*

| | Eigenvalue | % Total (variance) | Cumulative (Eigenvalue) | Cumulative (%) |
|---|---|---|---|---|
| 1 | 3.33 | 19.59 | 3.33 | 19.59 |
| 2 | 1.89 | 11.15 | 5.22 | 30.75 |
| 3 | 1.77 | 10.42 | 7.00 | 41.18 |
| 4 | 1.42 | 8.36 | 8.42 | 49.54 |
| 5 | 1.27 | 7.47 | 9.69 | 57.02 |
| 6 | 1.11 | 6.52 | 10.80 | 63.55 |

Table 10

*Factor scores for most important factors for cluster 1(a)*

| Parameters | Factor (1) | Factor (2) | Factor (3) | Factor (4) | Factor (5) | Factor (6) |
|---|---|---|---|---|---|---|
| Cadmium (mg/L) | -0.03 | -0.05 | 0.87 | 0.06 | 0.24 | 0.08 |
| Copper (mg/L) | 0.06 | 0.08 | 0.06 | -0.16 | -0.19 | 0.54 |
| Lead (mg/L) | 0.10 | 0.14 | 0.89 | 0.03 | -0.11 | -0.06 |
| Zinc (mg/L) | -0.08 | 0.57 | 0.24 | -0.03 | -0.02 | 0.08 |
| Alkalinity (mg/L) | -0.77 | -0.26 | 0.00 | -0.12 | 0.02 | 0.00 |
| BOD (mg/L) | 0.00 | 0.22 | -0.07 | 0.07 | 0.00 | 0.79 |
| COD (mg/L) | 0.12 | -0.14 | 0.13 | 0.47 | 0.43 | 0.53 |
| Fecal Coliform CFU/100 mL | 0.14 | 0.13 | 0.03 | 0.65 | -0.06 | -0.05 |
| Hardness (mg/L) | -0.81 | -0.19 | -0.07 | -0.18 | 0.07 | -0.01 |
| Nitrate Nitrogen (mg/L) | -0.10 | 0.09 | 0.09 | -0.11 | 0.70 | -0.22 |
| Nitrite Nitrogen (mg/L) | 0.03 | -0.06 | 0.00 | -0.04 | 0.71 | 0.04 |
| TDS (mg/L) | -0.77 | 0.27 | -0.03 | 0.05 | -0.10 | -0.07 |
| TSS (mg/L) | 0.10 | 0.84 | -0.08 | 0.20 | 0.00 | 0.16 |
| TKN (mg/L) | 0.12 | 0.01 | -0.01 | 0.70 | 0.06 | 0.10 |
| Total Phosphorus (mg/L) | -0.05 | 0.14 | 0.04 | 0.72 | -0.17 | -0.10 |
| Turbidity (ntu) | 0.27 | 0.84 | -0.05 | 0.15 | 0.01 | 0.07 |
| Conductivity (µmhos/cm) | -0.80 | -0.06 | 0.01 | -0.06 | 0.05 | -0.05 |

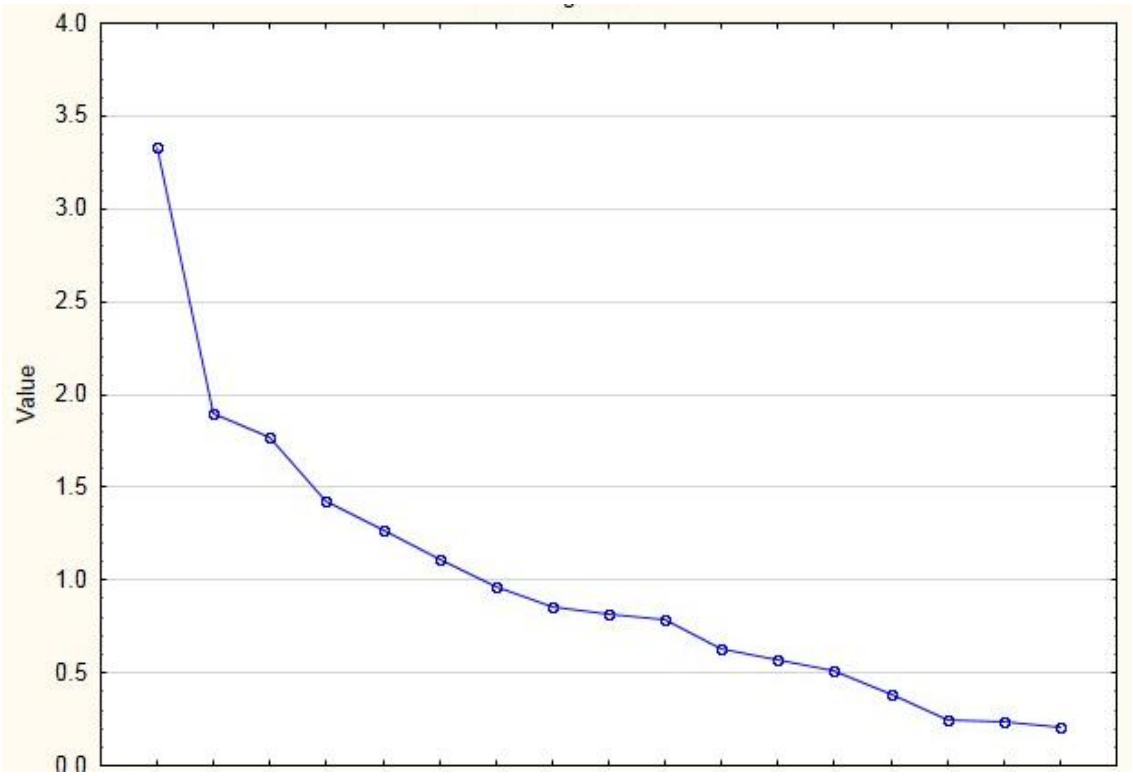


*Figure 30.* Screeplot of eigen values of principal factors for cluster 1(a)

Cluster 1(b): Cluster 1(b) consists of three water quality station located at Friendship St., Battleground Avenue and Bluff Run Rd. Cluster 1(b) had six factors (Figure 31) representing 63.45% of total variance (Table 13). The most important principal factors were TSS, turbidity, nitrite, cadmium, alkalinity, hardness, TDS, BOD, total phosphorus, lead, zinc, and TKN (Table 14). So, the number of parameters got reduced to twelve from seventeen. Figure 31 was a screeplot of eigen values.

Factor (1) represented 15.63% of variance which represented sediment with the eigen value of 2.66. Factor (2) contribute 12.49% of total variance but did not represent any particular aspect of water quality. Factor (2) had eigen value of 2.12. Factor (3) represented 11.82% of water quality with eigen value of 2.01. It represented physical component of water quality. Factor (4) represented mixture of nutrients and chemicals. Factor (4) represented 9.54% of total

variance with eigen value of 1.62. Factor (5) represented the heavy metal with 7.79% of water quality and eigen value of 1.33. Factor (6) had eigen value of 1.05 and represented the 6.17% of total water quality variance which represented nutrients. In total, there were total of twelve principal factors reducing the dataset by 29.34%. When one representative water quality station will be chosen, it would be reduced the size of datasets by 76.5%.

Table 11

*Eigen Values for cluster 1(b)*

| | Eigenvalue | % Total (variance) | Cumulative (Eigenvalue) | Cumulative (%) |
|---|---|---|---|---|
| 1 | 2.66 | 15.63 | 2.66 | 15.63 |
| 2 | 2.12 | 12.49 | 4.78 | 28.12 |
| 3 | 2.01 | 11.82 | 6.79 | 39.94 |
| 4 | 1.62 | 9.54 | 8.41 | 49.48 |
| 5 | 1.33 | 7.79 | 9.74 | 57.28 |
| 6 | 1.05 | 6.17 | 10.79 | 63.45 |

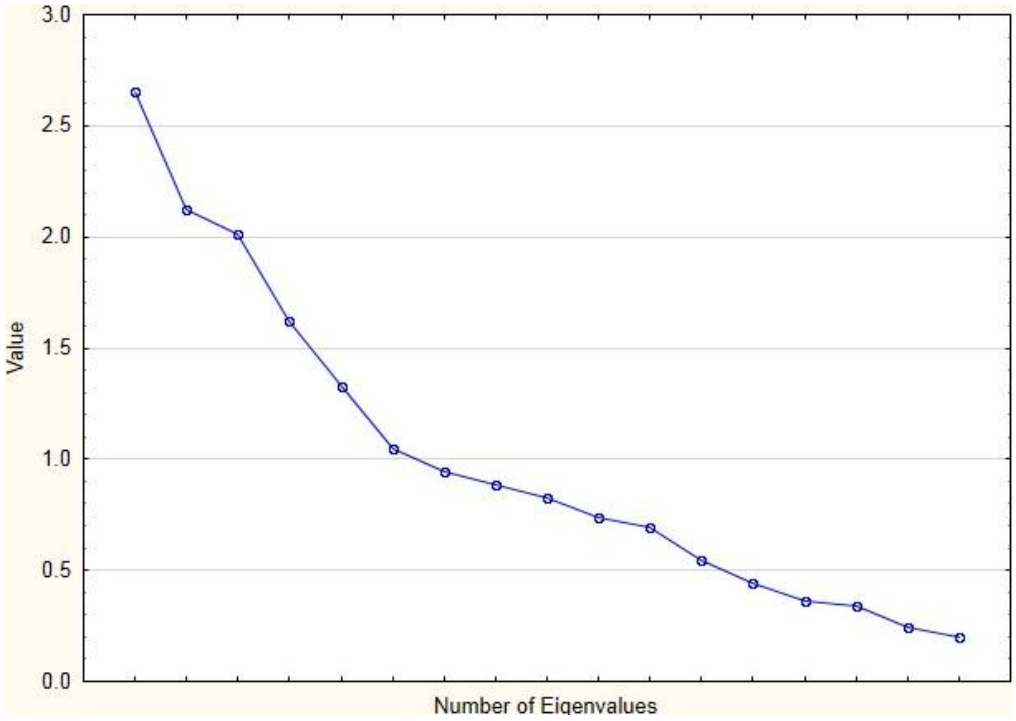


*Figure 31. Screeplot of eigen values of principal factors for cluster 1(b)*

Table 12

*Factor scores for most important factors for cluster 1(b)*

| Parameters | Factor (1) | Factor (2) | Factor (3) | Factor (4) | Factor (5) | Factor (6) |
|---|---|---|---|---|---|---|
| Cadmium (mg/L) | -0.01 | 0.78 | 0.13 | 0.01 | 0.43 | -0.03 |
| Copper (mg/L) | 0.53 | 0.18 | 0.04 | 0.08 | -0.05 | 0.01 |
| Lead (mg/L) | 0.13 | 0.34 | 0.17 | -0.07 | 0.77 | 0.07 |
| Zinc (mg/L) | -0.04 | -0.23 | -0.10 | 0.12 | 0.73 | -0.02 |
| Alkalinity (mg/L) | -0.17 | 0.17 | -0.76 | 0.05 | 0.03 | -0.16 |
| BOD (mg/L) | 0.05 | -0.04 | 0.01 | 0.86 | 0.04 | 0.07 |
| COD (mg/L) | -0.14 | 0.29 | 0.07 | 0.55 | -0.25 | 0.44 |
| Fecal Coliform CFU/100 mL | -0.01 | -0.11 | 0.10 | 0.03 | 0.10 | 0.60 |
| Hardness (mg/L) | -0.14 | 0.02 | -0.80 | -0.01 | -0.09 | -0.19 |
| Nitrate Nitrogen (mg/L) | 0.14 | 0.61 | -0.16 | -0.09 | -0.15 | 0.02 |
| Nitrite Nitrogen (mg/L) | -0.11 | 0.84 | -0.05 | 0.08 | -0.04 | -0.04 |
| TDS (mg/L) | 0.13 | -0.05 | -0.76 | 0.07 | -0.05 | -0.03 |
| TSS (mg/L) | 0.85 | -0.09 | 0.02 | 0.09 | 0.09 | 0.06 |
| TKN (mg/L) | 0.18 | 0.07 | 0.12 | 0.13 | -0.11 | 0.70 |
| Total Phosphorus (mg/L) | 0.28 | -0.05 | -0.07 | 0.79 | 0.11 | -0.01 |
| Turbidity (ntu) | 0.88 | -0.10 | 0.08 | 0.06 | 0.04 | 0.09 |
| Conductivity (µmhos/cm) | 0.00 | -0.01 | -0.48 | -0.21 | 0.14 | 0.38 |

Cluster (1a+1b): When all the water quality station in cluster 1 was analyzed together, whole dataset is divided into six factors which represented 61.75 % of total variance (Table 15). Factor (1) represented 16.74 percentage of the total variance with the eigen value of 2.85. Physical factors were its major contributors. They were alkalinity, hardness and TDS. Factor (2) represented 12.91% of total variance with eigen value of 2.19. Sediments were major contributor for this factor. They were TSS and Turbidity.

Factor (3) and (6) represented heavy metal with eigen value of 1.76 and 1.03 respectively. Total percentage variance represented by factor (3) and factor (6) were 10.33% and 6.08% respectively. Factor (4) represented nutrients, TKN, with eigen value of 1.49 and 8.76% total variance. Factor (5) did not have any principal factors but contributed 6.93% of total variance with eigen value of 1.18.

Table 13

*Eigen Values for cluster (1a+1b)*

| | Eigenvalue | % Total (variance) | Cumulative (Eigenvalue) | Cumulative (%) |
|---|---|---|---|---|
| 1 | 2.85 | 16.74 | 2.85 | 16.74 |
| 2 | 2.19 | 12.91 | 5.04 | 29.65 |
| 3 | 1.76 | 10.33 | 6.80 | 39.98 |
| 4 | 1.49 | 8.76 | 8.29 | 48.74 |
| 5 | 1.18 | 6.93 | 9.46 | 55.67 |
| 6 | 1.03 | 6.08 | 10.50 | 61.75 |

And the important principal factor were, cadmium, lead, TKN and copper (Table 16). So, the important parameters were reduced to nine reducing the size of dataset by 47%. If one representative water quality station is chosen for cluster 1, it will reduce the dataset size by 91%. Figure 32 was a screeplot of eigen values. In total, there were total of nine principal factors reducing the dataset by 47%. When one representative water quality station will be chosen, it would be reduced the size of datasets by 90.6%.

Cluster (2): Cluster (2) consists of only one water quality station located at Rankin Mill Rd. This cluster had five factors with eigen value greater than one representing 65.54% of total variance (Table 17). Factor (1) represented 24.81% of the total variance with the eigen value of 4.22. Sediments were the major contributor of factor (1) and they were TSS and turbidity. Factor (2) represented heavy metal with cadmium and lead. It contributed 14.17% of total variance with eigen value of 2.41.

Table 14

*Factor scores for most important factors for cluster (1a+1b)*

| Parameters | Factor (1) | Factor (2) | Factor (3) | Factor (4) | Factor (5) | Factor (6) |
|---|---|---|---|---|---|---|
| Cadmium (mg/L) | 0.03 | -0.10 | 0.78 | 0.06 | -0.39 | 0.05 |
| Copper (mg/L) | 0.02 | 0.17 | 0.02 | -0.22 | -0.05 | 0.70 |
| Lead (mg/L) | 0.08 | 0.08 | 0.89 | 0.03 | 0.01 | -0.07 |
| Zinc (mg/L) | -0.03 | 0.10 | 0.47 | -0.05 | 0.42 | 0.11 |
| Alkalinity (mg/L) | -0.83 | -0.19 | -0.01 | -0.06 | -0.07 | 0.04 |
| BOD (mg/L) | 0.00 | 0.03 | 0.00 | 0.43 | 0.11 | 0.66 |
| COD (mg/L) | 0.07 | -0.13 | -0.03 | 0.67 | -0.34 | 0.31 |
| Fecal Coliform CFU/100 mL | 0.17 | 0.00 | 0.04 | 0.38 | 0.15 | -0.09 |
| Hardness (mg/L) | -0.85 | -0.14 | -0.09 | -0.11 | -0.08 | 0.05 |
| Nitrate Nitrogen (mg/L) | -0.18 | 0.09 | 0.09 | -0.04 | -0.65 | -0.15 |
| Nitrite Nitrogen (mg/L) | 0.02 | -0.07 | 0.06 | -0.03 | -0.69 | 0.13 |
| TDS (mg/L) | -0.81 | 0.21 | -0.04 | 0.03 | 0.04 | -0.01 |
| TSS (mg/L) | 0.02 | 0.89 | 0.03 | 0.12 | 0.03 | 0.11 |
| TKN (mg/L) | 0.03 | 0.19 | -0.02 | 0.71 | -0.06 | -0.17 |
| Total Phosphorus (mg/L) | -0.21 | 0.24 | 0.07 | 0.53 | 0.16 | 0.18 |
| Turbidity (ntu) | 0.12 | 0.90 | 0.02 | 0.09 | 0.01 | 0.07 |
| Conductivity (µmhos/cm) | -0.70 | -0.02 | 0.03 | 0.00 | -0.04 | -0.08 |

Table 15

*Eigen Values for cluster (2)*

| | Eigenvalue | % Total (variance) | Cumulative (Eigenvalue) | Cumulative (%) |
|---|---|---|---|---|
| 1 | 4.22 | 24.81 | 4.22 | 24.81 |
| 2 | 2.41 | 14.17 | 6.63 | 38.98 |
| 3 | 2.00 | 11.79 | 8.63 | 50.77 |
| 4 | 1.40 | 8.22 | 10.03 | 58.99 |
| 5 | 1.11 | 6.55 | 11.14 | 65.54 |

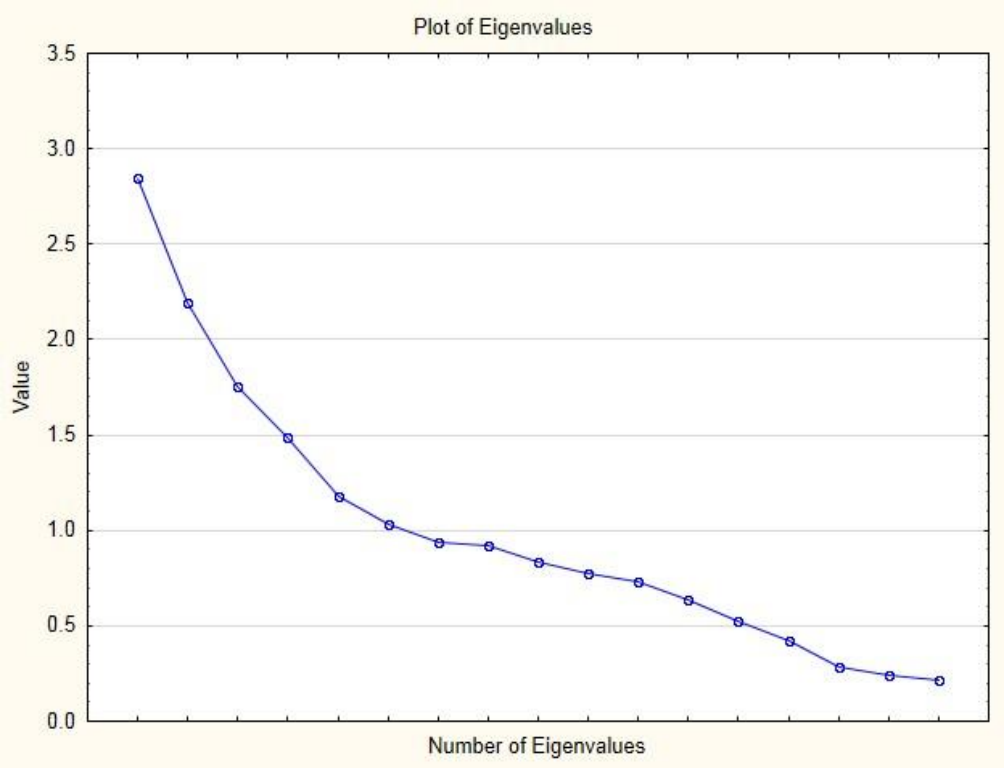


*Figure 32.* Screeplot of eigen values of principal factors for cluster (1a+1b)

Factor (3) and factor (5) did not represent any distinct component of water quality. Still, factor (3) contributed 11.79% of water quality with eigen value of 2. TP and zinc were its principal components. Factor (5) contributed 6.55% of total water quality variance with the eigen value of 1.11. Copper and fecal coliform were its principal Factor (4) had only one principal factor, BOD, which represented 8.22% of water quality with eigen value of 1.40.

Therefore, principal factors for cluster (2) were TSS, turbidity, cadmium, lead, zinc, total phosphorus, BOD, copper, and fecal coliform (Table 18). In total, there were nine principal factors reducing the data size by 47%. Figure 33 was a screeplot of eigen values.

Cluster 3(a): Three water quality stations were part of cluster 3(a), they were McConnell Rd., W.JJ Dr., and White St. Cluster 3(a) got reduced to four factors with fourteen principal factors which represented the 64.19% of total variance (Table 19).

Table 16

*Factor scores for most important factors for cluster (2)*

| Parameters | Factor (1) | Factor (2) | Factor (3) | Factor (4) | Factor (5) |
|---|---|---|---|---|---|
| Cadmium (mg/L) | -0.02 | 0.89 | 0.10 | 0.01 | 0.02 |
| Copper (mg/L) | 0.32 | 0.04 | 0.09 | 0.02 | 0.84 |
| Lead (mg/L) | 0.27 | 0.77 | -0.40 | -0.02 | 0.10 |
| Zinc (mg/L) | 0.38 | 0.03 | 0.71 | -0.04 | -0.06 |
| Alkalinity (mg/L) | -0.49 | 0.47 | 0.02 | 0.36 | -0.21 |
| BOD (mg/L) | 0.31 | -0.10 | -0.05 | 0.79 | 0.08 |
| COD (mg/L) | 0.13 | 0.43 | 0.27 | 0.32 | -0.07 |
| Fecal Coliform CFU/100 mL | 0.18 | -0.02 | -0.10 | 0.01 | 0.75 |
| Hardness (mg/L) | -0.49 | 0.03 | 0.29 | 0.15 | -0.20 |
| Nitrate Nitrogen (mg/L) | -0.27 | -0.17 | 0.55 | -0.48 | 0.21 |
| Nitrite Nitrogen (mg/L) | -0.03 | 0.24 | 0.13 | 0.50 | -0.27 |
| TDS (mg/L) | -0.49 | -0.10 | 0.69 | -0.07 | -0.11 |
| TSS (mg/L) | 0.80 | 0.18 | -0.03 | 0.10 | 0.22 |
| TKN (mg/L) | -0.31 | 0.06 | 0.00 | 0.65 | 0.29 |
| Total Phosphorus (mg/L) | -0.15 | 0.07 | 0.72 | 0.13 | 0.05 |
| Turbidity (ntu) | 0.86 | 0.10 | -0.13 | 0.23 | 0.21 |
| Conductivity (µmhos/cm) | -0.50 | 0.05 | 0.64 | 0.05 | -0.11 |

Table 17

*Eigen Values for cluster 3(a)*

| | Eigenvalue | % Total (variance) | Cumulative (Eigenvalue) | Cumulative (%) |
|---|---|---|---|---|
| 1 | 5.09 | 29.93 | 5.09 | 29.93 |
| 2 | 2.40 | 14.10 | 7.49 | 44.04 |
| 3 | 1.95 | 11.48 | 9.44 | 55.51 |
| 4 | 1.48 | 8.68 | 10.91 | 64.19 |

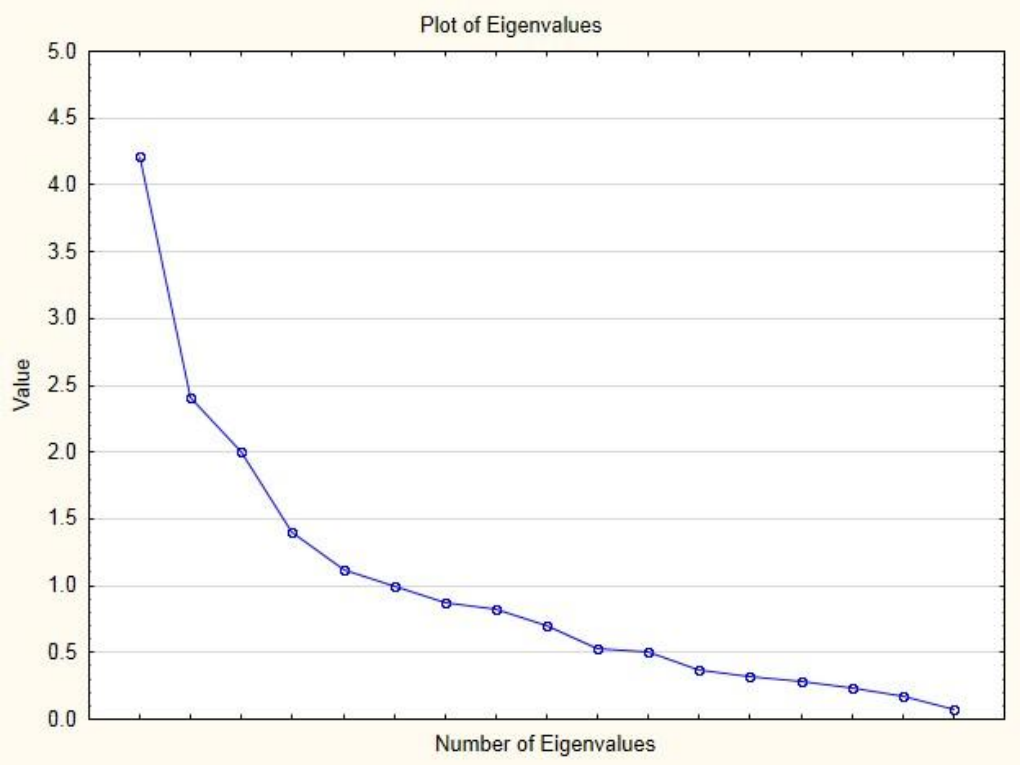


*Figure 33.* Screeplot of eigen values of principal factors for cluster (2)

Table 20 has Factor (1) was combination of sediment and heavy metal. It consisted of copper, lead, zinc, TSS and turbidity. It represented 29.93% of water quality with the eigen value of 5.09. Figure 34 was a screeplot of eigen values. Factor (2) represented physical component of water quality with alkalinity, TDS, hardness and conductivity as its principal components. It contributed 14.10% of total variance with eigen value of 2.40. Factor (3) consisted of nitrite and cadmium as principal component. It represented 11.48% of total variance with eigen value of 1.95. It was mixture of chemical component and heavy metal component of water quality. Factor (4) was mixture of different component of water quality. Its principal components were COD, fecal coliform and TKN. It represented 8.68% of total variance with the eigen value of 1.48. This only reduced the dataset by 17.64%. If one representative water quality station is chosen, then the data reduction would be 72.5%.

Table 18

*Factor scores for most important factors for cluster 3(a)*

| Parameters | Factor (1) | Factor (2) | Factor (3) | Factor (4) |
|---|---|---|---|---|
| Cadmium (mg/L) | 0.04 | 0.00 | 0.88 | -0.05 |
| Copper (mg/L) | 0.73 | 0.08 | 0.03 | 0.27 |
| Lead (mg/L) | 0.73 | 0.11 | 0.40 | -0.07 |
| Zinc (mg/L) | 0.79 | -0.16 | -0.06 | -0.03 |
| Alkalinity (mg/L) | -0.21 | -0.76 | 0.09 | -0.16 |
| BOD (mg/L) | 0.54 | -0.06 | -0.02 | 0.61 |
| COD (mg/L) | 0.02 | 0.13 | 0.34 | 0.76 |
| Fecal Coliform CFU/100 mL | 0.10 | 0.10 | -0.06 | 0.71 |
| Hardness (mg/L) | -0.19 | -0.84 | 0.08 | -0.11 |
| Nitrate Nitrogen (mg/L) | 0.09 | -0.16 | 0.47 | 0.15 |
| Nitrite Nitrogen (mg/L) | -0.07 | 0.00 | 0.88 | 0.01 |
| TDS (mg/L) | 0.05 | -0.75 | -0.06 | -0.06 |
| TSS (mg/L) | 0.82 | 0.19 | -0.06 | 0.18 |
| TKN (mg/L) | 0.27 | 0.21 | -0.01 | 0.72 |
| Total Phosphorus (mg/L) | 0.67 | 0.17 | 0.05 | 0.18 |
| Turbidity (ntu) | 0.72 | 0.35 | -0.05 | 0.23 |
| Conductivity (µmhos/cm) | -0.08 | -0.77 | 0.07 | -0.06 |

Note: factor values in "red" are principal components (factor score > 0.7, Liu et al. 2003)

Cluster 3(b): Four water quality stations were part of cluster 3(b), they were McConnell Rd., W.JJ Dr., and White St. Cluster 3(b) got reduced to five factors with fourteen principal factors. It represented 62.78% of total variance (Table 21). Factor (1) represented 24.42% of water quality with the eigen value of 4.15. It was a combination of the heavy metal and sediment. It consisted of copper, TSS and turbidity.

Factor (2) consisted of cadmium and nitrite. It represented the13.36% of total variance with eigen value of 2.27. Factor (3) represented 11.11% of total variance with the eigen value of 1.89. It represented physical component of water quality with hardness, TDS and conductivity as its principal components. Factor (4) represented the chemical component with COD. It contributed 7.22% of total variance with the eigen value of 1.23. Factor (5) did not have any

principal components but it contributed 6.66% total variance with the eigen value of 1.13.Reduction of parameters from seventeen to fourteen only reduced the dataset by 17.64% (Table 22). If one representative water quality station is chosen from the group, then the data reduction would be 72.5%. Figure 35 was a screeplot of eigen values.

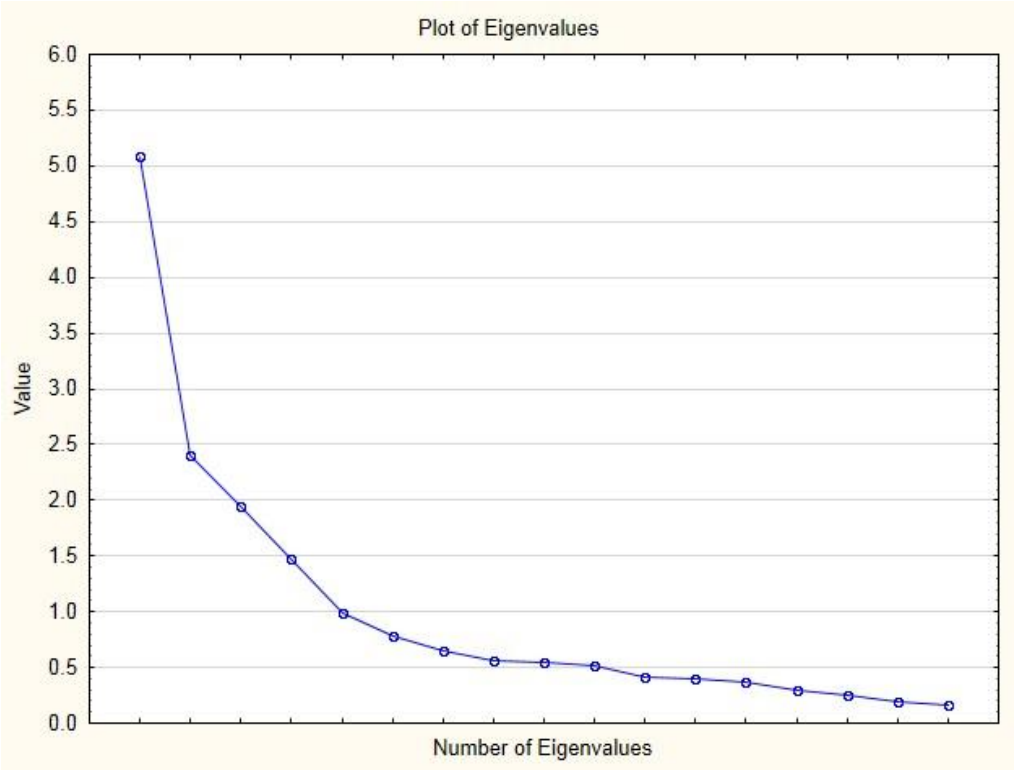


*Figure 34.* Screeplot of eigen values of principal factors for cluster 3(a)

Table 19

*Eigen Values for cluster 3(b)*

| | Eigenvalue | % Total (variance) | Cumulative (Eigenvalue) | Cumulative (%) |
|---|---|---|---|---|
| 1 | 4.15 | 24.42 | 4.15 | 24.42 |
| 2 | 2.27 | 13.36 | 6.42 | 37.78 |
| 3 | 1.89 | 11.11 | 8.31 | 48.89 |
| 4 | 1.23 | 7.22 | 9.54 | 56.12 |
| 5 | 1.13 | 6.66 | 10.67 | 62.78 |

Table 20

*Factor scores for most important factors for cluster 3(b)*

| Parameters | Factor (1) | Factor (2) | Factor (3) | Factor (4) | Factor (5) |
|---|---|---|---|---|---|
| Cadmium (mg/L) | 0.01 | 0.92 | 0.02 | -0.06 | -0.14 |
| Copper (mg/L) | 0.77 | 0.02 | -0.13 | 0.27 | 0.03 |
| Lead (mg/L) | 0.64 | 0.45 | -0.14 | -0.10 | -0.07 |
| Zinc (mg/L) | 0.50 | -0.13 | 0.08 | -0.15 | -0.10 |
| Alkalinity (mg/L) | -0.25 | 0.12 | 0.67 | -0.26 | 0.28 |
| BOD (mg/L) | 0.58 | -0.03 | 0.00 | 0.42 | 0.25 |
| COD (mg/L) | 0.04 | 0.33 | 0.02 | 0.72 | 0.25 |
| Fecal Coliform CFU/100 mL | 0.12 | -0.10 | -0.27 | 0.49 | -0.06 |
| Hardness (mg/L) | -0.30 | 0.10 | 0.76 | -0.16 | 0.23 |
| Nitrate Nitrogen (mg/L) | 0.13 | 0.28 | 0.09 | 0.16 | -0.67 |
| Nitrite Nitrogen (mg/L) | -0.09 | 0.91 | 0.05 | 0.11 | -0.02 |
| TDS (mg/L) | 0.01 | -0.07 | 0.84 | -0.03 | -0.22 |
| TSS (mg/L) | 0.79 | -0.02 | -0.13 | 0.05 | 0.15 |
| TKN (mg/L) | 0.04 | -0.05 | -0.10 | 0.63 | -0.17 |
| Total Phosphorus (mg/L) | 0.33 | 0.04 | -0.01 | 0.09 | 0.55 |
| Turbidity (ntu) | 0.72 | -0.03 | -0.28 | 0.17 | 0.01 |
| Conductivity (µmhos/cm) | -0.02 | -0.03 | 0.85 | -0.05 | -0.22 |

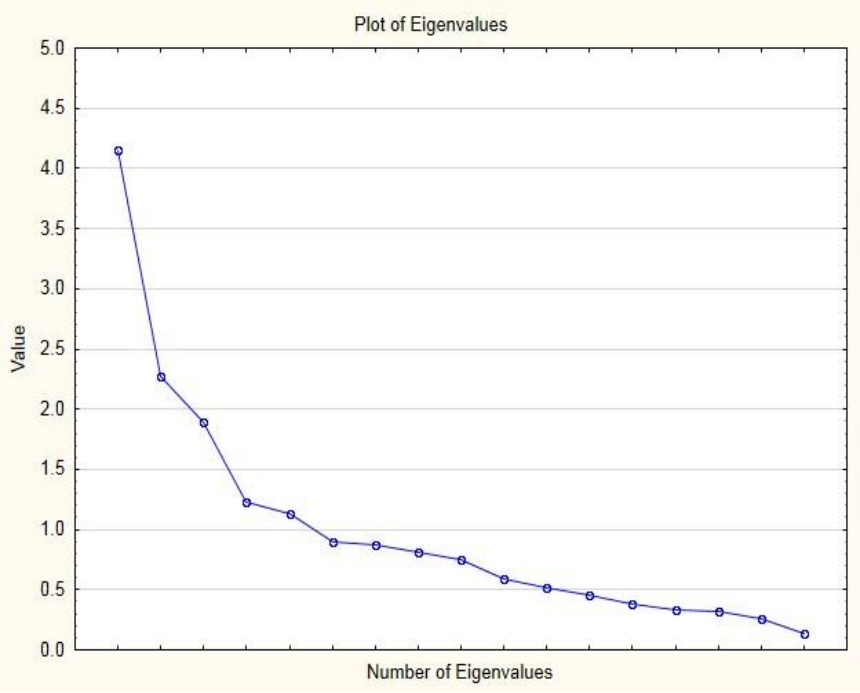


*Figure 35.* Screeplot of eigen values of principal factors for cluster 3(b)

Cluster (3a+3b): When all the water quality station in cluster (3) was analyzed together, whole dataset was divided into five factors which represented 63.99 % of total variance (Table 23). Factor (1) represented 25.29% of total variance with the eigen value of 4.30. It represented principal factor which were mixture of heavy metal and sediment. They were copper, TSS and turbidity. Factor (2) represented 13.57% of total variance with the eigen value of 2.31. It represented the physical component of water quality with TDS and alkalinity as its principal components.

Factor (3) had cadmium and nitrite as its principal component and it contributed 11.27% of total variance. It had eigen value of 1.92. Factor (4) represented 7.79% of total variance with the eigen value of 1.32. COD was the only principal component for the factor (4). Similarly, nitrate was the only principal component of factor (5). It contributed only 6.06% of total variance with the eigen value of 1.03.

Table 21

*Eigen Values for cluster (3a+3b)*

| | Eigen value | % Total (variance) | Cumulative (Eigen value) | Cumulative (%) |
|---|---|---|---|---|
| 1 | 4.30 | 25.29 | 4.30 | 25.29 |
| 2 | 2.31 | 13.57 | 6.61 | 38.87 |
| 3 | 1.92 | 11.27 | 8.52 | 50.14 |
| 4 | 1.32 | 7.79 | 9.85 | 57.93 |
| 5 | 1.03 | 6.06 | 10.88 | 63.99 |

In total the important principal factor were copper, TSS, turbidity, alkalinity, TDS, cadmium, nitrite, COD, and nitrate (Table 24). So, the important parameters were reduced to nine reducing the size of dataset by 47%. If one representative water quality station is chosen for cluster (3), it will reduce the dataset size by 92.43%. Figure 36 was a screeplot of eigen values.

Table 22

*Factor scores for most important factors for cluster (3a+3b)*

| Parameters | Factor (1) | Factor (2) | Factor (3) | Factor (4) | Factor (5) |
|---|---|---|---|---|---|
| Cadmium (mg/L) | 0.03 | 0.03 | 0.91 | -0.04 | 0.11 |
| Copper (mg/L) | 0.75 | -0.10 | 0.01 | 0.29 | 0.03 |
| Lead (mg/L) | 0.67 | -0.13 | 0.44 | -0.07 | 0.02 |
| Zinc (mg/L) | 0.60 | 0.08 | -0.15 | -0.11 | 0.24 |
| Alkalinity (mg/L) | -0.21 | 0.76 | 0.11 | -0.18 | -0.19 |
| BOD (mg/L) | 0.56 | 0.03 | -0.03 | 0.51 | -0.21 |
| COD (mg/L) | 0.03 | -0.02 | 0.33 | 0.74 | -0.12 |
| Fecal Coliform CFU/100 mL | 0.11 | -0.17 | -0.09 | 0.62 | 0.01 |
| Hardness (mg/L) | -0.22 | 0.83 | 0.08 | -0.12 | -0.12 |
| Nitrate Nitrogen (mg/L) | 0.12 | 0.08 | 0.30 | 0.14 | 0.71 |
| Nitrite Nitrogen (mg/L) | -0.07 | 0.05 | 0.89 | 0.08 | 0.06 |
| TDS (mg/L) | 0.03 | 0.77 | -0.09 | -0.06 | 0.23 |
| TSS (mg/L) | 0.79 | -0.15 | -0.03 | 0.12 | -0.18 |
| TKN (mg/L) | 0.11 | -0.15 | -0.06 | 0.63 | 0.18 |
| Total Phosphorus (mg/L) | 0.40 | -0.03 | 0.07 | 0.11 | -0.51 |
| Turbidity (ntu) | 0.72 | -0.30 | -0.03 | 0.20 | -0.05 |
| Conductivity (µmhos/cm) | -0.04 | 0.80 | 0.00 | -0.07 | 0.16 |

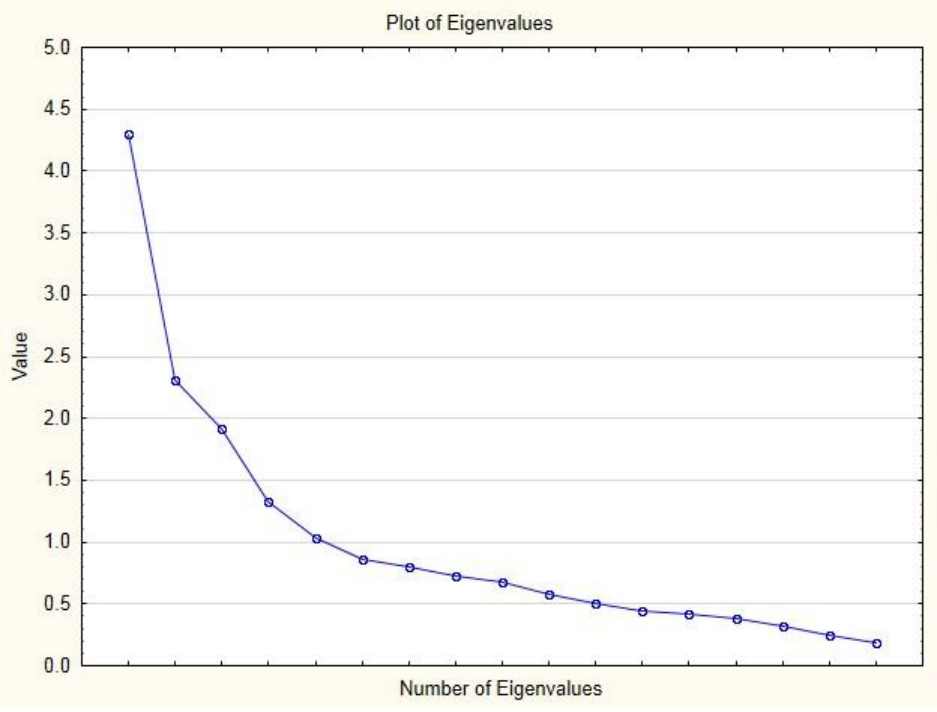


*Figure 36.* Screeplot of eigen values of principal factors for cluster (3a+3b)

Cluster (4): Cluster (4) consists of two water quality station located at 16th St and Fieldcrest Dr. Cluster (4) had five factors with eigen value greater than one (>1) representing 70.48% of total variance (Table 25). Factor (1) represented 24.34% of total variance with the eigen value of 4.14. It consisted of heavy metal and sediments (Table 26). They were lead, zinc, TSS and turbidity.

Factor (2) consisted of alkalinity, hardness, TDS and conductivity. It represented the physical component and contributed 16.38% of total variance with the eigen value of 2.79. Factor (3) contributed 11.92% of total variance with the eigen value of 2.03. It represented chemical component of water quality. It had BOD and COD as its principal component. Factor (5) had nitrate as its principal component contributed 6.67% of total variance with the eigen value of 1.13.

Table 23

*Eigen Values for cluster (4)*

| | Eigenvalue | % Total (variance) | Cumulative (Eigen value) | Cumulative (%) |
|---|---|---|---|---|
| 1 | 4.14 | 24.34 | 4.14 | 24.34 |
| 2 | 2.79 | 16.38 | 6.92 | 40.72 |
| 3 | 2.03 | 11.92 | 8.95 | 52.65 |
| 4 | 1.90 | 11.15 | 10.85 | 63.80 |
| 5 | 1.13 | 6.67 | 11.98 | 70.48 |

In total, there were thirteen principal factors reducing the data size by 23.52%. If one represented water quality station were to be chosen, there would be a reduction of 61.47% in the dataset size. Figure 36 was a screeplot of eigen values.

Table 24

*Factor scores for most important factors for cluster (4)*

| Parameters | Factor (1) | Factor (2) | Factor (3) | Factor (4) | Factor (5) |
|---|---|---|---|---|---|
| Cadmium (mg/L) | 0.05 | 0.05 | -0.02 | -0.95 | 0.03 |
| Copper (mg/L) | 0.38 | -0.03 | 0.23 | -0.03 | -0.04 |
| Lead (mg/L) | 0.83 | -0.13 | -0.04 | -0.33 | 0.10 |
| Zinc (mg/L) | 0.84 | 0.12 | 0.08 | 0.11 | -0.08 |
| Alkalinity (mg/L) | -0.13 | 0.82 | -0.04 | -0.13 | -0.14 |
| BOD (mg/L) | 0.28 | -0.05 | 0.86 | 0.02 | 0.08 |
| COD (mg/L) | -0.01 | 0.05 | 0.91 | -0.08 | 0.06 |
| Fecal Coliform CFU/100 mL | 0.01 | -0.05 | 0.66 | 0.05 | -0.09 |
| Hardness (mg/L) | -0.10 | 0.82 | 0.02 | -0.12 | -0.20 |
| Nitrate Nitrogen (mg/L) | 0.03 | 0.12 | -0.06 | -0.13 | 0.79 |
| Nitrite Nitrogen (mg/L) | -0.02 | 0.10 | 0.02 | -0.92 | 0.09 |
| TDS (mg/L) | -0.03 | 0.85 | -0.04 | 0.05 | 0.17 |
| TSS (mg/L) | 0.89 | -0.18 | 0.06 | 0.02 | 0.09 |
| TKN (mg/L) | 0.35 | -0.16 | 0.32 | 0.13 | 0.48 |
| Total Phosphorus (mg/L) | 0.48 | 0.30 | 0.09 | 0.13 | -0.43 |
| Turbidity (ntu) | 0.82 | -0.28 | 0.00 | 0.04 | 0.17 |
| Conductivity (µmhos/cm) | -0.07 | 0.86 | -0.05 | 0.00 | 0.13 |

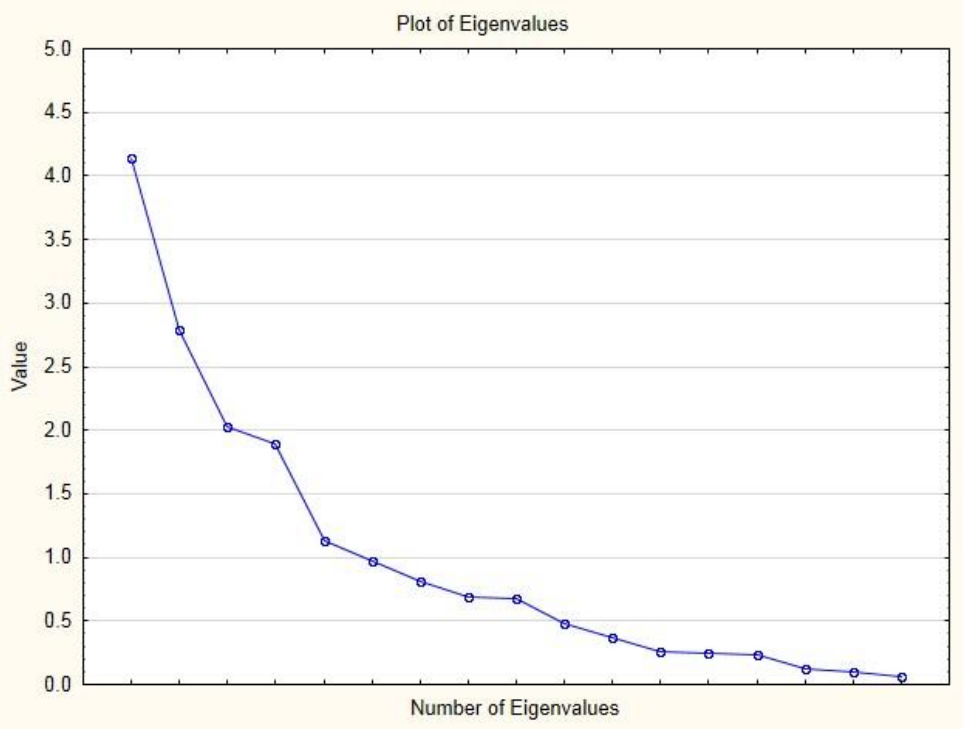


*Figure 37.* Screeplot of eigen values of principal factors for cluster (4)

**4.3.6 Cluster analysis of parameters.** First of all, all of the parameters were divided into different group according to their chemical, physical and biological properties. They were as follows:

1. Physical (P) - Conductivity, TDS, Hardness, Alkalinity
2. Chemical (C) – BOD, COD
3. Sediment (S) – Turbidity, TSS
4. Heavy Metal (HM) - Zinc, Copper, Lead, Cadmium
5. Nutrients (N) – TKN, TP, Nitrate, Nitrite
6. Bacteria (B) – Fecal Coliform
7. Mixed - Combination of different parameters

Then similarity between these groups of parameters was studied in different spatially similar water quality stations by the method of cluster analysis. At first, the data of entire water quality station was used for cluster analysis and parameters were clustered (Figure 38). They were grouped into four clusters for $(D_{link}/D_{max})$ *100 <50. First group consisted of conductivity, TDS, hardness, and alkalinity and was called as physical component of water quality. Second group consisted of eight parameters which had three distinct sub-divisions for $(D_{link}/D_{max})$ *100 < 42. First sub-division consisted of turbidity and TSS and was grouped under sediment. Second sub-division consisted of only zinc and third sub-division incorporated TKN, coliform, BOD, Total Phosphorus, and copper. So, third sub-division was a mixture of different component of water quality. Third cluster was divided into two distinct subgroups. First sub-group had nitrate, nitrite and COD as its elements and is mixture of nutrient and chemical component of water quality. Second subgroup consisted of lead and cadmium. It represented heavy metal component.

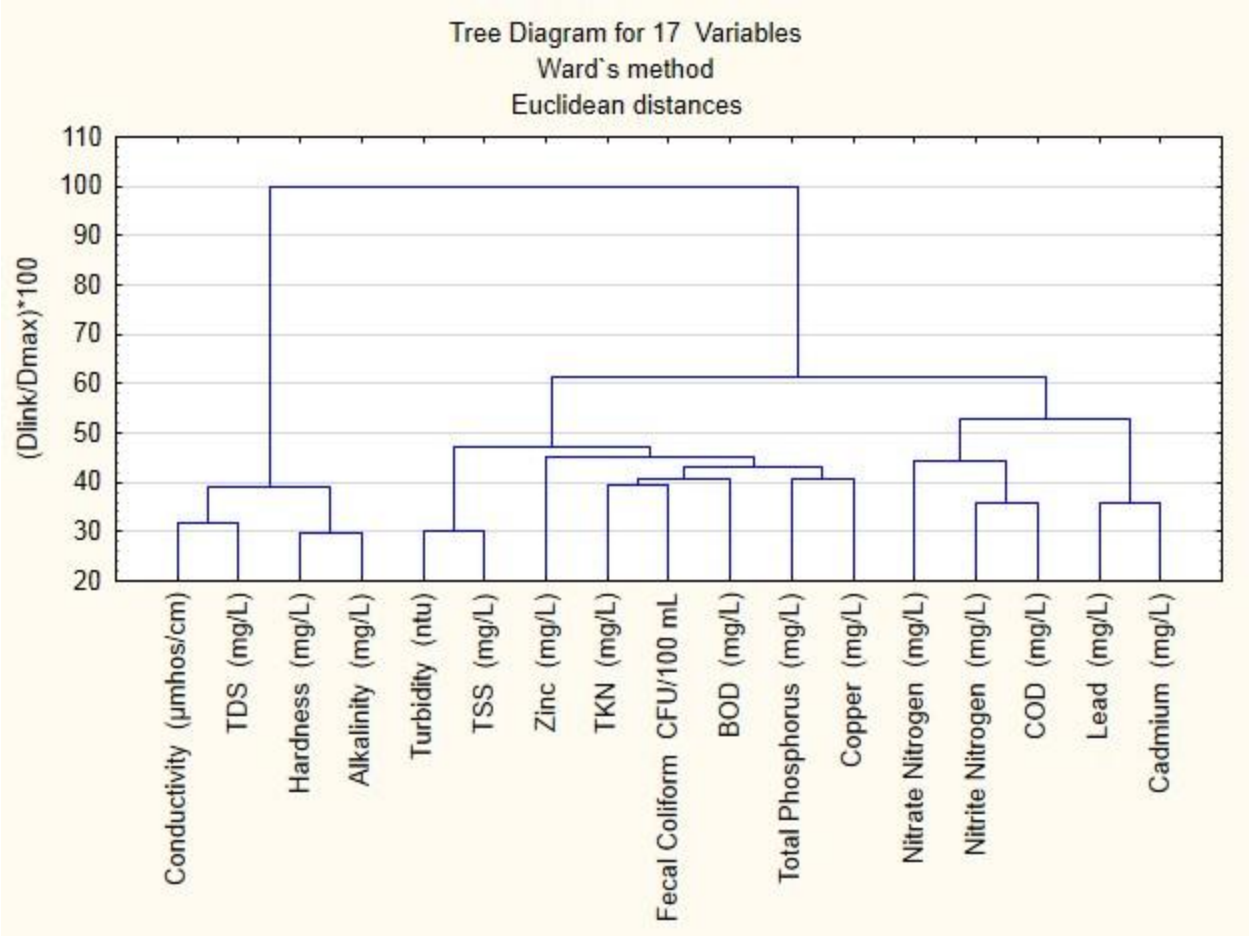


*Figure 38.* Cluster Analysis of all parameters for all water quality monitoring stations

Cluster 1(a): When CA was carried out on the dataset of water quality station from cluster 1(a), parameters were classified into four groups for $(D_{link}/D_{max})$ *100 <50 (Figure 39). First group consisted of conductivity, TDS, hardness and alkalinity and represented physical component of water quality. Second group consisted of nitrate, nitrite and COD. It was a mixture of N-C and did not represent any particular component of water quality. Third group consisted of TKN, fecal coliform, total phosphorus, BOD, turbidity, TSS and copper. It also did not represent any distinct water quality character and was a mixture of different components. Fourth group had zinc, lead and cadmium and represented heavy metal.

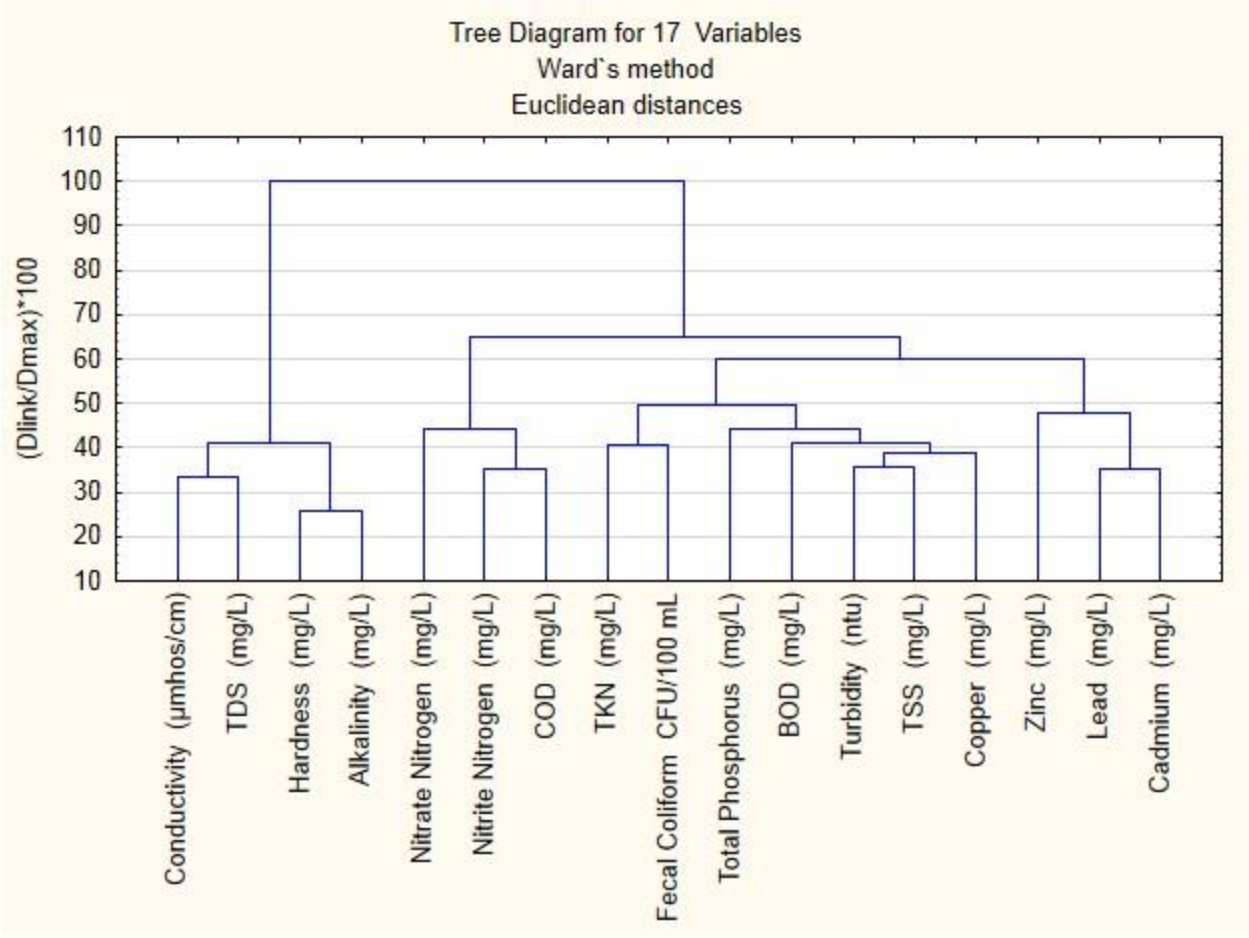


*Figure 39.* Cluster Analysis of all parameters for all water quality monitoring stations in cluster 1(a)

Cluster 1(b): Figure 40 showed the result of the CA on cluster 1(b). For $(D_{link}/D_{max})*100 > 60$, all the parameters were divided into five groups. First group represented the physical component of water quality and consisted of conductivity, TDS, hardness and alkalinity. Second group represented nutrient with nitrate and nitrite. Third and fourth clusters of parameters did not represent any particular component of water quality. It was a mixture of more than one group. Third group showed the correlation between TKN and COD whereas fourth group contained turbidity, TSS, Fecal Coliform, BOD, TP and copper. Fifth group showed the correlation between heavy metals; zinc, lead and cadmium. So, for Cluster 1(b) physical components, nutrients and heavy metals showed clear correlation among themselves.

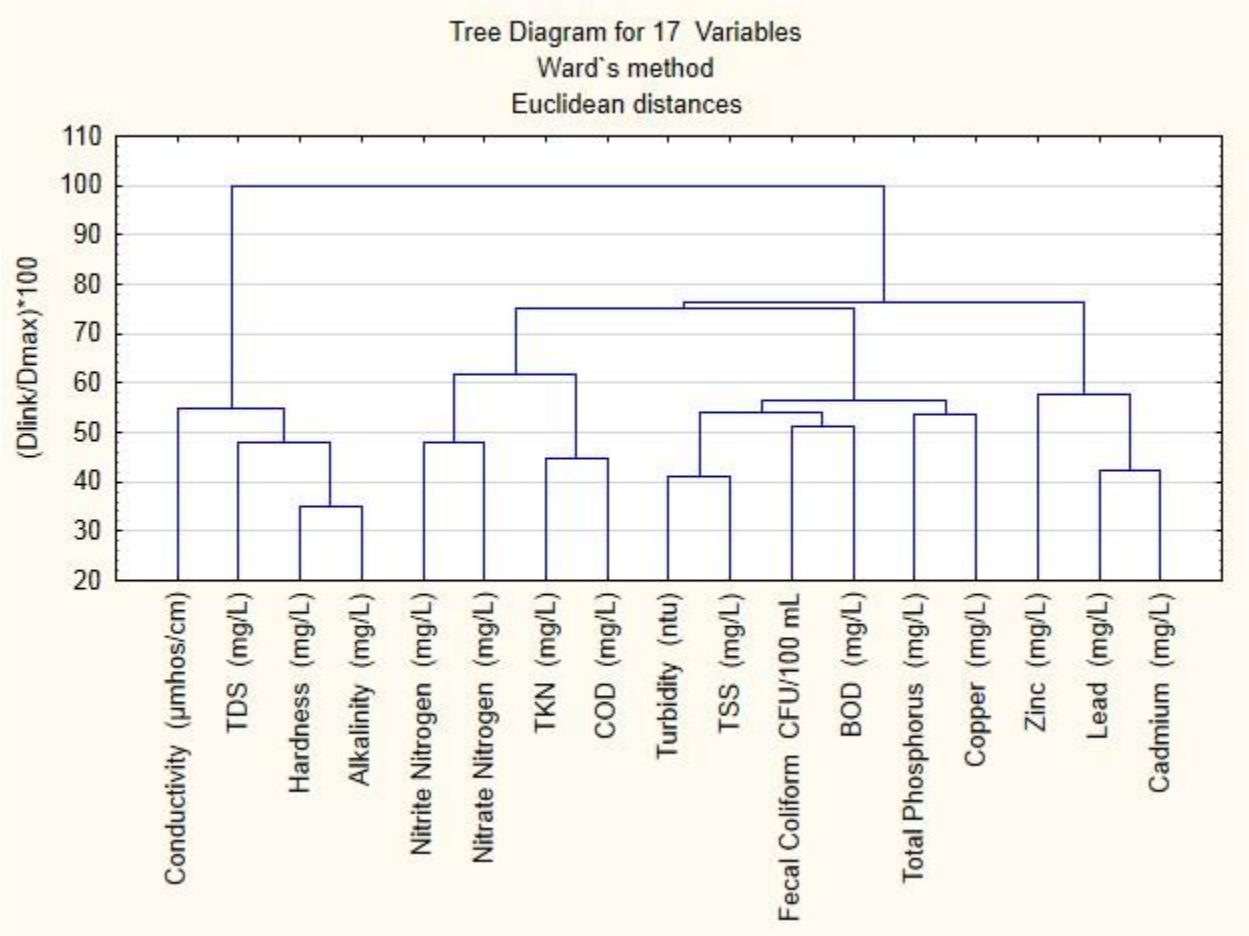


Figure 40. Cluster Analysis of all parameters for all water quality monitoring stations in cluster 1(b)

Cluster 1(1a+1b): Figure 41 shows the result of the CA of Cluster (1). For $(D_{link}/D_{max})*100> 50$, parameters were divided into six groups. These groups show the correlation between the parameters in the water quality monitoring stations in Cluster (1). By doing the CA of Cluster (1), effect of scale of data was also studied by comparing the results. Group one consisted of all the member of physical component of the water quality, which were true for both Clusters 1(a) and 1(b). Second group showed BOD was correlated with fecal coliform and TKN. So, it was mixture of different kind of parameters. Same was true for third

group which was a combination of TP, TSS, turbidity and copper. Group four represented the correlation between heavy metal and consisted of zinc, lead and cadmium.

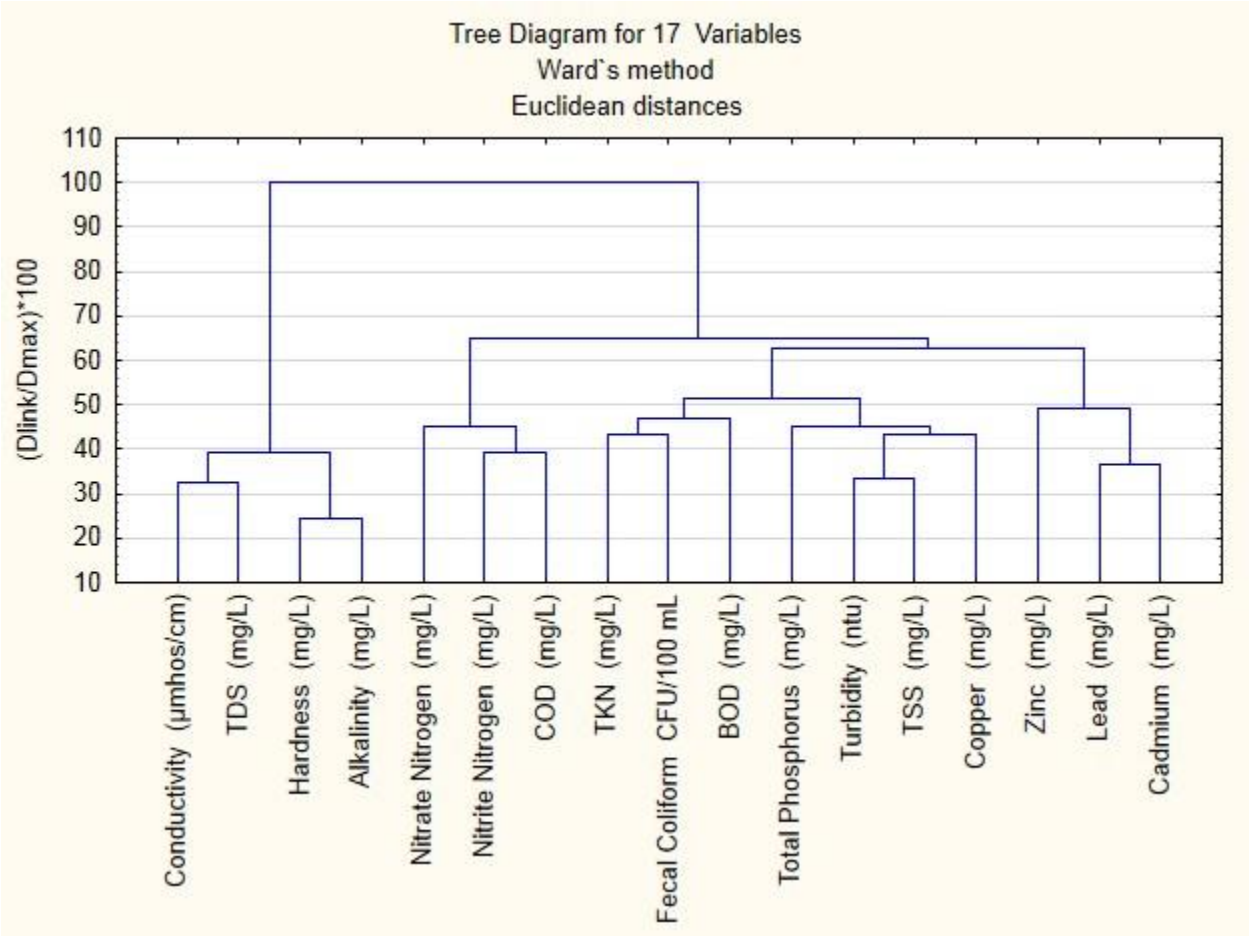


*Figure 41.* Cluster Analysis of all parameters for all water quality monitoring stations in the cluster (1a+1b)

Cluster 2: Figure 42 shows the result of the CA of Cluster (2). For $(D_{link}/D_{max})*100> 50$, parameters were divided into five groups. First group consisted of alkalinity, TDS, hardness and conductivity. It represented physical component of water quality. It showed these parameters were behaving in similar manner. Second group consisted of nitrate, TP and zinc. It was a mixture of heavy metal and nutrients. Turbidity and TSS were part of third group and represented sediment. Group four did not represent any distinct water quality class but was a

mixture. It consisted of nitrite, TKN and BOD. And fifth group was also mixture of parameters of different type's parameters. This group consisted of fecal coliform, copper, COD, lead and cadmium.

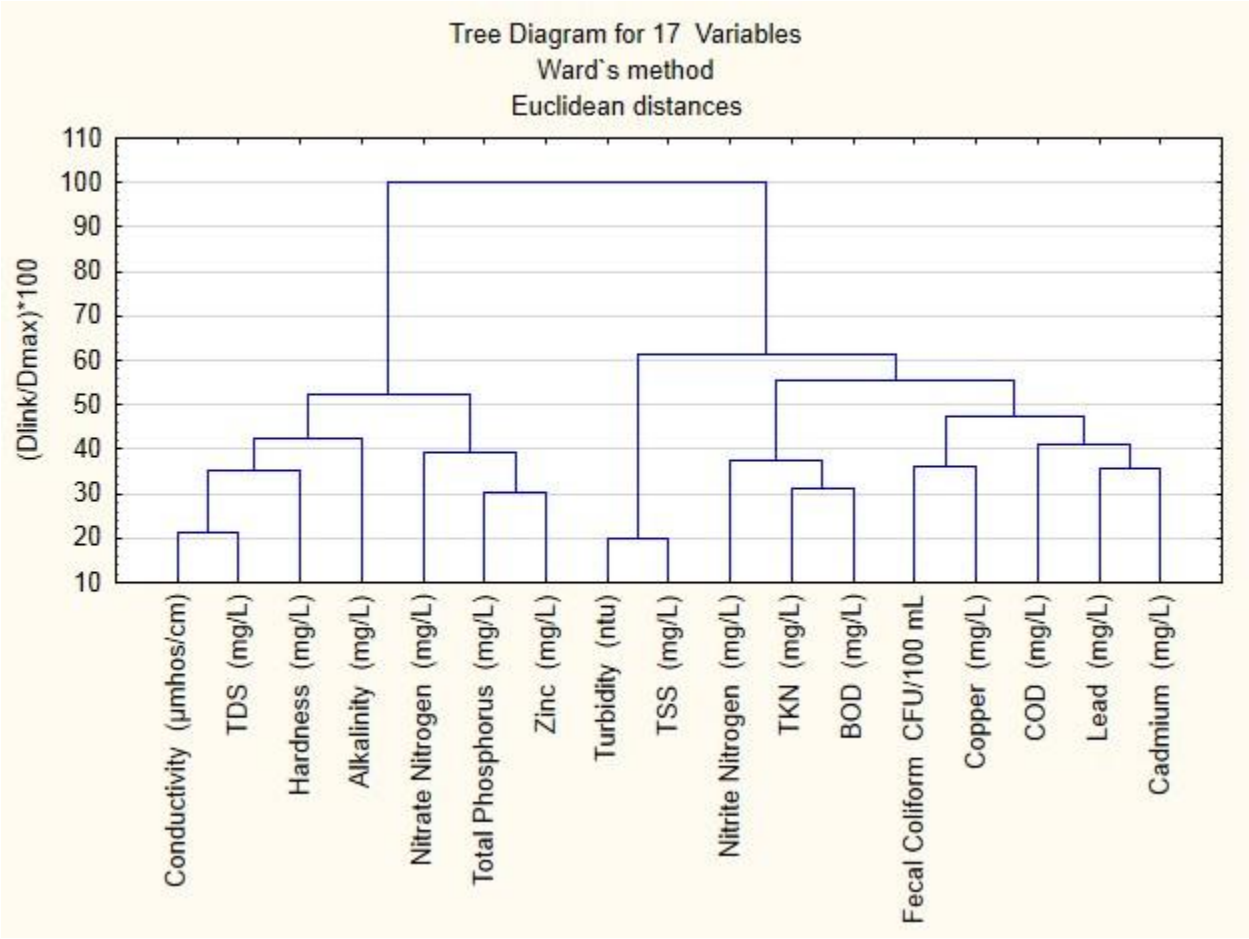


*Figure 42.* Cluster Analysis of all parameters for all water quality monitoring stations in cluster (2)

Cluster 3 (a): Figure 43 shows the result of the CA of Cluster 3(a). For $(D_{link}/D_{max})*100>$ 40, parameters were divided into five groups. First cluster represented physical component and consisted of alkalinity, TDS, hardness and conductivity. Second cluster was mixture of different types of parameter and consisted of fecal coliform, TKN, COD, TSS and BOD. Third cluster consisted of TP, turbidity and copper. Fourth cluster consisted of nitrate and zinc and fifth cluster consisted of nitrite, cadmium and lead.

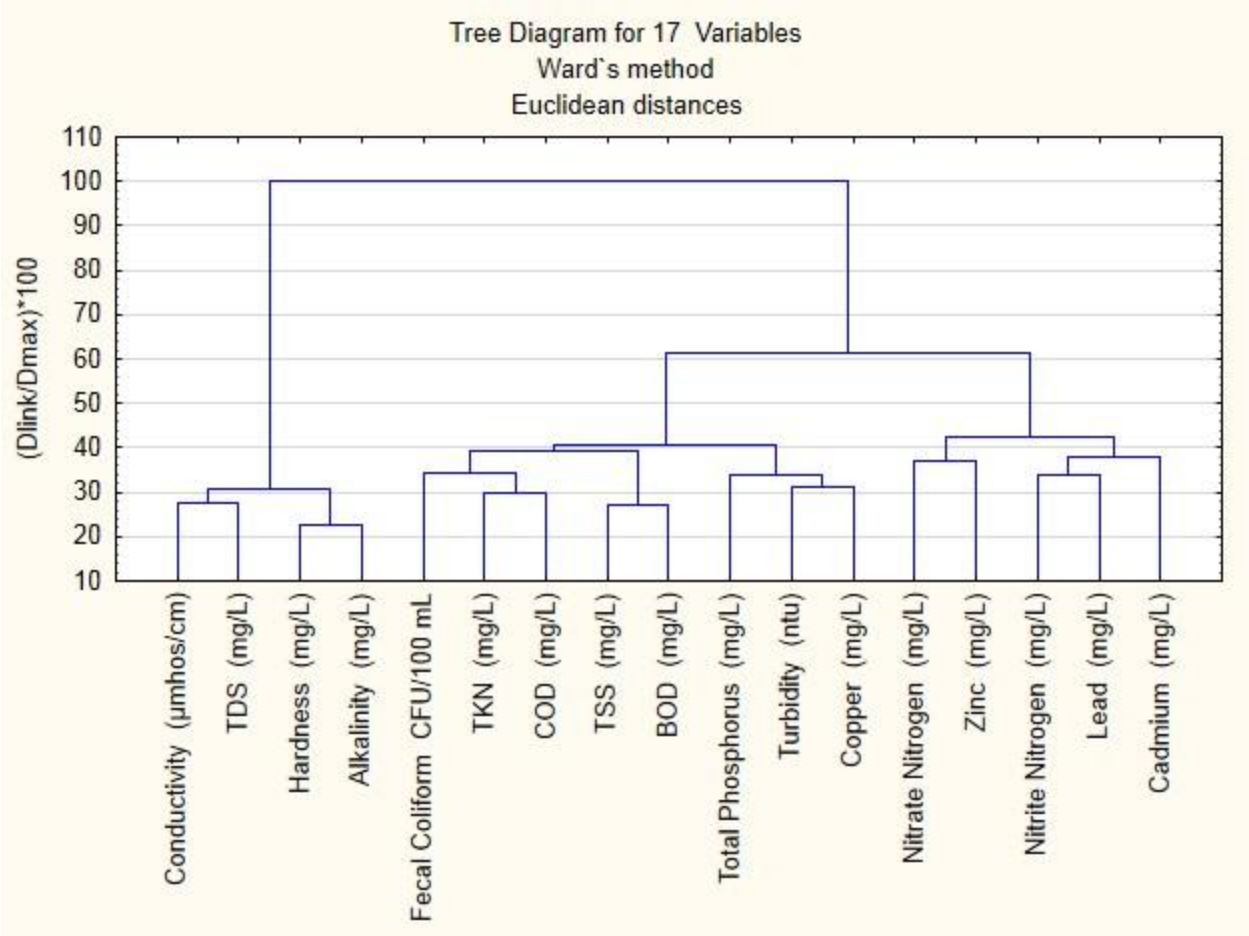


*Figure 43.* Cluster Analysis of all parameters for all water quality monitoring stations in cluster 3(a)

Cluster 3(b): Figure 44 shows the result of the CA of Cluster 3(b). For $(D_{link}/D_{max})*100>$ 40, parameters were divided into five groups. First cluster represented physical component and consisted of alkalinity, TDS, hardness and conductivity. Second cluster was mixture of different types of parameter and consisted of fecal coliform and TKN. Third cluster consisted of TSS, turbidity, zinc and lead. Fourth cluster consisted of TP, BOD and copper. And fifth cluster consisted of nitrite, nitrate, cadmium and COD.

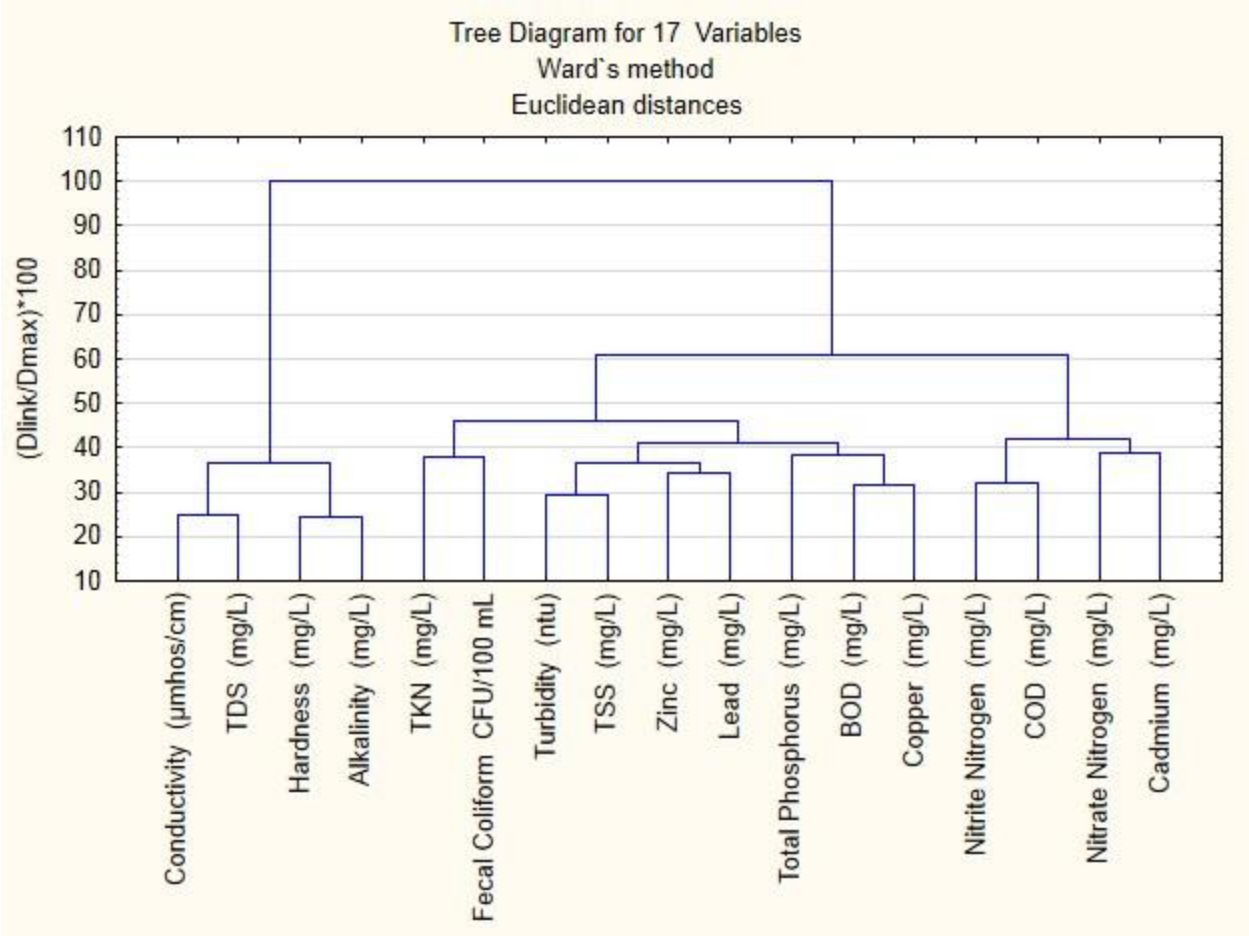


*Figure 44.* Cluster Analysis of all parameters for all water quality monitoring stations in cluster 3(b)

Cluster (3a+3b): Figure 45 shows the result of the CA of Cluster 3(a+b). For $(D_{link}/D_{max})*100 > 40$, parameters were divided into four groups. First cluster represented physical component and consisted of alkalinity, TDS, hardness and conductivity. Second cluster was mixture of different types of parameter and consisted of fecal coliform, COD, BOD and TKN. Third cluster consisted of TSS, TP, turbidity, zinc, copper and lead. It represented heavy metal and sediment. Fourth cluster consisted of nitrite, nitrate and cadmium.

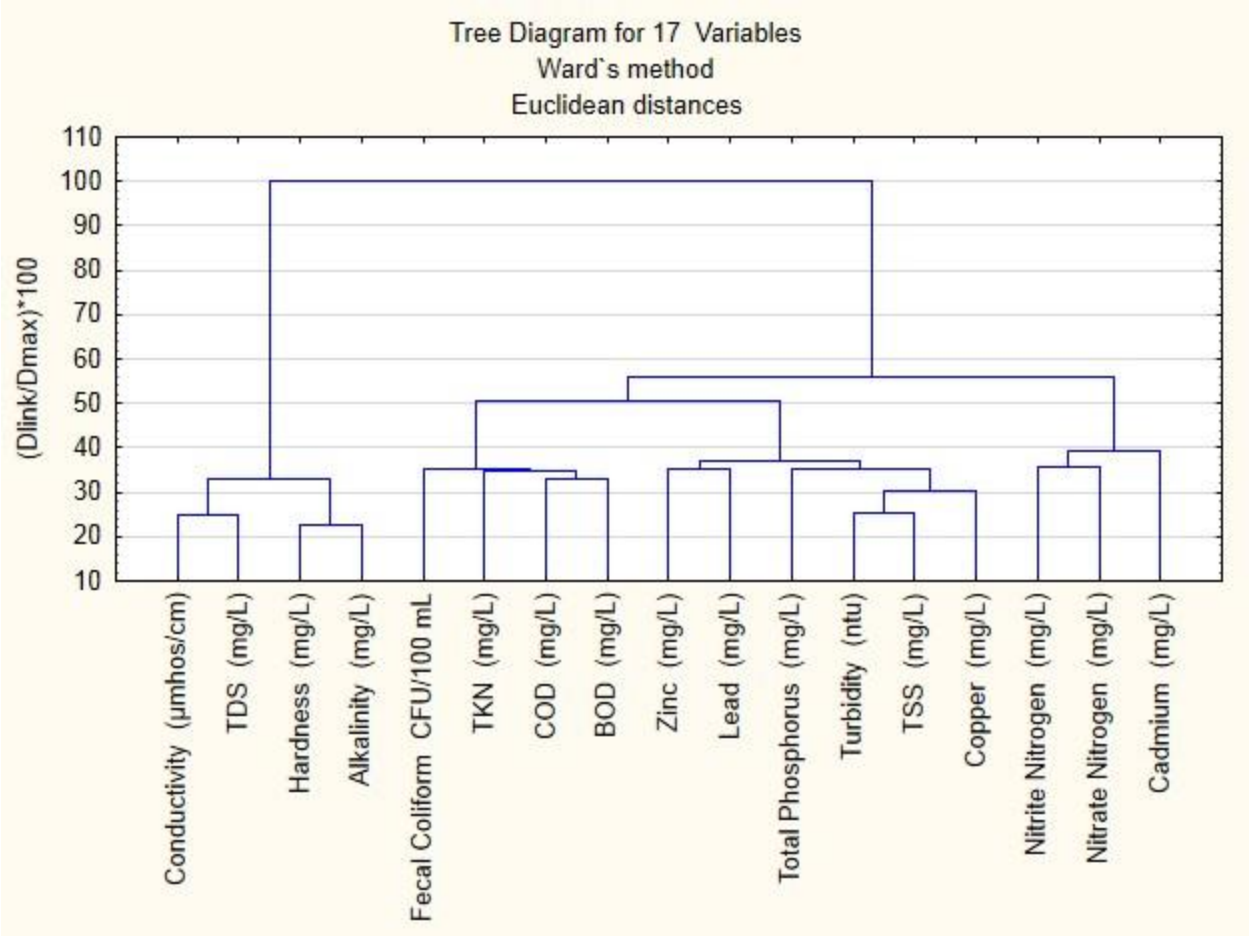


*Figure 45.* Cluster Analysis of all parameters for all water quality monitoring stations in cluster (3a+3b)

Cluster 4: Figure 46 shows the result of the CA of Cluster (4). For $(D_{link}/D_{max})*100> 50$, parameters were divided into six groups. First cluster represented physical component and consisted of alkalinity, TDS, hardness and conductivity. Second cluster consists of TSS and turbidity and represents sediment. Third cluster consisted of fecal coliform, zinc, TKN, BOD, TP and copper. Fourth cluster consisted of nitrite, nitrate, cadmium, lead and COD.

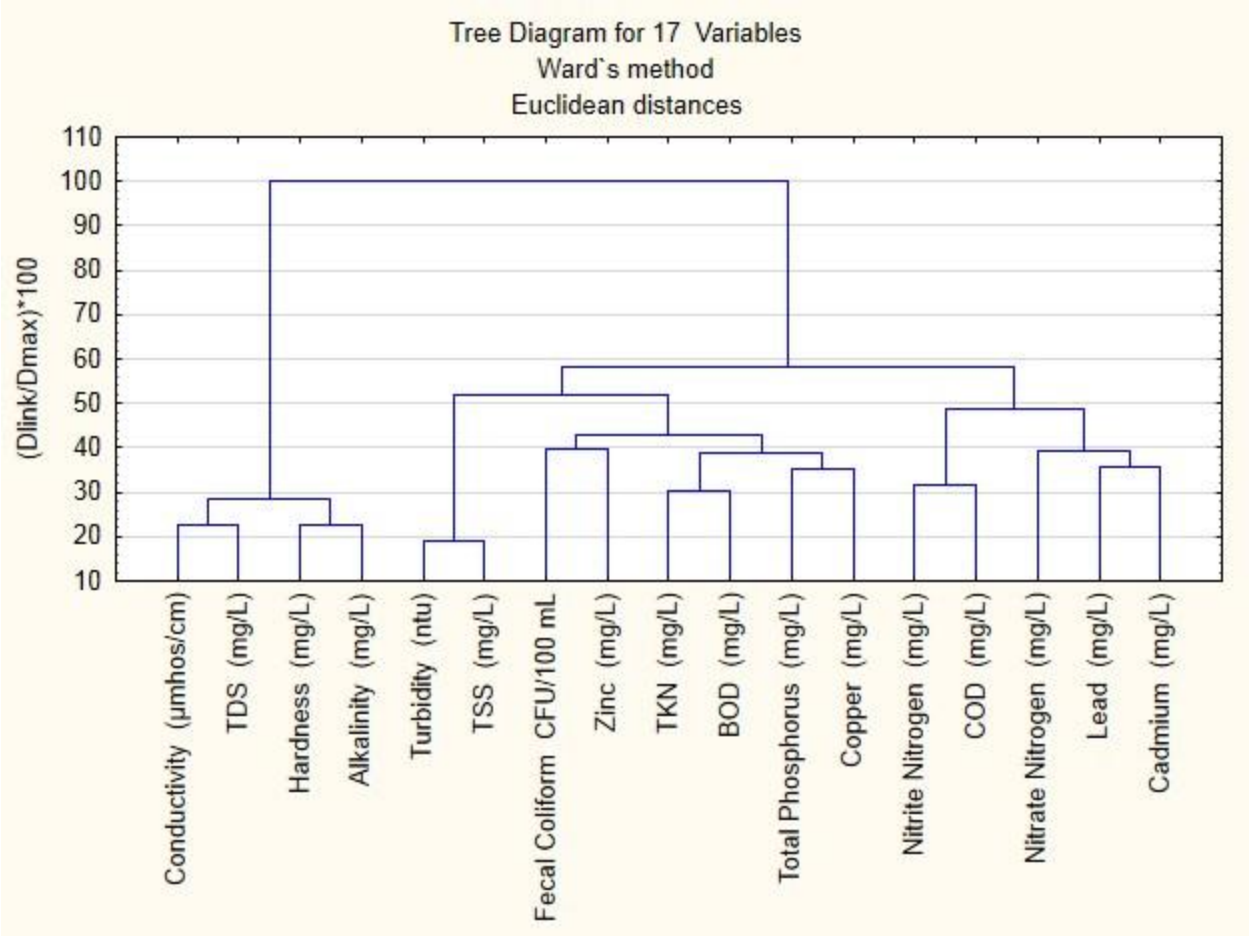


*Figure 46.* Cluster Analysis of all parameters for all water quality monitoring stations in cluster (4)

# CHAPTER 5

## Conclusions

The study presented here included the use of multiple statistical tools and techniques to better understand the variability of stream water quality parameters in the streams of City of Greensboro, North Carolina. The city regularly manages 16 monitoring stations within the city for 17 water quality parameters. The monitoring program used to sample on a bi-weekly basis since 1999, but since 2009 it started collecting it on a monthly basis. Various water quality parameters from all stations were analyzed statistically and learn the nature of pollution and data variability. The analyses were divided into 3 categories: rating curve development, comparison of 6 load estimation methods, and multivariate analyses (PCA and clustering analysis). The outputs of all methods were described below section wise.

### 5.1 Development of the Rating Curve

The LOADEST model was used for the construction of rating curve and load estimation. Three main statistical evaluators, coefficient of determination ($R^2$), partial load factor (PLF), and Nash- Sutcliffe Efficiency (NSE), were used to evaluate the model performance. Final results for each of the parameter:

1. Nitrate: the range of estimation varying from the "true" loading was from -6% to 16%. "True" loading represented the load calculated using simple flow-weighted method (Load = flow volume * concentration) extended for annual load comparison. It was overpredicted about 80% of the time (rest being underpredicted) representing a positive bias towards overprediction.. Regression between performance indicators found no correlation between PLF and $R^2$, poor correlation between PLF and NSE ($R^2$ = 0.35), and strong correlation between $R^2$ and NSF ($R^2$ =0.91).

2. Nitrite: Load estimation ranged between -14% to +12% from the true loadins. No systematic bias was found since 50% of the time load was underestimated and remaining 50% of the time it was overestimated. NSE value for nitrite was very poor for all of the water quality station except once. NSE and PLF showed strong correlation with $R^2$ value of 0.65. Correlation between PLF and $R^2$ was fairly weak ($R^2 = 0.34$), but there was almost no correlation between NSE and $R^2$. With increase in value of PLF resulted in decrease in value of NSE.
3. Total Dissolved Solids (TDS): Performance of LOADEST for TDS was exceptional with all performance evaluator performing very well. Estimated value differed from true value by -6% to 0%. Ninety percent of the time load estimated was less than true load, making the bias of estimation towards the underestimation. NSE and $R^2$ showed positive correlation ($R^2 = 0.46$).
4. Total Kjeldahl Nitrogen (TKN): Load estimation varied from true value by -2% and 9%. Model underestimated the loading 30% of the time and for the rest of 70% of time it overestimated the annual loading (positive bias). No correlation was found at all between the different performance indicators. All indicators were independent of each other with values of coefficient of determination being zero for them.
5. Total Phosphorus (TP): Load estimation varied from true loading by -22% to 9%. Model underestimated the value of TP 80% of the time (negative bias). PLF and $R^2$ showed average degree of correlation ($R^2 = 0.51$). There was no correlation found between NSE-$R^2$ and NSE-PLF.

6. Total Suspended Solids (TSS): Load estimation varied from the true loading by -51% to 23%. LOADEST in general underestimated the loading with 80% of time (negative bias). Performance indicators for TSS were found independent of each other (no correlation).

**5.2 Comparison of Load Estimation Methods**

A total of 6 interpolation methods/models (LOADEST, M1, M2, M3, M4 and M5) were constructed and used for annual load estimation from intermittent monitoring data. After analyzing the result for all the water quality stations, general patterns were established for each of the parameter, as follows:

1. For nitrate, TDS and TSS, the general trend of load estimated by different method followed the pattern: M3, M4 and M5 > LOADEST > M1 and M2
2. For nitrite, TKN and TP, the pattern was: LOADEST > M3, M4 and M5 > M1 and M2

**5.3 Multivariate Statistical Analysis**

Multivariate statistical methods were susceptible to different data pretreatment methods. Scaling pretreatment methods like auto-scaling and range-scaling produced well distributed dataset with no peaks. Vast-scaling and log-transformation methods produced datasets with few peaks without suppressing other parameters. Other pretreatment methods including centering, pareto-scaling, level-scaling and power transformations had one big peak suppressing all other parameters. Peaks in the pretreated data from centering, pareto-scaling, and power transformation were consistent with the original data. Pretreatment methods which produced bias data distribution (peak with suppressed data) ruled out themselves for further analyses. However, PCA analyses were carried out of for each of the methods for further examination.

In PCA analysis, the ranking of water quality parameters varied for each monitoring station according to different data pretreatment method. Ranking also differed among different

monitoring stations. Overall, the ranking of all parameters in all locations together found out that fecal coliform, conductivity and TSS are the most important parameter in the respective order and zinc, copper and cadmium in the same order as least important parameter. “Important” parameters directly refer to the degree of variability of parameters’ values.

Cluster analysis grouped water quality monitoring stations into spatially similar clusters in terms of similarity in variability of parameters (most to least important in PCA analysis). PCA/FA was applied on the entire dataset of the entire watershed as well as spatially similar stations. PCA/FA helped to reduce the size of the water quality dataset by reducing the parameter to be monitored whereas CA helped in reduction by reducing the number of the water quality station to be monitored. If one representative water quality station was chosen for each cluster, amount of data reduction achieved was 62.5%. When only FA/PCA was used, the amount of data reduced is 16.51%. Combination of FA/PCA and CA reduced the size of dataset by 71% and it represented the 64.47% of the total variance.

PCA/FA and CA was also used to see the correlation between different parameters. From these analyses we could see clear relationship between Conductivity, TDS, Hardness, and Alkalinity, and turbidity and TSS. Conductivity, TDS, Hardness, and Alkalinity represented the physical component of the water quality whereas turbidity and TSS represented sediment. Other water quality parameters appeared in mixed relationship with each other, which varied with different cluster of water quality stations.

The multivariate techniques applied in this study were very useful in data reduction as well as in interpreting the complex data. These techniques gave mathematical tools to plan and conduct effective and efficient water quality monitoring program by reducing in number of monitoring station and parameters to be monitored.

# References

Adams, S., Titus, R., Pietesen, K., Tredoux, G., & Harris, C. (2001). Hydrochemical characteristic of aquifers near Sutherland in the Western Karoo, South Africa. *Journal of Hydrology*, 241, 91-103. doi:10.1016/S0022-1694(00)00370-X

Alexander, R.B., Smith, R.A., Schwarz, G.E., Boyer, E.W., Nolan, J.V., & Brakebill, J.W. (2008). Differences in phosphorous and nitrogen delivery to the Gulf of Mexico from the Mississippi River Basin. *Environ. Sci. Technol*., 42, 822–830. doi:10.1021/es0716103

Aulenbach, B.T., & Hooper, R.P. (2006). The composite method: an improved method for stream-water solute load estimation. *Hydrol. Process*. 20, 3029–3047. doi:10.1002/hyp.6147

Boyacioglu, H., Boyacioglu, H., & Gunduz, O. (2005). Application of Factor Analysis in the Assessment of Surface Water Quality in BUYUK Menderes River Basin. *European Waters*, 9/10, 43-49, 2005.

Bradu, D., & Mundlak, Y. (1970). Estimation in lognormal linear models. *Journal of the American Statistical Association* 65(329), 198–211.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963–1987—Applications of quantile regression. *Econometrica*, 62(2), 405–458. Retrieved from http://www.cec.zju.edu.cn/~yao/lepp/QR3.pdf

Cohn, T.A. (1988). *Adjusted maximum likelihood estimation of the moments of lognormal populations from type I censored samples: U.S. Geological Survey Open-File Report 88-350*. Retrieved from http://pubs.usgs.gov/of/1988/0350/report.pdf

Cohn, T.A. (1995). Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers. *Reviews in Geophysics,* 33, 1117–1124.

Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., & Summers, R.M. (1992). The validity of a simple statistical model for estimating fluvial constituent loads: an empirical study involving nutrient loads in Chesapeake bay. *Water Resour. Res.,* 28 (9), 2353–2363. doi:10.1029/92WR01008

Cohn, T.A., Delong, L.L., Gilroy, E.J., Hirsch, R.M., & Wells, D.K. (1989) Estimating constituent loads. *Water Resources Research* 25(5), 937-942.

Cohn, T.A., Gilroy, E.J., & Baier, W.G. (1992b). Estimating fluvial transport of trace constituents using a regression model with data subject to censoring. 1992 *Proceedings of the Joint Statistical Meeting.*

Crawford, C.G. (1991). Estimation of suspended-sediment rating curves and mean suspended sediment loads. *Journal of Hydrology,* 129, 331–348. doi:10.1016/0022-1694(91)90057-O

Draper, N.R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley and Sons.

Duan, N. (1983). Smearing estimate—A nonparametric retransformation method. *Journal of the American Statistical Association,* 78, 605–610. doi:10.1080/01621459.1983.10478017

Efron, B. (1982). *The Jackknife, the Bootstrap, and other resampling plans (Cbms-Nsf Regional Conference Series in Applied Mathematics; 38).* Philadelphia: Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611970319.fm

Ferguson, R.I. (1986). River loads underestimated by rating curves. *Water Resources Research,* 22(1), 74–76.

Gilroy, E.J., Hirsch, R.M., & Cohn, T.A. (1990). Mean square error of regression-based constituent transport estimates. *Water Resources Research,* 26(9), 2069–2077. doi:10.1029/WR026i009p02069

Goolsby, D.A., & Battaglin, W.A. (2001). Long-term changes in concentrations and flux of nitrogen in the Mississippi River Basin, USA. *Hydrol. Process*., 15, 1209–1226. doi:10.1002/hyp.210

Goolsby, D.A., Battaglin, W.A., Aulenbach, B.T., & Hooper, H.P. (2000). Nitrogen flux and sources in the Mississippi River Basin. *Sci. Total Environ*., 248, 75–86. doi:10.1016/S0048-9697(99)00532-X

Guo ,Y., Markus, M., & Demissie, M. (2002). Uncertainty of nitrate-N load computations for agricultural watersheds. *Water Resourses,* 38(10), 1185. doi:10.1029/2001WR001149.

Helena, B., Pardo, R., Vega, M., Barrado, E., Ferna´ndez, J.M., & Ferna´ndez, L. (2000). Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. *Water Research,* 34, 807-816. doi:10.1016/S0043-1354(99)00225-0

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

Helsel, D.R., & Hirsch, R.M. (2002). Describing Uncertainty. In U.S. Geological Survey Techniques of Water-Resources Investigations, *Statistical methods in water resources* (pp-510). (http://water.usgs.gov/pubs/twri/twri4a3)

Hooper, R.P., Aulenbach, B.T., & Kelly, V.J. (2001). The national stream quality accounting network: a flux-based approach to monitoring the water quality of large rivers. *Hydrol. Process*., 15, 1089–1106. doi:10.1002/hyp.205

Jha, M.K., Wolter, C.F., Gassman, P.W., & Schilling, K.E. (2010a). Assessment of TMDL implementation strategies for nitrate impairment of the Raccoon River, Iowa. *J. Environ. Qual.*, 39, 4, 1317-1327, doi:10.2134/jeq.2009.0392

Jha, M.K., Schilling, K.E., Gassman, P.W., & Wolter, C.F. (2010b). Targeting land-use change for nitrate-nitrogen load reductions in an agricultural watershed. *J. Soil. Wat. Cons.*, 65, 2, 342-352, doi: 10.2489/jswc.65.6.342

Jha, M.K., Arnold, J.G., & Gassman, P.W. (2007). Water quality modeling for the Raccoon River Watershed. *Transactions of the ASABE*, 50, 2, 479-493.

Judge, G.G., Hill, R.C., Griffiths, W.E., Lutkepohl, H., & Lee, T.C. (1988). *Introduction to the theory and practice of econometrics (2nd edition).* New York: John Wiley.

Lee, J.Y., Cheon, J.Y., Lee, K.K., Lee, S.Y., & Lee, M.H. (2001). Statistical evaluation of geochemical parameter distribution in a ground water system contaminated with petroleum hydrocarbons. *Journal of Environmental Quality,* 30, 1548-1563. doi:0.2134/jeq2001.3051548x

Li, Z., Zhang, Y.-K., Schilling, K., & Skopec, M. (2006). Cokriging estimation of suspended sediment loads. *J. Hydrol*., 327, 389–398. doi:10.1016/j.jhydrol.2005.11.028

Likes, J. (1980). Variance of the MVUE for lognormal variance. *Technometrics*, 22(2), 253–258.

Littlewood, I.G., Watts, C.D., & Custance, J.M. (1998). Systematic application of United Kingdom river flow and quality databases for estimating annual river mass loads (1975–1994). *Sci. Total Environ*., 210–211, 21–40. doi:10.1016/S0048-9697(98)00042-4

Liu, C.W., Lin, K.H., & Kuo, Y.M. (2003). Application of factor analysis in the assessment of groundwater quality in Blackfoot Disease area in Taiwan. *Sci. Total Environ.*, 313, 77-89.

Maret, T.R., MacCoy, D.E., & Carlisle, D.M. (2008). Long-term water quality and

biological responses to multiple best management practices in rock creek, Idaho. *J. Am. Water Resour. Assoc*., 44 (5), 1248–1269. doi:10.1111/j.1752-1688.2008.00221.x

Moatar, F., & Meybeck, M. (2005). Compared performance of different algorithms for estimating annual nutrient loads discharged by the eutrophic River Loire. *Hydrol. Process*., 19, 429–444. doi:10.1002/hyp.5541

Powell, J.L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*,25, 303–325. doi:10.1016/0304-4076(84)90004-6

Preston, S.D., Bierman Jr., V.J., & Sillman, S.E. (1989). An evaluation of methods for the estimation of tributary mass loads. *Water Resour.*, 25, 1379–1389. doi:10.1029/WR025i006p01379

Reghunath, R., Murthy, T.R.S., & Raghavan, B.R. (2002). The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. *Water Research*, 36, 2437-2442. doi:10.1016/S0043-1354(01)00490-0

Robertson, D.M., & Roerish, E.D. (1999). Influence of various water quality sampling trategies on load estimates for small streams. *Water Resour.*, 35, 3747–3759. doi:10.1029/1999WR900277

Santhi, C., Arnold, J. G., Williams, J. R., Dugas, W. A., & Hauck, L. (2001): Validation of the SWAT model on a large river basin with point and nonpoint sources. *Journal American. Water Resources. Association*, 37(5), 1169–1188.

SÂRBU, C. & POP, H. (2005). Principal Component Analysis versus Fuzzy Principal Component Analysis A Case Study: The Quality of Danube Water (1985–1996). *Talanta*, 65, 1215-1220. doi:10.1016/j.talanta.2004.08.047

Schilling, K.E., & Zhang, Y. K. (2004). Baseflow contribution to nitrate-nitrogen export from a large agricultural watershed USA. *J. Hydrol.*, 295, 305–316. doi:10.1016/j.jhydrol.2004.03.010

Shenton, L.R., & Bowman, K.O. (1977). *Maximum likelihood estimation in small samples.* London: Charles Griffin and Co.

Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, & M., Kouimtzis, T. (2003). Assessment of the surface water quality in Northern Greece. *Water Research,* 37, 4119-4124. doi:10.1016/S0043-1354(03)00398-1

Singh, K.P., Malik, A., Mohan, & D., Sinha, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): a case study. *Water Research,* 38, 3980-3992. doi:10.1016/j.watres.2004.06.011

Turnbull, B.W., & Weiss, L. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, 34, 367–375. Retrieved from http://www.jstor.org/stable/2530599

Ullrich, A., & Volk, M. (2010). Influence of different nitrate-N monitoring strategies on load estimation as a base for model calibration and evaluation. *Environ. Monit. Assess*. doi:10.1007/s10661-009-1296-8.

USGS (2004). *Load Estimator (LOADEST): A Fortran Program for Estimating Constituent Loads in Streams and Rivers. Techniques and Models Book 4, Chapter 5, US Geological Survey, Reston, VA.* Retrieved from http://water.usgs.gov/software/loadest/doc/

USGS (2009a). *USGS Open-File Report 2007-1080 – Streamflow and Nutrient Fluxes of the Mississippi–Atchafalaya River Basin and Subbasins for the Period of*

*Record Through 2005, Methods Used to Estimate Nutrient Fluxes*. Retrieved from <http://toxics.usgs.gov/pubs/of-2007-1080/methods.html>.

USGS (2009b). *Application of Spatially Referenced Regression Modeling for the Evaluation of Total Nitrogen Loading in the Chesapeake Bay Watershed.* Retrieved from <http://md.water.usgs.gov/publications/wrir-99-4054/html/index.htm>.

Van Liew, M. W., Arnold, J. G., & Garbrecht, J. D. (2003). Hydrologic simulation on agricultural watersheds: Choosing between two models. *Trans. ASAE,* 46(6), 1539-1551.

Vega, M., Pardo, R., Barrado, E., & Deban, L. (1998). Assessment of seasonal and polluting effects on the qualityof river water byex ploratory data analysis. *Water Res.*, 32, 3581–3592. doi:10.1016/S0043-1354(98)00138-9

Vogel, R.M. (1986). The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distributional Hypothesis. *Water Resources Research*, 22(4), 587–590. doi:10.1029/WR022i004p00587

Wunderlin, D.A., Diaz, M.P., Ame, M.V., Pesce, S.F., Hued, A.C., & Bistoni, M.A. (2001). Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba, Argentina). *Water Research,* 35, 2881-2894. doi:10.1016/S0043-1354(00)00592-3

Xu, H. S., Xu, Z. X., Wu, W., & Tang, F.F. (2012). Assesment and Spatiotemporal Variation of Water Quality in the Zhangweinan River Basin, China. *Procedia Environmental Sciences.*, 13(2012), 1641-1652. doi:10.1016/j.proenv.2012.01.157

Zamyadi, A., Gallichand, J., & Duchemin, M. ( 2007). Comparison of methods for estimating sediment and nitrogen loads from a small agricultural watershed. *Can. Biosyst. Eng.*, 49, 127–136. doi:10.1016/j.jhydrol.2010.11.006

*Appendix*

Table 1.

*Ranking of parameters obtained by PCA conducted on data pretreated by different methods.*

*(a) 16$^{th}$ St.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 10 | 3 | 16 | 12 | 3 | 7 | 16 | 5 |
| 11 | 5 | 8 | 11 | 4 | 15 | 9 | 13 | 8 |
| 5 | 3 | 9 | 5 | 6 | 14 | 10 | 5 | 7 |
| 8 | 15 | 13 | 8 | 13 | 10 | 14 | 8 | 14 |
| 12 | 4 | 6 | 12 | 2 | 16 | 13 | 14 | 6 |
| 15 | 7 | 1 | 15 | 10 | 6 | 1 | 15 | 17 |
| 6 | 6 | 10 | 6 | 3 | 13 | 16 | 6 | 13 |
| 7 | 9 | 4 | 7 | 5 | 7 | 15 | 7 | 16 |
| 14 | 8 | 7 | 14 | 16 | 2 | 2 | 12 | 9 |
| 17 | 2 | 2 | 17 | 1 | 17 | 17 | 17 | 12 |
| 10 | 12 | 14 | 10 | 14 | 8 | 11 | 10 | 4 |
| 9 | 17 | 17 | 9 | 15 | 5 | 3 | 9 | 11 |
| 13 | 16 | 16 | 13 | 17 | 1 | 4 | 11 | 10 |
| 4 | 11 | 12 | 4 | 8 | 12 | 12 | 4 | 2 |
| 2 | 14 | 15 | 2 | 11 | 4 | 5 | 2 | 15 |
| 3 | 13 | 11 | 3 | 9 | 9 | 6 | 3 | 3 |
| 1 | 1 | 5 | 1 | 7 | 11 | 8 | 1 | 1 |

*(b) Aycock St.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 11 | 6 | 16 | 14 | 4 | 7 | 16 | 8 |
| 12 | 6 | 8 | 12 | 4 | 15 | 12 | 13 | 13 |
| 6 | 14 | 7 | 6 | 8 | 13 | 13 | 5 | 17 |
| 8 | 3 | 3 | 8 | 3 | 11 | 11 | 8 | 12 |
| 14 | 9 | 9 | 13 | 2 | 17 | 16 | 15 | 16 |
| 15 | 15 | 4 | 15 | 15 | 3 | 4 | 14 | 5 |
| 5 | 2 | 12 | 5 | 5 | 14 | 15 | 6 | 7 |
| 7 | 7 | 17 | 7 | 9 | 6 | 14 | 7 | 10 |
| 13 | 5 | 5 | 14 | 16 | 1 | 3 | 12 | 15 |
| 17 | 10 | 1 | 17 | 1 | 16 | 17 | 17 | 4 |
| 10 | 8 | 10 | 10 | 11 | 7 | 9 | 10 | 6 |
| 9 | 13 | 11 | 9 | 13 | 5 | 1 | 9 | 9 |
| 11 | 12 | 15 | 11 | 17 | 2 | 2 | 11 | 3 |
| 4 | 16 | 14 | 4 | 7 | 12 | 10 | 4 | 11 |
| 2 | 17 | 16 | 2 | 12 | 8 | 8 | 2 | 14 |
| 3 | 4 | 13 | 3 | 10 | 9 | 5 | 3 | 2 |
| 1 | 1 | 2 | 1 | 6 | 10 | 6 | 1 | 1 |

*(c) Battleground Avenue*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 16 | 11 | 16 | 14 | 4 | 5 | 14 | 8 |
| 14 | 1 | 1 | 14 | 4 | 14 | 11 | 15 | 17 |
| 5 | 7 | 10 | 5 | 6 | 11 | 7 | 5 | 7 |
| 8 | 13 | 16 | 8 | 10 | 9 | 10 | 9 | 13 |
| 13 | 4 | 5 | 13 | 2 | 16 | 14 | 16 | 14 |
| 15 | 11 | 2 | 15 | 17 | 1 | 2 | 13 | 5 |
| 6 | 2 | 4 | 6 | 5 | 15 | 17 | 6 | 10 |
| 7 | 12 | 15 | 7 | 11 | 6 | 13 | 7 | 12 |
| 12 | 8 | 7 | 11 | 15 | 2 | 3 | 10 | 16 |
| 17 | 9 | 13 | 17 | 1 | 17 | 16 | 17 | 9 |
| 10 | 14 | 17 | 10 | 12 | 8 | 8 | 12 | 6 |
| 9 | 5 | 12 | 9 | 13 | 5 | 1 | 8 | 11 |
| 11 | 10 | 14 | 12 | 16 | 3 | 4 | 11 | 2 |
| 4 | 3 | 6 | 4 | 3 | 12 | 9 | 4 | 3 |
| 2 | 15 | 3 | 2 | 9 | 7 | 15 | 3 | 15 |
| 3 | 17 | 9 | 3 | 7 | 10 | 6 | 2 | 1 |
| 1 | 6 | 8 | 1 | 8 | 13 | 12 | 1 | 4 |

*(d) Bluff Run Rd.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 1 | 1 | 16 | 5 | 4 | 5 | 16 | 7 |
| 10 | 9 | 15 | 10 | 6 | 10 | 12 | 14 | 8 |
| 5 | 6 | 6 | 6 | 10 | 12 | 10 | 5 | 13 |
| 8 | 12 | 14 | 8 | 11 | 8 | 9 | 8 | 16 |
| 13 | 11 | 5 | 12 | 2 | 16 | 17 | 15 | 12 |
| 15 | 16 | 9 | 15 | 16 | 1 | 1 | 11 | 14 |
| 6 | 8 | 13 | 5 | 3 | 17 | 14 | 7 | 5 |
| 7 | 15 | 8 | 7 | 13 | 5 | 7 | 6 | 9 |
| 14 | 7 | 7 | 14 | 15 | 3 | 3 | 12 | 3 |
| 17 | 10 | 4 | 17 | 1 | 15 | 16 | 17 | 11 |
| 12 | 14 | 12 | 13 | 8 | 13 | 15 | 13 | 10 |
| 9 | 4 | 2 | 9 | 9 | 11 | 4 | 9 | 17 |
| 11 | 13 | 10 | 11 | 17 | 2 | 2 | 10 | 15 |
| 4 | 2 | 17 | 4 | 12 | 7 | 8 | 4 | 4 |
| 2 | 17 | 16 | 2 | 14 | 6 | 13 | 3 | 6 |
| 3 | 5 | 3 | 3 | 4 | 9 | 6 | 2 | 1 |
| 1 | 3 | 11 | 1 | 7 | 14 | 11 | 1 | 2 |

*(e) Church St.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 12 | 6 | 16 | 15 | 4 | 5 | 16 | 8 |
| 11 | 4 | 3 | 11 | 6 | 12 | 14 | 12 | 7 |
| 6 | 3 | 1 | 6 | 1 | 13 | 9 | 5 | 17 |
| 8 | 13 | 5 | 8 | 7 | 8 | 6 | 8 | 9 |
| 13 | 9 | 4 | 13 | 2 | 17 | 16 | 14 | 13 |
| 15 | 14 | 7 | 15 | 14 | 3 | 2 | 15 | 12 |
| 5 | 2 | 11 | 5 | 5 | 15 | 15 | 6 | 5 |
| 7 | 6 | 14 | 7 | 10 | 5 | 10 | 7 | 16 |
| 14 | 17 | 8 | 14 | 16 | 2 | 4 | 13 | 6 |
| 17 | 7 | 12 | 17 | 4 | 16 | 17 | 17 | 10 |
| 10 | 15 | 15 | 10 | 12 | 7 | 11 | 10 | 3 |
| 9 | 10 | 13 | 9 | 13 | 6 | 1 | 9 | 11 |
| 12 | 11 | 17 | 12 | 17 | 1 | 3 | 11 | 14 |
| 4 | 5 | 2 | 4 | 3 | 14 | 12 | 4 | 2 |
| 2 | 16 | 10 | 2 | 9 | 10 | 13 | 2 | 4 |
| 3 | 8 | 16 | 3 | 11 | 9 | 7 | 3 | 15 |
| 1 | 1 | 9 | 1 | 8 | 11 | 8 | 1 | 1 |

*(f) Fieldcrest Dr.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 15 | 7 | 16 | 14 | 4 | 6 | 16 | 8 |
| 12 | 3 | 6 | 10 | 4 | 16 | 13 | 12 | 9 |
| 6 | 10 | 8 | 6 | 5 | 14 | 15 | 6 | 17 |
| 8 | 13 | 16 | 8 | 13 | 6 | 7 | 8 | 12 |
| 13 | 5 | 3 | 13 | 1 | 17 | 16 | 15 | 13 |
| 15 | 14 | 9 | 15 | 15 | 2 | 1 | 14 | 14 |
| 5 | 11 | 11 | 5 | 7 | 13 | 10 | 5 | 16 |
| 7 | 8 | 12 | 7 | 12 | 5 | 14 | 7 | 5 |
| 14 | 17 | 14 | 14 | 16 | 3 | 3 | 13 | 10 |
| 17 | 16 | 17 | 17 | 2 | 15 | 17 | 17 | 15 |
| 10 | 4 | 2 | 11 | 6 | 12 | 8 | 11 | 11 |
| 9 | 6 | 5 | 9 | 9 | 10 | 4 | 9 | 7 |
| 11 | 7 | 15 | 12 | 17 | 1 | 2 | 10 | 6 |
| 4 | 9 | 10 | 4 | 10 | 7 | 11 | 4 | 4 |
| 2 | 2 | 4 | 3 | 8 | 9 | 12 | 2 | 1 |
| 3 | 1 | 1 | 2 | 3 | 11 | 9 | 3 | 3 |
| 1 | 12 | 13 | 1 | 11 | 8 | 5 | 1 | 2 |

*(g) Friendship St.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 4 | 2 | 16 | 4 | 8 | 5 | 16 | 12 |
| 11 | 7 | 15 | 12 | 10 | 11 | 10 | 14 | 16 |
| 5 | 9 | 9 | 5 | 9 | 9 | 8 | 5 | 13 |
| 8 | 17 | 17 | 8 | 11 | 6 | 6 | 9 | 5 |
| 10 | 8 | 14 | 11 | 5 | 12 | 17 | 12 | 7 |
| 15 | 13 | 10 | 15 | 17 | 2 | 4 | 15 | 8 |
| 6 | 2 | 7 | 6 | 3 | 16 | 16 | 6 | 9 |
| 7 | 12 | 16 | 7 | 7 | 13 | 11 | 8 | 14 |
| 14 | 15 | 13 | 14 | 14 | 4 | 2 | 10 | 10 |
| 17 | 1 | 1 | 17 | 1 | 17 | 15 | 17 | 17 |
| 12 | 10 | 12 | 10 | 8 | 10 | 12 | 13 | 6 |
| 9 | 14 | 6 | 9 | 15 | 1 | 1 | 7 | 11 |
| 13 | 16 | 8 | 13 | 16 | 3 | 7 | 11 | 3 |
| 4 | 11 | 5 | 4 | 2 | 15 | 9 | 4 | 15 |
| 3 | 6 | 11 | 3 | 13 | 7 | 14 | 3 | 4 |
| 2 | 5 | 3 | 2 | 12 | 5 | 3 | 2 | 1 |
| 1 | 3 | 4 | 1 | 6 | 14 | 13 | 1 | 2 |

*(h) Kivett St.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 10 | 6 | 16 | 14 | 3 | 6 | 16 | 13 |
| 11 | 6 | 1 | 11 | 3 | 15 | 14 | 14 | 8 |
| 5 | 14 | 15 | 5 | 10 | 9 | 10 | 5 | 7 |
| 8 | 7 | 14 | 8 | 9 | 11 | 8 | 8 | 17 |
| 14 | 17 | 7 | 13 | 1 | 17 | 16 | 15 | 10 |
| 15 | 15 | 5 | 15 | 17 | 1 | 2 | 13 | 12 |
| 6 | 2 | 4 | 6 | 4 | 14 | 13 | 6 | 9 |
| 7 | 13 | 13 | 7 | 7 | 12 | 15 | 9 | 5 |
| 13 | 4 | 12 | 14 | 16 | 4 | 4 | 11 | 16 |
| 17 | 3 | 3 | 17 | 2 | 16 | 17 | 17 | 14 |
| 10 | 11 | 17 | 10 | 12 | 8 | 9 | 10 | 15 |
| 9 | 8 | 10 | 9 | 13 | 2 | 1 | 7 | 4 |
| 12 | 5 | 11 | 12 | 15 | 5 | 3 | 12 | 6 |
| 4 | 16 | 9 | 4 | 6 | 10 | 7 | 4 | 2 |
| 2 | 12 | 8 | 2 | 8 | 6 | 12 | 3 | 11 |
| 3 | 9 | 16 | 3 | 11 | 7 | 5 | 2 | 3 |
| 1 | 1 | 2 | 1 | 5 | 13 | 11 | 1 | 1 |

*(i) Mackay Rd.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 16 | 3 | 16 | 15 | 4 | 6 | 15 | 8 |
| 14 | 13 | 6 | 14 | 3 | 14 | 14 | 16 | 16 |
| 5 | 6 | 8 | 5 | 8 | 8 | 7 | 5 | 17 |
| 8 | 9 | 15 | 8 | 10 | 9 | 8 | 8 | 12 |
| 13 | 3 | 7 | 13 | 4 | 16 | 15 | 14 | 13 |
| 15 | 12 | 5 | 15 | 17 | 6 | 4 | 13 | 9 |
| 6 | 5 | 9 | 6 | 5 | 13 | 16 | 6 | 5 |
| 7 | 17 | 10 | 7 | 7 | 11 | 11 | 7 | 10 |
| 12 | 10 | 16 | 12 | 14 | 2 | 2 | 11 | 14 |
| 17 | 11 | 2 | 17 | 2 | 17 | 17 | 17 | 7 |
| 10 | 4 | 13 | 10 | 13 | 5 | 9 | 10 | 6 |
| 9 | 7 | 14 | 9 | 12 | 3 | 3 | 9 | 11 |
| 11 | 8 | 17 | 11 | 16 | 1 | 1 | 12 | 3 |
| 3 | 14 | 11 | 3 | 9 | 10 | 5 | 3 | 4 |
| 2 | 15 | 12 | 2 | 11 | 7 | 13 | 2 | 15 |
| 4 | 2 | 1 | 4 | 1 | 15 | 10 | 4 | 2 |
| 1 | 1 | 4 | 1 | 6 | 12 | 12 | 1 | 1 |

*(j) McConnell Rd.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 16 | 12 | 16 | 16 | 4 | 6 | 16 | 17 |
| 11 | 1 | 2 | 11 | 4 | 15 | 14 | 13 | 8 |
| 5 | 6 | 4 | 5 | 9 | 8 | 9 | 5 | 5 |
| 8 | 10 | 8 | 8 | 8 | 11 | 7 | 8 | 9 |
| 14 | 3 | 1 | 12 | 2 | 17 | 16 | 15 | 7 |
| 15 | 9 | 3 | 15 | 15 | 3 | 1 | 14 | 13 |
| 6 | 4 | 6 | 6 | 3 | 14 | 15 | 6 | 12 |
| 7 | 8 | 9 | 7 | 6 | 9 | 13 | 7 | 3 |
| 13 | 11 | 15 | 14 | 14 | 1 | 4 | 12 | 4 |
| 17 | 5 | 10 | 17 | 1 | 16 | 17 | 17 | 16 |
| 10 | 12 | 13 | 10 | 10 | 6 | 11 | 10 | 14 |
| 9 | 13 | 7 | 9 | 13 | 5 | 2 | 9 | 6 |
| 12 | 15 | 16 | 13 | 17 | 2 | 3 | 11 | 10 |
| 4 | 14 | 17 | 4 | 7 | 12 | 10 | 4 | 11 |
| 2 | 17 | 14 | 3 | 12 | 10 | 12 | 3 | 15 |
| 3 | 7 | 11 | 2 | 11 | 7 | 5 | 2 | 2 |
| 1 | 2 | 5 | 1 | 5 | 13 | 8 | 1 | 1 |

*(k) Merritt Dr.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 17 | 10 | 16 | 14 | 4 | 9 | 15 | 17 |
| 14 | 5 | 14 | 14 | 3 | 15 | 15 | 16 | 8 |
| 5 | 14 | 9 | 5 | 5 | 13 | 12 | 5 | 16 |
| 8 | 13 | 15 | 8 | 13 | 5 | 6 | 8 | 7 |
| 13 | 2 | 3 | 13 | 2 | 17 | 16 | 14 | 5 |
| 15 | 9 | 1 | 15 | 15 | 3 | 2 | 13 | 13 |
| 6 | 4 | 11 | 6 | 4 | 12 | 11 | 6 | 9 |
| 7 | 7 | 6 | 7 | 7 | 8 | 14 | 7 | 14 |
| 12 | 8 | 12 | 12 | 16 | 2 | 4 | 12 | 12 |
| 17 | 6 | 5 | 17 | 1 | 16 | 17 | 17 | 15 |
| 10 | 16 | 16 | 10 | 11 | 6 | 10 | 10 | 3 |
| 9 | 11 | 4 | 9 | 12 | 7 | 1 | 9 | 10 |
| 11 | 15 | 17 | 11 | 17 | 1 | 3 | 11 | 4 |
| 4 | 10 | 13 | 4 | 8 | 11 | 8 | 4 | 6 |
| 3 | 12 | 8 | 3 | 9 | 14 | 13 | 3 | 11 |
| 2 | 1 | 2 | 2 | 10 | 9 | 5 | 2 | 1 |
| 1 | 3 | 7 | 1 | 6 | 10 | 7 | 1 | 2 |

*(l) Old Oak Ridge Rd.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 15 | 1 | 4 | 14 | 4 | 16 | 14 | 15 | 9 |
| 5 | 14 | 8 | 5 | 5 | 15 | 13 | 5 | 8 |
| 8 | 12 | 5 | 8 | 7 | 12 | 9 | 9 | 13 |
| 14 | 2 | 3 | 15 | 2 | 14 | 16 | 16 | 1 |
| 16 | 17 | 2 | 16 | 15 | 3 | 5 | 14 | 17 |
| 7 | 4 | 1 | 7 | 1 | 17 | 17 | 7 | 14 |
| 6 | 7 | 17 | 6 | 12 | 6 | 11 | 6 | 10 |
| 13 | 11 | 12 | 12 | 17 | 1 | 2 | 13 | 6 |
| 17 | 16 | 6 | 17 | 3 | 13 | 15 | 17 | 11 |
| 11 | 9 | 14 | 11 | 11 | 11 | 10 | 11 | 15 |
| 9 | 15 | 7 | 9 | 9 | 8 | 1 | 8 | 16 |
| 12 | 8 | 13 | 13 | 16 | 2 | 3 | 12 | 7 |
| 4 | 13 | 15 | 4 | 8 | 10 | 7 | 4 | 5 |
| 2 | 6 | 9 | 2 | 10 | 5 | 12 | 3 | 12 |
| 3 | 3 | 10 | 3 | 14 | 4 | 4 | 2 | 4 |
| 1 | 5 | 11 | 1 | 6 | 9 | 6 | 1 | 3 |
| 10 | 10 | 16 | 10 | 13 | 7 | 8 | 10 | 2 |

*(m) Randleman Rd.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 6 | 6 | 16 | 14 | 5 | 7 | 16 | 8 |
| 11 | 11 | 7 | 11 | 6 | 12 | 11 | 13 | 10 |
| 6 | 14 | 2 | 6 | 3 | 14 | 12 | 5 | 17 |
| 8 | 9 | 8 | 8 | 8 | 8 | 6 | 9 | 13 |
| 13 | 12 | 1 | 12 | 2 | 16 | 16 | 12 | 9 |
| 15 | 7 | 3 | 15 | 15 | 4 | 2 | 15 | 16 |
| 5 | 4 | 12 | 5 | 4 | 15 | 15 | 7 | 7 |
| 7 | 16 | 16 | 7 | 11 | 10 | 14 | 6 | 5 |
| 14 | 3 | 14 | 14 | 16 | 3 | 4 | 14 | 12 |
| 17 | 2 | 4 | 17 | 1 | 17 | 17 | 17 | 11 |
| 10 | 10 | 17 | 10 | 10 | 6 | 8 | 10 | 3 |
| 9 | 13 | 9 | 9 | 12 | 2 | 1 | 8 | 14 |
| 12 | 8 | 15 | 13 | 17 | 1 | 3 | 11 | 6 |
| 4 | 5 | 13 | 4 | 7 | 11 | 10 | 4 | 15 |
| 2 | 15 | 11 | 2 | 9 | 9 | 13 | 3 | 2 |
| 3 | 17 | 10 | 3 | 13 | 7 | 5 | 2 | 4 |
| 1 | 1 | 5 | 1 | 5 | 13 | 9 | 1 | 1 |

*(n) Rankin Mill Rd.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 7 | 4 | 16 | 14 | 4 | 9 | 16 | 8 |
| 13 | 2 | 3 | 12 | 3 | 15 | 12 | 13 | 10 |
| 6 | 13 | 16 | 6 | 5 | 13 | 13 | 6 | 13 |
| 7 | 11 | 5 | 7 | 7 | 10 | 7 | 7 | 17 |
| 14 | 8 | 1 | 14 | 2 | 17 | 16 | 15 | 7 |
| 15 | 1 | 8 | 15 | 17 | 1 | 1 | 14 | 5 |
| 5 | 4 | 7 | 5 | 4 | 14 | 15 | 5 | 15 |
| 9 | 16 | 17 | 9 | 8 | 11 | 14 | 9 | 16 |
| 12 | 10 | 2 | 13 | 16 | 2 | 2 | 11 | 12 |
| 17 | 15 | 9 | 17 | 1 | 16 | 17 | 17 | 6 |
| 10 | 14 | 11 | 10 | 9 | 8 | 10 | 12 | 9 |
| 8 | 12 | 14 | 8 | 10 | 6 | 3 | 8 | 14 |
| 11 | 17 | 12 | 11 | 15 | 3 | 4 | 10 | 3 |
| 4 | 9 | 6 | 4 | 12 | 5 | 6 | 4 | 11 |
| 2 | 6 | 13 | 2 | 13 | 9 | 11 | 3 | 4 |
| 3 | 5 | 10 | 3 | 11 | 7 | 5 | 2 | 1 |
| 1 | 3 | 15 | 1 | 6 | 12 | 8 | 1 | 2 |

*(o) W.JJ Dr.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 15 | 1 | 2 | 15 | 11 | 4 | 8 | 16 | 8 |
| 12 | 5 | 7 | 12 | 5 | 14 | 15 | 13 | 7 |
| 5 | 15 | 13 | 5 | 7 | 13 | 11 | 5 | 9 |
| 8 | 10 | 11 | 8 | 13 | 9 | 7 | 9 | 13 |
| 11 | 6 | 6 | 11 | 4 | 15 | 16 | 14 | 16 |
| 16 | 17 | 1 | 16 | 9 | 5 | 3 | 15 | 12 |
| 6 | 2 | 5 | 6 | 2 | 16 | 14 | 6 | 14 |
| 7 | 4 | 17 | 7 | 8 | 10 | 13 | 7 | 15 |
| 14 | 8 | 16 | 14 | 17 | 2 | 4 | 12 | 5 |
| 17 | 16 | 9 | 17 | 1 | 17 | 17 | 17 | 10 |
| 10 | 12 | 15 | 10 | 15 | 8 | 9 | 10 | 17 |
| 9 | 9 | 3 | 9 | 14 | 3 | 1 | 8 | 3 |
| 13 | 7 | 10 | 13 | 16 | 1 | 2 | 11 | 11 |
| 4 | 11 | 12 | 4 | 6 | 7 | 6 | 4 | 6 |
| 2 | 14 | 4 | 2 | 10 | 11 | 12 | 3 | 4 |
| 3 | 3 | 14 | 3 | 12 | 6 | 5 | 2 | 1 |
| 1 | 13 | 8 | 1 | 3 | 12 | 10 | 1 | 2 |

*(p) White St.*

| Centering | Auto | Range | Pareto | Vast | Level | Log | Power | Cumulative |
|---|---|---|---|---|---|---|---|---|
| 16 | 16 | 9 | 16 | 14 | 3 | 12 | 16 | 17 |
| 11 | 3 | 6 | 11 | 3 | 16 | 13 | 12 | 8 |
| 5 | 11 | 4 | 5 | 9 | 11 | 10 | 5 | 9 |
| 8 | 7 | 14 | 8 | 11 | 8 | 5 | 7 | 5 |
| 12 | 1 | 1 | 12 | 1 | 17 | 16 | 13 | 12 |
| 15 | 15 | 8 | 15 | 15 | 4 | 2 | 15 | 7 |
| 6 | 6 | 11 | 6 | 4 | 12 | 14 | 6 | 16 |
| 7 | 2 | 12 | 7 | 6 | 9 | 15 | 8 | 13 |
| 14 | 12 | 15 | 14 | 16 | 2 | 9 | 14 | 14 |
| 17 | 10 | 5 | 17 | 2 | 15 | 17 | 17 | 10 |
| 10 | 14 | 16 | 10 | 13 | 5 | 7 | 10 | 11 |
| 9 | 8 | 10 | 9 | 12 | 6 | 1 | 9 | 6 |
| 13 | 17 | 17 | 13 | 17 | 1 | 3 | 11 | 4 |
| 4 | 13 | 3 | 4 | 8 | 14 | 11 | 4 | 15 |
| 2 | 4 | 2 | 2 | 7 | 13 | 8 | 2 | 2 |
| 3 | 9 | 7 | 3 | 10 | 7 | 4 | 3 | 1 |
| 1 | 5 | 13 | 1 | 5 | 10 | 6 | 1 | 3 |

Table *2*

*Summary of min, max and standard deviation of all parameters estimated by all methods for water quality station at:*

*(a) Battleground Avenue.*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 1507 | 1407 | 4031 | 4067 | 4040 | 1697 |
| | Min | 624 | 622 | 1625 | 1493 | 1490 | 2750 |
| Nitrate | Max | 4858 | 4858 | 17912 | 18103 | 18103 | 8800 |
| | Std. | 392 | 508 | 1280 | 1103 | 1104 | 1061 |
| | Min | 57 | 57 | 107 | 114 | 114 | 117 |
| Nitrite | Max | 1074 | 1776 | 3742 | 3621 | 3621 | 3109 |
| | Std. | 494022 | 451435 | 583797 | 548536 | 588316 | 478220 |
| | Min | 529080 | 526601 | 1103195 | 1136151 | 1044748 | 1000000 |
| TDS | Max | 2260333 | 2126550 | 3189268 | 3228332 | 3306108 | 2979676 |
| | Std. | 101269 | 186479 | 85147 | 98623 | 226747 | 197211 |
| | Min | 26701 | 23482 | 45391 | 55186 | 48534 | 115000 |
| TSS | Max | 285222 | 536702 | 390754 | 430188 | 872275 | 881172 |
| | Std. | 3465 | 4088 | 2659 | 2704 | 2801 | 2174 |
| | Min | 1446 | 1310 | 2889 | 3461 | 2618 | 4046 |
| TKN | Max | 12448 | 15074 | 11119 | 11351 | 12090 | 11848 |
| | Std. | 141 | 192 | 394 | 313 | 320 | 410 |
| | Min | 186 | 173 | 312 | 395 | 384 | 480 |
| TP | Max | 608 | 869 | 1620 | 1487 | 1437 | 2137 |

*(b) Church St.*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 3219 | 3484 | 4257 | 4067 | 3866 | 1220 |
| | Min | 278 | 291 | 1653 | 1902 | 1645 | 3530 |
| Nitrate | Max | 12548 | 13538 | 18100 | 14064 | 16710 | 7987 |
| | Std. | 458 | 472 | 1963 | 1294 | 1850 | 1373 |
| | Min | 19 | 19 | 123 | 130 | 130 | 209 |
| Nitrite | Max | 1743 | 1808 | 6937 | 4464 | 6774 | 4340 |
| | Std. | 1187358 | 1097858 | 1300498 | 1563360 | 1029934 | 478817 |
| | Min | 290205 | 283970 | 1848496 | 1920705 | 1740202 | 1675366 |
| TDS | Max | 5123422 | 4804008 | 6128633 | 6403352 | 5515754 | 3572308 |
| | Std. | 69654 | 107550 | 69826 | 52119 | 728716 | 813032 |
| | Min | 8771 | 7616 | 22760 | 18877 | 16400 | 179000 |
| TSS | Max | 274902 | 415893 | 301907 | 213271 | 2832462 | 2749995 |
| | Std. | 5686 | 5649 | 6677 | 7199 | 5809 | 3222 |
| | Min | 674 | 652 | 3431 | 4612 | 3538 | 6620 |
| TKN | Max | 20413 | 17190 | 25967 | 31745 | 24856 | 18833 |
| | Std. | 598 | 712 | 1113 | 430 | 1243 | 1326 |
| | Min | 87 | 76 | 630 | 413 | 645 | 1546 |
| TP | Max | 2356 | 2839 | 5057 | 2015 | 4516 | 5879 |

*(c) Fleming St.*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 570 | 498 | 861 | 245 | 332 | 137 |
| | Min | 216 | 147 | 1746 | 567 | 386 | 785 |
| Nitrate | Max | 1867 | 1589 | 4509 | 1403 | 1618 | 1187 |
| | Std. | 68 | 71 | 628 | 201 | 201 | 70 |
| | Min | 17 | 17 | 123 | 42 | 42 | 50 |
| Nitrite | Max | 217 | 217 | 1928 | 624 | 624 | 259 |
| | Std. | 278828 | 215732 | 196527 | 95891 | 99680 | 67763 |
| | Min | 154424 | 145635 | 1040708 | 381851 | 350405 | 326564 |
| TDS | Max | 962138 | 754668 | 1610884 | 626081 | 627430 | 514973 |
| | Std. | 56954 | 111660 | 125265 | 42024 | 67412 | 469205 |
| | Min | 17418 | 20056 | 120277 | 29784 | 26317 | 230347 |
| TSS | Max | 195833 | 358221 | 511871 | 154219 | 202653 | 1524682 |
| | Std. | 1824 | 1851 | 2729 | 1227 | 1354 | 699 |
| | Min | 270 | 283 | 2453 | 708 | 743 | 1450 |
| TKN | Max | 5812 | 5616 | 9639 | 4016 | 4401 | 3151 |
| | Std. | 500 | 154 | 1422 | 379 | 129 | 168 |
| | Min | 49 | 52 | 376 | 128 | 138 | 233 |
| TP | Max | 1536 | 481 | 4552 | 1262 | 549 | 663 |

*(d) McConnell Rd.*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 5466 | 5202 | 12889 | 12975 | 13387 | 4745 |
| | Min | 817 | 910 | 5326 | 4954 | 5501 | 7520 |
| Nitrate | Max | 20740 | 20749 | 52079 | 54629 | 56739 | 22235 |
| | Std. | 699 | 717 | 4095 | 4625 | 4641 | 3330 |
| | Min | 68 | 68 | 287 | 367 | 367 | 462 |
| Nitrite | Max | 2299 | 2154 | 13613 | 17018 | 15939 | 10057 |
| | Std. | 4098111 | 3281481 | 2894865 | 2922221 | 3189486 | 2183218 |
| | Min | 1077194 | 1104743 | 4428135 | 4642878 | 3954636 | 2600000 |
| TDS | Max | 16902891 | 13772584 | 14984545 | 16678104 | 16840909 | 12581257 |
| | Std. | 393706 | 1142813 | 421503 | 463632 | 1964075 | 1472952 |
| | Min | 10987 | 11445 | 77117 | 71735 | 74725 | 187000 |
| TSS | Max | 1373383 | 3303194 | 1566930 | 1739325 | 7551304 | 6461732 |
| | Std. | 16132 | 15671 | 11419 | 9426 | 12055 | 8161 |
| | Min | 2298 | 2066 | 8451 | 9228 | 9086 | 14500 |
| TKN | Max | 60267 | 55805 | 52722 | 37926 | 44397 | 44270 |
| | Std. | 1455 | 1808 | 1160 | 1157 | 1973 | 1669 |
| | Min | 215 | 146 | 941 | 1196 | 893 | 923 |
| TP | Max | 5831 | 6751 | 5223 | 5616 | 8330 | 8293 |

*(e) Merritt Dr.*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 690 | 638 | 2504 | 2340 | 2362 | 536 |
| | Min | 147 | 148 | 1373 | 938 | 951 | 1370 |
| Nitrate | Max | 2434 | 2007 | 11158 | 10507 | 10507 | 3545 |
| | Std. | 157 | 155 | 1136 | 946 | 953 | 719 |
| | Min | 15 | 15 | 89 | 86 | 86 | 84 |
| Nitrite | Max | 469 | 484 | 4223 | 3502 | 3613 | 2180 |
| | Std. | 428180 | 229100 | 522497 | 480422 | 537390 | 205496 |
| | Min | 164667 | 161499 | 797451 | 803129 | 666243 | 581000 |
| TDS | Max | 1766689 | 969191 | 2927042 | 2770326 | 2788504 | 1437407 |
| | Std. | 80948 | 388213 | 77610 | 91286 | 417490 | 3365385 |
| | Min | 4162 | 3919 | 20726 | 27360 | 25760 | 590437 |
| TSS | Max | 277226 | 1333421 | 240539 | 346633 | 1421003 | 13582052 |
| | Std. | 2145 | 2206 | 2840 | 2498 | 2611 | 2002 |
| | Min | 591 | 558 | 3121 | 2655 | 2505 | 3964 |
| TKN | Max | 8164 | 7891 | 13691 | 12608 | 12230 | 12453 |
| | Std. | 217 | 660 | 303 | 290 | 664 | 3426 |
| | Min | 49 | 42 | 260 | 210 | 214 | 1190 |
| TP | Max | 750 | 2335 | 1437 | 1261 | 2587 | 13945 |

*(f) Pleasant Ridge Rd.*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 1499 | 1488 | 4035 | 4080 | 4033 | 2936 |
| | Min | 1245 | 1248 | 2270 | 2113 | 2119 | 3549 |
| Nitrate | Max | 6906 | 6906 | 17925 | 18033 | 18033 | 13096 |
| | Std. | 512 | 625 | 1273 | 1117 | 1327 | 4190 |
| | Min | 63 | 63 | 136 | 106 | 106 | 163 |
| Nitrite | Max | 1381 | 2045 | 3741 | 3607 | 3948 | 15635 |
| | Std. | 266047 | 310056 | 353092 | 409317 | 454653 | 802280 |
| | Min | 452187 | 432420 | 942421 | 767618 | 734061 | 867893 |
| TDS | Max | 1372947 | 1448206 | 2136430 | 2139924 | 2191562 | 3476087 |
| | Std. | 235030 | 456931 | 168479 | 141908 | 155651 | 7268353 |
| | Min | 23497 | 25867 | 45482 | 39888 | 43911 | 444397 |
| TSS | Max | 905934 | 1726910 | 709854 | 576965 | 600070 | 23542684 |
| | Std. | 3532 | 4650 | 3320 | 4210 | 4910 | 22712 |
| | Min | 1346 | 1241 | 2745 | 2285 | 2107 | 5874 |
| TKN | Max | 11568 | 15488 | 11952 | 15441 | 19095 | 71273 |
| | Std. | 146 | 172 | 185 | 179 | 224 | 687 |
| | Min | 212 | 211 | 318 | 338 | 240 | 376 |
| TP | Max | 748 | 776 | 1024 | 962 | 1031 | 2605 |

*(g) Randleman Rd.*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 2060 | 1615 | 4287 | 4442 | 4645 | 1475 |
| | Min | 277 | 273 | 1696 | 1546 | 1596 | 3067 |
| Nitrate | Max | 8213 | 6158 | 18375 | 18546 | 18546 | 7922 |
| | Std. | 275 | 294 | 1877 | 1771 | 1904 | 2037 |
| | Min | 27 | 27 | 144 | 152 | 152 | 169 |
| Nitrite | Max | 987 | 1033 | 6237 | 6182 | 6468 | 6300 |
| | Std. | 1570545 | 1161471 | 956999 | 1062736 | 1107783 | 585139 |
| | Min | 378123 | 372767 | 2078994 | 1676108 | 1485877 | 1540000 |
| TDS | Max | 6466515 | 4894570 | 5432506 | 5495832 | 5487786 | 3917373 |
| | Std. | 151256 | 379436 | 203858 | 161835 | 731006 | 373333 |
| | Min | 4544 | 4301 | 31747 | 29321 | 29631 | 249049 |
| TSS | Max | 534017 | 1427588 | 808060 | 653552 | 2851287 | 1517200 |
| | Std. | 7415 | 5470 | 4738 | 3951 | 4882 | 2939 |
| | Min | 1228 | 985 | 4177 | 4555 | 3850 | 7229 |
| TKN | Max | 28941 | 19940 | 21198 | 16891 | 19007 | 16800 |
| | Std. | 425 | 657 | 495 | 1263 | 1326 | 590 |
| | Min | 107 | 86 | 657 | 649 | 514 | 1021 |
| TP | Max | 1676 | 2362 | 2089 | 5564 | 4749 | 2977 |

*(h) Rankin Mill Rd.*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 277116 | 93239 | 52883 | 141278 | 122963 | 45643 |
| | Min | 114820 | 114963 | 91781 | 183980 | 116645 | 175329 |
| Nitrate | Max | 1241975 | 460256 | 249541 | 714163 | 519627 | 315758 |
| | Std. | 4599 | 5731 | 5075 | 10445 | 12556 | 3918 |
| | Min | 434 | 442 | 510 | 1124 | 1143 | 1500 |
| Nitrite | Max | 15233 | 17652 | 18509 | 38224 | 41962 | 12616 |
| | Std. | 5394557 | 3664790 | 1912253 | 3662119 | 3941400 | 2839279 |
| | Min | 3329124 | 3323294 | 3335038 | 7132105 | 5573644 | 6428885 |
| TDS | Max | 24295515 | 16993466 | 10263452 | 20582868 | 20066101 | 14372756 |
| | Std. | 656705 | 2733564 | 201041 | 487410 | 2314083 | 975791 |
| | Min | 35131 | 36799 | 36805 | 89900 | 92838 | 634748 |
| TSS | Max | 2143152 | 10003857 | 685530 | 1937060 | 8841402 | 3607412 |
| | Std. | 33506 | 37769 | 19211 | 39228 | 41524 | 27355 |
| | Min | 17182 | 17187 | 19773 | 44455 | 44469 | 52780 |
| TKN | Max | 131410 | 142374 | 79241 | 160730 | 177579 | 121812 |
| | Std. | 17207 | 10915 | 7522 | 17808 | 17795 | 13398 |
| | Min | 5238 | 4050 | 1883 | 5596 | 3575 | 3529 |
| TP | Max | 71206 | 39849 | 26260 | 59033 | 58840 | 42477 |

*(i) W.JJ*

| Parameters | | M1(kg) | M2(kg) | M3(kg) | M4(kg) | M5(kg) | LOADEST(kg) |
|---|---|---|---|---|---|---|---|
| | Std. | 422 | 414 | 1065 | 804 | 934 | 271 |
| | Min | 91 | 78 | 399 | 238 | 319 | 574 |
| Nitrate | Max | 1646 | 1348 | 4617 | 2914 | 3766 | 1677 |
| | Std. | 114 | 138 | 441 | 288 | 365 | 366 |
| | Min | 7 | 7 | 31 | 36 | 34 | 44 |
| Nitrite | Max | 382 | 382 | 1352 | 863 | 1108 | 1105 |
| | Std. | 530580 | 320837 | 213055 | 936236 | 574645 | 170630 |
| | Min | 77310 | 77636 | 332063 | 216589 | 274326 | 331000 |
| TDS | Max | 2114084 | 1312669 | 1131824 | 3669644 | 2400734 | 1012107 |
| | Std. | 43686 | 45986 | 24262 | 17295 | 20779 | 76241 |
| | Min | 3910 | 3070 | 19812 | 11399 | 15606 | 46400 |
| TSS | Max | 172165 | 146496 | 101169 | 66538 | 83854 | 370607 |
| | Std. | 1930 | 1828 | 764 | 1967 | 1366 | 967 |
| | Min | 298 | 276 | 878 | 941 | 910 | 1996 |
| TKN | Max | 7775 | 7012 | 3685 | 7102 | 5394 | 5693 |
| | Std. | 133 | 115 | 107 | 135 | 121 | 122 |
| | Min | 27 | 23 | 112 | 38 | 75 | 189 |
| TP | Max | 511 | 397 | 501 | 471 | 486 | 665 |